We are reviewing Chi square tests and remember Chi square tests are used when we have to count the data and they are a multiple sample tests. We looked at a very simple example of observed an expected frequencies.

But now let's look at a slightly more complex example. This particular application of a Chi square test is called an association test. A Chi square test of association or a Chi square test of independence.

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

The example says that we want to understand the preference for a carbonated beverage by age and I want to understand if there is an association between brand preference and age. If I understand that the reason association then I might come up with the certain kind of differentiated marketing strategy that is based on age.

If I figure out based on my data that there is really know association between age and brand preference then age will not really be factor in my marketing strategy. How do I figure out whether or not there is an association? I can collect some data.

Let's say that I take a random sample of men and these men are between the ages of 15 and 55 and divide them into three groups. Males between the ages of 15-25, 26-40, and 41-55. If for each group I ask them to record their preference for Coke vs Pepsi vs Sprite. In the age group of 15-25, 49

people say they prefer Coke, 24 say they prefer Pepsi, and 19 say they prefer Sprite in the total of 92 people.

Let's say between the ages of 26-40, 50 say they prefer Coke, 36 say Pepsi, and 22 say Sprite. Between the ages of 41-55, 69 prefer Coke, 38 prefer Pepsi, and 28 prefer Sprite. I want to use this data to say does the preference for brand change by age. In other words, is there an association between brand preference and age group?

Remember this is all count to data. Therefore in order to figure out whether or not there is an association between age and brand preference I can run a Chi square test.

Why is this a hypothesis test? Because we want to check does the brand preference change by age or is changes in brand preference that we see in the data simply because of random chance variation. That is why we want to run a hypothesis test, because remember the percentages of preferences for Coke vs Pepsi vs Sprite are changing by age group. But is that change statistically significant?

Now if we want to run a Chi square test. What is our first step? We need to calculate expected values. We have observed values. We need to now calculate expected values. Remember expected values are calculated as what should I have seen in the sample if the null hypothesis was true. In our particular example, we want to test is there an association between brand preference and age.

In other words, the null hypothesis is there is no association between brand preference and age. So we need to calculate expected values and say if they really was no variation in brand preference because of age then what are the expected values should I see in the sample.

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

There is a mathematical calculation for these expected values. You can get them by doing

Expected Values = (Row Total * Column Total) / n

Where n is the sum of the observations in the sample. We can see these are the observed numbers and we need to calculate expected numbers. The way to do that is to do row total, times the column total divided by the sum across all the observations.

I have calculated that for the first set says that if the null hypothesis was true, I should have expected to see 46.1 preferences for Coke. What about for Pepsi? Remember we want to do row total which is 98, times column total which is 92 divided by the sum across all the observations which is 335. Similarly for Sprite row total is 69 times column total divided by 335.

We do that for the second column and it now becomes the (168 * 108)/335, (98 * 108) / 335, and (69 * 108) / 335. Finally the third column, (168 * 135) / 335, and of course we have always use an absolute reference cell formula, but I am just

showing you to calculations here. We can do it faster by using a formula.

Now that we have the expected values for all these brands. Notice that the row totals and the column totals for the observed and the expected all sum up to exactly the same thing and that should happen. Because remember if we are simply saying the null hypothesis was true and there was no association and this is how I would have expected to see the preferences for the same number of people.

One check to make sure that your expected values are calculated correctly is to look at a row total and a column total for the expected values. They should be exactly the same. Now of course we have observed values and expected values, we want to figure out are these differences in observed and expected statistically significant.

Remember we can simply do in excel. We can simply say chitest (observed values, expected values) and we get a p value of 0.824. What should be our conclusion? Given the p value of 0.824, can we conclude that there is an association between brand preference and age? We cannot reject the null hypothesis and therefore we conclude that based on this data there is no statistically significant association between brand preference and age. That was an example of what is called a test of association.

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

There is also another popular application of Chi squire test which is called a goodness of fit tests. What does the goodness of fit test do? It checks whether or not some data that you have follows and expected distribution. Here is an example.

Supposing that a gambler is playing a new game at the casino, which involves rolling three dice at a time. Winnings are directly proportional to the number of 6's rolled. If you look at the 100 rolls, this is the number of 6's that we have rolled. 48 times he has rolled zero 6's, 35 times out of 100 he has got an one six, 15 times out of 100 he has got two 6's, and 2 times out of 100 he has got three 6's.

What are we do with this information? We need to figure out are these valid outcomes or not. Is that gambler too lucky? Remember the winnings are directionally proportional to the number of 6's rolled. But how do we find that out?

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

When you role a dice whether or not you get a 6 follows a particular distribution. It follows a binomial distribution. If we know that the number of 6's rolled follow a binomial distribution, then we can actually calculate what should we expect to see in the number of 6's when we role three dices 100 times. Remember these are observed values but we can calculate expected values under a binomial distribution. How do we do that? I can use the binomial distribution formula to calculate

the probability that I will get zero 6's when I role three dice at a time.

Let's do that in excel. The expected number of 6's in 100 throws. Now we can use a binomial distribution to calculate that I will get zero successes which is zero 6 in three trials. I am throwing a dice three times, a probability of me getting a 6 on any one dice is 1/6 and false. So on any throw of three dice, there is a 57% probability that I will get zero 6's. Similarly what about one 6.

If I drag this down and I update the cell instead of the number I will get the probability of one 6, two 6's, and three 6's. Remember this is in one throw of three dice. If I am interested in the probability of the number of 6's in 100 throws of three dice then I can simply multiply this with 100. These numbers are the number of success I would expect to see because it follows a binomial distribution.

Now notice we have observed values which is what the gambler is getting and these are expected values. This is what we should have expected to see because when you throw a dice, the outcome of 6 follows a binomial distribution. Therefore now we just need to look at a difference between observed and expected values. For that we just need to do a Chi square test.

We have observed values, expected values and if we use the Chi test function we will get a p value of 0.001. What is this mean? Remember we are checking whether or not the distribution follows a binomial distribution. My null hypothesis is

distribution follows a binomial distribution and which distribution is that? It is the observed values. My alternate hypothesis is that this distribution does not follow a binomial distribution with a p value of 0.001.

We will reject the null hypothesis that this observed data distribution can be the output of a binomial distribution. In other words this is not something that we would expect to see from a binomial distribution. This data does not fit a binomial distribution. Therefore the gambler is probably playing with loaded dice. So this is an example of how we use Chi square to test for goodness of fit.

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

Of course we can check for goodness of fit in any distribution. For example, if I had some data that I wanted to check for normal distribution, I want to check does my data follow a normal distribution. I can still use the goodness of fit. How do I do that?

1. I take my data, I calculate the mean and the standard deviation of my data
2. Bin the data into sub-intervals
3. Based on the mean and the standard deviation I calculate expected probabilities of those sub-interval values using the normal probability function
4. Compare that frequency to what is observed in the data
5. Use the expected and the observed to construct the Chi square test and check

whether or not that data follows the normal distribution

Remember when we do goodness of fit test we are essentially checking does the data follow a certain distribution. We have some data. We use the parameters of the data to come up with the expected values and then we compare the expected values and the observed values and decide whether or not that data follows that distribution.

This is an application of a Chi square test on a goodness of fit test. But remember it is still your standard Chi square test that looks that count the data and the difference between observed and expected values.

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