We were reviewing how an ANOVA works and an ANOVA is used for a multiple sample mean testing.

>>

In an ANOVA we used variance calculations to reach a conclusion about sample means. We calculate what are called within group variation and between groups variation.

Remember within group variation is the variation of each unit in a group against its group mean and between group variation is the variation of each group mean from the overall mean.

>>

Within group variation is calculated in the following manner.

- Calculate the mean for each group
- Subtract each sample mean from every unit in that group
- Square the difference
- Add up the squared differences

Which mathematically can be written us

$$SSW = \sum_{K} \sum_{I} (Y_{ik} - \bar{Y}_{K})^2$$

>>

Between groups variation is a simpler calculation. What we do here is

- Calculate a grand mean which is simply the mean of all observations across all the groups
- Subtract each group mean from grand mean
- Square the difference
- Multiply each squared score by the sample size
- Add it all up

Notice what is means is that your sample size for each group doesn't necessarily needs to be same. Meaning that we could have 5 groups and each group could have a different sample size but that not would be a problem to running an ANOVA.

>>>

What we now have is sum of squares within and sum of squares between. We now need to divide each quantity by its appropriate degrees of freedom. Remember we had said degrees of freedom for within is n-k, where k is the total number of groups and n is the total number of sample points. Degrees of freedom for between is k-1 where k is the total number of groups.

>>>

Once we do that we will now be able to calculate an f-statistic or a test statistic for an ANOVA. In our retail example, we had looked at shelf height impacting sales. We said lets place a product at different shelves for ten days each and look at the sales every day. So when we did this, these are the sales that we observed for ten days on each shelf.

The average is highlighted in orange and we can say that the averages are different for each group. What we are trying to do is we are trying to check whether or not these differences in averages are because of shelf height or are they simply because of random chance variation. So let's do the calculations in excel.

This is our data. Let's start with sum of squares with them. The first thing that I am going to do is to calculate the averages for each group, which is simply the average of ten data points. So these are my group averages. Then I am going to take a difference of each unit from its group average. So (210.5-179.92). Similarly (198.1-179.92). Again for each unit and its group mean.

So these are the differences for group A, similarly differences for the second group, third group and so on. Then we square the differences. Because this is a variance calculation. So we take the differences and we square the differences. The sum of all these differences across all the groups is my total sum of squared differences within. So this is simply the sum of squared differences of each unit from its group mean. So the sum of squared differences is 34735.

What about sum of squares between? For sum of squares between we need to calculate a grand mean. A grand mean is simply the average across all the data points in all the samples that we are testing. So the grand mean is 181.54. Now we take the difference of each group mean from the grand mean. So each group mean (179.92-181.540), (185.09-185.54). so the difference of each group mean from the grand mean then we square it up and we multiplied by sample size.

When we take the differences remember that we are squaring the differences then we multiply that with a sample size. In this case my sample size is 10 for each group. But remember you could have a situation where your sample size could be different across groups. Then we sum all of that numbers and we get 250.71.

>>

Therefore our sum of squares within is 34,735, our sum of squares between is 250. Now we have to divide these by the degrees of freedom. So degrees of freedom within are n-k. So in this example, 50, total is my sample size and 5 is my groups. So 45. So mean square within is 771.88, simply 34735/45.

>>

Similarly sum of squares between is 250.71
Degrees of freedom between is (5-1), (k-1) = 4
Mean square between is 62.7.

>>

Now we simply take a ratio of mean square between and mean square within and we will get a value of 0.08. You could look this up in an ANOVA table. This is an example of an ANOVA table.

There are numerator degrees of freedom at the top and vertically running down the page are denominator degrees of freedom.

In our example the numerator degrees of freedom is 4 and the denominator degrees of freedom is 45. There are two sets of numbers here. The bold numbers are for α=0.01 and the light numbers are for α=0.05. If we use an α of 0.05 and we look at the intersection of 4 numerator degrees of freedom and 45, which we don't have so 46, denominator degrees of freedom. We get a critical value of 2.57.

We compare this critical value to our distance. Remember the F-statistic here is the distance measure. It is not a p value. So we compare 0.08 to 2.57. We will reject the null hypothesis only if our calculated distance is greater than 2.57. In our example we failed to reject the null hypothesis. Because our calculated distance is much closer to the mean than the critical distance.

What does it mean to say we fail to reject the null hypothesis? Essentially we cannot conclude that height hasn't impact on sales. The variation that we see is simply random chance variation. Of course instead of doing all these calculations manually, which are tedious and of course prone to error, what we could do is to simply use the tool to do the calculations for us.

>>

For example in excel, there is an option in data analysis called Anova: single factor. I will explain why we are using single factor a little later. But for now, let's just run this in excel. Go to data, data analysis, and choose Anova single factor.

>>>

If you click ok, you will get a menu like this. Excel asks us for an input range. For us the input range is these data points. Now grouped by. Sometimes your samples could be in row and sometimes in columns. In our particular example, each sample is one column. Therefore I have chosen group by columns. Because I have labels in the first row, I have used a tick mark for labels in the first row and I have left the default alpha to 0.05. If you say ok you will get ANOVA output in a new work sheet.

>>>

This is what the ANOVA output looks like. The first table is simply summary statistics. It gives us the averages and the variances for each group. The second table is the actual ANOVA calculations. Of course like in every hypothesis test what we are really interested in is the p value. The p value is 0.98. So with the p value of 0.98, are we going to reject the null hypothesis? No, we failed to reject the null hypothesis.

All the calculations that we had done are reflected here. The within group variation 34735, between group variation 250.1, degrees of freedom between, mean square between, and the f statistic. We had calculated all of these manually. But of course we don't really need to do that. Because if you have access to a tool, the tool will do all the calculations for you. But of course it's the validation of the calculation that we had just in and the p value now is 0.98. Therefore we failed to reject the null hypothesis.

Let's done this in excel. We can see this in practice. Data, data analysis, Anova single factor, and you say ok. This is my data. Remember we don't want to include the means in the data. Grouped by columns, labels in the first row, alpha is 0.05, and say ok, you will get an output that looks like this.

Remember what we are interested in is the p value of 0.98. When we look at a p value of 0.98, we know we cannot reject the null hypothesis.

>>

Our business conclusion finally is that we cannot reject the null hypothesis that all the means are equal. In other words we cannot conclude based on this data for this product that shelf height has an impact on sales. We will conclude that shelf height does not have a statistically significant impact on sales.

>>>