

## *One-Way Analysis of Variance*

*Thomas W. Pierce  
Department of Psychology  
Radford University*

Analysis of Variance (ANOVA) is one of the foundation tools of data analysis for researchers in the behavioral sciences. Let's see how it works...

Let's say that I show you a set of 15 scores.

X  
---  
15  
14  
13  
12  
11  
10  
9  
8  
7  
6  
1  
2  
3  
4  
5

There are 15 raw scores in the set. They represent the achievement test scores of 15 seventh graders. The lowest possible score is a zero and the highest possible score is seventeen. My question to you is this: why didn't all the students get the same score? That's right. Why didn't all 15 students get the same score? Fifteen students took the same test and we got fifteen different scores. Let's say that all of the students grew up in the same small town. They all had the same school books and the same teachers. They all come from roughly the same socio-economic background and have access to the same TV shows. So why do some of the kids do really well and some get totally blown away? There must be some explanation for it. Some kids studied and some kids didn't – that might explain some of it. Some kids got a good night's sleep and others didn't – that might explain some of it. Some kids were sick and others weren't. All right, those all sound like possibilities.

The question I asked isn't stupid at all. In fact, as psychologists it's your job to be answering that type of question. It's your job to notice ways in which people are different from each other. It's your job to try to measure this variability and to explain it.

## IT'S YOUR JOB TO TRY TO EXPLAIN WHY PEOPLE ARE DIFFERENT FROM EACH OTHER!

Professional psychologists are paid to explain why people differ from each other on variables like depression, sustained attention, job performance, and aggressiveness. We take it for granted that there's variability in the measures we study in psychology. But, in fact, it would look a look more odd or suspicious if the fifteen achievement test score looked like this: 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8. Variability in raw scores is normal. Lack of variability in raw scores is not normal. The job of psychology is to generate explanations for why people are different from each other. In fact, that is just what a **theory** in psychology does. Each theory represents a proposed explanation for why the scores on some measure of interest are not all the same. An **experiment** is often a careful and systematic test of a theory.

Experiments in psychology are tests of proposed explanations for why the scores for some variable are not all the same. The **dependent variable** in an experiment is the variable that needs explaining. Every **independent variable** in an experiment represents the proposed explanation for why the scores on the dependent variable are not all the same. The amount of sleep the night before an exam, the amount of time spent studying, and whether or not the student ate breakfast are all possible explanations for why some scores on the achievement test are higher than others. They all represent possible independent variables that might be examined by an experiment.

Starting out, I only know one thing about every student in that class – I know their score on the test. Now let's say that there is additional thing that I know about each of the fifteen students. Let's say that I know that six weeks before they took the test each of the students was assigned to one of three groups. The five students who got scores of 15, 14, 13, 12, and 11 had been randomly assigned to a group that got a lot of tutoring in how to do well on the test. The five students who got scores of 10, 9, 8, 7, and 6 got a moderate amount of tutoring. The five students who got scores of 1, 2, 3, 4, and 5 got no tutoring on how to take the test.

It turns out that I have an experiment in which there are three groups of students that were all recruited from the same population (i.e., from the same pool of potential subjects). At the beginning of the experiment there's nothing that I know of that makes the students in one group any different from the students in either of the other two groups. Then I do something to change the conditions in just one way – some students get a lot of tutoring, some get a moderate amount of tutoring, and some don't get any tutoring. The independent variable in the experiment is the Amount of Tutoring. After I change the conditions I get to see what happens. I get to see what every student looks like in terms of their achievement test score. Achievement Test Scores represent the dependent variable in the experiment. That's the way an experiment works. You change the conditions in just one way and then you see if anything happens. ***Because your independent variable is the only way in which the groups differ from each other, any differences that you see among the means for these groups must be because of the one***

*difference that we know about – the one difference that the experimenter put there, the independent variable!*

So now we know two things about each student. We know their score on the test and we know which group they were in. So how do we get a sense of whether the independent variable actually had an effect on the achievement test scores? Did changing the amount of tutoring cause a change in the scores on the test? Well, we already know that the mean score for everyone who took the test was 8.0. Let's call this mean the Total Mean and use the symbol  $\bar{X}_T$  to stand for it. We also know the mean score for the students in each group. The mean score for the students in the first group – the group that got a lot of tutoring – is 13.0. The mean score for the students in the second group – the group that got a moderate amount of tutoring – is 8.0. The mean score for the students in the third group – the group that didn't get any tutoring – is 3.0. Let's use the symbol  $\bar{X}_G$  to represent the mean of a group. The table below represents the information we have so far.

	X
	---
Lot of Tutoring	15
$\bar{X}_1 = 13$	14
	13
	12
	11
	-----
Moderate Amount	10
$\bar{X}_2 = 8.0$	9
$\bar{X}_T = 8.0$	8
	7
	6
	-----
No Tutoring	1
$\bar{X}_3 = 3.0$	2
	3
	4
	5

Eventually we want to be able to know how good a job the IV does at explaining the variability of everyone's scores on the DV. We want to see if the Amount of Tutoring can explain at least a little bit of way those 15 raw scores weren't all the same – why they weren't all equal to that Total Mean of 8.0. But let's say that for right now we'd just like to see how good a job the IV can do at explaining just one person's score. If we look at the person who got the score of 15. **Why didn't that one person get a score equal to the Total Mean of 8.0?** Can knowing the amount of tutoring that person got help us to answer this question?

If we want to see if we explain why that person had a score that was different from the mean of everybody, the best place to start is to figure out exactly how much of a deviation from this Total Mean we have to explain. And that's easy. The person's raw score is 15 and the mean of everybody is 8.0. That means that the person had a score that was 7 points higher than the mean of everybody. The total deviation for that person that needs to be explained is a deviation of 7 points. In equational-type terms that deviation can be expressed as

$$\text{Total Deviation} = (X - \bar{X}_T) = 15 - 8.0 = +7$$

So the Total Deviation for that person that needs to be explained is +7. Out of that Total Deviation of +7, how can we account for if we take the amount of tutoring they got into account? Look at it this way. Let's say that you've been asked to provide your best guess about what the person's score is. You don't know what score they got and you don't know which group they were in. All you know is that they were one of the 15 students in the class. **If you don't know which group they were in your best guess would have to be the mean of all fifteen students – the Total Mean of 8.0.** And how far off would your best guess be? You'd be off by 7 points. That's the Total Deviation we talked about a couple of minutes ago. Now let's say that you're given one additional piece of information. You find out that that person had been in the group that got a lot of tutoring. Now what's your best guess? Are you still going to go with the 8.0, the mean of all 15 students? No! Of course not. You'll go with the best information you've got. You'll use the mean of their group as your best guess. And the mean of that person's group is 13.0. Think about it. Your best guess when you don't know which group they're in is 8.0. Your best guess when you do know which group they're in is 13.0. How much more accurate is your best guess on the basis of having this one additional piece of information – on the basis of knowing the amount of tutoring they got. It's the difference between these two best guesses of course. It's the difference between the mean of their group (13) and the Total Mean (8.0), which is a difference of 5.0 points. The equational-type way of representing this deviation is...

$$\bar{X}_G - \bar{X}_T = 13 - 8 = +5$$

Knowing which group the person is in takes you 5 points close to their actual score. That deviation of 5 points is a deviation we can account for because we can explain where that deviation comes from. It comes from the fact that people who get a lot of tutoring have scores that are 5 points higher, on average, than the mean of all the students who took the test.

So at this point we can say that out of the Total Deviation of that person's score from the mean of everybody of 7 points we can explain 5 points out of that total of the 7. That's seems pretty good. So if the IV can account for 5 units out of the total of 7, what can the IV not account for? Well, duh! It must be the remaining two points. So where does it come from? Look at it this way. You know that the person is in the group that got a lot of tutoring. And you know that everyone in that group was treated exactly alike. They all had the same instructor, for the same amount of time, using the same materials, at the

same time of day, etc. You get the picture. So why didn't the five students in that group get exactly the same score? Why did our student get a score that was two points higher than the mean of their group? We don't know! There must be some explanation for it, but we don't have it. The deviation between the person's raw score and the mean of their group is something that we don't have an explanation for. It's a deviation that is unaccounted for by the IV. The equational-type way of representing this deviation is...

$$X - \bar{X}_G = 15 - 13 = +2$$

So we end up with an interesting relationship. The total deviation that there is to explain (7 points) looks like it equal to a deviation that the IV can account for (5 points) plus a deviation that the IV cannot account for (2 points). In equational-type terms the relationship looks like this...

Total Deviation	=	Deviation	+	Deviation Not
		Accounted For		Accounted For
$X - \bar{X}_T$	=	$\bar{X}_G - \bar{X}_T$	+	$X - \bar{X}_G$
+7	=	+5	+	+2

And the neat thing is that this relationship holds for any of the fifteen raw scores that you might want to look at.

So we figure out how good a job the IV does at explaining the deviation of one person's score from the Total Mean. Now we're ready for the next step. How good a job does the Amount of Tutoring do at accounting for the deviations of all fifteen scores from the Total Mean? This is the same thing as asking how good a job does the IV do at accounting for the variability of the fifteen raw scores.

Where should we start if we want to see how good a job the IV does when we take all the scores into account, not just one score? Well, how about we start in the same place we did before. If we want to explain the variability of a set of raw scores we first have to ask "variable around what?" The answer is "variable around the Total Mean". We need to measure the total amount of variability that needs to be explained. So if, at the level of one person's raw score, we subtracted the Total Mean from the raw score, we need to do this same thing for all fifteen raw scores. This will give us 15 deviations of raw scores from the Total Mean. It'll give us 15 Total Deviations that need to be accounted for. Here's how it looks...

	X	$X - \bar{X}_T$
	---	-----
Lot of Tutoring	15 - 8	+7
$\bar{X}_1 = 13$	14 - 8	+6
	13 - 8	+5
	12 - 8	+4
	11 - 8	+3
-----		
Moderate Amount	10 - 8	+2
$\bar{X}_2 = 8.0$	9 - 8	+1
$M_T = 8.0$	8 - 8	0
	7 - 8	-1
	6 - 8	-2
-----		
No Tutoring	1 - 8	-7
$\bar{X}_3 = 3.0$	2 - 8	-6
	3 - 8	-5
	4 - 8	-4
	5 - 8	-3

Now, instead of having 15 separate deviations that need to be accounted for, we need to generate one number that measures the total amount of variability in the entire data set that needs to be accounted for. We've got a bunch of deviation scores. How do we get a measure of variability out of them. Obviously we can't just calculate the mean of these deviation scores because the mean of any set of deviation scores is zero, but we can get rid of the negative signs by squaring each of the deviations. Now we've got fifteen squared deviations of raw scores from the Total Mean. Here's what they look like...

	X	$X - \bar{X}_T$	$(X - \bar{X}_T)^2$
	---	-----	-----
Lot of Tutoring	15 - 8	+7	49
$\bar{X}_1 = 13$	14 - 8	+6	36
	13 - 8	+5	25
	12 - 8	+4	16
	11 - 8	+3	9
-----			
Moderate Amount	10 - 8	+2	4
$\bar{X}_2 = 8.0$	9 - 8	+1	1
$M_T = 8.0$	8 - 8	0	0
	7 - 8	-1	1
	6 - 8	-2	4
-----			
No Tutoring	1 - 8	-7	49
$\bar{X}_3 = 3.0$	2 - 8	-6	36
	3 - 8	-5	25
	4 - 8	-4	16
	5 - 8	-3	9
			-----
			280

Now, if you add all of these squared deviations up you get 280. The sum of these squared deviations is a perfectly good measure of variability. Because **280 represents all of the variability in this set of 15 raw scores that needs to be accounted for** we refer to it as the **Sum of Squares Total**.

So out of this total of 280, how much variability in these achievement test scores is actually accounted for by knowing how tutoring they got? To figure this out the best place to start is to think how we calculated the *Deviation Accounted-For* for just one subject. The Deviation Accounted-For for one person was based on how much better your best guess would be about their real score when you know which group they're in compared to when you don't know which group they're in. It's based on the deviation between the mean of their group and the Total Mean. If we can do this for one person, we can do this for all 15 scores. So for all 15 students let's say that we take the mean of their group and then subtract the mean of everybody. In effect, we're calculating a Deviation Accounted-For for each of the 15 students. Then we take these 15 Deviations Accounted-For and then square them to get rid of the negative signs. Then we add these Squared Deviations Accounted-For up. Here's how it looks...

	X	$\bar{X}_G - \bar{X}_T$	$\bar{X}_G - \bar{X}_T$	$(\bar{X}_G - \bar{X}_T)^2$
	---	-----	-----	-----
Lot of Tutoring	15	13-8	+5	25
$\bar{X}_1 = 13$	14	13-8	+5	25
	13	13-8	+5	25
	12	13-8	+5	25
	11	13-8	+5	25
	-----			
Moderate Amount	10	8-8	0	0
$\bar{X}_2 = 8.0$	9	8-8	0	0
$M_T = 8.0$	8	8-8	0	0
	7	8-8	0	0
	6	8-8	0	0
	-----			
No Tutoring	1	3-8	-5	25
$\bar{X}_3 = 3.0$	2	3-8	-5	25
	3	3-8	-5	25
	4	3-8	-5	25
	5	3-8	-5	25
				-----
				250

...And we end up with a sum of squared deviations of 250. We can say that the **Sum of Squares Accounted-For** is 250. When we take all of the scores into account, the IV accounts for 250 units out of 280. Another way of looking at it is that out of all the reasons for why the students could differ from each other on achievement test scores, our

one proposed explanation – the Amount of Tutoring – accounts for 250 units out the total of 280.

So how much variability in these 15 achievement test scores is not accounted for? It seems like it must be a sum of squares of 30. And it is. So where does this number come from? Well, when we measured the *Deviation Not-Accounted-For* for that one person we took the person's raw score and subtracted the mean of their group. That was because we didn't have any explanation for how the scores within a group could be different from each other when everyone in that group was treated exactly alike as far as the independent variable goes. If that's what we did at the level of a single person, that's what we ought to do for everyone's scores. So we take all 15 raw scores and subtract the means of each of their respective groups. And we end up with 15 deviations between raw scores and the mean of their groups. We end up with 15 Deviations Not-Accounted-For. Then, to get a measure of the variability that is not accounted for by the independent variable, we take all 15 of these deviations and square them. Now we've got 15 squared deviations that are not accounted for. When we add them all up we get the sum of squared deviations that are not accounted for. Here's what it looks like...

	X	$X - \bar{X}_G$	$(X - \bar{X}_G)^2$
	---	-----	-----
Lot of Tutoring	15 – 13	+2	4
$M_1 = 13$	14 – 13	+1	1
	13 – 13	0	0
	12 – 13	-1	1
	11 – 13	-2	4
	-----		
Moderate Amount	10 – 8	+2	4
$M_2 = 8.0$	9 – 8	+1	1
$M_T = 8.0$	8 – 8	0	0
	7 – 8	-1	1
	6 – 8	-2	4
	-----		
No Tutoring	1 – 3	-2	4
$M_3 = 3.0$	2 – 3	-1	1
	3 – 3	0	0
	4 – 3	+1	1
	5 – 3	+2	4
			-----
			30

The **Sum of Squares Not-Accounted-For** ends up being 30. Which is what we thought it would be.

We saw before that the Total Deviation for one person's score is equal to a deviation that is accounted for by the IV plus a deviation that is not accounted for by the IV. You've probably already noticed that the same kind of relationship holds when you take all of the scores into account. The Sum of Squares Total is equal to the Sum of Squares Accounted-For plus the Sum of Squares Not-Accounted-For.

$$\text{SS Total} = \text{SS Accounted-For} + \text{SS Not-Accounted-For}$$

$$\begin{aligned} \Sigma(X - \bar{X}_T)^2 &= \Sigma(\bar{X}_G - \bar{X}_T)^2 + \Sigma(X - \bar{X}_G)^2 \\ 280 &= 250 + 30 \end{aligned}$$

It turns out that the Sum of Squares Total is something that we can take apart. A statistician would say that we can *partition* the Sum of Squares Total into the sum of Squares Accounted-For and the Sum of Squares Not-Accounted-For.

Now that we've seen how to quantify the degree to which the IV can account for variability in the DV, let's talk about these sums of squares a little bit more. ***What would the scores in a data set have to look like in order for the Sum of Squares Total to be equal to zero?*** Could that really happen? No, probably not. That would be a situation where there was nothing for the IV to have to explain. It would be a situation where every time you took a raw score and subtracted the Total Mean you'd always get zero. The only way for that to happen would be if all of the raw scores were the same number. This would be a situation where there was no variability at all in the data set.

***What would the data have to look like to get a SS Total that is greater than zero and a SS Accounted-For that is equal to zero?*** Well, you know that the scores in the data set aren't all the same because the SS Total is greater than zero. The only way for the SS Accounted-For to be equal to zero is if every time you took a person's group mean and subtracted the Total Mean, you always got a value of zero. The only way for this to happen is if group means are always equal to the Total Mean. And the only way for this to happen is if all of the group means are equal to each other. So why does this make sense? If tutoring didn't have anything to do with achievement test scores – if tutoring has no effect on achievement test scores – what you expect these group mean to be? Well, if tutoring doesn't do anything to achievement test scores is there any reason to think that one group should do any better than another group? No! So if tutoring has no effect of achievement test scores the group means should all be the same. Because the average of all the group means has to turn out to be the mean of all subjects in the study, this would mean that the group means should also turn out to be equal to the mean of everyone. In this case there would be no variability between the groups. In fact, statisticians refer to the Sum of Squares Accounted-For as the **Sum of Squares Between-Groups**.

***What would the data have to look like for the SS Total to be greater than zero, but the SS Not-Accounted-For to be equal to zero?*** Well, the Sum of Squares Not-Accounted-For is based on deviations between a person's raw score and the mean of their group. For this sum of squares to be equal to zero the deviation between a raw score and the group mean would always have to be zero. The only way for this to happen would be if all of the scores within each group were the same. In this case there would be no variability within the groups. In fact, statisticians refer to the Sum of Squares Not-Accounted-For as the **Sum of Squares Within-Groups**.

## The F-ratio

So how do you know whether the Amount of Tutoring accounts for a significant amount of variability in achievement test scores? The place to start is to recognize that this is a yes or no question. Either the IV has a significant effect on the DV or it doesn't. The researcher has to choose between these two options. These two options are the same two options that we worked with in conducting z-tests and t-tests. The *null hypothesis* for this question is that there is no significant effect of tutoring on achievement test scores. The *alternative hypothesis* is that there is a significant effect of tutoring on achievement test scores.

Now, if we lived in a perfect world, it would be easy to see if the IV had an effect on the DV or not. For the null hypothesis to be false, all you have to be able to say is that the IV had some effect on the DV. It doesn't have to have a large effect or even a noticeable effect. It's a matter of whether it had any effect. All you'd have to do is to see if the SS Accounted-For was greater than zero or not. If it's equal to zero the null hypothesis must be true and the IV has no effect at all on the DV. If it's greater than zero the null hypothesis must be false. Any differences between the means would indicate that, at least to some small extent, changing the conditions in terms of the IV result in changes in the scores on the DV.

To understand why we don't live in this perfect world you have to remember what we've got to work with here. We've got three samples of participants. These samples were all drawn from the same population. Then the experimenter did something to make the groups different from each other in just one way – different groups got different amounts of tutoring. And now we've got to use the mean scores for the people in these different groups to make a decision about whether changing the amount of tutoring caused a change in the scores. All we have to go on are these sample means. And what's the job of a sample mean? -- to give you an unbiased estimate of the mean of a population. What should these sample means look like if the null hypothesis is true? If the null hypothesis is true, the reality is that changing the amount of tutoring had no effect on the scores the students got. If the null hypothesis were true then what really happened in the experiment is that we'd started out by drawing three samples of students from the same population. Then our IV didn't do anything to make the students in these groups any different from each other. So we end up with three samples drawn from the same population. We end up with three sample means that are all estimates of the same population mean. If the null hypothesis is true these sample means are all supposed to be the same number, which would result in the SS Accounted-For being equal to zero.

BUT, this assumes that these sample means are giving us perfect estimates of that one population mean. The job of a sample mean is to give you an unbiased estimate of the mean of a population. There's no guarantee that a sample mean is a perfect estimate, only that it is no more likely to be too low than it is to be too high. So, even when the null hypothesis is true, it's not fair to expect that these three sample means are all going to be the same number. ***When the null hypothesis is true the sample means can be different from each other just by chance.*** This means that even when the IV has no

effect at all, the SS Accounted-For can still be some number greater than zero. In fact, it almost certainly will be. We can't see whether the IV had an effect on the DV by looking to see whether the SS Accounted-For is greater than zero or not. It could be different from zero just by accident – by chance alone.

Darn. I thought we were almost done with this stuff...

Because the sample means could be different from each other just by chance – because the sample means aren't giving me perfect estimates, I can't just look at the SS Accounted-For to see if it's zero or some number greater than zero. The question now really becomes one of deciding whether the SS Accounted-For is **enough** greater than zero to be confident that it's not just greater than zero by chance alone. Now we're in exactly the same kind of situation that we were in when doing z-tests and t-tests. We're going to be forced to make a decision based on some odds. Just like with a t-test it'll turn out that the only thing we'll be able to know for sure are the odds of making a mistake if we decide to reject the null hypothesis.

The idea behind both a z-test and a one-sample t-test was that there was one number we were making our decision about, the mean of a sample. The question was whether we were willing to believe that our one sample mean was a member of a collection of other sample means that were obtained when the null hypothesis was true.

The idea behind an independent samples t-test was that there was one number we were making our decision about, the difference between two sample means. The question was whether we were willing to believe that the one difference between means obtained from our experiment was a member of a collection of other differences between means that were collected when the null hypothesis was true.

So what should we do in this latest situation? The same kind of thing. Decide what kind of number we're going to make our decision about. Then figure out what these numbers would look like if we repeated this same experiment thousands and thousands of times when the null hypothesis was true. The question then would be just a matter of deciding whether we were confident that our number did not belong in this collection. If we decide our number doesn't belong in a collection of numbers obtained when the null hypothesis is true, it must have been obtained when the null hypothesis was false – we'd decide that the Amount of Tutoring had an effect on Achievement Test scores.

So what kind of number should we use? The SS Accounted-For would not be a good choice because this number varies along with the number of participants in the study. Two studies could have the identical means, but the study with the larger sample size will end up with a larger SS Accounted-For. So we've got to go after a number that is not influenced by the sample size.

The number that Ronald Fisher latched onto in the 1930s was based on the ratio of variability accounted-for to variability not-accounted-for. Or...

$$\frac{\text{Variability Accounted-For}}{\text{Variability Not-Accounted-For}}$$

Now, the tempting thing to do at this point would be to take the SS Accounted-For and divide it by the SS Not-Accounted-For, giving us a ratio of 8.33.

$$\frac{\text{SS Accounted-For}}{\text{SS Not-Accounted-For}} = \frac{250}{30} = 8.33$$

Unfortunately, we already know that you can't compare one sum of squares to another sum of squares when they're based on different numbers of values. It wouldn't make much sense to compare the sum of 20 numbers to the sum of only 10 numbers. It's the same thing with the sum of squares. It's a SUM. It's what you get when you add up a bunch of numbers. In this case those numbers just happen to be squared deviations from the mean. Okay, smartie, so what do we do? Well, you know that you can't compare one sum to another sum, but you **can** compare the mean of one set of numbers to the mean of another set of numbers, even though the sample sizes for the two sets of scores might be very different from each other.

And that's what we need to do here. Instead of comparing the SUM of squared deviations accounted-for to the SUM of squared deviations not-accounted-for, we need to compare the MEAN of the squared deviations accounted-for to the MEAN of the squared deviations not-accounted-for. What's another name for the mean of a bunch of squared deviations? The **variance**! The number we need will come from taking the Variance Accounted-For and dividing it by the Variance Not-Accounted-For. The name for this number is the **F-ratio** and is represented by the letter **F**.

$$F = \frac{\text{Variance Accounted-For}}{\text{Variance Not-Accounted-For}}$$

So how do we get this number? That's easy. You remember how to calculate the variance for a set of raw scores. You take the sum of squares and then divide it by the appropriate number of degrees of freedom.

$$S^2 = \frac{(X - \bar{X})^2}{N-1}$$

The steps needed to calculate the F-ratio for our experiment are contained in a table referred to as an **ANOVA Table**. This table is basically just a way of organizing all of the steps needed to generate the F-ratio. The steps needed to calculate the F-ratio are

presented from the left side of the table to the right. The ANOVA table starts by listing the two sources of variability needed to calculate the F-ratio – Accounted-For and Not-Accounted-For.

Source
-----
Accounted-For
Not Accounted-For

Just to the right of each source we can now list the one piece of information we have about each of them, the sum of squares associated with them.

Source	SS
-----	----
Accounted-For	250
Not Accounted-For	30

Now we know that we need the variances that correspond to each of the two sources of variability. To get them we need to divide each sum of squares by the appropriate number of degrees of freedom. “df” refers to the number of degrees of freedom. “MS” refer to the Mean Square for each source of variability. A Mean Square is the same thing as a variance. Remember, the variance is nothing more than the mean of a bunch of squared deviations. For the moment, I’ve simply listed the degrees of freedom for both sources of variability. Once we’ve gotten our F-ratio I’ll go back and explain where these numbers came from. For the Accounted-For row, a sum of squares of 250 divided by 2 degrees of freedom gives us a Mean Square of 125. The number of degrees of freedom accounted-for is equal to the number of groups – 1. For the Not-Accounted-For row, a sum of squares of 30 divided by 12 degrees of freedom gives us a Mean Square of 2.5. The number of degrees of freedom not accounted for is equal to the number of groups multiplied by the number of people in each group – 1, or (#groups)(n-1).

Source	SS	df	MS
-----	----	----	----
Accounted-For	250	2	125
Not Accounted-For	30	12	2.5

#groups - 1

(#groups)(n-1)

The F-ratio is simply the Mean Square Accounted-For of 125 divided by the MS Not-Accounted-For of 2.5. This gives us the number 50.0

Source	SS	df	MS	F
-----	----	---	-----	----
Accounted-For	250	2	125	50
Not Accounted-For	30	12	2.5	

The F-ratio for this data set is 50.0. It came from taking one variance and dividing it by another variance. This is where the term **Analysis of Variance** comes from. *So what does this number mean?* Well, it basically means that the ratio of variance accounted-for to variance not-accounted-for is 50:1. Another way of saying the same thing is that the variance accounted-for is 50 times larger than the variance not-accounted-for. *Is an F-ratio of 50 good?* It sounds pretty good. The question really is, is this F-ratio large enough for us to be confident that the IV really did have an effect in the DV. Is it large enough for us to be confident that it's not that large just by chance alone? Unfortunately, it will always be possible that the F-ratio we got was that large just by chance alone. Because of this we will never be able to know for sure if we're making a mistake if we decide to reject the null hypothesis. But, just like in a t-test we will be able to know the **odds** of making a mistake if we decide to reject the null hypothesis. If we use an alpha level of .05 we're saying that we'll reject the null hypothesis if we can show that the odds are less than 5% that the null hypothesis is true. We're saying that we're willing to take on 5% of risk of making a Type I error.

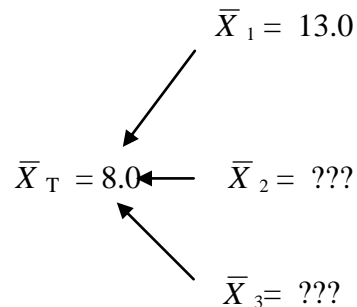
*So how do we figure out whether the odds are less than 5% that the null hypothesis is true? How do we figure out if the odds are less than 5% that we got an F-ratio that large just by chance?* Well, how did we figure out whether we got a value for **t** that was far enough above or below zero for us to be confident that it wasn't that far away just by chance? We had to have something to compare that number to. We compared it to the distribution of other values for **t** that you'd get when the null hypothesis was true – when the values for **t** were different from zero just by chance. So in this case, we've got to do the same thing. We have to compare our F ratio to a collection of other F-ratios to see if it looks like it belongs with them or not. We can then use that collection of other F-ratios to determine the lowest F-ratio we could have and still have it belong in that set. We need a **critical value for F**.

Pragmatically, we can find out what this critical value for F is by looking it up in a **Critical Values for F Table**. There are three things that you need to know in order to look up the critical value for F. You need to know the alpha level you want to use. In this case we want to use an alpha level of .05. You need to know which column to look in. That's determined by the number of degrees of freedom for the denominator of the F-ratio – 2 in this example. You need to know which row to look in. That's determined by the number of degrees of freedom for the numerator of the F-ratio --12 in this example. Once you've located this number, you see that you've zeroed in on a value of 3.89. In order to say that our F-ratio is significant it has to be greater than or equal to 3.89. Because our F-ratio of 50 is obviously greater than the critical value of 3.89 our decision is to reject the null hypothesis. Our conclusion is that "Tutoring has a significant effect on Achievement Test scores,  $F(2,12) = 50.0, p < .05$ ."

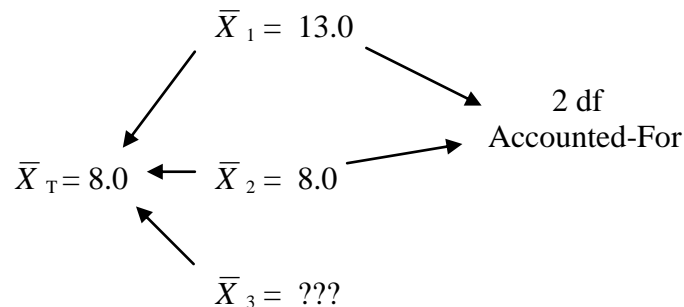
### Degrees of Freedom in ANOVA

Now, let's go back and talk about that degrees of freedom thing. Why 2 degrees of freedom for the Accounted-For row and 12 degrees of freedom for the Not-Accounted-For Row? It seems like it should be 14 degrees of freedom for both of them. After all, in each case we added up 15 squared deviations to get the sums of squares. *Shouldn't the number of degrees of freedom be equal to the number of values minus one or  $15-1 = 14$ ?* When working with the Sum of Squares Total that's exactly what it is. The total number of degrees of freedom for the study is equal to the total number of participants (15) minus one degree of freedom, giving you 14 degrees of freedom. But for the Accounted-For and Not-Accounted-For sources of variability there's something else you have to keep in mind. The number of degrees of freedom is always equal to the number of **independent values** that are free to vary. That "independent values" part is important in this case.

As far as the number of degrees of freedom accounted-for goes, think of it this way. The Sum of Squares Accounted for is based on deviations between group means and the Total Mean. In effect, you can think of this situation as one where the three group means are being used to calculate the Total Mean. If you know that the Total Mean is equal to 8.0 and you know that one of the three group means is equal to 13.0, are the other two group means free to vary? Do these last two group means have to be any particular number? No. As long as you pick numbers for these last two means that get the Total Mean to come out to 8.0, this will work out fine.



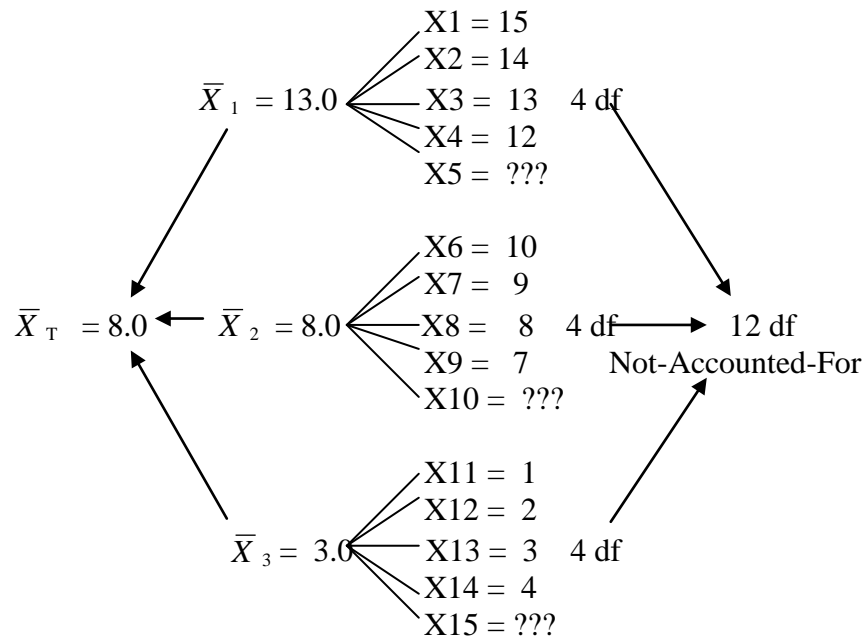
Now, let's say that you know that a second group mean in the set is 8.0. Is the last group mean fixed? Does it have to be a particular number? Now the answer is YES! If the first two group means are 13.0 and 8.0 the last group mean has to be 3.0 to get the Total Mean to be 8.0.



Out of the three group means that are being used to calculate the Total Mean, only two of them are free to vary. Once you know two of them the last one is fixed. It is not free to vary. That's one way of describing why the number of degrees of freedom for the Accounted-For term is equal to two.

So how about the number of degrees of freedom not-accounted-for? Where does this number come from? Think of it this way. The Sum of Squares Not-Accounted-For is based on deviations between the raw scores and the group means. In effect, you can think of this situation as one where the raw scores are being used to calculate the group means. If we look at the scores for the first group, the mean of the five scores in this group is 13.0. If we know that one of the five scores in this group is 15, are the other four scores in the group free to vary? Yes. You could change those numbers around, as long as the mean of the group came out to 13.0. How about if you know that three of the five scores in that group are 15, 14, and 13? Are the remaining two scores free to vary. They are. Now if you know that four of the five scores in the group are 15, 14, 13, and 12, is the last score in that group free to vary? NO! To get the mean of that group to come out to 13, that last score in that group has to be 11. That last score is fixed. It is not free to vary. So when five raw scores are being used to calculate that one group mean only **four** of them are free to vary.

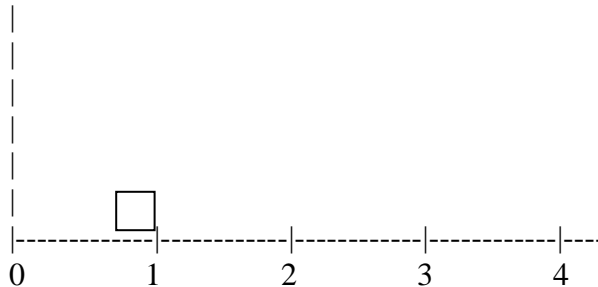
The same thing goes for each of the three groups. There are four degrees of freedom not-accounted-for for the first group, four degrees of freedom not-accounted-for for the second group and four for the third group. In all, taking every group into account, this brings us up to a total of 12 degrees of freedom not-accounted-for.



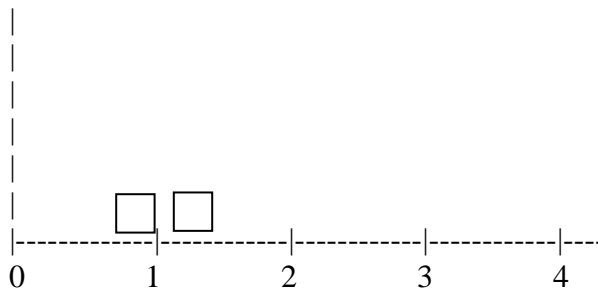
### Origin of Critical Values for F

Okay, so now you know how to look up critical values for F in the Critical Values for F table. Where do these numbers come from? I mean, what did someone have to do to figure out what that number ought to be? The critical value for F comes from the same kind of place that critical values for t come from. You know that you're willing to reject the null hypothesis if you can show that the odds are less than 5% that the null hypothesis is true. To know whether these odds are less than 5% or not you have to imagine that you do the experiment when you could know for sure that the null hypothesis is true – when you could know for sure that the only thing making the groups means different from each other is chance.

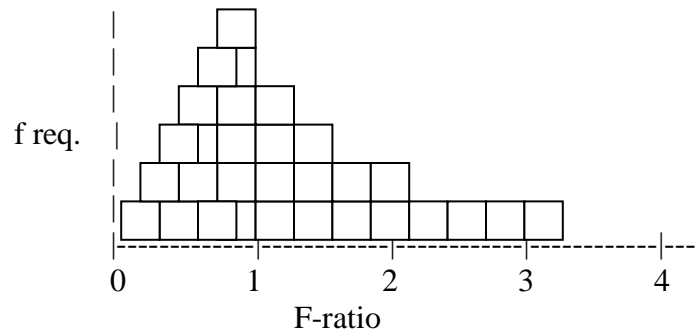
Given that, you imagine that you do the experiment when the null hypothesis is true. The design of the experiment is the same. There are three groups and there are five participants in each group. What value for F would you expect to get? Zero? If sample means gave you perfect estimates you ought to get an F-ratio of zero when the null hypothesis is true. But you can't expect these sample means to give you perfect estimates. When the group means are just different from each other by chance, the sum of squares accounted for is going to be greater than zero. This, in turn, will produce an F-ratio that is greater than zero. Let's say that this F ratio turns out to be 0.80. Let start a graph of a frequency distribution that shows us where this F ratio belongs on the scale of possible F ratios. The little box represents the location of that one F-ratio.



Now, let's say that the null hypothesis is true and you do the same experiment a second time. There's no reason, other than chance for the F-ratio to be greater than zero, but now the F-ratio turns out to be 1.20. Here's where that second F-ratio goes on the frequency distribution.



Now, you pretend that you do the same experiment over and over and over. Every time you have the same design: the IV is the Amount of Tutoring, the DV is the Achievement Test score, there are three groups, and five participants in each group. An every time you do the experiment the null hypothesis is true. Now we can get a sense of what F-ratios look like when the null hypothesis is true. The name for this collection of numbers is the *sampling distribution of the F-ratio*.



Obviously, even though we've conducted the same experiment, we don't always get the same F-ratio. Even when the null hypothesis is true, and in reality the IV has no effect on the DV, every once in a while you can get F-ratios that are fairly substantial. Using the laws of probability statisticians can tell us what all of these F-ratios would look like if someone actually went to the trouble of doing the same experiment over and over and over again and collected the F-ratio from each of these separate experiments. What these mathematical descriptions of this collection of F-ratios tells us is that exactly 5% of the F-ratios that we would collect are at 3.89 or higher. Exactly five percent of the area under this curve is found above the value of 3.89. Knowing that, we can say that if we get an F-ratio from our experiment that is greater than or equal to 3.89 we'll know that it might be possible that we got an F-ratio that large just by chance, but we'll also know that the odds of that happening are less than 5%. It's the shape of this distribution of F-ratios that determines the critical value for F. It turns out that the shape of this distribution looks different for every combination of a number of degrees of freedom for the numerator of the F-ratio and the denominator of the F-ratio. The more degrees of freedom in the denominator the more the values on the long tail of the distribution are clustered closer to the center of the curve (giving you smaller critical values). The more degrees of freedom you have in the numerator the more the values on the long tail of the distribution are clustered closer to the center of the curve (also giving you smaller critical values).

### Conceptual definition of the F-ratio

So what F-ratio should you expect to get if the null hypothesis is true? Zero? We said in the verbage above that it's just not fair to expect sample means to give you perfect estimates of population means. When the null hypothesis is true, and the sample means are all estimates of that one population mean, you're going to get a number in the numerator of the F-ratio that is greater than zero. There's no reason for it being greater than zero other than chance. And in statistics, anything that you don't have an

explanation for is referred to as Error. As far as the denominator of the F-ratio is concerned, that number is based on the deviations between the raw scores and the group means. Those are all deviations that we don't have an explanation for. So, for a statistician, the Mean Square Not-Accounted-For is attributed to Error. Therefore, conceptually, when the null hypothesis is true the F-ratio is equal to Error divided by Error, or...

$$\frac{\text{Error}}{\text{Error}} = 1$$

Whatever forces of chance or error are operating to give you a number in the top part that is greater than zero tend to be equal to the forces of chance or error that give you a number in the bottom part that is greater than zero. So if the error in the top part is roughly the same as the error in the bottom part, you ought to get an F-ratio that is right around 1.0. When the null hypothesis is true, you should expect to get a F-ratio right around 1.0. The critical value for F isn't telling you how far above zero your F-ratio has to be in order to reject the null hypothesis. It's telling you how far above 1 your F-ratio has to be in order to reject the null hypothesis.

When the null hypothesis is false, there is something in addition to chance that's making the group means different from each other. When the alternative hypothesis is true the IV has an effect on the scores. This means that the IV is actively driving the group means away from each other. The group means aren't just different from each other by chance. They're different from each for a reason. They're different from each other because of the effect of the treatment the experimenter provided. When the null hypothesis is false the numerator of the F-ratio is equal to Error plus the effect of the treatment. The denominator is still just due to Error.

$$\frac{\text{Error} + \text{Treatment}}{\text{Error}} > 1$$

When there is an effect of the treatment present in the numerator there's a reason for the F-ratio to be greater than 1.0. The critical value for F tells us how far a F-ratio has to be above 1.0 to get to the point where only 5% of F-ratios are that far above 1.0 just by chance.