In this session we will take a look at how we compute probabilities when a data follows t-distribution. We will also take a look at how we do a single sample test of mean which is a hypothesis test as well as we will take a look at how we do two-sample test of mean and a two-sample test of mean could be either independent sample test of mean or paired sample test of mean. So let's get started.

The relevant function in R which computes the cumulative probability for a data which has the t-distribution is called pt. Let's take a look at the arguments for this function. There are two arguments. One is the q value and the other is the degrees of freedom. The other three arguments are seldom used.

Let's suppose that we want to find the cumulative probability of observing a t-value up to 1.65 with 29 degrees of freedom. This is the command or this is the syntax we will use. Since you want to observe a t-value up to 1.65, the first parameter is 1.65. Since the degrees of freedom are 29, so we are supplying degrees of freedom as 29.

Always keep in mind that for t-distribution and when we are talking about the univariate data or a single distribution, the degrees of freedom are always n-1 where n is the sample size. Let's execute this command. This is the probability of observing a t-value up to 1.65.

Let's take a concrete example. Let's assume that we take a sample of 28 items and we find that the sample mean is 30, the sample standard deviation is 5, and the sample we assume is coming from a population whose mean is 35. Now the relevant question here is what are the chances that we will observe a sample mean of at most 30. So what you want to find out is the probability that x would be <= 30.

Since our sample size is 28, which is less than 30, we should ideally use a t-distribution to model probabilities. Now in order to use the distributions, the first thing you need to do is to find out the t-value corresponding to the data value of 30. This t-value can be found out through this data transform. So what I am doing here is I am subtracting my sample mean from the population mean and dividing it by the standard error. Since this is the sampling distribution I should divide it by the standard error which should be the sample standard deviation dividing by the square root of sample size.

If I do this computation I figure out that my t value is -5.291. So these two statements would be equivalent. It would be the same thing to say that I want to find the probability of observing x up to 30 or observing t up to -5.2915. Because this t value corresponds to the data value of 30. So I will execute this command.

A degrees of freedom here is 27. The reason is our sample size is 28 and as I told previously the degrees of freedom is always 1- the sample size. Let's execute this command. The excel counterpart for this command would be t.dist (the relevant t values, the degrees of freedom, and true). We are supplying true in the excel counterpart, the reason is because we are finding out t cumulative probability.

Let’s now take a look at how we can do a hypothesis test of a single sample mean using t-distribution. The first thing we will do is we will simulate some data. What I am going to do is, I am going to simulate 16 data points from a population whose mean is 2 and whose standard deviation is 1. These 16 simulated data points are stored in the vector x.

Let’s take a look at the mean of this sample of 16 observations. Now since this is a sample, the mean is little bit different from the population from which the sample has been taken. If we want to do a hypothesis test our null hypothesis would be that our mean is equal to 2. Since we are observing a number greater than 2, our alternate would be that our mean is greater than two.

In order to the hypothesis test we will use the R function t.test. The first parameter would be the vector of samples, the next parameter is the alternative. Since here the alternative is greater than, so I will write here greater. The last parameter is the hypothesized population mean.

Let’s execute this command. As you can see, the p value is greater than 5%. The p-value here is 37% which makes sense. Because we had already sampled this data from a population which had a mean of 2. In the previous example our alternate hypothesis was directional. We were doing a one-tailed test. We can also do a two-tailed test.

The only way in which our syntax would changes in place of greater or less, we will say two-sided for the alternative parameter. Let's execute this. As you can see the p value again remains very, very high. It is substantially higher than 5%. We cannot reject our null hypothesis.

Let's now take a look at how we can do two-sample t-test. Whenever we talk of two-sample t-test they can either be independent sample t-tests or paired sample t-tests. We will first of all discuss the hypothesis test in case of two samples for two independent samples.

First of all what we will do is we will simulate data and store this data in two separate vectors. Here I am simulating 20 data points and storing them in vector x1. These 20 data points are being sampled from a normal distribution with mean of two and standard deviation of 1. Similarly for the second vector x2, we are sampling 20 points which come from a population with mean 3 and standard deviation of 1.5.

Let's execute this command and get our two samples. Let's take a look at the mean of these two samples. Both the means are different. Let's do a two-sample hypotheses test. The way we will do that is we will use the t.test which we previously also used. But instead of supplying just a single vector, we will be supplying both the vectors corresponding to the samples that we have.

We are doing a two-tailed test. So the alternative would be two.sided. Also mu here is supplied as zero. The reason for this is in our null and alternate hypothesis, we say that means of both the samples are different or not different. When in my null hypothesis, I say that means of both the samples are different or not different. When in my null hypothesis we say that the means of both the samples are different. It implies that the mean difference is zero. So that's why I am supplying zero here.

Let's execute this command. As you can see that the p value is very, very low implying that both the means are different. It also makes sense because we had sampled both of these means or both of these vectors from different populations. If we have to do a paired sample t-test, the only way in which our syntax would changes. We will add this parameter paired and supply with a value of two.

Let's execute this command. Again we can see that the p value is very, very low, and lower than 5% implying again that we should reject our null hypothesis.

>>

In this section we have taken a look at how we do probabilistic computations once you know that our data follows t-distribution and we have figured out that the relevant cumulative probability distribution function is pt. we also discussed how we do hypothesis tests for single sample mean as well as two sample means when a data follows t-distribution.

>>>