We have been reviewing hypothesis test and we have looked at many hypothesis tests which are single sample or two sample. In a single sample test, we look at a sample outcome and compare that against an expected population outcome. In a two sample test, we compare means between two samples. But many times in real life you may end up having to compare means across more than two samples. Those are called multiple sample tests. When we run a test, a hypothesis test to compare means across multiple samples.

>>>

We will review two kinds of multiple sample tests. One is ANOVA and the other is called Chi square. In ANOVA what we will review? A one way ANOVA, two way ANOVA, and what are called Post hoc tests.

>>>

When we have multiple samples one way to test the hypothesis for difference in means across multiple samples is to run multiple t-test. For example, supposing we had three samples, and we wanted to test whether the sample means are the same. One way to do it could be run multiple t-test. I could check

Mean 1 = Mean 2, one t-test;

Mean 2 = Mean 3, second t-test;

Mean 1 = Mean 3, third t-test.

This is possible, but there are two issues with this approach.

1. You will have to run multiple tests
2. Come up with a single outcome which is all the means equal across multiple samples

There is a lot of work involved. Every time you run a hypothesis test, there is some chance of error. Remember error is the chance that you reach the wrong conclusion. In any hypothesis test we can never be 100% sure about our conclusion. There is always some chance that we are reaching the wrong conclusion.

Therefore the more test you run, the more the chances are that we are introducing error in the outcome that we finalize. Therefore it is better for us to figure out a test where we can check all the sample means at one time instead of running multiple tests and an ANOVA is one way of doing that.

>>>

To understand how ANOVA works let's look at an actual business situation. Imagine that you work for a big box retailer and you want to understand does shelf height have an impact on sales. Most of us intuitively know that if for example a product is stocked at high level, it may be more easily accessible to customers, they see it and therefore they buy more.

If the product on the other hand is stored at the highest shelf or the lowest shelf, the customers may not see it as much and therefore may be sales at lower. Now of course this varies from product to product, category to category. Some products people are going to buy because they need them irrespective of where they are stored and sometimes heavy products are stored at the bottom shelf.

So it is not necessary that the impact is same for every category or that there is an impact for every category. But certainly for some products, you could argue that where the product is placed in terms of height may have an impact on sales.

Supposing you are a retailer and you have a particular product, and you want to test this theory. One way to do that would be, let's say you have isles where you store your product and each isle has five shelves. One way to test this would be, take a same product, store it at different shelf heights for ten, ten days each and look at the sales.

If the sales are different then maybe you can conclude that height has an impact on sales. If the sales are same then maybe you conclude that height does not have an impact on sales.

>>>

Let's say we run this for a particular product. We stored that product at different shelf heights for ten days of each and we look at the total sales achieved per day. Let's say that this is what the data looks like. Of course there are other things that we must take care of. The ten days that we stock the product in each shelf must be similar. Meaning if we are starting on Monday and ending on Wednesday or Thursday then for every shelf you must to do the same.

There should be no other promotions are happening at the same time. All the other things that impact sales of this product must take constant. So assuming that all of that is taken care of and the only difference in this ten days is simply the product placement at different shelf heights. Then what we can do is we can look at these average sales at different shelf heights and you can see that there are differences in the average sales and we can test whether or not the observed differences are significant.

Remember sales is a random variable which means you will see variation in sales even if there is no underlying cause simply because of random chance variation. So what we will do with an ANOVA is run a test, a hypothesis test to check if the variation that we see across these five samples is significant or not.

>>>

Analysis of variance (ANOVA) is a mean test. Even though it is called as analysis of variance, remember we are testing differences in sample mean outcomes. Essentially we use variance calculations to reach a conclusion about means. That is why this is called analysis of variance but the outcome is still a test about sample means.

There are two variances that we calculate in an ANOVA. Something called a within group variance and something called a between groups variance. The total variance in all the observations that we use for an ANOVA is called the overall variance and the overall variance is calculated just like a variance calculation. Take a total mean of all the observations and sum the square differences of each observation from the total mean. That sum is the overall variance, the total sum of squares and also called SST.

Within group variation is the sum of square differences between each observation and the mean for the group that it belongs to. So in our example, there are five groups, five samples. If you take each group mean and then you calculate the variance of each unit in a group against its group mean that is within group variation.

Between groups variation is the sum of squared differences between each group mean and the overall mean. So essentially what we are doing is we are dividing overall variance into two types of variance. We are saying overall variance is within group variance + between groups variance.

>>

It turns out intuitively that how an ANOVA works is mathematically there are two ways of estimating the standard error of the mean which is essentially a measure of variance in a sample. One method uses sum of squares within and the other method uses sum of squares between. Remember both these methods are estimating the same number which is the standard error of the mean.

It turns out that if the sample means are similar then whether you use approach one or approach two, you should get very similar numbers. If however the sample means are very different then when you use approach one, you will get a different estimate of the standard error of the mean then when you use approach two.

What an ANOVA does is that it looks at a ratio of sum of squares within and sum of squares between. If they are similar the ratio is close to 1, then we conclude that the means are not different. If the ratio is not close to one, then we conclude that the means are different. So that's intuitively how an ANOVA works.

>>>

Another way of looking at an ANOVA intuitively is to say any observation in an experiment can be broken down into

- The overall mean of all the observations.
- + (or -) how far the average of the group is from the overall mean
- + (or -) how far is an observation from the average of the group

If the sample means are not different because the independent variable are the factor that we are using in this case shelf height has no impact. Then the within group variation and between group variation should be similar and any differences will be small and because of randomness. However if the variations are not similar it means that the factor that we are changing in this case shelf height has an impact on the outcome.

So this is intuitively how an ANOVA works.

>>>

The test statistic for an ANOVA is essentially the ratio of mean square between/ mean square within where mean square between is simply sum of squares between/degrees of freedom between. Mean square within is sum of square within/degrees of freedom within.

To run an ANOVA therefore, we need to calculate four quantities.

Sum of squares between. You can see the formula here.

$$SSB = \sum_{k=1}^{K} N_k (\overline{Y}_k - \overline{Y})^2$$

This is between groups variation.

Within group variation,

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

Which is each group summed over all the groups.

Then we also need degrees of freedom between and degrees of freedom within. Degrees of freedom between is simply k-1, where k is the number of groups. In our example we have five groups, so degrees of freedom between will be 4.

Degrees of freedom within is n-k, n is the total sample and k is the number of groups. So total sample in our example is 50 because we had five groups, 10 days each. So 50 observations – 5 groups. So n-k for us will be 45.

>>>

The test statistics follows an F-distribution and F-distribution is a continued distribution which depends on two degrees of freedom. There is a numerated degrees of freedom and a denominator degrees of freedom. In fact an F-distribution was the two Chi square distributions.

A Chi square distribution again is a continuous distribution which is generated as a square of a normal. Remember for practical purposes what we need to know is that the test statistic that is generated from an ANOVA calculation follows what is called as an F-distribution. F-distribution is a continuous distribution that requires two degrees of freedom, a numerated degrees of freedom and a denominator degrees of freedom.

>>

In an ANOVA, what is the null hypothesis? The null hypothesis will be that all my sample means are actually equal and that is any variation is simply because of randomness. What about my alternate hypothesis? The alternate hypothesis is that at least one pair of means are unequal. This is important to understand. The alternate hypothesis does not require that all the sample means be different from one another.

Remember the alternate is the negation of the null. The null says all my means are equal. The negation will be even if one of the pair of the means are unequal that would be negated. So the alternate says at least one of my means is unequal. The test statistic is the ratio of within group variation to between groups variation that we have just say.

>>

What we will look at in the next step is how we actually calculate the test statistic for ANOVA.

>>