We have reviewed how to run hypothesis test using a single sample. One of the special cases of a hypothesis test is when the distribution is not normal.

>>

Because we know the Central Limit theorem, we do understand that when we take multiple random samples from an underlying population, the distribution of the population does not have to be normal. So many times, we end up using normal distributions to calculate P values in a hypothesis test. But it is not necessary that every hypothesis test must use a normal distribution. If you know the underlying population distribution, you may be able to use directly the right the distribution for the hypothesis test.

For example, let us say a manufacturer claims in a TV ad that 2 out of 5 people prefer their washing powder to any other brand. A random sample of 25 people resulted in 4 people expressing preference for this brand. Is the manufacturer's claim justified? Test at 95% level of confidence. Now let us think about this for a minute. We want to test the hypothesis that 2 out of 5 people prefer this washing powder to any other brand. If you look at the data itself, the outcome preference is a random variable. This random variable has two outcomes. Either someone prefers this washing power to other brands or they do not. Therefore, the distribution of this random variable outcomes is actually binomial.

So if I wanted to calculate how likely is it that in a random sample of 25 people, I will see 4 people with this preference when I was expecting 2 out of 5 people with the preference. I can actually use binomial distribution to calculate a P value. What would be the null hypothesis here?

Null hypothesis - H0: Brand preference is 40% - 2 out of 5
Alternate hypothesis - H1: Brand preference is less than 40%

We have been asked to use a 95% level of confidence; therefore, the significance level is 5%. Now we have to calculate the P value.

In order to calculate a P value, we need test distribution and we can use a binomial distribution here.

If I calculate the P value using a binomial distribution, I can say what is the probability of seeing <=4 preferences out of 25 when I was expecting a 40% preference rate.

So P = Binomial distribution(4,25,0.4,TRUE)
Where,

- 4 - Outcome that is observed
- 25 - Number of binomial trials
- 0.4 – probability of success

Remember 40% expected success rate in the population (2/5) and TRUE to give us the cumulative probabilities.

In hypothesis testing, we are looking typically for outcomes as extreme or more extreme than the observed outcome. If we do this calculation, this P value is 0.009. Given this P value and the level of significance of 5%, what should be our conclusion? We should reject the null hypothesis, because the calculated P value is lesser than the level of significance. Therefore, we should conclude that the preference rate is not 40%; the preference rate is unlikely to be 40%, it is lot lower than the 40%.

>>>

We could, of course still use a normal distribution, if we wanted to do that. It turns out you can approximate a binomially distributed random variables to normal distribution with the mean that is equal to n times p and standard deviation, equal to npq raised to the power 0.5.
So the mean will be 10 and the standard deviation is 2.44. We could use a normal distribution formula and come up with a P value.

1. This will be an approximation – it will NOT be exact
2. Because our sample sizes are less than 30, the normal distribution will not be very close

However if we did do this using the approximation to a normal distribution, we would get a P value of 0.006 and we would still reject the null hypothesis. So remember when we run a hypothesis test, many times we will just directly use a normal distribution because we may not know what the underlying population distribution is. But it doesn't mean that you have to use a normal distribution for a hypothesis test. If you know the population distribution, it may be better to directly use the appropriate population distribution to calculate the probability of observing an outcome as extreme or more extreme in the sample because of random chance variation.

>>>

As we reviewed the different types of hypothesis test, let us now look at very important type of hypothesis test, which is a hypothesis test when you sample sizes are low.

>>>

Remember that the Central Limit theorem says that as long as your sample sizes are sufficiently large, the distribution of sample averages will be normal irrespective of the underlying population distribution. But what happens if your sample size is less than 30?

>>>

For example, Let us say we want to test the hypothesis that college students sleep less a lot less than the general population. So the average sleep hours for the population is, say 8 hours. We take a random sample of 10 college students and you ask them how many hours they slept in the last night and this is the data that you get. The average turns out to be 6.64.

Based on the sample, can we conclude that students sleep less than the general population? Some of you may say why I should restrict my sample size to 10; why can't I just ask 30 students or 40 students. Certainly, you can. But many times, there may be a budget constraint; there may be a time constraint. It may be expensive for you to ask 40 students and so on. So there could be situations where you may not be able to increase your sample size. You still have to make a conclusion from your small sample size. In this example, how do I conclude whether this is random chance variation or it tells me that students really sleep a lot less than the general population?

>>

In order to compute the probability of an observed outcome when sample size is <30, it turns out that the sample means don't follow a normal distribution, but they do follow what is called T-Distribution or Student T-Distribution. We should not use the CLT normal approximation here because CLT holds only for sample sizes of at least 30.

So what is the T-Distribution? This is what it looks like. A T-Distribution is a continuous probability distribution that explicitly depends on sample size. The smaller the sample size, the less it looks like a normal distribution. The larger the sample size, the more closely it resembles a normal distribution. It turns out that when your sample sizes are 30 or greater, the T-Distribution and a normal distribution are very similar. This is what a T-Distribution looks like.

>>

We are going to use T-Distribution to calculate the P value. It turns out that T - Distribution of sample averages taken from samples of less than 30 from an underlying population has the property that the sample averages will follow the T-Distribution with n-1 degrees of freedom. As sample size increases and approaches 30, the T-Distribution will approximate a normal distribution.

The test it takes has calculated like this.

$$t = \frac{\overline{x} - \mu}{s/\sqrt{n}}$$

>>

Remember we want to do a hypothesis test and every hypothesis test is set up using the same framework.

- in the null hypothesis, students sleep same as regular population meaning, they also sleep 8 hours
- In the alternate hypothesis, students sleep less than the general population
- Significance level – 5%

Now we have to calculate a P value.

Remember the P value has to be calculated using a T-Distribution. If I want to use a distance measure, I will first calculate a test statistic.

$$t = \frac{\overline{x} - \mu}{s/\sqrt{n}}$$

(6.44 – 8)/(1.1/(10^0.5)) = -3.90

What is -3.90? Remember the negative value on the test distance is that your sample outcome is to left of the population mean. When we use distance measures, either we can calculate a P value looking at the distance in a T table or we can compare a distance to a critical distance. What is critical distance? Remember we want to use a level of significance of 5%, so we will reject a null hypothesis if the P value of the sample outcome is less than 5%. Equivalent to the 5%, there will a distance, which is critical. So If I look at 5%, 9 degrees of freedom, one tailed test.

What is 9 degrees of freedom? Every time you do a T test, you have to use what is call degrees of freedom. Degrees of freedom is nothing but n-1 which is the sample size. Sample size is 10; degrees of freedom is 9 - We are doing a one tailed test. We will talk about one tailed test in a little while. For now, let's just say we are doing a one tailed test. 0.05 is our critical significance level. So the critical distance has now become 1.833

>>>

Just to understand what this 1.833 is. Remember we have a sample outcome, which is -3.90. This is how far the sample outcome is away from the mean. We will reject the null hypothesis if this sample outcome probability is less than 0.05. So let us say this is the point where the P value is 0.05. Now for this 0.05 P value, we have horizontal distance as well. This horizontal distance is -1.83. It is -1.83 distance to the left of the mean.

Essentially, if you have a sample outcome, which is somewhere to left of -1.83, we will reject the null hypothesis. Because we know that its P value has to be less than -0.05. If your distance is to right of -1.83, we will not reject the null hypothesis. So remember the critical distance is simply the equivalent distance related to the significance level. The significance level is in terms of probability and the critical distance is in terms of distance from mean. So the significance level is measuring in Y axis and the distance is on the X axis.

In our particular example, since we have an outcome, which is to the left of the critical distance further away from the mean than the critical distance, we will reject the null hypothesis.

>>

What if I wanted to calculate the T-Distribution P value directly using a distribution formula? In excel, it is a two-step process. First you have to calculate the distance and then you have to use the distance measured in excel in the T-Distribution formula.

So what is the distance? The distance was -3.90; the degree of freedom are n-1, which is 9. So what we will put in excel is T Distribution (-3.90, 9, TRUE), which gives me 0.00181.

>>>

Just to demonstrate this in excel, remember we need a two-step process. First step is to calculate the distance.

We have an observed sample outcome, which is 6.64 minus the population means, which is 8. This is the numerator. We need to divide this by the standard deviation and the square root of sample size. The standard deviation in the sample is 1.1 divided by the square root of sample size, which is 10. We will get a distance of -3.90. Now we will use this distance measure in the T-Distribution formula, applying the calculated distance of -3.90, degrees of freedom of 9 and TRUE and we get a P value of 0.001783 or 0.0018.

Based on this P value, what should be our conclusion? We reject the null hypothesis and conclude that students sleep less than the general population.

>>>