

Chapter 9

Analysis of Variance

Researchers do not restrict their thinking to the comparison of only two groups. Often there are a number of treatment conditions that could logically be compared. For example, let us return to the example involving low achieving math students. At present we have considered only two possible sequences of the material; a traditional sequenced approach and one offering a revised ordering. Additional possibilities exist in making the curriculum more integrated by spiraling through the material and showing the relationships between different areas of mathematics. An integrated approach could take intact sections of different topics and continually cycle through progressively more advanced topics, which we will call a moderately integrated approach, or make each of the sections of the curriculum deal with concepts viewed from two or more mathematical perspectives. This will be called a radically integrated system. Table 9.1 contains data that we will use in this example¹.

Our example problem now is a comparison of the relative efficacy of four different strategies for sequencing the material in the schools mathematics curriculum with special attention to the performance of low achieving students. The question is how are we going to approach answering the question of whether “treatment” makes a difference. In particular we will be concerned with how we can obtain reasonable answers while maintaining control over our chance of making inferential errors.

In Chapters 6 and 7 the concept of Type I error was relatively straightforward. We had a single decision to make and we wanted to retain strict control over ‘coming to the conclusion that the treatments worked when all that really happened was an unusual random event’ (a Type I error). We wanted the chance of making a Type I error to be 5% so we required that the observed value of t or z be one of the most extreme 5% when the null hypothesis was true that in order for us to reject the null hypothesis. If we had only a 5% chance of making this type of error on this decision we had only a 5% chance of making it in the experiment because the experiment involved only one decision. Now we are faced with a number of treatment groups and potentially many questions on which we might make errors. If there are many questions to be asked should we focus on the chance that we will make an error on a question or should we worry about the chance that we will make an error

¹ The experiment we outline here would be extremely difficult to actually conduct since it would require random assignment of low achieving students across four major types of curriculum organization. One of the advantages of examples is that you get to make up data that would take years to collect.

any place within the set of questions. That will be an important decision influencing all other aspects of the experiment.

Table 9.1

	Old Seq	New Seq	Mod Int	Rad Int
	23	30	27	30
	21	36	31	26
	30	28	28	24
	26	28	35	32
	27	39	35	34
	29	31	30	31
	32	34	32	28
	31	27	31	20
	22	24	27	33
	26	28	29	36
	27	34	33	31
	29	28	35	29
	20	35	32	30
	28	31	26	24
	23	24	28	29
	30	32	32	33
	26	27	35	35
	28	29	28	30
	31	28	26	28
	30	30	32	30
Sum of scores	539	603	612	593
Means	26.95	30.15	30.60	29.65
<u>Sum Squares</u>	<u>238.95</u>	<u>290.55</u>	<u>182.80</u>	<u>296.55</u>
d.f	19	19	19	19
Variance	12.58	15.29	9.62	15.61

Think about some risky behavior, something where there is a chance that things will go wrong that is totally beyond your control (a random event). Now imagine engaging in this risky behavior several times. How does the chance that something will go wrong on a specific event relate to the chance that something will go wrong anywhere within the set of risky events? Clearly the chance that something goes wrong somewhere within the complete set of events is much greater than the chance on one specific event. On one event you may escape fine but if you keep doing whatever it is that you are doing, you will greatly increase your chance of a problem. In statistics we call the chance that you will make an error on a specific decision the “Type I error rate per comparison” and the chance that you

will make an error anywhere within the set of questions the “Type I error rate experimentwise”. When we say that we are willing to tolerate a 5% chance of a Type I error in a multiple question experiment we have to specify which of these two types of type I error we mean. If we mean that each separate question has a 5% chance of an error then the Experimentwise rate will be quite large. On the other hand if we have only a 5% chance that an error will occur anywhere within the multiple questions we ask then we will have to have a very small chance of error on each question, a very small per comparison rate. One of the major moderating factors that effects the relative size of the per comparison and experimentwise error rates will be the number of questions that are asked.

Types of multiple contrasts

How shall the experimenter in the present example explore the difference in the four different approaches? If we polled a number of different experts in the field they would suggest a number of different approaches. The differences between the approaches in part reflect differing levels of theoretical underpinnings for the study. Here are three different possibilities.

Planned Contrasts: The experimenter may have a small number of theoretically relevant issues that the experiment was designed to address. In this study, for instance, I made up the treatment groups with three specific questions in mind that I believe would address the issues present. The first question deals with the difference between curriculum that have a specific sequence as opposed to those that have an integrated approach. Thus I would be interested in comparing the average of the first two treatment groups to the average of the last two groups (those introduced in this chapter). This comparison would provide information about this basic distinction between mathematical curriculum approaches. Secondly I would be interested in the relative efficacy of the two sequences independent of any comparison with integrated groups. This question is exactly the same as was developed in chapters 6 and 7. The only difference is that now it is enveloped in a larger experiment that has additional questions. The last question would be about the difference between the moderate and radically integrated programs. Like the second question this one ignores some of the groups but we remember that those that were ignored were considered in other questions. While there are lots of additional questions that might be asked, these three would be all that would be of interest in this approach to the analysis.

Pairwise Contrasts: This approach is quite different from the previous. In it the experimenter has no specific questions posed but rather wishes to know about the differences between each pair of means. A test of significance is to be performed for each

pair of means to determine if they are statistically significantly different. In general there are $k(k-1)/2$ pairs of means where k is the number of treatments. In the present example there are six different pairs and thus there would be six separate tests of significance. In this approach there is no interest in combining groups such as we did in question one of the planned contrasts.

Any and All Contrasts: In this approach the experimenter wants to wait until the results have been perused to determine what questions seem to explain the results. It is rather like the game show Jeopardy where the answers are given and the experimenter then decides on the questions to ask. Combinations of treatment groups as well as pairwise contrasts are allowed. Clearly theory has little to do with the questions asked.

Regardless of overall strategy, every question asked could be formulated as a t -statistic similar to those that we constructed in chapter 7. In Formula 9.1 we have used Formula 7.7 as the model to construct the standard error of the difference between means and Formula 7.9 as the model for the t -statistic. The combination of the two formulas have four major

parts: $(\bar{X}_1 - \bar{X}_2)$ the observed difference between the means, $(\mu_1 - \mu_2)$ the null hypothesis, S_p^2 the pooled sample variances, and n_1 and n_2 the sample sizes. In adapting these formulas for general usage minor modifications will be required.

$$t = \frac{(\bar{X}_a - \bar{X}_b) - (\mu_a - \mu_b)}{\sqrt{MS_w \left(\frac{1}{n_a} + \frac{1}{n_b} \right)}} \quad 9.1$$

The sample means being compared may be straightforward or complex. If the specific question of interest were the difference between treatment group 1 and 2 then the sample means used would be simply those for treatment groups 1 and 2. If the question is about combined groups, the means will have to be generated from the existing information. For example, if we were interested in testing whether the average of the first two treatment groups (those with set sequences) differed from the last two treatment groups (those with an integrated structure) we would have to generate both means to be compared. The first mean would be obtained by adding all of the scores for subjects in either group 1 or 2 together and dividing by the combined sample size. Likewise the second mean would be obtained by combining the information from treatments 3 and 4. Think about how you would find the means to be compared if the question of interest had been formulated as the difference between the old sequence and the average of all of the new treatments. In this question one of the means is simply that for Treatment group 1 while the second group would be

obtained by summing all of the scores in all three “new” programs and dividing by the combined n.

In all tests that we conduct on specific contrasts the null hypothesis is that the real underlying difference between the two means being compared is zero. Thus the null hypothesis is that $\mu_a - \mu_b = 0$. In the test of sequenced methods (Treatment groups 1 and 2) versus integrated methods (treatment groups 3 and 4), the null hypothesis is that the underlying population mean for sequenced treatments is the same as the underlying population mean for integrated treatments. Note that this does not mean that all four groups are the same. Rather that when the groups are combined as described the resulting averages are the same. This null hypothesis should be retained even if the means were as varied as $\mu_1 = 10$, $\mu_2 = 20$, $\mu_3 = 13$, and $\mu_4 = 17$ because the average of μ_1 and μ_2 $(10+20)/2 = 15$ is equal to the average of μ_3 and μ_4 $(13+17)/2 = 15$. Note that other questions asked about the groups would likely reflect the difference in the four hypothetical population means.

In this presentation we will assume that the variances within the treatment groups are homogeneous. Remember from Chapter 7 that when the variances are assumed to estimate a common underlying population variance they can be combined to provide a single, best estimate of that variance. Instead of having only two groups to pool together to find an estimate of the variability of scores we have k different groups, each of which provides an independent estimate of the variability of scores within the population. Since we are willing to assume these variance estimates are all homogeneous we simply pool the information together. The sum of squares from each group is combined into a accumulated Sum of Squared Deviations from Within Groups and the degrees of freedom are accumulated to form the Degrees of Freedom from Within Groups. The variance estimate is obtained by doing the single division of the Sum of Squared Deviations from Within Groups by the Degrees of Freedom Within Groups. The resulting estimate of the population variance that we will use is going to undergo a name change. As we get into more advanced statistics we find many things are called variance and the term has a specific, somewhat restrictive meaning. So we are going to substitute a more neutral term instead. If we think back to our introduction to the variance we defined it as a sum of all the squared deviations divided by the number of free deviations that were present (called the degrees of freedom). In normal usage we call a sum that has been divided by how many things were summed, a mean. We are dealing with squared deviations so when we take the Sum of Squared Deviations and divide it by the Degrees of Freedom we should legitimately be able to call the resulting term a Mean Squared Deviation. We shorten the phrase slightly to Mean Square and add a notation of the source of the information. In the present example the information comes

from subject differences from within the samples. This is shortened to Mean Square Within. The interpretation of the Mean Square Within is identical to that of the pooled variance.

The last terms we need to consider are sample sizes. In general we will assume that the number of individual within each treatment group is the same. All of our considerations become more complex though not impossible if we have unequal ns. The sample means that we are comparing may be composites of several groups and thus the number of subjects for one of our generated means may be some multiple of the n for a single group. In the comparison of Sequenced Groups versus Integrated Groups each of the means that are compared would be based on two separate groups of subjects. Since each group had 20 subjects both n_a and n_b would be 40. In the question of whether the old sequence differed from the average of the three new treatments n_a would be 20 while n_b would be 60.

Three Examples

To see how we would find the t-statistic for tests of significance of increasing complexity let's work three examples on the current data. The first will simply test the null hypothesis that $\mu_1 - \mu_2 = 0$; the second $(\mu_1 + \mu_2)/2 - (\mu_3 + \mu_4)/2 = 0$; and the last that $\mu_1 - (\mu_2 + \mu_3 + \mu_4)/3 = 0$. These are tests of the group 1 versus group 2; the average of the first two versus the average of the last two; and the difference between the first group and the average of the last three groups, respectively.

For all three questions we will need to find the Mean Square Within (MS_w). The formula for MS_w is:

$$MS_w = SS_w / df_w = \frac{\sum(x_1 - \bar{x}_1)^2 + \sum(x_2 - \bar{x}_2)^2 + \dots + \sum(x_k - \bar{x}_k)^2}{(n_1 - 1) + (n_2 - 1) + \dots + (n_k - 1)}$$

With equal ns the denominator reduces down to $k(n-1)$. From Table 9.1 we obtain the following values from the last step in finding the separate variances to plug into the formula for MS_w .

$$\begin{aligned} MS_w &= (238.95 + 290.55 + 182.80 + 296.55) / (19 + 19 + 19 + 19) \\ &= 1008.85 / 76 \\ &= 13.27 \end{aligned}$$

If you compare this value to the four separate variances you will find that it is simply the average of the four separate variances. This will be true whenever the ns are equal. With unequal ns the value will be shifted off slightly toward the variances that are based on the greater number of degrees of freedom.

Example 1 $\bar{X}_a = \bar{X}_1 = 26.95$ and $\bar{X}_b = \bar{X}_2 = 30.15$ $n_a = n_1 = 20$ and $n_b = n_2 = 20$

The t-statistic would be

$$t = [(26.95 - 30.15) - (0)] / \sqrt{13.27(1/20 + 1/20)}$$

$$t = -3.2 / 1.152$$

$t = -2.78$ which would be from the t distribution with 76 degrees of freedom if the null hypothesis is true. In all of the remaining formulas the -0 is omitted.

Example 2 $\bar{X}_a = (\bar{X}_1 + \bar{X}_2) / 2 = (26.95 + 30.15) / 2 = 28.55$ and $\bar{X}_b = (\bar{X}_3 + \bar{X}_4) / 2 = (30.60 + 29.65) / 2 = 30.125$. $n_a = n_1 + n_2 = 40$ and $n_b = n_3 + n_4 = 40$

$$t = (28.55 - 30.125) / \sqrt{13.27(1/40 + 1/40)}$$

$$t = -1.575 / 0.815$$

$t = -1.93$ which would be from the t distribution with 76 degrees of freedom if the null hypothesis is true.

Example 3 $\bar{X}_a = \bar{X}_1 = 26.95$ and $\bar{X}_b = (\bar{X}_2 + \bar{X}_3 + \bar{X}_4) / 3 = (30.15 + 30.60 + 29.65) / 3 = 30.13$. $n_a = n_1 = 20$ and $n_b = n_1 + n_2 + n_3 = 60$.

$$t = (26.95 - 30.13) / \sqrt{13.27(1/20 + 1/60)}$$

$$t = -3.18 / 0.94$$

$t = -3.39$ which would be from the t distribution with 76 degrees of freedom if the null hypothesis is true.

Notice that we have only calculated the values of t. We have not made any decisions about whether the differences are significant or not. To do that we have to talk about what kind of control we want to exercise over type I error. The hardest part of this topic is coming up with the appropriate critical values.

Remember from Chapters 6 and 7 that control over Type I error is one of the primary concerns of statistics. The threat of meaningless information being shared as important dictates that we require differences to be believable before being declared real. Because of this concern we will generally adopt Experimentwise control over Type I error. Thus when we say the type I error rate is 5% we mean that even if all treatments come from a common population, for the set of questions asked in the experiment the chance that we will falsely reject any of the questions when we analyze the data is 5%. This control is based on all of the questions being asked having true null hypotheses.

The simplest way to conceptualize all the contrast null hypotheses being true is to think of all the groups as having been drawn from the same population. If this were true

any contrast you could come up with would represent a true null hypothesis (since everything is from the same underlying population). This overall statement about the true population means is called the omnibus null hypothesis since it covers all of the more specific contrast null hypotheses. It is represented by the formula $\mu_1=\mu_2=\dots=\mu_k$. If the omnibus null hypothesis is true then all of the specific null hypotheses are also true. On the other hand if the omnibus null hypothesis is false it does not mean that all of the specific null hypotheses are false. If any of the values of μ is different than the rest the omnibus null hypothesis would be false. For example, imagine that the truth is that $\mu_1=\mu_2=\mu_3 \neq \mu_4$. Our specific test of $\mu_3-\mu_4=0$ would be false but null hypotheses such as $\mu_1-\mu_2=0$ would still be true. Our control of experimentwise Type I error is achieved by adopting the worst possible case for making Type I errors and then adjusting the error rate so that even in that worst case we would have no more than 5% chance of making a Type I error. Because we have a greater chance of making a Type I error when we make many decisions where a Type I error is possible, the worst case (the case most likely to lead to a Type I error) would be if all of the contrast null hypotheses were true. If we can adjust the error rates for each comparison so that when all of the tested null hypotheses were true the chance that any of them would actually result in an error is 5%, then we will have more than adequate coverage should only a subset of the tested hypotheses be true.

The most straightforward way to control Type I error rate at some specific alpha level such as .05 is to use a mathematical inequality called Bonferroni's inequality. The inequality states that the experimentwise error rate will always be less than the sum of the contrast error rates. If there are three contrasts and each is conducted so that the per comparison error rate is $.05/3=.0167$ then Bonferroni's inequality states that the experimentwise error rate (EW) will be: $EW < .0167 + .0167 + .0167$. What this requires is that we be able to generate the t-value with the required degrees of freedom such that only .0167 of the total area exceed the absolute value of t. In our t-table the columns we wondered about earlier now make more sense. The .025, .0167, .0125, .01 and .0083 values are exactly what we will need if we divide .05 into 2, 3, 4, 5 and 6 parts, respectively. Thus the required critical value with three (3) contrasts to maintain Type I error rate at less than .05 would be the value in the column .0167 (two tailed) for the appropriate degrees of freedom. In our problem there are 76 degrees of freedom since our table doesn't have 76 df we would use the next smaller number of degrees which is 60. The t-value for 3 contrasts with an experimentwise error rate of .05 based on Bonferroni's inequality would be 2.463. In order to be declared significant the absolute value of the calculated t would have to exceed 2.463. If the null hypothesis for a contrast is true only .0167 of the t values

will exceed 2.463. Being this rigorous on each contrast is what keeps the experimentwise error rate down to less than .05.

Returning to our example let us calculate the three t-values for the three contrasts outlined in the section on planned contrasts. We already did the first and second contrasts in the example problem above. The third question we asked in the earlier discussion was the difference between the moderately integrated curriculum and the radically integrated curriculum, Groups 3 and 4. Following the same strategy presented above I calculate the t-value as $t=(30.60-29.65)/1.152=.82$. Note that the standard error is the same as for the other question involving the difference between two of the original groups. Thus we have:

$$H_0: (\mu_1 + \mu_2)/2 - (\mu_3 + \mu_4)/2 = 0 \quad t = -1.93 \text{ less than } 2.463 \text{ so retain } H_0$$

$$H_0: \mu_1 - \mu_2 = 0 \quad t = -2.78 \text{ greater than } 2.463 \text{ so reject } H_0$$

$$H_0: \mu_3 - \mu_4 = 0 \quad t = 0.82 \text{ less than } 2.463 \text{ so retain } H_0.$$

The result indicate that the difference between the original sequence and the revised sequence are statistically, significantly different. The new sequence is superior to the old sequence. The other two null hypotheses tested did not differ sufficiently to result in the difference being declared statistically significant.

The questions regarding pairwise differences that is described above is usually tested using a test which is specifically designed for these types of test but beyond the level that we have time for in this course. There are several of these special techniques. Some of the more popular are called Tukey's test, REGWA, and Newman-Keuls. We could also use Bonferroni's inequality if we wanted to. The advantage of the special techniques is that the critical value that is required is smaller with these techniques that with Bonferroni. Thus we would not have to obtain as big a difference in order to reject the null hypothesis. This means we would have slightly more power. Rather than introduce an entirely new approach we will analyze the present pairwise contrasts using Bonferroni's inequality. Remember that there are four groups so there are 6 different pairs of means to be tested. This means that we would conduct each test at the $.05/6=.0083$ alpha level. Again we will use 60 degrees of freedom. The required critical value is 2.729. (The corresponding critical value for Tukey's test is 2.64.) In order to reject the null hypothesis associated with any difference the value of t must be one of the most extreme .00833 which starts at an absolute value of 2.729. As before when we tested a null hypothesis involving just two of the original groups the standard error of the difference between means will be 1.152 (note: see first example problem to check this value). The six mean differences and associated t values are:

Pair of Means	Difference	t	decision
1 versus 2	-3.20	-2.78	reject
1 versus 3	-3.65	-3.17	reject
1 versus 4	-2.70	-2.34	retain
2 versus 3	-0.45	-0.39	retain
2 versus 4	0.50	0.43	retain
3 versus 4	0.95	0.82	retain.

The analyses indicate that two of the new treatment groups, the revised sequence of Treatment 2 and the moderately integrated program of Treatment 3, are statistically significantly different from the original sequence of topics but that none of the new treatments are sufficiently different from one another to reject the respective null hypothesis.

Bonferroni's inequality is not useful if the experimenter plans to wait until the data are collected before formulating the questions. In essence the experimenter using this strategy has an infinite (or at least very large) set of possible questions. Bonferroni's inequality would handle this problem by dividing .05 by infinity. We would end up arguing that no difference could be big enough to convince us there was a significant difference. Clearly some other approach is needed. Again we will only allude to the answer because it is really beyond the scope of this course. A test called Scheffe's test, in honor of its developer, can be used to maintain type I error experimentwise even when you ask questions after looking at the data and mold the question to what you have seen there. It achieves this control by demanding a quite large test statistic before you are allowed to declare something significant, but at least it is not infinity. In the present study the critical value is 2.88. Before a difference can be declared significant it must have a t-value greater than 2.88.

None of the differences we have considered in the last two sections would be declared significant if the experimenter had wanted to wait until the data were in before thinking about the analysis. However, one major comparison would reach significance and that is the comparison of the old sequence to the sum of all the new treatments. This t-value was calculated in example 3 above and has a t of -3.39. In addition the test of the old sequence versus the average of the two integrated programs has a t of -3.18 and would be declared significant. Other contrasts of treatment one versus some combination of treatments 2, 3, and 4 would also produce t values greater than 2.88 but the experimenter might be satisfied with the two that we have calculated. They suggest that in general the

present sequence (treatment 1) is inferior to innovative treatments, and that integrated strategies are significantly better than the existing sequence.

A course in experimental design such as EDPSY 593 at the University of Washington Educational Psychology program is really the appropriate place to learn about obtaining the critical values presented above. They are presented here simply to let you know that there exist several options for how the information in an experiment might be packaged for consumption.

Omnibus Tests

A common analysis performed in addition to the specific questions described so far in this chapter is a test of the omnibus null hypothesis. Remember that the omnibus null hypothesis is the general hypothesis that $\mu_1 = \mu_2 = \dots = \mu_k$. It is the hypothesis that all treatment groups came from a common population, none of the treatments had a differential effect. Clearly if any test of a specific questions null hypothesis is rejected then the omnibus null hypothesis is not true. What then could be gained from testing the omnibus null hypothesis separately? The answer becomes simpler if we reverse the order in which we think of performing the tests. The common strategy is to first address the omnibus null hypothesis. We will develop a strategy to provide a test that will incorrectly reject the omnibus null hypothesis only .05 of the time (when the null hypothesis is completely true). If we reject this null hypothesis then we have the belief that somewhere within the set of means there are some effects to be found. We have reason to do the more specific test of significance. Even better we have a simple understanding of the Type I error rate associated with this omnibus test since it is a single test of the combined differences that might exist within the treatments. There is no confusion about whether this applies to a single decision or to the whole set since there is only a single decision (and it is about the whole set). If all the treatments have the same effect only 5% of the time will we see differences in the means that are so large that they make us believe that treatments had differential effects.

If the omnibus test is retained we conclude that the differences between the means is not of sufficient magnitude to convince us that there were differential treatment effects. The natural outcome of this conclusion is that we are finished with the analysis. There is nothing to be found by more specific analyses. Thus the omnibus test is sometimes used as a gateway to prevent or invite the asking of more specific questions. A rejected omnibus null hypothesis invites additional analyses because we believe that differences exist in the set while a retained omnibus null hypothesis indicates that the evidence does not support differential treatment effects. Strategies that use the omnibus test in this manner are called protected tests. The omnibus test protects against increased Type I error rate. Most of the

techniques for performing specific analyses already adequately control type I error to 5% experimentwise and thus do not need to be “protected”. But, because of the ubiquitous nature of the omnibus test we will develop the strategy by which it works.

Test of the Omnibus Null Hypothesis

The omnibus null hypothesis is that $\mu_1 = \mu_2 = \dots = \mu_k = \mu$. If this hypothesis is true then the sample means $\bar{X}_1, \bar{X}_2, \dots, \bar{X}_k$, are all estimates of the same underlying population mean.

The sample means would not all be the same but the variability of the sample means would simply be due to random variability. Way back in chapter 3 around formula 3.3 we first learned about how sample means differ from one another when drawn from a common population. We found that how much sample means vary is simply a reflection of how much the scores themselves differ from one another, and how many people a mean was based upon. Formula 3.3 was

based upon. Formula 3.3 was $\sigma_{\bar{x}}^2 = E(\bar{X} - \mu)^2 = \frac{\sigma^2}{n}$ and a comparable formula

could be written with sample estimate $s_{\bar{x}}^2 = \frac{S^2}{n}$ Throughout chapters 3 to the present

chapter we have talked about how much sample means will vary ($s_{\bar{x}}^2$) based on how much the scores differ from one another (s^2). Now we want to look at it from a different perspective and acknowledge that in the present situation, if the null hypothesis is true we have a set of sample means that randomly differ from one another. If we want an estimate

of $\sigma_{\bar{x}}^2$ we can get it by simply calculating how much the k different sample means we have observed differ from one another. As absurdly simple as it sound, we can estimate how much sample means vary from one another by looking to see how much a collection of sample means vary from one another. We could calculate the variance of a set of sample means by the same rules that we used for finding the variance of scores. The only difference would be that each of the numbers inserted would not be an original observation but a mean score based on n observations. While simpler computational formulas exist the required formula for the variance of sample means could be written as:

$$s_{\bar{x}}^2 = \frac{\sum (\bar{X}_k - \bar{X})^2}{(k-1)} \quad 9.2$$

where \bar{X}_k is one of the set of sample means and \bar{X} is the average of all of the sample means. The number that is obtained is an estimate of the variance of sample means with k-

1 degrees of freedom. That is, the estimate is based on k values and operates about the mean of the values used, thus $k-1$ degrees of freedom.

The actual test of significance for the omnibus null hypothesis introduces an entirely new theoretical distribution to add to the normal, t , and χ^2 that we have met thus far. This is the F distribution. The F distribution describes the variability of the ratio of two estimates of a common population variance. For instance, if you took two samples from a common population and found the variances of the scores within each sample these variances would not be exactly the same. Because they are based on the specific scores in that sample they would randomly differ from one another. But, because they estimate the same parameter they shouldn't be too different. We find the ratio of the two variances. This ratio should be somewhat near to 1.0 since both numerator and denominator estimate the same parameter but sometimes the first variance might be a little bigger and sometime a little smaller. The distribution of this ratio about its modal value of 1 is what is described by F . The distribution of the ratio will depend in part on how stable each of the estimates in the ratio is. If the numerator or denominator is very unstable because it is based on very little information then from one pair of samples to the next the ratio could change greatly. On the other hand if both variances are based on many degrees of freedom (lots of information) then the variances themselves are very good estimates of the population variance and the ratio should be very close to 1.0. Thus when we use the F distribution we will have no consider the number of degrees of freedom on which the numerator and denominator are based. Different F distributions will be appropriate depending on the number of degrees of freedom. This is closely related to the changes in t as the number of degrees of freedom changed.

The above paragraph seems far afield from the problem we have been considering in the rest of the chapter. We are worried about the means of k different groups while the paragraph concerned looking at the distribution of two variance estimates when they both estimate a common population variance. The trick is to work our problem until it fits the information required of F . First, lets remember that the MS_w is the pooled estimate of the population variance. It is the best overall estimate of how much scores randomly vary from one another. It is an estimate of σ^2 based on $k(n-1)$ degrees of freedom. If the treatment have no effect the MS_w estimates random variability and when the treatments do have an effect, if it is an additive effect, the MS_w still estimates σ^2 . When we construct our F test the denominator will be the MS_w which clearly is an estimate of random variability.

Now we return to our discussion of the variability of sample means. Formula 3.3 and the comparable formula using S^2 in place of σ^2 indicate a relationship between the

random variability of sample means and the random variability of the scores on which the samples were defined. Sample means vary one- n th as much as scores vary or, conversely, scores vary n times more than sample means. The connector between the two values is “ n ”. If I have an estimate of the variability of scores and want an estimate of the variability of means I divide the score variance by n . If I have an estimate of the variability of means and I want an estimate of the variability of scores I multiply by n . In equation 9.2 we have obtained an description of the variance of sample means from the set of k means. To convert this to a description of the variance of scores we only have to multiple our variance of sample means by the number of subjects in a sample. There is, however, an important contingency here if we are to talk about these descriptions as being random variances. The argument just mounted will provide estimates of random variability only if the differences between the means within our set reflecting only random differences, that is, that the omnibus null hypothesis be true. If the omnibus null hypothesis is false indicating that some treatments are superior to others, then the differences among the sample means do not reflect just the random difference in the population but also reflects the treatment differences. We call the description obtained by multiplying the variability of means by n the MS_t where the t stands for treatment. The earlier change in terminology from variance to Mean Squares is particularly important for MS_t . If the null hypothesis is true it would be appropriate to use the phrase *variance* in talking about this value since, when H_0 is true the MS_t does just describe random variability. However, when the null hypothesis is false and some treatments systematically differ from others, it would be inappropriate to use the word variance as now there are systematic as well as random factors that influence the size of MS_t . The phrase Mean Square for Treatments is a simple description of the source of the information. MS_t is based on the mean of the squared deviations for the treatment groups. It does not beg the question of whether the null hypothesis is true or false. It is appropriate in either case.

The MS_t will be used in the numerator of the F statistic that we calculate making it:

$$F = \frac{MS_t}{MS_w}$$

If the omnibus null hypothesis is true then the numerator, MS_t estimates σ^2 and the denominator also estimates σ^2 . Thus the expected distribution of the F ratio calculated should be the F distribution with $k-1$ and $k(n-1)$ degrees of freedom. Or, more simply, if the omnibus null hypothesis is true then both MS_t and MS_w estimate random variability and ought to be fairly comparable. How far from 1.0 the ratio might get would be described by the F distribution. On the other hand, if the omnibus null hypothesis is false then the MS_w

still estimates σ^2 but MS_t reflects both random variability (σ^2) and the real treatment differences. Thus MS_t should be larger than the MS_w . If the omnibus null hypothesis is false then the F ratio should not be around 1.0 but should be greater than 1.0. How much greater it will be depends on how much the treatments changed scores, the number of people on which the means are based and the luck of the draw. The trick, of course, is in deciding whether an observed F is simply one of the events that randomly happened when the null hypothesis is true, or whether it reflects a systematically larger numerator, which, in turn, means real treatment effects. This is the exact same dilemma we had with any test of significance and our rules will be identical.

The first rule is that the null hypothesis is kept until it is overwhelmed by the data. If the value observed is at all reasonable when we examine what would happen with a true null hypothesis then we will retain the null hypothesis. In order to overwhelm the null hypothesis the observed F ratio must be so large that it almost never would get that large if the null hypothesis were true. The level set for “almost never” is the rate at which we are willing to make a Type I error, our alpha level. In this problem then we will retain the null hypothesis unless the Mean Square that came from differences between the *treatments* is so much larger than the Mean Square that came from *within* the groups that it would occur only 5% of the time when the null hypothesis is true. In the example that we have been working throughout this chapter the Mean Square Treatments (MS_t) is based on four means and thus has $4-1=3$ degrees of freedom. The Mean Square Within (MS_w) is based on four different sample variances each of which has 19 degrees of freedom, a total of 76 degrees of freedom. The F ratio will be the description of what the ratio between variances behaves like when the numerator has 3 df and the denominator 76 df. The theoretical distribution of F ratios from an F distribution with 3 and 76 degrees of freedom is shown in Figure 9.1. This figure comes to us from statistical theory much as the z and t distribution were provided for us. We note that the shape of the curve is a positive or right skewed distribution with an elongated tail extending toward larger values. If the null hypothesis for the experiment we have been considering is true then the F ratio we calculate will simply be one of the values that comes up randomly from this distribution. However, if the null hypothesis is false then we would expect that the F ratio is larger than that expected randomly from this distribution. How large? Well, it has to be so large that it doesn't look like the sort of number that comes from the curve drawn. Actually we are not so demanding that it has to be totally outside the range of values that occur in Figure 9.1 but it has to at least be so large that only 5% of the F distribution pictured has values of that magnitude. If it turns out that the null hypothesis is really true and we just happened to get very unusual

sample such that the calculated F ratio is one of these top 5% then we will be misled. We will reject the null hypothesis incorrectly, we will make a Type I error. But, this is the same story that we heard in chapters 5, 6, and 7. It is the risk we must take if we are ever going to be able to detect real differences. In the curve drawn below it an F-ratio of 2.725 marks the point where the top 5% of the F distribution begins. We would require that the observed F-ratio be 2.725 or larger before we will reject the null hypothesis.

Figure 9.1 F distribution with 3 and 76 degrees of freedom

0 .5 1.0 1.5 2.0 2.5 3.0 3.5 4.0

For the example problem the grand mean is 29.34. The squared differences of the four means from the grand mean are:

$$(30.15-29.34)^2+(26.95-29.34)^2+(30.60-29.34)^2+(29.65-29.34)^2=8.0519$$

and $MS_t = (20)(8.0519)/(4-1)=53.68$.

From early in the chapter we had $MS_w=13.274$ so our observed F ratio is:

$$F=53.68/13.274=4.04$$

where the numerator is based on 3 df and the denominator on $4(19)=76$ df. While we know what the critical value is from Figure 9.1, in applied problems we would now have to consult the F table in the text to obtain a critical value. The table is set up so that we find the degrees of freedom in the numerator by going to the appropriate column and then go to the row with the denominator degrees of freedom. Again we have to use 60 since 76 is not one of the rows listed. The critical value with 3 and 60 degrees of freedom is 2.76 which is only slightly greater than the value of 2.725 which we found when we used the actual number of degrees of freedom. Space in tables makes this slight overstatement necessary. Our observed F-ratio of 4.04 is larger than the critical value of 2.76 so we reject the null hypothesis. The variability present in the sample means is too large for us to believe that it

is just due to random differences. The means vary so much that it suggests that the treatments really did change people performance.

Some researchers use F tests when k is equal to 2. It works fine producing an F with $k-1=1$ degree of freedom. We learned in Chapter 7 that with only two groups we could use the two sample t-test to test for differences. If you analyzed the same data using the analysis of variance F test described here you would reach exactly the same conclusion as with the t test of Chapter 7. In fact the calculated value of F would be exactly the square of the t value calculated ($F=t^2$). Whenever you see an F reported with 1 degree of freedom in the numerator you should think of the information provided as identical to what would have been provided with a t-test.

A rejected null hypothesis when $k>2$ does not say anything specific about which treatments had different effects. The null hypothesis tested is that $\mu_1=\mu_2=\dots\mu_k$. If we reject this hypothesis we only know that not all of the signs are “equal” signs. One or more of these should be “unequal” signs. We know that the groups appear to differ. This is when we must return to the topics covered at the beginning of this chapter and ask specific questions about the treatments. What questions we ask depends on how we have conceptualized the problem at hand. We may have planned contrasts, do all possible pairs, or simply explore the data. Thus we have come full circle back to multiple contrasts. Is it necessary to do the omnibus test as well as specific multiple contrasts? Probably not, almost all of our multiple contrast procedures already control type I error to 5% so inserting the omnibus test as an added condition will make the test have an effective type I error rate of even less than 5% which will reduce your power to detect treatment effects. In reading research you will constantly come upon F tests like we have described in this section so, whether you believe they are needed or not, it is important to understand their rationale.