Let's now look at what are called two-sample tests. It turns out very often in business situations, what we end up doing a checking the differences in the average of two samples.

>>>

The examples that we have looked at so far have been what are called single sample test. Essentially we have a single sample from an underlying population. We compare the sample mean to the population mean. But actually if you think about it very often we don't have access to the population, what we have is samples.

For example, supposing that there is a company that has been facing a lot of customer complaints about how long calls take to be resolved when the clients contact the customer care center. So a Senior Executive has been asked to figure out ways of reducing call resolution time when customers call.

So may be like a six sigma project where there is a look at all the processes, all the steps that happen, when a customer calls the customer care center and there has been a project that has determine here are some ways where we can cut time and make the calls more efficient when customers call.

In order to make sure that this approach is working, what we may want to do is look at an average of the call resolution time before project has been implemented and look at the average of the resolution time after an improvement project has been implemented.

If you are think about it we are taking a sample calls before we do an improvement project and then we take another sample of calls after the project has been implemented. We want to make sure that we see a reduction in call resolution time in the second sample relative to the average in the first sample. So that is a two-sample test.

>>>

Supposing that this is the data that we collected. We took a random ten calls and we looked at how much time each call took before the project was implemented. We get an average of 8.75 minutes across these ten calls. Then we take another random ten calls after the project was implemented and we look at an average and we see 7.3.

Now what we want to do is to figure out is the reduction that we see in the second sample statistically significant? Remember the call time for any random customer is a random variable and so when you see a reduction in a sample it may not necessarily be because there is a difference in or an improvement based on the project. It could be simply random chance variation. Before we conclude that yes, this project has worked we may want to be sure about how confident are we that this is not because of random chance variation.

In most business situations, typically this is the kind of hypothesis test that we will implement. We will be looking at differences across samples rather than really checking a sample against a population outcome.

How do we run a hypothesis test in a situation like this? Again the framework of hypothesis test is the same. But obviously because we are looking at the difference in sample means across two samples we have a different test statistics.

>>

The test statistic in this case, this is a two-sample t-test and the test statistic is captured mathematically as the difference in the sample outcomes.

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Where n1 and n2 are the sample sizes.

This assumes that the variance between the samples are equal. Sometimes the sample variances are unequal is which case the test statistic looks a lot more complex. However we really don’t have to remember these test statistics because we will implement the two-sample t-test in whatever tool that we are using.

>>

We could mathematically calculate the test statistic assuming equal variance by putting this values in the formula that we just saw. But what we will do is, we will run this in excel and see what the test statistic is.

This is our data. Average call time before implementation for ten calls. Average call time post implementation for ten calls. To run this in excel, we need the data analysis add in. If we go to data, data analysis and choose t-Test Two-Sample Assuming Equal Variances. At this point we don't know if the variances are equal or not. We could of course calculate them and find out.

But let's just run the test assuming equal variance because excel automatically calculates it when we run a t-test. When you do this you get a menu that ask for the range of variable one. Variable one is sample one. The range of variable two which is sample two. Notice I am not including the averages in the range. Because the average is the calculated measure not the data.

Now excel is asking for a hypothesized mean difference. Remember in a hypothesis test we have a null hypothesis and an alternate hypothesis. What will be the null hypothesis in this particular example? It will be that there is no difference in the average call time post the implementation. So the difference that we are testing is actually a difference of zero between sample1 and sample2. Therefore the hypothesized mean difference is zero.

We have labels. So I am going to use labels and we are going to leave alpha which is our level of significance to the standard 5%. If you say ok you can see excel has generated a lot of output. It has generated summary statistics for the samples. The averages, the variances, and the number of observations etc.

What we are really interested in a hypothesis test is a p value. In this particular example, what is the alternate hypothesis? The null hypothesis is that there is no differences in the call time after the project is implemented. The alternate hypothesis most likely will be a one-tailed test that there is a reduction in the call time post the implementation.

Therefore if you are doing a one-tailed test, we are interested in the p value of a one-tailed test. As you can see here, excel has calculated a p value of a one-tailed test. So this is our p value. Now what do we do? We compare the p value to the level of significance. The level of significance is 5%, 0.05. So what conclusion should we reach? Because the calculated p value is greater than the level of significance we cannot reject the null hypothesis.

Remember we will reject the null hypothesis only if the p value is less than alpha. We wanted to be 95% confident that there is a difference. With this calculated p value we can only being 92% confident and therefore we will not reject the null hypothesis. What does that mean from the business perspective? We cannot be 95% confident that there is a statistically significant reduction in the call time post the implementation of the project.

In other words, the project has reduced call time but not to the level of confidence that we required which is 95%.

>>>

If you look at the variance calculation the variances are not equal. Therefore what we can do is instead rerun this t-test assuming unequal variances. If you go back to the data in the data analysis tab, notice that for t-test there are two options, t-test assuming equal variances and t-test assuming unequal variances. If the difference in your variance between the two samples is more than 15% you ideally should be rerunning the test with unequal variances. So if you rerun this, it is the same menu.

Just to be precise, we will rerun this with unequal variance t-test. Variable 1 range, Variable 2 range, hypothesis mean difference is zero, we have labels, and we say ok. Again what we are really interested in is the p value of a one-tailed test. Even when we rerun this, the p value is not changing very much and therefore we will not reject the null hypothesis and conclude that the difference in the call time is not statistically significant at a 95% level of confidence.

>>>

In the previous example, we compared two samples that had equal observations. However it is not necessary that the two samples have to have equal observations. They could be unequal observations between the samples. For example, the first sample could be 10 calls and the second samples could be 15 calls.

We are allowed to use samples with unequal observations. But we have to make sure that the degrees of freedom that we use is one less than the sample size of the smaller sample. We assume similar variance across the two samples and run an equal variance test. But if the variance is not similar and you can get the output in the test then just simply rerun the test using unequal variance in excel.

>>>

Let's now look at one other kind of a t-test. In the previous example, when we looked at the two samples and the calls in the two samples it was not the same customer's. We looked at ten random calls before the project was implemented and ten random calls after the project was implemented but it is not the same people calling us before and after. In some situations it may be the same person.

For example, let's look at a test for a weight loss drug. Now if we think about that how we would implement a hypothesis test to test whether or not the drug leads to weight loss. Ideally the way we would do it is select the group of people, check their weights when they enter the test, give them the drug for whatever the assigned protocol time is 2 months, 3 months etc., and then check their weight after they take the drug for that specified length of time.

It cannot be any ten people before and any ten people after. It has to be the same people. In other words the sample units are the same in both the sample. In such a situation we are doing what is called a paired difference t-test and by design you have to have the same number of sample units in both the sample.

>>>

The test statistic for a paired difference t-test is

$$t = \frac{\bar{d}}{\sqrt{s^2/\ n}}$$

Something that is easily implemented and calculated using a tool. In this particular example we want to test the hypothesis that drug leads to weight loss and we want to be 95% confident about the outcome.

This is our data. To run this test in excel, go the data analysis tab and choose paired two sample for means. Remember this is a paired t-test. Again you will get a menu variable one range which is the weight-pre and variable two range which is the weight-post, and the hypothesis mean difference. In this example what will be the null hypothesis? That the drug does not have any impact on weight loss and therefore the difference between the averages is zero. I have labels. Alpha is 0.05 and say ok. Again we had a lot of output.

What should be the alternate hypothesis? If you really have no idea about the drug at all, we may want to look at a two-tailed test. But most often by the time we get to a drug testing stage there must be some reason for believing that a particular drug has the certain kind of impact.

If we run a one-tailed test that drug does have an impact on reducing weight then I would look at a one-tailed p value. In this example, the one-tailed p value is 0.03 and significance level is 0.05. Therefore since the calculated p value is less than the level of significance we can conclude that the drug leads to weight loss at a level of confidence of 95%.

Remember there is a 3% probability that what we are seeing is driven by random chance variation and 3% is less than the 5% cut-off probability that we are willing to work with. Therefore we will reject the null hypothesis and conclude that taking the drug leads to loss in weight.

>>>

One final thing to think about. There is an option called hypothesized mean difference when we do the two-sample test. We have used a hypothesized mean difference of zero. But sometimes you may be able to test the hypothesized mean difference of greater than zero.

>>

For example you may say that in my call center example, my project improvement plan will reduce call resolution time by more than 15%. So it is not necessary that we are not always checking the mean difference to be zero. You may say that my call resolution time is reduced by more than one minute on average. In that case, the hypothesized mean difference could be one.

Depending on the business situations, it is possible that the hypothesize that we are checking may be for a difference greater than zero. In that case you will simply use the right difference in the hypothesized mean difference option. That is how two-sample test for. We looked at independent sample t-test which is the sample units are not the same between the tests and paired difference t-test which is that the sample units are not the same between the two tests. Both of them are t-tests and there are test statistics that we can calculate and use distribution which is a t-distribution. We can of course do this test manually but most often it is just easiest to implement this test in tools that are available to us.

>>