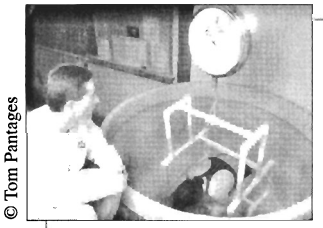
© Tom Pantages

A team of researchers (exercise scientists and nutritionists) interested in the relative effects of diet and exercise on body composition conducted a study with four randomly selected groups: (1) A control group received no treatment, (2) a diet group followed a diet but no exercise program, 3) an exercise group followed an exercise program but no diet, and 4) a combined group followed both a diet and an exercise program. Percent body fat of the subjects in all four groups was measured at the end of the 6-week program.

If the researchers wanted to analyze the data using *t* tests, six tests would be required to make all possible group comparisons of the groups. This procedure is both cumbersome and invites a type I error. Is there a better way? In this chapter we discuss the disadvantages of multiple t tests, and we learn how to compare three or more means with a single analysis of variance test.

Analysis of variance (ANOVA) is a parametric statistical technique used to determine whether significant differences exist among means from three or more sets of sample data. In a *t* test, the differences between the means of two groups are compared with the difference expected by chance alone. Analysis of variance compares the variability among three or more group means, the between-group variability, with the variability of scores within the groups, the within-group variability. This produces a ratio value called *F* (*F* = average variance between groups divided by average variance within groups). The symbol for analysis of variance (*F*) is named after the English mathematician Ronald Aylmer Fisher (1890–1962), who first described it (Kotz & Johnson, 1982, Vol. 3, p. 103).

If the between-group variability exceeds the within-group variability by more than would be expected by chance alone, it may be concluded that at least one of the group means differs significantly from another group mean. The null hypothesis (H_0) for an *F* test is designated as

$$\mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k.$$

The null hypothesis assumes that the means of two or more populations are equal or that a single population is the parent of the several untreated samples drawn from it. Therefore, untreated means of samples randomly drawn from the same population(s) should not differ by more than chance. When at least one sample mean is significantly different from any other, *F* is significant and we reject the null hypothesis.

Like the *t* test, the theoretical concepts of analysis of variance are based on random samples drawn from a population and the characteristics of the normal curve. If a large number of samples are randomly drawn from a population, the variance among the scores of all subjects in all groups is the best estimate of the variance of the population. When the variance among all the scores in a data set is known, it may be used to determine the probability that a deviant score is not randomly drawn from the same population. This argument may be expanded to infer that if randomly drawn scores are randomly divided into subgroups, the variance among the subgroup means may be expected to be of the same relative magnitude as the variance among all of the individual scores that comprise the several groups.

With untreated data, when only random factors are functioning between the group means and within the scores of the groups, the between-group and within-group variances should be approximately equal. The *F* ratio, the ratio of the average between-group variance divided by the average within-group variance, is expected to be about 1.00. When the value of the *F* ratio exceeds 1.00 by more than would be expected by chance alone, the variance between the means is judged to be significant (i.e., the difference was caused by a factor other than chance), and H_0 is rejected.

The *F* ratio is analogous to the *t* ratio in that both compare the actual, or observed, mean differences with differences expected by chance. When this ratio exceeds the limits of chance at a given level of confidence, chance as a cause of the differences is rejected. In the *F* test, the actual differences are the variances between the group means, and the expected differences are the variances within the individual scores that make up the groups. The *t* test is actually a special case of analysis of variance with two groups. Because *t* uses standard deviations and ANOVA uses variance (*V*) to evaluate mean differences, and $SD^2 = V$, when there are only two groups in ANOVA, $t^2 = F$.

Analysis of variance is one of the most commonly used statistical techniques in research. But students often ask, Why is this new technique needed when a *t* test could be used between each of the groups? For example, in a four-group study, why not conduct six *t* tests—between groups A and B, A and C, A and D, B and C, B and D, and C and D? There are three reasons why multiple *t* tests are not appropriate.

1. **There is a greater probability of making a type I error (rejecting the null hypothesis when it is really true) when one conducts multiple *t* tests on samples taken from the same population.** When a single *t* test is performed, the findings are compared to the probability of chance. If a confidence level of 95% is set, we are willing to refute chance if the odds are 19 to 1 against it. When multiple *t* tests are conducted on samples randomly drawn from the same population, the odds of finding the one deviant conclusion that is expected by chance alone increase.

When 20 *t* tests are performed at $p = .05$ on completely random data, it is expected that one of the tests will be found significant by chance alone (1 to 19 odds). Therefore if we perform 20 *t* tests on treated data, and 1 of the 20 tests produces a significant difference, we cannot tell if it represents a true difference due to treatment or if it represents the one deviant score out of 20 that is expected by chance alone. If it is due to chance but falsely declared to be significant, a type I error has been made. This dilemma is sometimes referred to as the **familywise error rate** (the error rate when making a family of comparisons).

Keppel (1991, p. 164) reports that the relationship between the single comparison error rate (α) and the familywise error rate (FW_α) is

$$FW_\alpha = 1-(1-\alpha)^C, \tag{9.01}$$

where C is the number of comparisons to be made. If we conduct six t tests, comparing all possible combinations of four groups (A, B, C, and D), at $\alpha = .05$, then

$$FW_\alpha = 1-(1-.05)^6 = .26.$$

In this example, conducting multiple t tests raises the probability of a type I error from .05 to .26. Keppel suggests that FW_α may be roughly estimated by the product of C and α. In this example, $FW_\alpha \cong 6 \times .05 = .30$. This method will always overestimate FW_α but is fairly close for small values of C and α. ANOVA eliminates this problem by making all possible comparisons among the means in a single test.

There may be occasions when multiple tests on samples from the same population are required. When this is the case, a commonly used modification of the alpha level called a **Bonferroni adjustment** is recommended. To perform the adjustment, divide the single-test alpha level by the number of tests to be performed. If five tests are to be made at $p = .05$, the adjusted alpha level to reject H_0 would be .01 (.05/5 = .01).

2. **The *t* test does not make use of all available information about the population from which the samples were drawn.** The t test is built on the assumption that only two groups have been randomly selected from the population. In the t test, the estimate of the standard error of the difference between means is based on data from two groups only. When three or more groups have been selected, information about the population from three or more samples is available that should be used in the analysis, yet t considers only two samples at a time. Analysis of variance employs all of the available information from all groups simultaneously.

3. **Multiple *t* tests require more time and effort than a simple ANOVA.** It is easier, especially with a computer, to conduct one F test than to conduct multiple t tests.

Because of these reasons, analysis of variance is employed in place of multiple t tests when three or more groups of data are involved. Analysis of variance can determine if a significant difference exists among any of the groups represented in the experiment, but ANOVA does not identify the group or groups that differ. A significant F value only indicates that at least one significant difference exists somewhere among the many possible combinations of groups. When a significant F is found, additional **post hoc** (after the fact) tests must be performed to identify the group or groups that differ. If F is not significant, no further analysis is needed because we know that there are no significant differences among any of the groups.

Assumptions in ANOVA

The F test is based on the following assumptions:

- The population(s) from which the samples are drawn is normally distributed. Violation of this assumption has little effect on the F value among the samples (Keppel, 1991, p. 97). The F test produces valid results even when the population is not normally distributed. For this reason it is considered to be robust.
- The variability of the samples in the experiment is equal or nearly so (homogeneity of variance). As with the assumption of normality, violation of this assumption does not radically change the F value. However, as a general rule, the largest group variance should not be more than two times the smallest group variance.
- The scores in all the groups are independent; that is, the scores in each group are not dependent on, not correlated with, or not taken from the same subjects as the scores in any other group. The samples have been randomly selected from the population and randomly assigned to conditions. If there is a known relationship among the scores of subjects in the several groups, use repeated measures analysis of variance (see chapter 10).
- The data are based on a parametric scale, either interval or ratio. (For nonparametric data analysis, see chapter 13.)

The F test, like the t test, is considered robust. It provides dependable answers even when there are violations of the assumptions. Violations are more critical when sample sizes are small or Ns are not equal. If violations are committed that cannot be controlled and that the researcher thinks may increase the possibility of a type I error, a more conservative p value should be used to compensate for the violations (i.e., use $p = .01$ rather than $p = .05$).

Sources of Variance

The computation of F is simple in theory and does not involve difficult mathematics. When several groups of data are compared, each group has a mean (the group mean) and the entire data set, all groups combined, has a mean (the grand mean).

The grand mean is computed by summing the scores from all groups and dividing by the total number of subjects in all groups combined (N). Each of the group means may differ from the grand mean. The variance, or deviation, of the group means from the grand mean is called the **between-group variance.**

In addition to variance between the means, each individual score deviates from the mean of its group by a certain amount. This source of variance is called

within-group variance. These two sources of variance, between-group and within-group, are the basic components used to compute the analysis of variance, F.

A third source of variance, the **total variance,** may be computed by determining the deviation of each score in each group from the grand mean. Total variance is equal to the sum of between-group variance and within-group variance:

$$\text{Variance}_{\text{total}} = \text{Variance}_{\text{between}} + \text{Variance}_{\text{within}}.$$

Total variance and between-group variance are relatively easy to calculate. But calculating within-group variance, although not mathematically difficult, is tedious. Because total variance is the sum of between- and within-group variance, within-group variance may be determined by subtracting between-group variance from total variance. Total variance is not essential for determining F, but it is useful in calculating within-group variance.

Sum of Squares and Mean Square

The **sum of squares** (SS) is the sum of the squares of the deviations of each score from a mean:

$$SS = \Sigma (X - \bar{X})^2.$$

It is computed by subtracting each score from the mean, squaring the deviation scores, and adding them up. The sum of squares *within* any group can be computed from the individual scores and the group mean:

$$SS_W = \Sigma (X - \bar{X}_{\text{group}})^2.$$

The sum of squares between groups can be computed by subtracting the grand mean from each group mean:

$$SS_B = \Sigma (\bar{X}_{\text{group}} - \bar{X}_{\text{grand}})^2.$$

The sum of squares for the total can be computed by subtracting each individual score and the grand mean:

$$SS_T = \Sigma (X - \bar{X}_{\text{grand}})^2.$$

The total sum of squares is always equal to the between-group sum of squares plus the within-group sum of squares:

$$SS_T = SS_B + SS_W. \tag{9.02}$$

In ANOVA, the size of the sum of squares between groups is compared to the size of the sum of squares within groups. The size of the sum of squares is dependent on (a) the number of scores summed and (b) the size of the squared deviations from the mean, or the variance. To account for the differences in the number of scores that make up the between-group (number of groups) and within-group (number of individual scores) sums of squares, the **mean square** (MS) is computed by dividing each SS by the appropriate degrees of freedom ($MS = SS/df$). This process makes the mean, or average, variabilities for SS_B and SS_W comparable.

Degrees of freedom within are determined by subtracting the number of groups (k) from the total number of subjects in all groups (N):

$$df_W = N - k. \tag{9.03}$$

Degrees of freedom between are determined by the number of groups (k) minus 1:

$$df_B = k - 1. \tag{9.04}$$

To determine mean square within, divide SS_W by the degrees of freedom within:

$$MS_W = SS_W / df_W. \tag{9.05}$$

To determine mean square between, divide SS_B by the degrees of freedom between:

$$MS_B = SS_B / df_B. \tag{9.06}$$

These mean square values are now directly comparable and can be used to calculate F:

$$F = \frac{MS_B}{MS_W}. \tag{9.07}$$

The mean square within (MS_W) is the denominator of the F test. Like SE_D in the t test, it represents the amount of variance that can be expected due to chance occurrences alone. The F ratio compares MS_W to the differences between the means represented by mean square between (MS_B).

In ANOVA, MS_W is often referred to as mean square error and may be denoted MS_E. The term *error* does not mean a mistake; it means the variance that can occur by chance alone, or the variance that is unaccounted for by the effects of treatment. From this point on, we will use the representation MS_E, but it is synonymous with MS_W.

When MS_E is large, it tends to mask small treatment effects. When MS_E is small, it is easier to identify treatment effects. *The intent of all research designs is to keep MS_E as small as possible.* The value of MS_E depends on the variability in the data and on the number of scores (N) that make up the data. Of these two factors, only N can be controlled by the researcher. Therefore, studies with large Ns are more powerful in detecting mean differences.

Calculating *F*

We can calculate *F* using the formulas presented thus far. The concepts for this calculation of *F* are easy to understand because they are based on the definition of terms we have already introduced. But, like the definition formulas for standard deviation and for Pearson's correlation, the definition method for ANOVA is tedious to use when raw scores contain decimal values and *N* is large. A second computational method, the raw score method, is easier to use. The definition method is almost never used in practical applications but is presented here to explain the concepts that underlie ANOVA. The raw score method should be used to solve actual problems. Both methods, when correctly applied, produce the same answers.

The Definition Method

Table 9.1 presents hypothetical strength measurements on five groups (X_1 to X_5) of 7 subjects each. The numerical values do not represent any particular unit of measure. They are purposely small and discrete so that the calculations will be easy. The subjects were randomly selected from a population and randomly assigned to groups. Each subject completed 6 weeks of strength training. In this table, *n* is the number of subjects in each group, while *N* is the total number of subjects or scores in all groups combined. We shall test the null hypothesis (H_0) that none of the treatments had an effect. H_0 predicts no significant differences among any of the group means.

Table 9.1 Data for Simple ANOVA

	X_1	X_2	X_3	X_4	X_5 (control)
	4	5	5	8	5
	5	7	4	4	4
	6	9	6	6	3
	7	8	5	8	4
	4	9	5	5	6
	6	7	6	6	4
	5	10	4	7	5
ΣX	37	55	35	44	31
n	7	7	7	7	7
$\bar{X}$	5.29	7.86	5.00	6.29	4.43

$\Sigma X_T = 37 + 55 + 35 + 44 + 31 = 202$
$N = 7 + 7 + 7 + 7 + 7 = 35$
$M_G = 202/35 = 5.77$

Groups 1, 2, 3, and 4 participated in different strength training programs; group 5 was the control group. We want to know whether any differences exist between the mean scores of the groups after the differential training. In this case, using ANOVA to solve the problem takes the place of 10 separate *t* tests.

The steps for calculating *F* by the definition method are as follows:

1. Calculate the sum of each group (ΣX_1, ΣX_2, ΣX_3, . . .), the mean of each group ($\bar{X}_1$, $\bar{X}_2$, $\bar{X}_3$, . . .), the grand sum (ΣX_T), and the grand mean (M_G). These values can be found at the bottom of table 9.1.
2. Calculate the within-group sum of squares (SS_W) for each group.
   - Determine the deviation scores of each raw score from its group mean ($d = X - \bar{X}_{group}$), as shown in table 9.2.
   - Square each deviation and find the sum of d^2 for each group, as shown in table 9.3.

Table 9.2 Calculation of Within-Group Deviations

d_1	d_2	d_3	d_4	d_5
–1.29	–2.86	0.00	1.71	.57
–.29	–.86	–1.00	–2.29	–.43
.71	1.14	1.00	–.29	–1.43
1.71	.14	0.00	1.71	–.43
–1.29	1.14	0.00	–1.29	1.57
.71	–.86	1.00	–.29	–.43
–.29	2.14	–1.00	.71	.57

Note: These values are derived by subtracting the group mean from table 9.1 from the individual score (for the first score (X) in group X_1: 4.00 – 5.29 = –1.29).

Table 9.3 Squaring and Summing Within-Group Deviations

$(d_1)^2$	$(d_2)^2$	$(d_3)^2$	$(d_4)^2$	$(d_5)^2$
1.66	8.18	0.00	2.92	.33
.08	.74	1.00	5.24	.18
.50	1.30	1.00	.08	2.05
2.92	.02	0.00	2.92	.18
1.66	1.30	0.00	1.66	2.46
.50	.74	1.00	.08	.18
.08	4.58	1.00	.50	.33
$\Sigma(d_1)^2 = 7.40$	$\Sigma(d_2)^2 = 16.86$	$\Sigma(d_3)^2 = 4.00$	$\Sigma(d_4)^2 = 13.40$	$\Sigma(d_5)^2 = 5.71$

Table 9.4 Calculation of Between-Group Deviations

	$\bar{X}$	M_G	$d_B = (\bar{X} - M_G)$	$(d_B)^2$
Group 1	5.29	5.77	–0.48	.23
Group 2	7.86	5.77	2.09	4.37
Group 3	5.00	5.77	–.77	.59
Group 4	6.29	5.77	.52	.27
Group 5	4.43	5.77	–1.34	1.80
				$\Sigma(d_B)^2 = 7.26$

- Sum the squared deviations from each group. This value is the sum of squares within:

$$SS_W = \Sigma\Sigma(X - \bar{X}_{\text{group}})^2 = 7.40 + 16.86 + 4.00 + 13.40 + 5.71 = 47.37.$$

3. Calculate the between-group sum of squares (SS_B).
   - Find the deviation (d) of each group mean from the grand mean, square each deviation, and sum these deviations (Σd^2_B) (table 9.4).
   - Because there are n times more values contributing to SS_W than to SS_B, multiply Σd^2_B by n (7) to make SS_B directly comparable with SS_W:

$$SS_B = 7.26 \times 7 = 50.82.$$

4. Determine the degrees of freedom between and within:

$$df_B = 5 - 1 = 4, \quad df_E = 35 - 5 = 30.$$

5. Determine the mean square between (MS_B) and the mean square error (MS_E) by dividing the SS values by the appropriate df:

$$MS_B = 50.82/4 = 12.71, \quad MS_E = 47.37/30 = 1.58.$$

6. Determine the ratio (F) between MS_B and MS_E:

$$F = 12.71/1.58 = 8.04.$$

The Raw Score Method

The definition method for calculating F is quite tedious to perform by hand on data containing decimal values and large N. Alternative calculation formulas for SS_B and SS_W that use raw scores rather than deviation scores make the computation much easier. Recall from equation 9.02 that $SS_T = SS_B + SS_W$. Because SS_T and SS_B are easier to calculate than SS_W, SS_W may be determined by the equation $SS_W = SS_T - SS_B$. Once SS_B and SS_W are found, the remainder of the calculation of F is the same as described in the definition method.

The computational formulas to calculate SS_B, SS_T, and SS_W are as follows:

$$SS_B = \frac{(\Sigma X_1)^2}{n_1} + \frac{(\Sigma X_2)^2}{n_2} + \frac{(\Sigma X_3)^2}{n_3} \ldots + \frac{(\Sigma X_k)^2}{n_k} - \frac{(\Sigma X_T)^2}{N} \quad (9.08)$$

$$SS_T = \Sigma X^2 - \frac{(\Sigma X_T)^2}{N} \quad (9.09)$$

$$SS_W = SS_T - SS_B \quad (9.10)$$

The raw score method requires that each individual raw data value (X) be squared and the sums of squares for each group be calculated. The raw data from table 9.1 are squared and summed in table 9.5.

Table 9.5 Raw Data (X) Squared and Summed

$(X_1)^2$	$(X_2)^2$	$(X_3)^2$	$(X_4)^2$	$(X_5)^2$
16	25	25	64	25
25	49	16	16	16
36	81	36	36	9
49	64	25	64	16
16	81	25	25	36
36	49	36	36	16
25	100	16	49	25
$\Sigma(X_1)^2 = 203$	$\Sigma(X_2)^2 = 449$	$\Sigma(X_3)^2 = 179$	$\Sigma(X_4)^2 = 290$	$\Sigma(X_5)^2 = 143$

$\Sigma X^2 = 203 + 449 + 179 + 290 + 143 = 1{,}264$

Applying the data from tables 9.1 and 9.5 to equations 9.08, 9.09, and 9.10 yields

$$SS_B = \frac{37^2}{7} + \frac{55^2}{7} + \frac{35^2}{7} + \frac{44^2}{7} + \frac{31^2}{7} - \frac{202^2}{35}$$

$$SS_B = 1{,}216.57 - 1{,}165.83 = 50.74$$

and

$$SS_T = 1{,}264 - \frac{202^2}{35} = 98.17$$

and

$$SS_W = 98.17 - 50.74 = 47.43.$$

These values differ slightly from the values calculated by the definition method, because the definition method requires rounding and squaring of rounded values. The values calculated by the raw score method are more accurate.

The computations of degrees of freedom, mean square values, and F are the same in both methods:

$$df_B = 5 - 1 = 4$$

$$df_E = 35 - 5 = 30$$

$$MS_B = \frac{50.74}{4} = 12.69$$

$$MS_E = \frac{47.43}{30} = 1.58$$

$$F = \frac{12.69}{1.58} = 8.03$$

Determining the Significance of *F*

The significance of F is determined by referring to tables A.4, A.5, and A.6 in appendix A, which show the values of F for appropriate degrees of freedom between and within groups. In these tables df_B are read across the top of the table, and df_E are read down the left-hand side.

Table A.4 shows the values of F for $p = .10$. When $df_B = 4$ and $df_E = 30$, the critical F value is 2.14. Because the obtained F (8.03) exceeds this value, we look in table A.5, which lists the F values for $p = .05$. For $p = .05$, critical $F = 2.69$. The obtained F is still larger than critical F, so we look in table A.6, which has F values for $p = .01$. For $p = .01$, critical $F = 4.02$. The obtained F (8.03) also exceeds this critical F value, so the obtained F is declared significant at $p < .01$. This means that the odds are less than 1 in 100 that an F larger than 4.02 would be obtained by chance alone. Therefore, the level of confidence reached by the obtained F value (8.03) is greater than 99%. We conclude that the treatment had an effect on at least one of the groups and reject H_0.

The results of analysis of variance are usually reported in table form. Table 9.6 is an example of a typical ANOVA table.

Table 9.6 Tabular Report of Analysis of Variance

Source of variance	Sum of squares	*df*	Mean square	*F*	*p*
Between groups	50.74	4	12.69	8.03	<.01*
Within groups (error)	47.43	30	1.58		
Totals	98.17	34			

* This value is only approximate from table A.6. Computer programs can calculate the exact p value for a given F.

Post Hoc Tests

As we discussed earlier, a significant F alone does not specify which groups differ from one another. It only indicates that there are differences somewhere among the groups. To identify the groups that differ significantly from one another, a post hoc test must be performed.

A post hoc test is similar to a t test, except post hoc tests have a correction for familywise alpha errors built into them. Some are more conservative than others. Conservative means that the tests are less powerful because they require larger mean differences before significance can be declared. Conservative tests offer greater protection against type I errors, but they are more susceptible to type II errors.

Several post hoc tests may be applied to determine the location of group differences after a significant F has been found. Two of the most commonly used tests, **Scheffé's confidence interval (I)** and **Tukey's honestly significant difference (*HSD*)**, are described here. Scheffé permits all possible comparisons, while Tukey permits only pairwise comparisons. Tukey is easier to apply, and is appropriate in most research designs. For information on other post hoc tests, see Keppel (1991, p. 170–177) or other advanced statistical texts.

Scheffé's Confidence Interval (I)

The Scheffé test is the most conservative post hoc test. It should be used if all possible comparisons—that is, more than just pairwise comparisons—are to be made. Pairwise comparisons contrast one group mean against another. All possible comparisons include pairwise comparisons plus comparisons of combinations of groups to single groups or other combinations. For example, the average of two treatment effects may be compared with the average of two other treatment effects, or the average of several treatments may be compared to the control group. Scheffé places no restrictions on the number of comparisons that can be made.

Scheffé's confidence interval (I) permits us to compare the raw score means for any two groups or combinations of groups. The formula is

$$I = \sqrt{(k-1)(F_\alpha)\left(\frac{2MS_E}{n}\right)}, \tag{9.11}$$

where k = number of groups, F_α = the value of F from tables A.4 to A.6 for a given p value and the df_B and df_E values used in ANOVA, MS_E = mean square error from ANOVA, and n = size of the groups.

If groups are not equal in size, equation 9.11 may be modified as follows to accommodate any two groups with unequal values of n:

$$I = \sqrt{(k-1)(F_\alpha)(MS_E)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}. \tag{9.12}$$

Because the ns in our example are equal, we apply equation 9.11 to the data from table 9.6. We compute I at $p = .01$ and $p = .05$ as follows:

$$I = \sqrt{(5-1)(4.02)\left(\frac{(2)(1.58)}{7}\right)} = 2.69, \quad p = .01$$

$$I = \sqrt{(5-1)(2.69)\left(\frac{(2)(1.58)}{7}\right)} = 2.20, \quad p = .05.$$

This interval size (I) is the raw score value by which any two means must differ to be considered significant. Constructing a mean difference table makes it easy to identify the groups that differ from one another.

Table 9.7 shows the differences between all pairwise combinations of means. Scheffé's I requires a mean difference of 2.20 for $p = .05$ and 2.69 for $p = .01$. The table identifies groups 2 and 1 as significantly different at $p < .05$, and groups 2 and 3, and 2 and 5 significantly different at $p < .01$.

The ordered values of the means provide additional insight.

Group	Mean
2	7.86
4	6.29
1	5.29
3	5.00
5 (control)	4.43

Table 9.7 Mean Difference Analysis

	Group 1	Group 2	Group 3	Group 4	Group 5
Group 1	0.00	2.57*	.29	1.00	0.86
Group 2		0.00	2.86**	1.57	3.43**
Group 3			0.00	1.29	0.57
Group 4				0.00	1.86
Group 5					0.00

Note: Values above are calculated by taking the absolute value of the difference between two means in table 9.1 (i.e., $|\bar{X}_1 - \bar{X}_2| = |5.29 - 7.86| = 2.57$).
* $p < .05$.
** $p < .01$.

It is clear that the treatment given to group 2 had a significant effect. Group 2 differs significantly from all other groups except group 4. Groups 1, 3, and 4 do not differ significantly from control.

An alternate method of identifying specific group mean differences using Scheffé's method may be developed by solving equation 9.12 for F. This results in the following:

$$F_{\text{Scheffé}} = \frac{(\bar{X}_1 - \bar{X}_2)^2}{(k-1)(MS_E)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}. \tag{9.13}$$

$F_{\text{Scheffé}}$ is an F value calculated for any two specified groups (in the equation, $\bar{X}_1$ and $\bar{X}_2$ are specified, but any two means could be used). It is interpreted differently than I. I represents the raw score mean difference between any two groups that must be attained for declaration of significance. But $F_{\text{Scheffé}}$ is an actual F value that must be compared to tables A.4 to A.6 for df_B and df_E in the ANOVA analysis to determine if the two compared means differ significantly.

Equation 9.13 is sometimes used by computer programs to calculate the F values for all possible pairwise comparisons. If the computer does not also print the p values, the $F_{\text{Scheffé}}$ values produced by the computer must be compared with critical values in tables A.4 to A.6 to determine the significance of the difference between any two groups.

Tukey's Honestly Significant Difference (HSD)

Tukey's honestly significant difference (*HSD*) test, like I, calculates the minimum raw score mean difference that must be attained to declare significance between

any two groups. But Tukey's test does not permit all possible comparisons; it only permits pairwise comparisons: Any single group mean may be compared to any other group mean. The formula for *HSD* is

$$HSD = q_{(k,df_E)}\sqrt{\frac{MS_E}{n}}, \tag{9.14}$$

where q = a value from the Studentized range distribution (see tables A.7, A.8, and A.9 in appendix A) for k and df_E at a given level of confidence (note that k is used, not df_B), MS_E is the mean square error value from the ANOVA analysis, and n is the size of the groups.

Equation 9.14 assumes the *ns* in each group are equal. It may be modified to compare any two groups with unequal values of n as follows:

$$HSD = q_{(k,df_E)}\sqrt{\frac{MS_E}{2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}. \tag{9.15}$$

Because *n*s are equal in our example, equation 9.14 is applied for both $p = .01$ ($q = 5.05$) and $p = .05$ ($q = 4.10$) as follows:

$$HSD = 5.05\sqrt{\frac{1.58}{7}} = 2.40,\ p = .01$$

$$HSD = 4.10\sqrt{\frac{1.58}{7}} = 1.95,\ p = .05.$$

These values (2.40 and 1.95) represent the minimum raw score differences between any two means that may be declared significant.

Tukey, a more liberal test, confirms Scheffé but also finds that groups 2 and 1 differ at $p < .01$ (see table 9.7). The values at $p = .01$ and $p = .05$ are both lower in Tukey's *HSD* test than for Scheffé's *I*. This makes *HSD* more powerful (i.e., more likely to reject H_0) than *I*.

Because we started with the null hypothesis and are not making any comparisons other than pairwise (i.e., we are not interested in the combined mean of two or more groups compared to other combined means), Tukey's test is appropriate. Scheffé may be too conservative for this research design. Based on the analysis by Tukey, group 2 is significantly different from groups 1, 3, and 5 at $p < .01$. Figure 9.1 presents the results in bar graph form. The T symbol above each bar represents standard deviation.

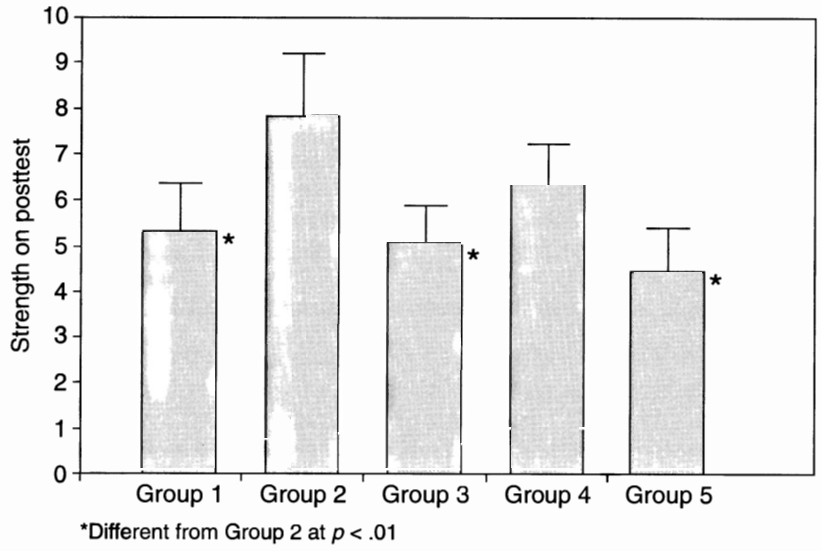

Figure 9.1 ANOVA on strength scores. Mean plus or minus standard deviation.

Concluding Statement Regarding Post Hoc Tests

When testing the null hypothesis with simple ANOVA, a researcher should first conduct an F test to determine if there are any differences among any of the groups and to determine the value of MS_E. If F is not significant, no further calculations are needed. If F is significant and comparisons other than pairwise comparisons are needed, Scheffé should be used to investigate differences among various combinations of groups. If F is significant, and only pairwise comparisons are to be made, Tukey should be used to contrast mean differences.

The Magnitude of the Treatment (Size of Effect)

A significant F is not the only factor considered when evaluating the importance of a finding. The researcher should also consider the size of the effect of the treatment. A significant F indicates the probability that the differences occurred by chance. ANOVA is actually a test of the reliability of the differences. When F is significant, the odds are good that similar mean differences would be found if the experiment were repeated. However, when N is large and MS_E is small, ANOVA may produce a significant F when the mean differences are so small that they have little or no practical value. Determining the size of the effect is a method of comparing treatment effects, independent of sample size.

For example, two groups of 200 subjects each participate in aerobic exercise programs for 4 days a week and 5 days a week. We may find that $\dot{V}O_2$max values in milliliters per kilogram per minute are 47.6 and 49.1, respectively, and that the difference is significant at $p < .05$. Although this difference may be real, or not due to chance, it is probably not large enough (1.5 milliliters per kilogram per minute) to be worth the effort of an additional workout per week.

R^2 (eta squared)

The most simple measure of the effects of treatment in ANOVA is R^2:

$$R^2 = \frac{SS_B}{SS_T}. \tag{9.16}$$

R^2 is the ratio of the variance due to treatment and the total variance. It produces a rough estimate of the size of the effect. For the data from the strength training experiment (table 9.6), $R^2 = 50.74 / 98.17 = .52$. This means that 52% of the total variance can be explained by the treatment effects. The remaining 48% is unexplained.

The value of R^2 will vary between 0.00 and 1.00, depending on the relative size of the treatment effects. It is a measure of the proportion of the total variance that is explained by the treatment effects. In this case, a little more than half of the total variance is explained. This is a fairly large proportion and confirms the conclusion that at least one treatment was effective in improving strength. In some statistics books (see Tabachnick & Fidell, 1996, p. 53), R^2 is called **eta squared.**

Omega Squared (ω^2)

Another method of determining the size of the effect is **omega squared** (ω^2):

$$\omega^2 = \frac{SS_B - (k-1)(MS_E)}{SS_T + MS_E}. \tag{9.17}$$

For the data in table 9.6, ω^2 is calculated as follows:

$$\omega^2 = \frac{50.74 - (5-1)(1.58)}{98.17 + 1.58} = .45.$$

Omega squared is a more accurate measure of effect size because it attempts to account for the unexplained variance (MS_E) and usually produces a smaller value than R^2. The value of $\omega^2 = .45$ is smaller than $R^2 = .52$, but it still indicates that a large proportion of the total variance may be attributed to treatment effects. This would still be considered a large effect size. Keppel (1991, p. 66) suggests that for ω^2 a value of .15 is large, .06 is medium, and .01 is small in the behavioral sciences.

The size of the effect is a relatively new concept in experimental design, but it is an important one to calculate and report. Thomas, Salazar, and Landers (1991, p. 344) state:

> Authors need to be convinced that they should report the magnitude of effects, as small differences can be easily declared significant based on some combination of small variances and large *N*s (Thomas and Nelson, 1996), or the reverse can occur—large differences can be declared nonsignificant due to large variances and small *N*s.

An Example From Leisure Studies/Recreation

Researchers in the effective use of leisure time have postulated that play activity may be used to reduce stress. It is proposed that play activity is most effective when it is perceived by the subject to be free play, or not directed by others. To test this hypothesis, Finney (1985) randomly divided male and female college-age subjects into four groups: high perceived control of the play experience (i.e., low structure), moderate perceived control, low perceived control, and a control group, who performed what they considered to be work, not play. The groups had 19, 20, 20, and 20 subjects, respectively.

All subjects performed a 30-minute stress-producing task—they worked 12 pages of math problems while listening to periodic bursts of 95 decibels of noise delivered through headphones. Upon completing this 30-minute stress period, the subjects engaged for 10 minutes in one of four play activities that varied in the amount of perceived control that the subjects had over their own play behavior. Following this play period, the subjects attempted to solve four geometric puzzles. The subjects were unaware that two of the puzzles were not solvable. Persistence on the two unsolvable puzzles (measured by time in total seconds spent on the two puzzles before giving up) was the dependent variable used to assess the effectiveness of the play period in reducing the stress created by the work task. Whereas H_1 was used to justify the study, the null hypothesis H_0 was tested statistically.

A simple ANOVA applied to the data yielded the information in tables 9.8 and 9.9. To identify specific mean differences, Finney applied $F_{\text{Scheffé}}$ (equation 9.13) to the comparisons between each group. This produced a matrix of F values contrasting each group with every other one. Table 9.10 demonstrates this analysis.

Table 9.8 Effects of Play on Reduction of Stress

Source of variance	Sum of squares	*df*	Mean square	*F*	*p*
Between groups	1,193,210.000	3	397,736.667	4.189	<0.01
Within groups (error)	7,121,448.000	75	94,952.640		
Totals	8,314,658.000	78			

Table 9.9 Group Means (Time in Seconds) and *SD* by Level of Structured Play

	High	Medium	Low	Control
Mean	705.90	466.30	427.90	385.05
Standard deviation	507.64	251.93	235.31	108.80

Note: Data are in total seconds.

Table 9.10 $F_{Scheffé}$ Values for Pairwise Comparisons

	High	Medium	Low	Control
High	0.000	1.964	2.643	3.521*
Medium		0.000	0.052	0.232
Low			0.000	0.064
Control				0.000

* $p < .05$.

The findings of $F_{Scheffé}$ declare that only the group with perceived high control of the play period differs from the control group at $p < .05$. (Table A.6 indicates that critical F for $df_B = 3$ and $df_E = 75$ is 4.13 at $p = .01$, and table A.5 indicates critical F is 2.76 at $p = .05$.)

Thinking that perhaps $F_{Scheffé}$ may be too conservative for this type of data, Finney then applied Tukey's post hoc test to analyze specific raw score mean differences (see table 9.11). Using equation 9.15, Finney found Tukey's *HSD* was 316.95 at $p = .01$ and 257.79 at $p = .05$. Tukey confirms Scheffé by indicating that the high group differs from the control group, but with Tukey, the difference is found to be significant at $p < .01$. The Tukey test also found that the high group differed from the low group at $p < .05$. But the low and medium groups still did not show significant differences between each other or from the control group.

Table 9.11 Mean Difference Values for Pairwise Comparisons

	High	Medium	Low	Control
High	0.00	239.60	276.00*	320.85**
Medium		0.00	38.40	81.25
Low			0.00	42.85
Control				0.00

** $p < .01$; * $p < .05$.

The value of ω^2 was .1080, revealing a moderate effect size. About 11% of the differences between the groups can be attributed to the play treatment.

Summary

ANOVA compares the means of three or more groups. When F is significant, it indicates that somewhere among the several means there is a significant difference. However, the specific mean differences are not identified. A post hoc test must be used to identify which means differ. After a significant F is found, some measure of the size of the effect of the treatment should be computed to ascertain the relative proportion of the differences that can be attributed to the treatment effects.

Remember that when two means differ significantly at some p value (for example, $p = .05$), the p value indicates the odds (5 in 100) that the means really are not different—that is, the odds that the null hypothesis is true. This is the same as stating that we are 95% sure that the null is false (95% level of confidence). We typically start with the assumption that there is no difference (H_0), and then we test that assumption. If we find significant F values, the null is rejected at the given level of confidence. We never really know if we are correct in accepting or rejecting the null hypothesis, we only know the odds, or probability, that we are correct in our conclusion.

Problems to Solve

Solve problem 1 by hand with a calculator, then check your hand calculated answers against computer-generated results. Use a computer to solve problems 2 and 3.

1. A biomechanics researcher wanted to test whether good, average, and poor sprinters differed in their horizontal foot speed. She classified the sprinters into three groups based on their sprint times. The horizontal foot speed at touchdown in feet per second was then analyzed with the following results:

   Good 4, 5, 8, 6, 7, 6
   Average 7, 8, 9, 6, 7, 10
   Poor 10, 13, 12, 8, 11, 12

   A. What are the mean values for each group?
   B. What are the sum of squares values?
   C. What are the degrees of freedom and mean square values?
   D. What is the F value? Is it significant? If so, what is the p value?

E. Set up a mean difference table.
F. Apply Scheffé's *I* to the data. Do any of the means differ?
G. Apply Tukey's *HSD* test. Do any of the means differ?
H. Does Tukey differ from Scheffé in the interpretation of the mean difference table?
I. What is the size of the effect? Is it small, moderate, or large?

Hint: If you use a computer to solve this problem, create two columns in the software database. Column one, the group column, will contain the numbers 1 (rows 1-6, good), 2 (rows 7-12, average), or 3 (rows 13-18, poor). Column two (the score column) will contain the foot speed scores.

2. A volleyball coach noted that players who practiced jumping during practice time seemed to be able to jump higher in the games. He wondered if jumping practice was more or less effective than lower extremity weight training to increase vertical jump height. To test this phenomena, he proposed the null hypothesis, that there are no significant differences between players who practiced jumping, players who participate in lower extremity weight training, and players who do neither exercise.

   For six weeks prior to the start of the season, he randomly divided his team into three groups of 10 players each. Group 1 (Control) practiced regularly without any special jumping or weight training. Group 2 (Jumping) spent the last 20 minutes of each practice session in jumping exercises, and Group 3 (Weight Training) spent the last 20 minutes of each practice session doing high resistance leg presses. Following are the results of a vertical jump test (in inches) taken two days before the first game.

Control	Weight training	Jumping
23	26	33
21	25	32
25	28	36
26	31	29
31	34	27
27	35	34
32	29	35
34	27	37
25	28	35
21	25	28

   A. What are the means and standard deviations for the three groups?
   B. What is the value of *F*? Is it significant? Is so, what is the *p* value?
   C. Did either of the experimental group means significantly exceed the control group? (Use Tukey's *HSD*.) Which one(s)?
   D. Do you accept or reject the null hypothesis?

   Review the hint given in problem 1 above for help on setting up the database. Problem 2's database will be similar to problem 1, with three groups in column 1 designated by numbers 1, 2, or 3, and vertical jump scores in column 2.

3. A physical education teacher wanted to know the effects of various activity levels on body composition. She surveyed her physical education classes and categorized 30 students into five activity levels: inactive, semiactive, normal, active, and very active, based on the amount of daily exercise in which the students participated. She measured percent body fat using skinfold calipers on all subjects with the following results.

Inactive	Semiactive	Normal	Active	Very active
30.2	29.4	22.9	17.6	10.9
29.6	17.6	25.4	13.4	13.7
35.2	26.4	19.6	20.3	12.8
19.1	25.3	18.7	19.6	14.7
26.3	22.5	21.8	15.1	9.3
22.4	28.6	24.9	10.7	12.7

   Use a computer to solve this problem.

   A. What is the independent variable in this study? What is the dependent variable?
   B. What are the means and standard deviations for each group?
   C. Is there a significant difference among any of the groups? What is the probability that the null hypothesis is true?
   D. Create a mean difference table and indicate the significance of the differences among the groups.

   Review the hint given in problem 1 above for help on setting up the database. Problem 3s database will be similar to problem 1s, but you will have five groups in column 1 designated by numbers 1, 2, 3, 4, or 5 and body fat values in column2.

See appendix C for answers to problems.

Key Words

Analysis of variance
Between-group variance
Bonferroni adjustment
Eta squared (R^2)
Familywise error rates
Mean square
Omega squared
Post hoc
Scheffé's confidence interval *(I)*
Sum of squares
Total variance
Tukey's honestly significant difference *(HSD)*
Within-group variance