

Essentials of Data Science with R Software- 1
Professor Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology Kanpur
Lecture 51
t - Distribution

Hello friends, welcome to the course Essential of Data Science with R Software- 1, in which we are trying to understand the basic concepts related to the probability theory and statistical inference. You can recall that in the last lecture we started a discussion on sampling distribution and we had understood what is chi square distribution, so continuing on the same lines in this lecture we are going to talk about one more sampling distribution which is t distribution, that is just t right and in case if you have understood the chi square distribution, understanding t or even F in the next lecture will not be difficult for you.

As you have seen that in the case of chi square what we had done? We had taken the sum of squares of the standard normal variates and that is going to follow a chi square distribution and if there are n number of random variables then the degrees of freedom is going to be n and the sampling distributions are characterized by the degrees of freedom and you have seen that by changing the degrees of freedom the probability density functions, their curves they all actually change. So, for each value of the degree of freedom you will get a new probability function, the probability function will remain as only t but that will have different types of characteristics.

(Refer Slide Time: 1:45)

t - Distribution:

Let X and Y be two independent random variables where $X \sim N(0, 1)$ and $Y \sim \chi_n^2$. Then the ratio

$$\frac{X}{\sqrt{Y/n}} \sim t_n$$

follows a t-distribution (Student's t-distribution) with n degrees of freedom. This is central t-distribution.

A random variable X has a t-distribution if the PDF of X is given as

$$f_X(x) \equiv f(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{x^2}{n}\right)^{-\frac{(n+1)}{2}}; -\infty < x < \infty.$$

We write $X \sim t_n$.

So, let us now begin this lecture and we try to understand this t distribution in this lecture. So, now you can see here that once again I am going to consider a function of random variables, so now we have to understand that how are we going to construct that function of random variable which is going to follow a t distribution.

So, suppose there are two random variables X and Y and both are independent, very important condition that you have to keep in mind that X and Y are two independent random variables, and X is following a $N(0, 1)$ distribution and Y is following a chi square distribution with n degrees of freedom.

So, I can say here X follows $N(0, 1)$, Y follows chi square with n degrees of freedom and both of them are independent. Now, we define a function of these random variable, see how, X divided by square root of Y by n, so this is written here is $\frac{X}{\sqrt{Y/n}}$, so this is something like $N(0, 1)$ divided by square root of chi square divided by its degrees of freedom n.

So, the distribution of this statistics will follow a t distribution with n degrees of freedom, and this is also called as students t distribution and this is a central t distribution, in the last lecture we had understood what is the difference between a central distribution and non central distribution and based on that we have central chi square and non central chi square. Similarly, we will have a central t and non central t.

One question comes here either this is chi square or t, we are trying to consider the function of random variable how do you get that whether this function has chi square or this function as t? Well just for the information that we have some statistical technique, and we try to employ those things and then we try to find out the distributions of the functions of random variable. Well, I am not considering them here because they are the part of the statistics courses that are taught, but if you wish you can look into any of the standard book and you will find such methodologies.

So, now let me try to give you here the probability density function of this t but once again I would say just like in the last lecture that the form of the distribution may look

complicated but definitely, we are not going to use the form anywhere, we are simply going to use the information that this is t distribution and what are its degrees of freedom.

So, a random variable X has a t distribution with n degrees of freedom if the PDF of X is

given by like this, $f(x) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi} \Gamma(\frac{n}{2})} \left(1 + \frac{x^2}{n}\right)^{-\frac{(n+1)}{2}}$; $-\infty < x < \infty$, t is here and n is in

the subscript, that is about the standard notation.

(Refer Slide Time: 5:20)

Chi square (χ^2) Distribution:

Let X and Y be two independent random variables where $X \sim N(\mu, \sigma^2)$
 and $\frac{Y}{\sigma^2} \sim \chi_n^2$. Then the ratio

$$\frac{X}{\sqrt{Y/n}} \sim \text{noncentral } t_n$$

follows a noncentral t-distribution with n degrees of freedom which has one more parameter – noncentral parameter.

So, this is the central t distribution, so once there is a central t distribution there will always be a non central t distribution also, so just to give you an information that what is this non central t distribution and how it is obtained. So, if X and Y are two independent random variables, so that X is following a normal distribution but definitely here the mean is not 0 and variance is not 1, but now we can define our Y here as $\frac{Y}{\sigma^2} \sim \chi_n^2$.

So, what will happen? That the mean is not going to be 0, then in case if I try to consider this ratio $\frac{X}{\sqrt{Y/n}}$, then this will follow a non central t distribution with n degrees of freedom, and when we are talking of non-central distribution then there will be one more parameter this is called as non centrality parameter and if this non-centrality parameter is 0, the non central distribution becomes central distribution. And once again I would say that


whenever we are going to use this t distribution or whenever we say t distribution we will mean the centrality distribution unless and until we specify that we are going to consider the non central t distribution. (Refer Slide Time: 6:52)

t - Distribution:

- The mean and variance of a random variable $X \sim t_n$ distribution is

$$E(X) = 0, n > 1$$

$$Var(X) = \frac{n}{n-2}, n > 2.$$
- The t- distribution is symmetric.
- The t- density has thicker "tails," indicating greater variability, than does the normal density.
- The "degrees of freedom" specify the shape of the distribution. When the degrees of freedom are more than 30, the shape of t and normal distributions are almost the same.



Now, in case if you try to find out the mean and variance of a random variable following a t distribution with n degrees of freedom, the mean will come out to be expected value of X to be 0 when n is greater than 1 and variance of X will come out to be $n/(n - 2)$. So obviously, the variance has to be positive so there is a condition n greater than 2 has to be imposed on it and some properties of this t distribution that t distribution is symmetric and I will show you that it is very similar to normal curve, but the difference is that that the t density has thicker tails in the comparison to the normal distribution and this indicates the greater variability than this in the normal density.

So, it would be like this if a normal here is like this then your t is going to be here like this, I will try to show you but it will be like this so these are the actually the tails, these are here the tails, and the degrees of freedom they specify the shape of the distribution when the degrees of freedom are more than 30 then the shape of t and normal distribution they are almost the same, that is an very important result and I will try to address it once again after couple of slides but definitely you always have to keep in mind that this is going to happen that when you are trying to take the degrees of freedom in the t

distribution to be greater than or equal to 30 the probability curves of normal and t they will become almost the same.

(Refer Slide Time: 8:30)

t - Distribution: Student's theorem

Let X_1, X_2, \dots, X_n are identically and independently distributed random variables with $X_i \sim N(\mu, \sigma^2)$. Then

$$\frac{\sqrt{n}(\bar{X} - \mu)}{s} \sim t_{n-1}$$

is then t-distributed with $(n - 1)$ degrees of freedom where

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2.$$

(Handwritten notes in the slide: $\bar{X} - \mu$ and s/\sqrt{n} are circled in green. The formula for s^2 is also circled in green with a checkmark.)

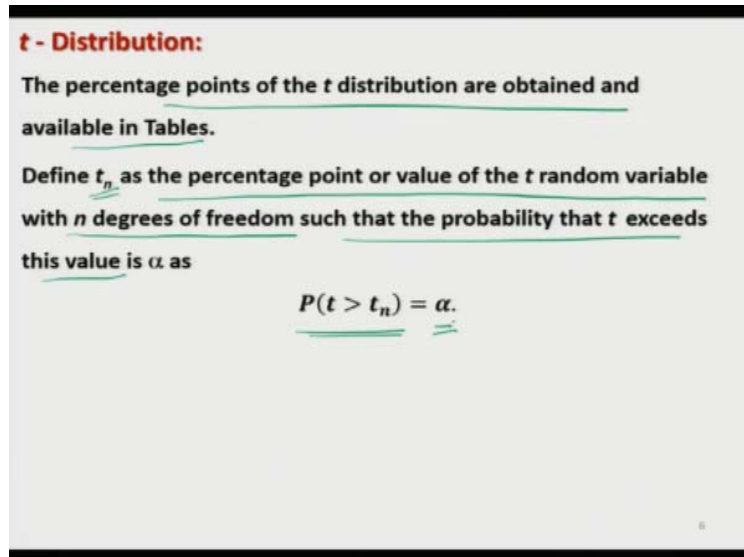
Now, there is one very important result that we will be using many times in the statistical inference when we want to draw a statistical conclusion from the given set of data. So, suppose the s square is the sample variance, that you have the observations X_1, X_2, \dots, X_n which are identically and independently distributed and then you try to compute this sampling variance from these observations and here this X_i 's are assumed to be normal μ sigma square.

Then in case if I try to take here this statistic $\frac{\sqrt{n}(\bar{X} - \mu)}{s}$ and then whole quantity multiplied by square root of n, so this is actually $\frac{\sqrt{n}(\bar{X} - \mu)}{s}$ like this, that is easy to remember. And if you try to see what I have done here that we have simply standardized the sample mean, then the probability distribution of this statistics that is $\frac{\sqrt{n}(\bar{X} - \mu)}{s}$ this will be a t distribution with $n - 1$ degrees of freedom.

So, you can see here that these degrees of freedoms are changed, and it is reduced by 1 and 1 simple reason I can give you here that because in this case the sigma square is unknown to us so we are trying to estimate it on the basis of given sample of data and

because of that this degree of freedom is reduced by 1. Well, there are different types of interpretation of these degrees of freedom, but I just thought that I should inform you here this thing.

(Refer Slide Time: 10:15)

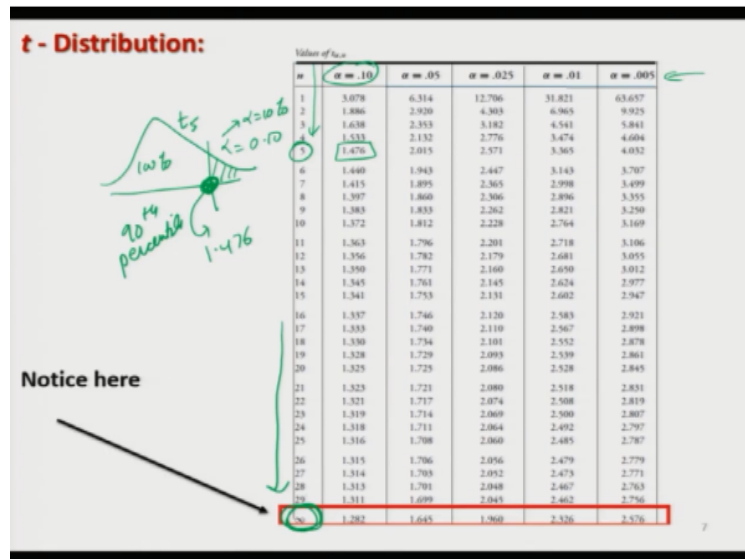


t - Distribution:
The percentage points of the t distribution are obtained and available in Tables.
Define t_n as the percentage point or value of the t random variable with n degrees of freedom such that the probability that t exceeds this value is α as

$$P(t > t_n) = \alpha.$$

Now, as you had tabulated the probabilities in the chi square distribution case similarly, we can tabulate the probabilities under the t distribution also. So, the percentage points of the t distributions are obtained, and they are available in the tables but now we are going to use the R software but just for information that you must know these things what were happening in the past and that how these probabilities are defined and how they are used. So, if we define this quantity here t_n as the percentage point or value of the t random variable with n degrees of freedom such that the probability that t exceeds this value is alpha that means probability that t is greater than t_n that is equal to α .

(Refer Slide Time: 10:57)



So, now for that actually these probabilities are computed, and they are available in most of the statistics books in the appendix and these tables will look like this, I am just trying to give you here one simple example. Suppose you want to find out the value of t with 5 degrees of freedom and alpha is equal to 0.10, so this value will come out to be here 1.476. So, if you try to create this curve over here, so if this value here is α is equal to 0.10 then this value here this will come out to be 1.476 if this curve is t distribution with 5 degrees of freedom and you can see here on the first column you have here degrees of freedom and in the first row they are given different values of alpha.

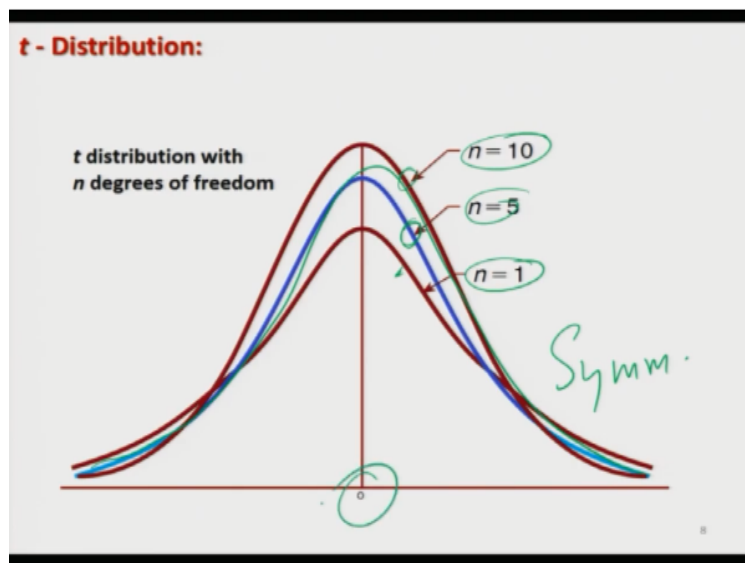
Well these probabilities were computed earlier only for a given values of alpha because these calculations were made by hand, people manually calculated these integrals, so that is why they are restricted only for particular choices of alpha but now with the help of software you can choose any value of alpha and if you try to recall your definition of percentile you are trying to take here that this whole area is 100 percent and you are trying to take this alpha to be only here ten percent, what is this thing don't you think that this is equivalent to some percentile, this is simply your here 90th percentile.

So, now you can see that whatever you had done earlier now they are coming together and the same thing you can interpret in different ways so you are essentially trying to say

here that you can compute any percentile and that would be related to the value of your alpha.

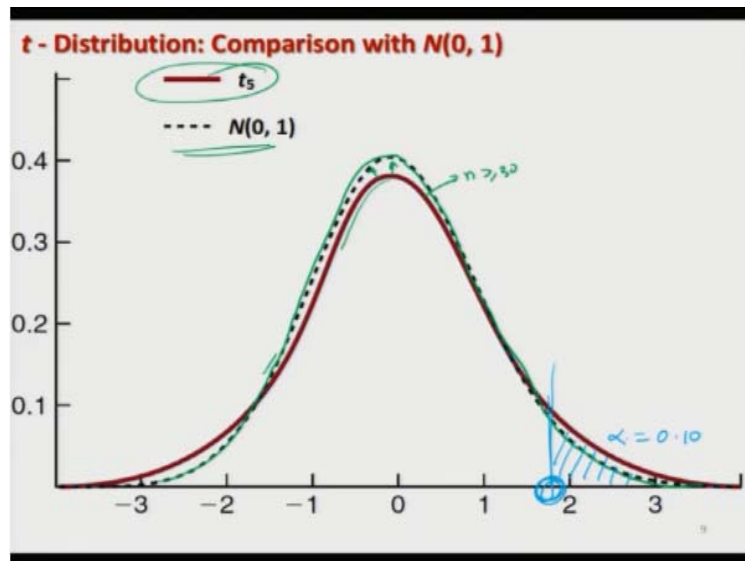
One important thing you will see in these tables that these degrees of freedom are increasing and they will come up to 29 after that they take it to be infinity, do you know why? Because as soon as you reach n equal to 30 or more the probability density function of normal and t they will become almost the same, so instead of computing the probabilities from the t distribution, you can estimate them from the normal distribution and that is why we always say that n greater than equal to 30 can be considered as a large sample in the context of statistical inference and in the context of the tool test of hypothesis, but we will use it later on.

(Refer Slide Time: 13:25)



But just to show you that how this deep distribution will look like for different degrees of freedom so you can see here if I try to take here n equal to 1 this curve will be here like this and if your n is equal to here 5 this curve will become here like this blue and if n is equal to 10 the curve will become here like this one. So, you can see here that this is here a symmetric curve, this is symmetric, and this is here 0.

(Refer Slide Time: 13:58)



t - Distribution:

Value of $t_{\alpha, n}$

n	$\alpha = .10$	$\alpha = .05$	$\alpha = .025$	$\alpha = .01$	$\alpha = .005$
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.474	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.133	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.695	2.040	2.456	2.750

Notice here

Handwritten notes on the table include: 't = 1.476' for n=5, '90% percentile' with an arrow pointing to the t5 row, and 'alpha = 0.10' with an arrow pointing to the first column header.

Now I try to show you here the comparison of normal and t distribution, so you can see here with this black dotted line I am trying to indicate the $N(0, 1)$ and with this line in red color I am trying to indicate the t_5 distribution that means t distribution with 5 degrees of freedom and if you try to increase this degree of freedom as soon as you make it here $n > 30$ then this distribution will simply overlap the normal distribution here like this and you have seen that this time I have made the smooth curve.

So, now as soon as you try to make this degrees of freedom to be here n greater or equal to 30 this curve will shift and this curve will become nearly the same as $N(0, 1)$. Now suppose, if you want to compute certain probability, suppose your n is greater than 30 and now suppose you want to compute here some probability here, suppose this is here α is equal to 0.10 as I took the example earlier.

Now, when both these curves for n greater or equal to 30 are becoming the identical then you try to compute the value here or this probability by using the t distribution or say normal distribution they will give you the same value and that is the basic concept when we try to say that why there are no values after 30, here you can see.

And now you can see here whether this curve is here t distribution or say $N(0, 1)$ if your n is greater than equal to 30 how does it make any difference whether you compute this point from the tables of normal distribution or from the tables of t probability or equivalently you try to use the R command for the normal distribution or for the t distribution and that is why many times we say that n greater than equal to 30 can be considered as a large sample, why? The reason I will tell you later on but that is the reason because t distribution and normal distribution for degrees of freedom more than equal to 30 they are identical.

(Refer Slide Time: 16:13)

t - Distribution: R Commands

Usage

dt(x, df) gives the density,

pt(q, df, lower.tail = TRUE) gives the distribution function,

qt(p, df, lower.tail = TRUE) gives the quantile function,

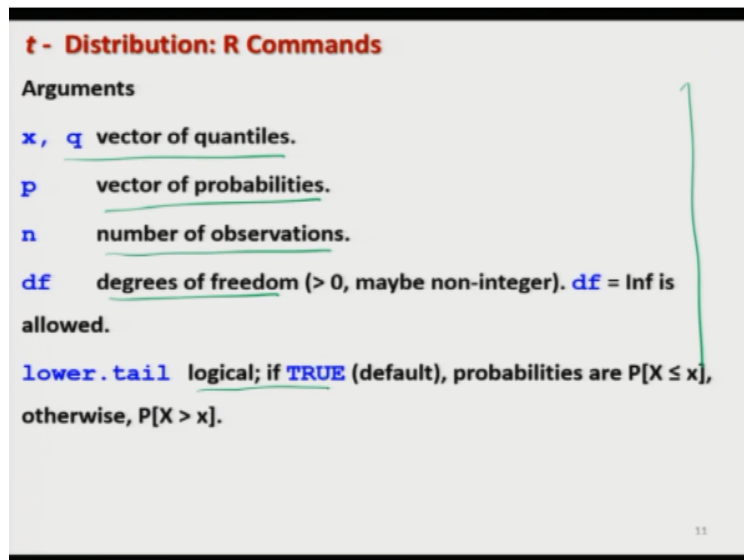
rt(n, df) generates random deviates.

10

Now, how to execute this t distribution in the R software. So, we have a command here different types of command for getting here density, quantile, CDF, random number just like any other distribution, so dt here is the command that will give you the density, so d means here density and t is indicating the t distribution. So, you simply have to give here the value and specify the degrees of freedom.

And similarly if you want to find out the CDF then you have to give here pt and then you have to give here the value of q and then the value of degree of freedom and then you can use here lower dot tail is equal to TRUE or FALSE depending on your requirement and similarly if you want to find out the quantile you have to use the function here qt there you have to give the value of p, the value of degrees of freedom as df and then you can use the option lower dot tail is equal to TRUE or FALSE depending on your requirement. And similarly, if you want to generate the random number from the t distribution the command here is rt and then you have to give here n that how many random numbers you want to generate and then you have to specify the degrees of freedom.

(Refer Slide Time: 17:25)



t - Distribution: R Commands

Arguments

- x, q** vector of quantiles.
- p** vector of probabilities.
- n** number of observations.
- df** degrees of freedom (> 0, maybe non-integer). **df = Inf** is allowed.
- lower.tail** logical; if TRUE (default), probabilities are $P[X \leq x]$, otherwise, $P[X > x]$.

11

So, let me try to just take some examples and these are the details which I just shown you. So, here x, q they are the vector of quantile, p is the vector of probabilities, n is the

number of observations, df is the degrees of freedom and lower dot tail that is a logical variable taking value TRUE or FALSE that you know how to use it.

(Refer Slide Time: 17:44)

t - Distribution: R Commands

`pt(q, df, lower.tail = TRUE)` calculate the CDF $F(q) = P(X \leq q)$ at any point q .

Suppose we want to find the probability from t_{10} .

$P(X \leq 5) = F(5) = \int_0^5 f(x) dx$, then

`> pt(q=5, df=10)`
`[1] 0.9997313`

or equivalently

`> pt(q=5, df=10, lower.tail = TRUE)`
`[1] 0.9997313`

```
R Console
> pt(q=5, df=10)
[1] 0.9997313
> pt(q=5, df=10, lower.tail = TRUE)
[1] 0.9997313
> |
```

Now, let me try to take here some examples and we try to compute different types of probabilities from the t distribution. So, suppose we want to compute the probability that X is less than equal to 5, where X is following a t distribution with here 10 degrees of freedom. So, now you know that this is simply the value of CDF at X equal to 5. So, this probability is $\int_0^5 f(x) dx$ where f(x) is your t distribution.

So, now if you want to compute the CDF you know the command here is pt, so I try to use the command here pt and q is equal to here 5 which is coming from here this 5 and then degrees of freedom equal to 10 which are coming from the specification of the distribution and if you try to see this value will come out to be 0.9997313 and if you want to use here the option lower dot tail is equal to TRUE that is the default option this will also give you the same value.

(Refer Slide Time: 18:38)

t - Distribution: R Commands

Suppose we want to find the probability from t_{10} .

$P(X > 6) = 1 - P(X \leq 6) = 1 - F(6)$, then

```
> 1 - pt(q=6, df=10)
```

```
[1] 6.605443e-05
```

or equivalently

```
> pt(q=6, df=10, lower.tail = FALSE)
```

```
[1] 6.605443e-05
```

R Console

```
> 1 - pt(q=6, df=10)
[1] 6.605443e-05
> pt(q=6, df=10, lower.tail = FALSE)
[1] 6.605443e-05
> |
```

Now similarly, if I want to compute in the same distribution probability X greater than 6, then this is going to be 1 - probability that X less than equal to 6 which can be written as 1 - F(6) so this can be obtained exactly in the same way as you did earlier that 1 - CDF at X equal to 6 which is specified by pt q equal to 6, 6 is coming from here, 1 is coming from here and the degrees of freedom they are coming from the specification of the distribution and if you try to compute it this will come out to be here like this.

And similarly, if you do not want to use this concept of 1 minus CDF you can use here an option lower dot tail is equal to FALSE and this will give you the same value here and this is here the screenshot I will try to show you all these calculations on the R console, but I know that these are very simple things for you.

(Refer Slide Time: 19:29)

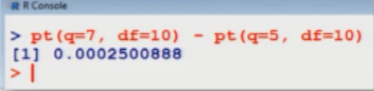
t - Distribution: R Commands

Suppose we want to find the probability from t_{10}

$$P(5 \leq X \leq 7) = \int_5^7 f(x) dx = F(7) - F(5).$$

This is obtained as $F(7) - F(5)$ in R as

```
> pt(q=7, df=10) - pt(q=5, df=10)
[1] 0.0002500888
```



14

Now similarly, if you want to compute the probability that X is lying between 5 and 7, then you have to find out the $\int_5^7 f(x) dx$ where $f(x)$ is the t distribution and this probability can be written as $F(7) - F(5)$ which F is your here CDF. So now you know how to compute this $F(7)$ and $F(5)$, $F(7)$ can be computed by the command here pt, q equal to 7 df equal to 10 and 5 can be computed by pt q equal to 5 and df equal to 10 and if you try to solve it in the R console you will get this value, and this is here the screen shot.

(Refer Slide Time: 20:08)

t - Distribution: R Commands

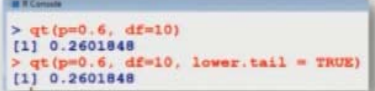
$qt(p, df, lower.tail = TRUE)$ gives the quantile function and calculates the quantile which is defined as the smallest value x such that $F(x) \geq p$, where F is the CDF $F(x) = P(X \leq x)$ at any point x from t_{df} .

For example, suppose we want to determine the 60% quantile q which describes that $P(X \leq q) \geq 0.6$ from t_{10} can be obtained by the command

```
> qt(p=0.6, df=10)
[1] 0.2601848
```

or equivalently

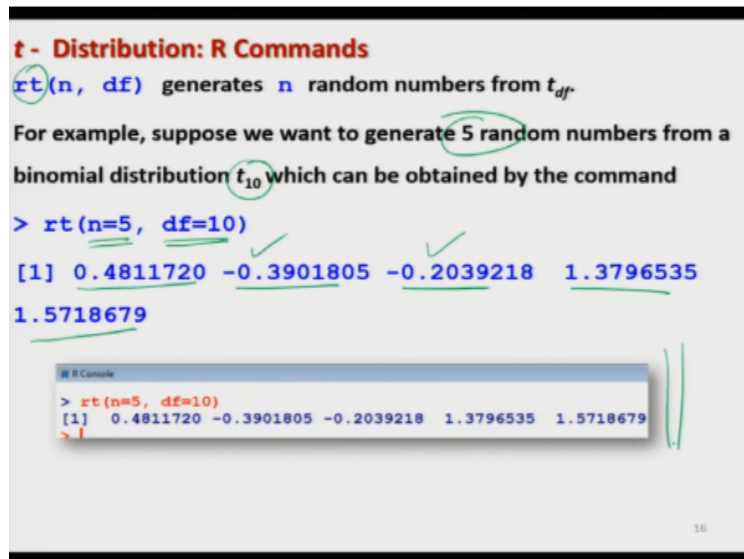
```
> qt(p=0.6, df=10, lower.tail = TRUE)
[1] 0.2601848
```



15

Similarly, if you want to find out here the quantile then my command in the t distribution is qt. So, suppose, I want to find out the 60 percent quantile from the same distribution that is t with degrees of freedom 10, so that can be obtained by writing qt p is equal to 0.6 and df equal to 10, so the 0.6 is coming from here and df equal to 10 that is coming from the degrees of freedom of the t distribution and if you try to see this value come out to be 0.26. And similarly if you want to use here the option lower dot tail is equal to TRUE then it will give you the same value that is the default option.

(Refer Slide Time: 20:48)



t - Distribution: R Commands
`rt(n, df)` generates `n` random numbers from t_{df} .
For example, suppose we want to generate 5 random numbers from a binomial distribution t_{10} which can be obtained by the command
`> rt(n=5, df=10)`
[1] 0.4811720 -0.3901805 -0.2039218 1.3796535
1.5718679

```
R Console
> rt(n=5, df=10)
[1] 0.4811720 -0.3901805 -0.2039218 1.3796535 1.5718679
>
```

And similarly, if you want to generate the random numbers from this t distribution, then my command here is rt. So suppose, I want to generate 5 random numbers from the t distribution with 10 degrees of freedom, so that can be obtained here is rt n equal to 5 df equal to 10 and this will give you, you can see here 1, 2, 3, 4 here 5 random numbers and you can see here they are lying between $-\infty$ and $+\infty$, these two numbers are here negative and remaining are here positive and this is the screenshot here but definitely I would like to show you these operations on the R console also.

(Refer Slide Time: 21:39)

```
> pt(q=5, df=10)
[1] 0.9997313
> 1 - pt(q=6, df=10)
[1] 6.605443e-05
> pt(q=6, df=10, lower.tail = FALSE)
[1] 6.605443e-05
> pt(q=7, df=10) - pt(q=5, df=10)
[1] 0.0002500888
> qt(p=0.6, df=10)
[1] 0.2601848
> rt(n=5, df=10)
[1] -0.5751701 2.2059282 0.3599131 0.1465156 0.9095759
> rt(n=15, df=10)
[1] 0.1641751 0.3195443 -0.8802413 -1.6292165 -1.1471424
[6] 0.3396289 -0.3827036 -0.3890492 -0.9686813 0.3901570
[11] 1.8416455 -0.4853459 -0.1195517 1.5263014 0.3111501
> rt(n=15, df=5)
[1] -0.33618542 -0.43552490 0.22271909 0.31395065 -0.15175315
[6] 0.67394909 -1.08093881 0.03806681 -0.65096961 -0.08513128
[11] -1.14725676 0.48651204 0.57000027 -0.11669977 -0.41026561
> |
```

So, let me try to show you how can you compute these different types of probabilities on the R console. You can see here package is available inside the base package of R you simply have to write down here $pt(q=5, df=10)$ and you can get here the same value that you have reported here.

Similarly, if you want to compute the probability X greater than 6, then you can use this command $1 - f(x)$ and this is giving you this value and if you try to use here option lower dot tail is equal to FALSE then you do not need to use the subtraction like $1 - f(x)$ but it will give you the directly the value of $1 - f(x)$ which is here like this, you can see this value of this value they are the same.

Similarly, if you want to compute the probability of X between 5 and 7 then this value comes out to be here same what you have reported in the slides. And similarly, if you want to compute the 60th quantile then you can see here that is the command qt can be used p is equal to 0.6 and this is going to give you the value of the quantile and similarly if you want to generate the random numbers, well that is going to generate the random numbers but they are not going to be the same what you have written here in the slide because they are random, so every time you generate, they will be different.

So, you can see here that these are the five random numbers which are obtained by specifying n equal to 5 and if you try to increase this random number from 5 to 15 there are fifteen values here like this and if you try to change the value of degrees of freedom here make it here 5 then these values will also change.

So, now we come to an end to this lecture and you can see here that was a pretty simple lecture and exactly on the same lines as in the chi square distribution, one of the most important result which you have to remember that how this t distribution has been obtained, so there are two random variable one is normal 0,1 another is chi square and then you have to create a statistics by $N(0, 1)$ divided by square root of chi square divided by degrees of freedom and this is going to follow our t distribution.

And another important result what you have to keep in mind that is square root of $\frac{\sqrt{n}(\bar{X}-\mu)}{s}$ that is the standard error this will follow our t distribution also. So, these two results we are going to use many times in the forthcoming lectures, so I would request you one thing that you try to practice them in the R software, try to look into the details about this t distribution in the books and I will see you in the next lecture with details on F distribution till then good bye.