We have just reviewed how to run an ANOVA in excel. We can simply use the data analysis tool back and choose single factor for us to get ANOVA output without having manually calculate the between group variation, the within group variation and so on. We are using single factor because in our example there was one factor which we think influence sales in a store, which is height of the shelf. Any time we do a hypothesis test, the most important piece of the output that we will be reviewing is the p value and in this example the p value is 0.98.

What does that imply? It means it's not likely that we are going to reject the null hypothesis. We are not going to reject the null hypothesis because the variation that we see in the sample means is simply very, very likely in fact 98% likely because of random variation. What was the null hypothesis in our example? Remember that all the sample means are equal. In other words the shelf height has no impact on the sales of this particular product.

>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>

One way to validate our calculations, remember that we had said that the total variation in the sample is being split into between group variation and within group variation. We can actually calculate total variation in our sample simply by calculating variance. Remember the variance calculation.

- Take a mean
- Subtract each value from the mean

- Square it up
- Sum it all up

If we take the variation of all the sample points and when I say sample points this is across the five groups and we sum it up. We will get total variation and ideally that should add up to between group variation and within group variation.

How do we check this? These are all our sample points. If I calculate an average across all the sample points, this is my grand mean. Now all I need to do is I need to take a difference of each point from the grand mean. Let's do that here. I am going to simply take a difference of every sample point from the grand mean and I am going to square the difference.

I am going to absolute reference this cell. This is the grand mean. This is not going to change. If I just drag this all the way across and then down I am going to get the individual differences, the sum of square differences. If I sum this all up this is going to be the total variance in the sample that we have. This should add up to between group variation and within group variation.

If we have calculated between group variation in the between group variation tab and we have calculated within group variation here. So if I add it up, this number and this number are exactly the same. Because remember in an ANOVA all we have done is we have partitioned the total variation into

between group variation and within group variation.

>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>

Let's quickly review some assumptions for us to apply or use an ANOVA correctly.

1. The populations from which the samples were obtained must be normally or approximately normally distributed
2. The samples must be independent of each other
3. The variances of the population must be equal

If the variances of the population are not equal, the outcome, the p value may not be very reliable.

Remember an ANOVA is used as a multiple sample test when the dependent variable is essentially nothing but the outcome variable. What we are interested as the outcome is a continuous variable and the independent variable are the factors are discrete. We will using the term dependant and independent extensively in this course. But in our example of the shelf height against the sales, the outcome variables that we are interested in was the sales variable. That was the continuous variable. The independent variable are the factor that was influencing the sales was shelf height and that is a discrete variable with five levels. Shelf height can have 1, 2, 3, 4, or 5 heights.

>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>

Of course in real life, it is not always the case that they will be a single factor that influences the outcome variable.

>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>

Let's look at a more complex example, where they may be two factors influencing the outcome. So if we extend the same example, we were looking at it could be that the sales of the product don't depend just on the shelf height, but may be also the aisle placement. In an organized retail store, there are many aisles and many products stocked on an aisles. You could say that depending on which aisle the product is in and, which shelf height the product is in, it sales could be different.

Imagine that we had a store that had five aisles and five shelf heights. Then the total number of combinations potentially that you could test are 25. If we had three aisles and five shelf heights, then the combinations that we would test are 15 different placements for brand A. So how do we determine if the mean sales rate are different between the groups.

>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>

What we have run is the example of two-way ANOVA. A two-way ANOVA is useful when we want to compare the effect of multiple levels of two factors and we have multiple observations at each level. One of the interesting things about a two-way ANOVA is that we will actually be testing

three null hypothesis in a two-way ANOVA. How is that?

- The first null hypothesis says the impact of factor one on the sales is the same or there is no difference in the sample means because of factor one
- The second null hypothesis says there is no difference in the sample means because of factor two. These are very straightforward.

Where does the third hypothesis come from?

- The third hypothesis says that the combination of the first factor and the second factor has no impact on the sample means. This is called an interaction hypothesis

Remember when you have two factors, we actually have three effects. There is the effect of factor 1, there is the effect of factor 2, and then there is the combination of the effect of factor 1 and factor 2 not having an impact on your dependant variable.

>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>

How do we interpret the output? Since we have three null hypothesis, we will get three p values. Usually you want start by looking at the interaction p value. If the interaction p value is significant it implies that the combination of factor one and factor two are important. Therefore even if the individual p values are not significant, the combination is significant.

If the interaction p value is not significant you may want to rerun the ANOVA, but not include the interaction term in your analysis.

>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>

Let's do this using an example. This example we are going to look at is there a difference in calories burned in exercising based on whether you stretch before the exercise and whether you wear weights during the exercise. So the calories burned which is being captured in the third column here is your dependent variable and we have two factors here. Stretching and ankle weights.

Each factor has two levels. Stretching has no stretch and stretch and the ankle weight has no weights and weights. Essentially we want to look at combinations of each of these two factors on whether or not there is a difference in the calories burned during exercise. If I don't stretch and I don't wear weights what sort of calories do I burn? If I do stretch but I still don't wear weights what are the amounts of calories that I have burned and so on.

This is an example also of a two-way ANOVA with replication. Replication essentially means that you have multiple observations for each combination of factors. So notice here that for every combination of factors, no stretch and no weights, stretch and no weights, no stretch and weights, stretch and weights. We have four observations. So this is an ANOVA with replication. You could also do an ANOVA sometimes without replication where

you have only one observation for every combination of factors. But we are going to run an ANOVA with replications.

>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>

The calculation for a two-way ANOVA also uses between group variation and within group variation. But obviously the calculations will be even more tedious. What we will simply do is to use excel to run a two-way ANOVA. In excel, we have to go to data analysis, and choose ANOVA two factor with replication.

Another thing that you have to remember is that to run a two-factor ANOVA in excel, the data has to be arranged in this manner. So you have one factor in the first column: no stretch and stretch, you have the second factor as a row options, and then you have the combination of the observations appropriately listed in this manner.

This is our data here. What I have done is I have listed the data in this manner. No stretches and no weights are these observations that I have listed here. No stretch and weights are these observations that I have now put here. You have to arrange the data in this manner. Once the data is arranged all we have to do is go to data, data analysis, and choose Anova two-factor with replication. Because we have multiple observations for every combination of factors. Say ok. We have specified our input range here. Rows per sample, this is the number of observations you have per sample.

Remember in our example, we have five. Because we have five observations for every sample and the sample is essentially the combination of factors. So I am going to leave alpha at 0.05 and I am going to say ok, then we get output. If you look at the output, as usual in the beginning we get the summary statistics for each sample.

So for every combination of weight, no-weight, stretch etc., we get summary statistics and finally we get the ANOVA output. Remember this is a two-way ANOVA which means we are going to get three p values because we are testing three null hypothesis.

The first null hypothesis sample refers to stretch or no-stretch column. The second null hypothesis columns refers to no weights or weights. The third null hypothesis interaction talks about the combination of the first factor and the second factor. You can see here that eventually we are interested in the p values. The combination is not significant but the individual factors are significant.

What is our conclusion based on this ANOVA output? Whether or not you stretch before an exercise, has a significant impact on the calories burned during the exercise and whether or not you wear ankle weights while exercising has a significant impact on the calories burned during exercise?

>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>

However the combination of stretching and ankle weights does not seem to have additional impact on calories burned during exercise. That is how we would interpret the output.

>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>

This is how we run a two-way ANOVA in excel and interpret the output. One final thing to remember at part ANOVA. For example when we interpret the output here and we say that stretching before exercising has the significant impact on calories burned. Now we are rejecting the null hypothesis here. However the ANOVA is simply says that there is a difference in the sample means. But it does not tell us which of these is significantly different?

If we look at our source of variation like stretch, no-stretch, weights, and no-weights, The ANOVA is telling us that there is a difference in the sample means. But it is not telling us whether no stretch is different, stretch is different, or weights are different. We don't really know which of these sample means is significantly different. The ANOVA simply tells us that there is a difference.

>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>

But very often we want to know which ones are really different and for that you have to do what are called Post Hoc tests.

>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>

In an ANOVA, remember if you reject the null hypothesis the only thing that we understand from that rejection is that at least one group mean is unequal. We really can't tell from an ANOVA output which of those group means and how many are unequal. For that we have to run Post Hoc tests.

There are many post Hoc tests like Tukey tests and LSD tests etc. and essentially what they are used to do is if an ANOVA we find has statistically significant outcome, we run the post hoc test to figure out to which of the group means are unequal.

>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>

We looked at what is ANOVA, how to do the basic calculations of an ANOVA, how to run one-way ANOVA, and two-way ANOVA in excel. We also understood that sometimes we have to run post hoc test on the ANOVA output. If we are interested in figuring out which of those group means is really unequal.

What we will do next is look at Chi square test which is another example of a multiple sample test.

>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>