CHAPTER 22

INTRODUCTION TO ANALYSIS OF VARIANCE

Chapter 18 on inferences about population means illustrated two hypothesis testing situations: for one population mean and for the difference between two population means. Of course, the fact that I stopped at just two means was simply for my convenience. We could have extended the examples to situations involving more than two means. For example, what if you did an experiment where there were 3 different conditions and each group or condition produced a sample mean. How do the differences in the 3 sample means reflect on the differences amongst the three population means? The reason why I have waited until now to extend the examples of hypothesis tests for more than two population means is that I wanted to introduce you to "analysis of variance". Analysis of variance (**ANOVA**) is a versatile procedure in statistical analysis that will handle the case of "two **OR MORE** means". In fact, analysis of variance will do much more but we will restrict ourselves to cases of simple extensions of multiple population means. For other applications of analysis of variance, please see more advanced statistics textbooks.

What if you conducted a study comparing three different methods of teaching statistics: Programmed, Computer, and Control. The Programmed method involves programmed instruction, the Computer method involves using computer delivery of instructional materials, and the Control method is the old lecture-discussion method. In doing your study, you select 30 students at random from the population of college students and then **randomly assign** 10 to each of the 3 methods. You conduct the study, collect data on a criterion posttest, and the data look something like the following.

DATA FROM A 3 GROUP EXPERIMENT

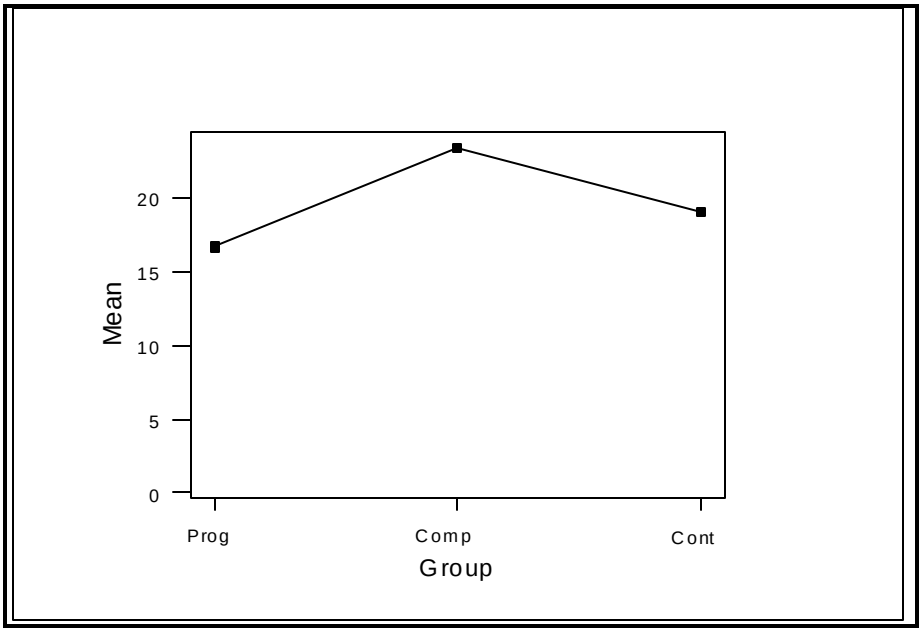
Prog	Comp	Cont
16	30	18
21	19	10
17	25	12
10	20	29
10	30	24
25	18	28
12	13	14
25	28	10
20	24	27
11	27	18

Some descriptive statistics are shown at the top of the next page.

N	MEAN	STDEV

Prog	10	16.70	5.89
Comp	10	23.40	5.70
Cont	10	19.00	7.51

A simple graph of the data to show possible differences in the means of the three groups would look like the following. On the surface, the data from the experiment indicate that the Computer method seems to produce the highest level of performance, the Control method the next best, and the Programmed method the lowest. But, since this is simply one of many different experiments that could have been done in the same way, the question is: would other experiments similar to this produce the same pattern of results? That is, if I randomly select 30 more students, randomly assign 10 to each method, redo the study, and then analyze the data, will I find similar results?



In all other inferential situations that we have examined, we have encountered the concept of sampling error. From sample to sample, the statistics will vary around some parameter value or parameter values. In the present case, we actually have 3 parameters to deal with: population mean for Programmed method, population mean for Computer method, and population mean for Control method. Shown as a null or H(0) hypothesis, we could state it as follows.

$$H(0): \mu_{Prog} = \mu_{Comp} = \mu_{Cont}$$

What if we actually had the situation where all of these methods were equally

effective (or equally ineffective if you want to view it from that perspective!). That means that the means from all 3 populations are equal. So, theoretically speaking, when we sample from each of the populations and look at the same means for each sample, all three sample means should be the same too. But, because of sampling error, the means will vary from sample to sample even though the population means are identical. The factors that will impact on the amount of sampling error are the same ones we saw previously for sample means: sample size and population variability. With smaller sample sizes, there will be more sampling error just as there will be more sampling error if the populations from which the samples are drawn are very heterogeneous. Without spending more time on this fact, it will be necessary however to explore the concept of variation more before we can continue with our discussion of analysis of variance.

Partitioning Sources of Variation

At the heart of analysis of variance is the concept that we can partition or subdivide the variability into several parts. This is very similar to the discussion of partitioning criterion variance into predictable and error parts in Chapter 11. You might want to go back and quickly review that material; there, we partitioned into a good or predictable part, and a bad or error part. Something of a similar nature occurs here too. In the example above with 3 conditions or groups, and 10 students in each group, there will be three parts or ways in which we can subdivide the overall variability. To show you this, I have created several demo sets of data, with only 3 numbers in each of 3 groups. Look at the first set of data. To properly develop the concepts of analysis of variance, we need not only the data from the 3 separate groups but also a newly created combination set of data where **all 3 groups have been combined into one**. I let Minitab help me do the work.

In this first case, notice that the newly stacked or combined set of data looks just like each of the separate groups, except that it is longer. Then notice that the overall "Combo" group has 0 variability. Also notice that there are no differences amongst the 3 groups. Finally, note that there is no variation within any of the separate groups either.

```
A   B   C     Combo
5   5   5       5 |
5   5   5       5 |   <------ From Group A
5   5   5       5 |
                5   |
                5   |   <------ From Group B
                5   |
                5     |
                5     |    <------ From Group C
                5     |
```

DESCRIPTIVE STATISTICS ON THE 3 GROUPS AND COMBO DATA

	N	MEAN	STDEV

Group A	3	5.0000	0.0000
Group B	3	5.0000	0.0000
Group C	3	5.0000	0.0000
Combo	9	5.0000	0.0000

The generic version of the analysis of variance model looks as follows.

VAR COMBO GROUP = VAR AMONGST GROUPS + VAR WITHIN GROUPS

The "Combo" group in analysis of variance is sometimes called the "composite" group; the one with the data from all separate groups stacked or combined together. The **TOTAL** variability of the "Combo" or composite group is subdivided or partitioned into two components: the variability reflecting differences **BETWEEN** or amongst the means across the groups, and the variability of scores around the means **WITHIN** each of the separate groups. The "Combo" group is usually called "Total", the variability amongst the groups is usually called the "Between" groups or BG, and the variability within the groups is usually called the "Within" group variability or WG. Thus, the formula can be written with a little more precision as follows.

TOT VAR = BG VAR + WG VAR

Inorder to attach specific quantities to each of these components, we need a method by which to calculate the "variability" values. To do this, we will use the numerator part from the variance formula. This is called the sums of squares value (SS) since the numerator of the variance formula represents the sum of the squared deviations around some mean. Thus, to be more specific, the variability components formula looks as below.

SS(TOT) = SS(BG) + SS(WG)

To calculate each SS value, we need to find the sum of the squared deviations around the appropriate mean. For the SS(TOT), the deviations will be of the scores in the "combo" set of data around the mean of the stacked or composite group; ie, set of data with all the separate groups combined together. In the case of the SS(BG) value, we need to look at the deviations of the **group** means around the overall **composite** mean (ie, the mean of the stacked or composite set, is sometimes called the "grand mean"). Finally, the SS(WG) value will be found by looking at the deviations of the scores within each group from the mean of each separate group. The computational formulas we need are shown and explained on the next page. .

$$SS(TOT) = \sum_{j} (X_{ij} - \overline{\overline{X}}_{c})^2$$

$$SS(BG) = \sum_{g} [n_g(\overline{\overline{X}}_{j} - \overline{\overline{X}}_{c})^2]$$

$$SS(WG) = \sum_{g} (X_{ij} - \bar{X}_{j})^2$$

For the SS(TOT), the subscript **ij** represents any value in the combined stacked data set from the "composite" column of data. The subscript **.** will be used to indicate the mean of the composite set of data or the grand mean. If you want, you can verbalize each symbol as: **X sub ij and X bar dot**. For the SS(BG), the **n sub g** represents the number of observations (n) in each group. **X bar sub j** represents the mean of each particular group: group = j = 1, etc. You have to sum these values for each group. For the SS(WG), the summation over (**sub g**) simply means that you need to use this formula for each group separately and then add together the values from each group. For our first set of "demo" data, where all the numbers are the same values of 5's, it should be apparent that there is 0 variability anyway you look at it. In the stacked set of data, the SS(TOT) component, all the deviations of the scores around the grand mean of 5 are 0 and therefore the sum of the squared deviations around the grand mean will be 0. For the deviations of the group means around the grand mean, each group mean also deviates 0 from the grand mean. Therefore, the SS(BG) value will also be 0. Finally, since all scores within each group are 5's, there will be 0 variation within each group. Therefore, the SS(WG) value will also be 0. Therefore, we have the following.

SS(TOT) = SS(BG) + SS(WG)

0 = 0 + 0

Now look at a second "demo" set of data.

D	E	F	NewCombo	
5	6	7	5	|
5	6	7	5	| <------ From Group D
5	6	7	5	|
			6	|
			6	| <------ From Group E
			6	|
			7	|
			7	| <------ From Group F
			7	|

DESCRIPTIVE STATISTICS ON SECOND DEMO SET

	N	MEAN	STDEV
Group D	3	5.0000	0.0000
Group E	3	6.0000	0.0000
Group F	3	7.0000	0.0000

NewCombo	9	6.000	0.866

Notice in this set of data, that all the data within each separate group are the same: 5's in D, 6's in E, and 7's in F. So, since all the numbers within each group are the same, what do you think the value will be for the SS(WG)? If you said 0, then you have hit the jackpot. So, in our model of the components of variability, the component of SS(WG) will = 0. But, notice that while there is 0 variability within the groups, there are differences across the groups as reflected by differences among the separate group means. How do these group means differ from the overall or grand mean of the "NewCombo" composite set of data? The first group mean differs by (5 - 6) = -1 in that it is 1 less than the overall mean. What about group E? The mean of group E is the same as the overall mean and therefore, it deviates (6 - 6) = 0 from the grand mean. Finally the mean of group F deviates (7 - 6) = 1 since it is 1 above the grand mean. Also, notice that there is some variation in the stacked or "NewCombo" set of data. So, without having actual numbers, you should see that the formula will look like the following.

$$SS(TOT) = SS(BG) + 0$$

Whatever variability there is overall in the composite set is due to differences amongst the groups and has nothing to do with variations within the groups, since there are none. To find the specific numbers for our model, we need to obtain the sum of the squared deviations around the grand mean, and the sum of the squared deviations of each group mean around the grand mean.

To find the SS(TOT) value, we need to sum up the squared deviations of the scores in the composite group from the composite or grand mean. Since the grand mean is 6, we would subtract the grand mean of 6 from each of the 9 values (5 - 6, 5 - 6, 5 - 6, 6 - 6, etc.) and then square those deviations. For this set of data, the squared deviations for the 5's would sum to 3, the squared deviations for the 6's would sum to 0, and the squared deviations for the 7's would also sum to 3. Thus, the SS(TOT) value = 6.

To obtain the SS(BG), we need to square the deviations of each group mean from the grand mean and weigh each of these totals by n in the group. Since the mean in each group stands for each and every value in the group, we need to multiply each group value by the n for that group. To do this, we would take the first group mean of 5 and subtract from it the grand mean of 6, square the deviation, and weight the result by 3, or n. In this case, the value is 3. For the second group, 0 is the difference between the group mean and the grand mean so we have 0 for this second value. For the third group, the deviation of the group mean from the grand mean is again 1, and the squared value is 1, and weighing it by n of 3 yields a value of 3. So, for the SS(BG), we would have 3 for the first group, 0 for the second, and 3 for the third group = 6.

Since the SS(WG) value is 0, we now have all our model components and they are as follows.

SS(TOT) = SS(BG) + SS(WG)

6 = 6 + 0

Again, in this example, all the total variation [SS(TOT)] is due to variations between groups [SS(BG)].

Now look at a third set of demo data. First of all, note that each group has the same set of numbers and therefore the means in each group are the same. Thus, there should not be any variation amongst the groups since the means are equal. But, there is variation within the groups and also notice that the "Combo3" composite set of data has variability.

THIRD SET OF DEMO DATA

G	H	I	Combo3	
6	6	6	6	|
5	5	5	5	| <------ From Group G
4	4	4	4	|
			6	|
			5	| <------ From Group H
			4	|
			6	|
			5	| <------ From Group I
			4	|

DESCRIPTIVE STATISTICS ON THIRD DEMO SET OF DATA

	N	MEAN	STDEV
Group G	3	5.000	1.000
Group H	3	5.000	1.000
Group I	3	5.000	1.000
Combo3	9	5.000	0.866

To obtain the SS(TOT), we would need to find the sum of the squared deviations around the grand mean. Since the grand or composite mean is 5, we would be subracting 5 from each of the 9 values, squaring the deviations, and adding them up. We would have (6 - 5) squared, and (5 - 5) squared, and (4 - 5) squared, etc. for a total of the sum of the overall squared deviations being 6. To find the SS(WG), we need to square the deviations around the mean of each group and then add together these 3 values; one for each group. Since each group has the same set of data in this case, we can do the first group and then simply add that value together 3 times. The squared deviations in Group G would be (6 - 5) squared, plus (5 - 5) squared, plus (4 - 5) squared to equal 2. But, since there are 3 identical groups (Groups H and I are the same), the value of 2 must be added 3 times = 6.

Thus, the SS(WG) is equal to 6.

Thus, we have already seen that the SS(BG) is 0 since the means of the 3 groups are identical. But, there is a SS(TOT) value of 6 and there is a SS(WG) value of 6. So, what we have for our model in this third demo case is as follows.

SS(TOT) = SS(BG) + SS(WG)

6 = 0 + 6

Now look at the fourth and final set of "demo" data.

FOURTH SET OF DEMO DATA

J	K	L	Combo4	
6	7	8	6	|
5	6	7	5	| <------ From Group J
4	5	6	4	|
			7	|
			6	| <------ From Group K
			5	|
			8	|
			7	| <------ From Group L
			6	|

DESCRIPTIVE STATISTICS ON DEMO SET 4

	N	MEAN	STDEV
Group J	3	5.000	1.000
Group K	3	6.000	1.000
Group L	3	7.000	1.000
Combo4	9	6.000	1.225

Here is what I would consider a more typical case in that there is variability amongst the groups (since the means of the groups are different), variability within the groups (since the values within each group vary), and there is overall variability in the composite or "Combo4" set of data. Thus, the normal case is that there is both variability between groups and within groups.

Again, to find the SS(TOT), we need the sum of the squared deviations around the composite mean. Since the composite or grand mean = 6, we would be finding deviations of each score from 6 like (6 - 6) squared, (5 - 6) squared, (4 - 6) squared, (7 - 6) squared, and so on. If you complete all 9 values, the SS(TOT) will be equal to 12. To obtain the SS(BG), we need the squared deviations of the group means from the grand mean, weigh each squared deviation by n, and then add them up across the 3 groups. So, we would

have: [(5 - 6) squared * 3] + [(6 - 6) squared * 3] + [(7 - 6) squared * 3] for a total of 6 for SS(BG). Finally, for the SS(WG), we need to find the sum of the squared deviations around the mean of each group and then add them together. For Group J, the deviations would be (6 - 5), (5 - 5), and (4 - 5), and if you square these and add them up, you have 2. For Group K, the same pattern occurs with 2 being the squared deviations around the mean of 6, and the same thing happens for Group L with the sum of the squared deviations around the mean of 7 being 2. Thus, the 3 groups add up to having a SS(WG) value of 6.

Therefore, what we have for this final analysis of variance variability model is as follows.

SS(TOT) = SS(BG) + SS(WG)

12 = 6 + 6

As I said before, generally speaking, some of the SS(TOT) is split between the BG term and the WG term.

Before letting you off the hook on this "partitioning of variance" idea, since it is crucial, let's try to think our way through the breakdown for the original set of data from the experiment. Remember that there were 3 instructional method groups: Programmed, Computer, and Control. Again, the descriptive statistics look as follows.

DESCRIPTIVE STATISTICS ON METHOD EXPERIMENT

	N	MEAN	STDEV
Prog	10	16.70	5.89
Comp	10	23.40	5.70
Cont	10	19.00	7.51
Combo	30	19.70	6.81

Since there are differences amongst the sample means, the SS(BG) value will not be 0. Also, from the standard deviations for each of the 3 groups (since they too are not 0's), there will be some variability for the SS(WG) term in our model. Here again we have the case where some of the SS(TOT) will be partitioned or subdivided into SS(BG) and some partitioned into SS(WG).

To find the SS(TOT), we would subtract the grand mean of 19.7 from each of the 30 values (16 - 19.7), and (21 - 19.7), and (17 - 19.7), etc. and square each of these and then add up these squared deviations for a total of 1344.3. Thus, **SS(TOT) = 1344.3.**

To find the SS(WG) value, we will need to find the squared deviations in each group and then add these 3 values together. Thus, for the first group, we would find the deviations of (16 - 16.7), and (21 - 16.7), etc. and square these for a total of 312.1. Then, for the second group, we would find deviations of (30 - 23.4), and (19 - 23.4), etc. and

square these for a total of 292.4. Finally, for the third group, we would find deviations of (18 - 19), and (10 - 19), and square these deviations and add them to a total of 508. Thus, for SS(WG), we would add together the values of 312.1 + 292.4 + 508 = 1112.5. Thus, the **SS(WG) = 1112.5.**

Therefore, the full model for this set of experimental data would be:

SS(TOT) = SS(BG) + SS(WG)

1344.3 = 231.8 + 1112.5

Now that I have made you go through each step to find the SS(TOT), the SS(BG), and the SS(WG) values, I want to show you how Minitab could calculate these values in a more efficient manner. One of the several analysis of variance commands is called "AOVO" for analysis of variance one way. The reason it is called "one way" is the type of experiment we were referring to above is one where we are manipulating one independent variable: type of instructional method. On this single independent variable, there are 3 levels that refer to the 3 kinds of methods we used. Since we have varied the type of instructional method, we are obviously interested in what impact this manipulation of instructional method had on our criterion measure (dependent variable) of performance. So, with the command aovo, we can do the appropriate analysis of variance for our experimental design which is called a one factor design. To use the aovo command in Minitab, you need to put the data in three columns. So, assume that the data are placed in columns 1 to 3 (the Prog, Comp, and Cont groups respectively). See below for the command.

MTB > **aovo c1-c3**

The output from that command is at the top of the next page.

RESULTS OF ANALYSIS OF VARIANCE AOVO COMMAND

```
SOURCE      DF       SS        MS        F        p

FACTOR       2      231.8     115.9     2.81    0.078
ERROR       27     1112.5      41.2
TOTAL       29     1344.3

                                        INDIVIDUAL 95 PCT CI'S FOR MEAN
                                        BASED ON POOLED STDEV
LEVEL     N     MEAN    STDEV  -----+---------+---------+-------
Prog     10   16.700    5.889  (-------*--------)
Comp     10   23.400    5.700                (--------*-------)
Cont     10   19.000    7.513       (-------*-------)
                                  -----+---------+---------+-------
POOLED STDEV = 6.419              15.0        20.0        25.0
```

Notice that the output consists of several pieces of information about each group like mean and standard deviation plus a confidence interval for each group for estimating where the true population mean might be. However, at the moment, the main interest I have in the aovo output is the column called SS. Note that the SS column is our sums of squares values for BG, WG, and TOT, and the values there were the ones we calculated previously. The names above of Factor, Error, and Total correspond to BG, WG, and TOT. More on this later.

So, what have we learned by going through several sets of demo data? First, the composite variability (for all groups stacked together) can be subdivided into two components: between groups and within groups. Second, we saw that all components would be 0 if all the numbers in all the data sets were identical. But, since there is usually some variability within the groups and the means are different, then some of the SS(TOT) will be partitioned to the SS(BG) term and some to the SS(WG) term. One particular principle that we will be utilizing soon is the idea that there will tend to be (relatively speaking) more variation between the groups than within the groups as the means become more and more separated. And of course, the more the means differ from one another in our experiment (taking into account sampling error), the more we should be willing to reject the null hypothesis of equal population means. But, for the moment, we need to draw this "partitioning" section to a close so we can move on to other topics.

One Factor Analysis of Variance

In simple experiments, one independent variable is manipulated such as type of teaching method, amount of dosage of drug, or amount of reinforcement in an animal learning study. There will be two or more levels on this independent variable such as 3 methods of teaching, 4 different dosage amounts, or 3 different quantities of food as reinforcement. Regardless of the number of levels there are though, it is only one independent variable the experimenter is manipulating. After manipulation of the independent variable, one collects data on some outcome measure like test scores for the teaching methods study, or reaction times for the dosage of drug study, or amounts of time to learn a task in the animal learning study. In any case, the outcome measure is called the dependent variable. In short, the purpose of a simple experiment is to see if the single independent variable has an impact on the single dependent variable. This is a cause and effect model in that if one manipulates the independent variable and sees some reliable impact on the dependent variable, it is assumed (but not proven) that it was the independent variable that caused the differences (amongst the groups) on the dependent variable. There are many aspects to experimental design that we do not have time here to discuss but, in a nutshell, the above is the basic logic and strategy. So, what about a simple experiment and the methods by which we make some decision to conclude whether or not the independent variable does impact on the dependent variable? Here is where the full analysis of variance procedures can help us.

What if we have done a study where we have developed 3 different processes for employees to assemble an electronic component. The first method involves using diagram

books, the second method involves using a helper, and the third method involves using a computer-aided design. In the experiment, 18 new employees are randomly assigned (6 each) to one of the three methods. After the study is completed, data are collected on a performance test. The data look like the following.

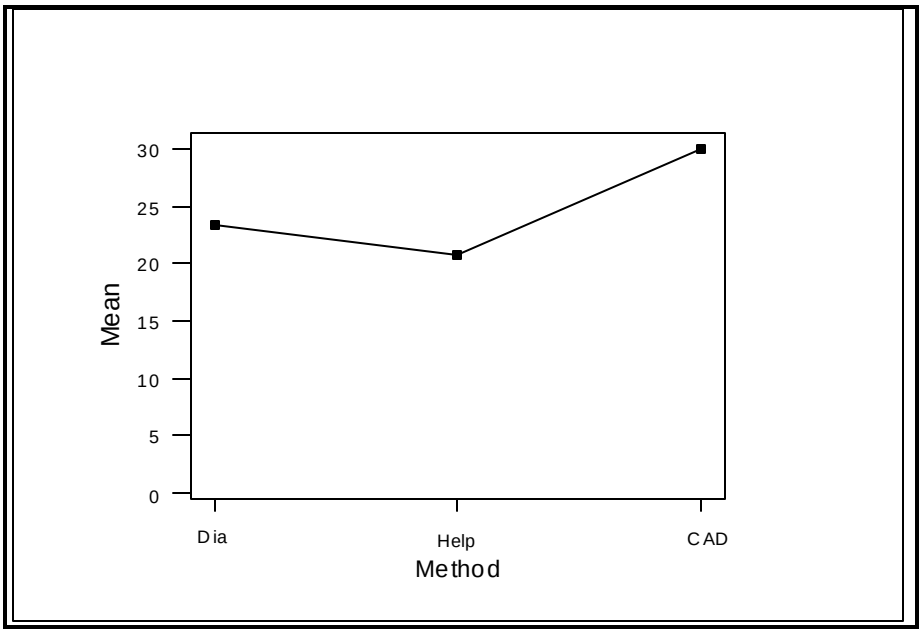
DATA FROM AN EXPERIMENT OF 3 ASSEMBLY METHODS

Dia	Help	CAD
27	24	36
25	22	32
25	21	30
24	21	28
20	20	27
19	17	27

DESCRIPTIVE STATISTICS FROM THE EXPERIMENT

	N	MEAN	STDEV
Dia	6	23.33	3.14
Help	6	20.833	2.317
CAD	6	30.00	3.52

Looking at the data, it appears that the third method produces the highest level of performance. But, are there really any true differences amongst the three methods? How likely is it that the differences we observe (with the third method looking the best) are really due to sampling error? To find out, we will do a one factor analysis of variance and then make some decision with respect to the null hypothesis. In this case, the null hypothesis is that the three methods are equally effective. In all likelihood, we are hoping to be able to reject the null hypothesis and be able to show that one method is superior to the others. A graph of the data is shown below.



In doing the ANOVA, we first need to partition our sums of squares values into the TOT, BG, and WG. Let's think about that first and then let the aovo command from Minitab help us with that. To find the SS(TOT), we would need to put all the data (18 values) in one combo group and then find the sum of the squared deviations around the grand mean. Well, it happens that the **grand mean is 24.72**. So, we would subtract and square and add up the values from (27 - 24.72) and (25 - 24.72) ... down to the last value of (27 - 24.72). This would be the SS(TOT). To find the SS(BG), we would subtract the grand mean of 24.72 from each group mean (23.33 - 24.72, etc.), square, and then weigh by n = 6 for each calculation. Finally, for the SS(WG), we would subtact the mean in each group from each score in that group (27 - 23.33, 25 - 23.33, etc.), square, and add up this total for all 3 groups. Now, look at the Minitab output from the AOVO command that will do that.

ANALYSIS OF VARIANCE OUTPUT

SOURCE	DF	SS	MS	F	p
METHOD	2	269.44	134.72	14.63	0.000
ERROR	15	138.17	9.21		
TOTAL	17	407.61			

LEVEL	N	MEAN	STDEV
Dia	6	23.333	3.141
Help	6	20.833	2.317
CAD	6	30.000	3.521

POOLED STDEV = 3.035

INDIVIDUAL 95 PCT CI'S FOR MEAN
BASED ON POOLED STDEV

```
----+---------+---------+---------
     (-----*----)
(-----*----)
                     (----*----)
----+---------+---------+---------
20.0       25.0       30.0
```

The data we concentrated on in the "partitioning" section above was the fact that the total sums of squares could be subdivided into the between groups sums of squares and the within groups sums of squares. In this case, the SS(TOT) of 407.61 is partitioned into 269.44 for the BG part and 138.17 for the WG part.

However, the procedure we are examining is called analysis of variance. The sums of squares values are just the numerator parts of the variance formula. So, what we need to do is to convert the sums of squares values (SS for TOT, BG, and WG) into variance like expressions. And, how would you do that? Recall that the variance is an average of the squared deviations. So, since the SS values are the squared deviations or numerator part, we need to divide them by an n type factor so that the division converts the SS values into variance values. Usually, when we compute a variance value, we divide by the n that represents the number of numbers or things that were involved in the sums of squares calculations. For example, if we have 10 numbers and want the variance, we find the sums of squares and divide by 10. But, since we are usually using sample data to estimate the population variance, instead of dividing by 10, we would divide by n - 1 = 9 in this case. So, to make a sums of squares value into a variance value, we divide a value of n - 1. Look back at the aovo output above.

When we calculate the SS(BG) value, the number of "things" that are involved are the "3" groups. Recall that we find the squared deviation of each group mean around the grand or composite group mean. So, the number of things involved in this sums of squares calculation is the number of groups. So, if the n is the number of things and the number of things is the number of groups, then n - 1 in this case is the number of groups - 1 = 2. This 2 value in this case is called the degrees of freedom value associated with the BG term. We have encountered degrees of freedom in several places before. So, the degrees of freedom associated with the BG component is 2. For the within groups sums of squares component or SS(WG), the appropriate n - 1 value is the number of things within each group. In our case, there are 6 things or observations within each group and therefore the degrees of freedom or n - 1 value in each group would be 6 - 1 = 5. But, there are three groups here so there are 6 - 1 = 5 in each of 3 groups or 5 + 5 + 5 = 15 total. The degrees of freedom for the SS(WG) component is 15. For the SS(TOT) term, the number of things involved in that calculation would be all the data in the composite set of observations and this would involve all 18 of the data points (combined data set from all 3 groups) and therefore the degrees of freedom value for the SS(TOT) component would be 18 - 1 = 17. Note that the column df puts the appropriate n - 1 value for each of the three components: SS(BG), SS(WG), and SS(TOT).

With the sums of squares values in the SS column and the appropriate n - 1 type factor in the df column, all we need to do to convert SS values into variance type values is to divide the SS value by the appropriate n - 1 or df value. In the analysis of variance output table above, notice that there is a column titled MS which stands for mean square. By dividing the SS value by the df value, we are making a variance value which in analysis of variance jargon is called a "mean square" since a variance is an average of the sums of

squares or a mean (of the) square(s). The MS or mean square column gives us the variance components that we need to finish off the analysis of variance and make a decision about the null hypothesis.

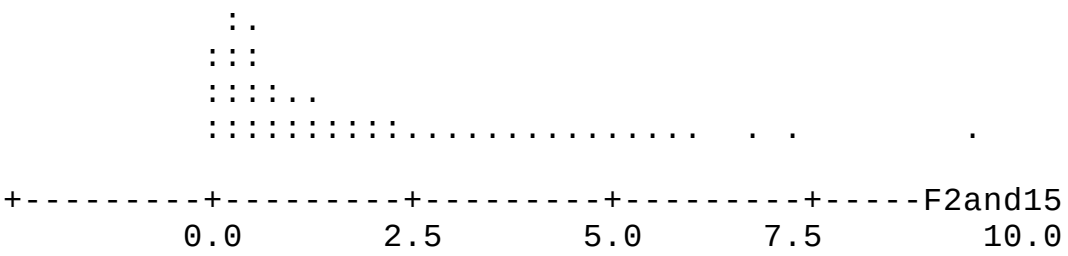
But, before completing the ANOVA, it is important to understand what each of these MS values represents. Remember, the TOT is being partitioned into the BG and WG parts. The MS BG variance value, 134.72 in this example, represents the variations amongst the means **across** the different populations. Although we are dealing with sample data, the MS BG gives us our best estimate of the variations or differences amongst the population means. Relatively speaking, the larger this term is, the more it appears that there are variations across the different populations with respect to their means and therefore the more likely it is that the population means are different. The MS WG represents our best estimate, from sample data, of how much variation there is **within** each of the populations. Relatively speaking, if the MS WG term is small, then it suggests that populations are very homogeneous but if the MS WG term is relatively large, then it suggests that the populations are very heterogeneous. Recall I said in Chapter 16 on sampling error of means that the greater the variability in the population (ie, the more heterogeneous it is), the greater the sampling error. Therefore, what we have to contend with here in our final stages of the analysis of variance is the problem that if the MS WG term is relatively large, and that means more sampling error, then it will be harder to detect any true difference in population means, if any differences really exist.

But, in any case, since the MS BG term reflects on possible differences in population means and the MS WG term reflects on sampling error, a comparison of the MS BG term to the MS WG term should help us decide whether to retain or reject the null hypothesis. Generally speaking, it should make sense that **if the MS BG term is relatively large and the MS WG is relatively small**, then we are in a better position to **reject the null** and say that the differences in means component (BG) outweighs the sampling error component (WG) and therefore it looks like there really are differences somewhere amongst the population means. And, this is exactly what we do: we compare the MS BG term to the MS WG term and then we make a decision about the null hypothesis on the basis of this comparison.

Also notice in the aovo output table, a column called F. This is called the F ratio and represents the ratio of the MS BG term to the MS WG term. **THE RATIO OF THE MS BG TERM TO THE MS WG TERM IS CALLED THE F RATIO** in analysis of variance and it is on this F ratio that we base our decision to retain or reject the null hypothesis. In our example above, the MS BG = 134.72 and the MS WG = 9.21 and therefore our F ratio is 134.72 / 9.21 = 14.63. Since the F ratio is larger than 1, it says that the BG factor (that reflects differences in population means) is much more dominant than the WG factor (that reflects sampling error). The question is: how much larger than 1 (ie, how much more dominant) will the F ratio need to be before we will feel comfortable enough to reject the null hypothesis?

What we are working with here is an F ratio and an F ratio is simply an F

distribution that we simulated in the Chapter 15 on statistical distributions. In addition, we used the F ratio previously when we tested an H(0) hypothesis about the difference between two population variances. Well, we have two variances here in our MS BG and our MS WG. So, it seems like what we need to help us make a decision is a critical F value from an F distribution with the appropriate degrees of freedom. In our problem, we have 2 degrees of freedom associated with the numerator or the BG term and 15 degrees of freedom associated with the denominator or WG term. Therefore, we are working with an F distribution with 2 and 15 degrees of freedom. To refresh your memory, here is a simulated F distribution with 2 and 15 degrees of freedom. Recall that F distributions tended to be quite seriously positively skewed.



In using the F distribution in this case, we need to use it in a "one tail" fashion. Recall in the discussion on testing frequency data hypotheses with chi square, I indicated that we used a one tail test in that case since it was only large chi square values that represented departures from the null hypothesis. The same thing is true here with our application of the F ratio. We can state this as a principle as follows.

THE F RATIO (MS BG TO MS WG) IS EXPECTED TO BE A VALUE OF 1 WHEN THE NULL HYPOTHESIS IS TRUE.

THE F RATIO (MS BG TO MS WG) IS EXPECTED TO BE LARGER THAN 1 WHEN THE NULL HYPOTHESIS IS NOT TRUE.

Thus, when we do the analysis of variance, we should see an F ratio close to 1 if the null is true but the F ratio **should be much larger than 1 if there really are some differences amongst the population means**. Since we only want to reject the null hypothesis when the F ratio is sufficiently larger than 1, we are not interested in F ratios that are less than 1. Hence, it is on the larger side of 1 that we are interested in, not the smaller side of 1. Therefore, we have a one sided or one tailed hypothesis. The implication of this is that instead of finding critical points for both the 2 1/2 percentile rank and the 97 1/2 percentile rank, we will concentrate all 5 percent at the upper end and therefore we will use the 95th percentile rank. For an F distribution with 2 and 15 degrees of freedom, the 95th percentile rank could be found with Minitab as follows.

```
MTB > invcdf .95;
SUBC> f 2 15.
```

0.9500 **3.6824**

Thus, the critical baseline point on our F distribution is 3.6824. This is what we compare our calculated F ratio to; in our case, the F ratio was 14.63. Since the F ratio we calculate from our study is much larger than the critical point, we reject the null hypothesis. As an alternative to comparing the F ratio from the aovo output to the critical F value, we could also look at the p value that is part of the aovo output (like we did before when doing t tests). Again, if the p value is larger than .05, we would retain the null hypothesis. But, if the p value is equal to or less than .05, we would reject the null hypothesis. By rejecting the null hypothesis, we are simply saying that the sample experimental data suggest to us that there are some differences amongst the 3 population means. In rejecting the null hypothesis, all we are saying is that it appears that there are some differences out there amongst the population means. What are some of the possibilities? Here are some possibilities based on our experiment.

Mean 1 not = to Mean 2 not = to Mean 3

This says that all 3 means are different from each other. But, what about the following?

(Mean 1 = to Mean 2) but both different from Mean 3

This says that the first two populations have the same means but both are different (higher or lower) than the mean in population 3. And, what about this one?

Mean 1 different than both (Mean 2 = Mean 3)

This says that populations 2 and 3 have the same means but both of these are different (higher or lower) than the mean of population 1. The bottom line here is that it only takes two means to be different to enable us to reject the null hypothesis. Thus, when we reject the null hypothesis, it only implies for sure that at least two population means will be different, not necessarily all of them. So, if we reject the null hypothesis, how can we tell where and how many population means are really different? There are procedures called follow-up tests that enable you to better "pinpoint" where the differences are and these are briefly described in the next section of this Chapter.

However, to get some idea about where the differences may be, there are two preliminary possibilities. First of all, we can look back at the original graph where the group means were plotted. The pattern certainly suggests that the mean from the third group (which is estimating the mean of population 3) is different than the other two. Whether the means for population 1 and population 2 are different also is not quite so clear. Secondly, we can also look at the graph of the confidence intervals from the aovo output as another source of evidence. Notice that the confidence intervals for the first two groups overlap with one another suggesting that there may be no difference between those two population means. But, neither the confidence interval for the first nor the second group overlaps with

the confidence interval for group 3 which suggests that the mean of the third population is different than both the means in populations 1 and 2. My best guess here, without doing further follow-up tests, is that means 1 and 2 could be the same but both are different from mean 3. In our study, therefore, the data suggest that the third strategy of working on electronic components seems to be superior to the first two strategies but the first two seem to be equally effective.

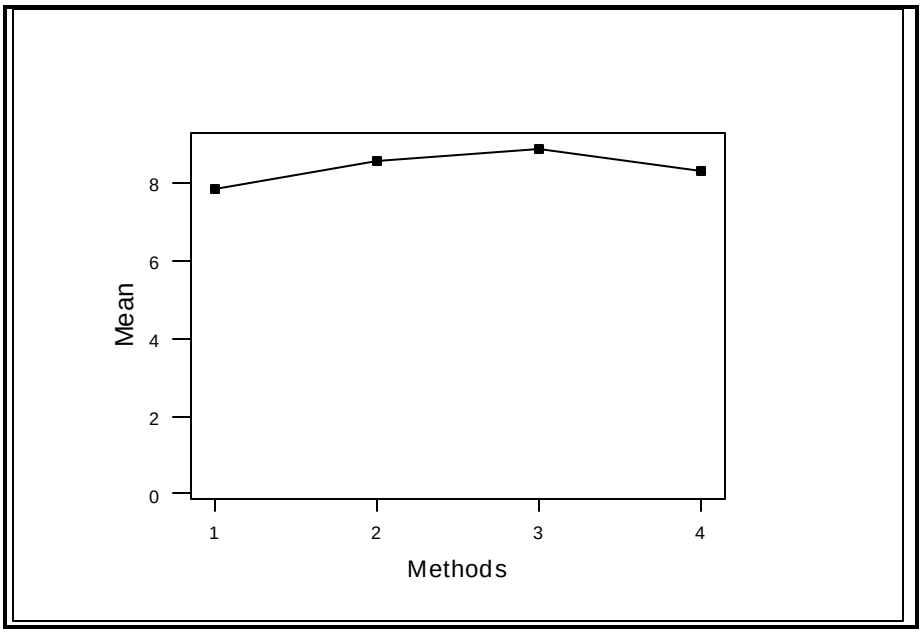
Before showing you the more formal follow-up procdures, let's do one more quick example. What if you are doing a learning experiment where you are trying 4 different types of reinforcement to see which ones, if any of them, produce better learning. The study is completed and produces the data as follows.

Meth1	Meth2	Meth3	Meth4
10	12	13	10
8	10	10	7
11	8	6	8
7	9	8	9
6	10	8	10
6	6	9	7
7	5	8	7

DESCRIPTIVE STATISTICS FROM 4 GROUP REINFORCEMENT STUDY

	N	MEAN	STDEV
Meth1	7	7.857	1.952
Meth2	7	8.571	2.440
Meth3	7	8.857	2.193
Meth4	7	8.286	1.380

A graph of the data is as follows.



From the means and the graph, it does not look like there is much difference in the means. The analysis of variance summary table gives us the following information.

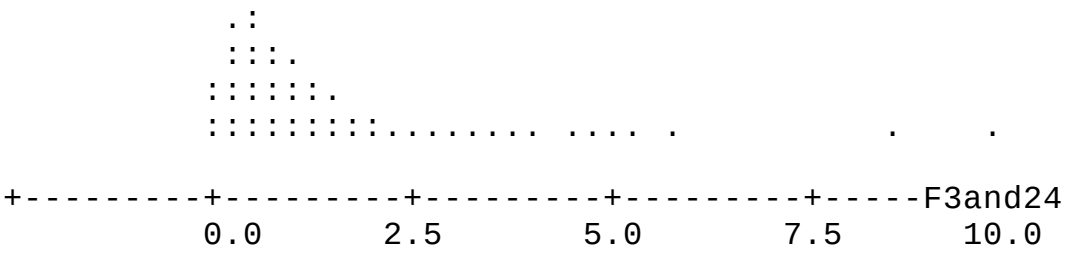
ANALYSIS OF VARIANCE

SOURCE	DF	SS	MS	F	p
METHOD	3	3.82	1.27	0.31	0.818
ERROR	24	98.86	4.12		
TOTAL	27	102.68			

INDIVIDUAL 95 PCT CI'S FOR MEAN
BASED ON POOLED STDEV

LEVEL	N	MEAN	STDEV	---------+---------+---------+---
Meth1	7	7.857	1.952	(---------*---------)
Meth2	7	8.571	2.440	(---------*---------)
Meth3	7	8.857	2.193	(---------*---------)
Meth4	7	8.286	1.380	(---------*---------)
				---------+---------+---------+---
POOLED STDEV =		2.030		7.5 9.0 10.5

First you should note that the 95% confidence intervals all overlap with one another which suggests that all population means could be the same. In addition, note the F ratio is less than one. Since we are only interested in rejecting the null hypothesis when the F ratio is much larger than one, we clearly must retain the null hypothesis. But, just to run you through the full cycle, we would need a critical value from an F distribution with 3 and 24 degrees of freedom. A simulated F distribution is shown below.



The critical F value we need from that distribution is below.

MTB > **invcdf .95;**
SUBC> **f 3 24.**
0.9500 **3.0088**

Our calculated F value is much smaller than the value needed (3+) to reject the null hypothesis and therefore we retain the null hypothesis. You do not even need to bother checking what the critical value is if the calculated F ratio is less than 1, like it is in the current example. In addition, if you check the p value, you see that it is much greater than .05 and therefore that is another way in which you know to retain the null hypothesis.