Essentials of Data Science with R Software - 2
Sampling Theory and Linear Regression Analysis
Prof. Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology Kanpur

Sampling Theory with R Software
Lecture - 36
Bootstrap Methodology
Bootstrap Confidence Interval

Hello friends, welcome to the course Essentials of Data Science with R Software 2 where we are trying to understand the topics of Sampling Theory and Linear Regression Analysis and in this lecture, we will continue with our modules Sampling Theory with R Software and we are going to handle the topic of Bootstrap Methodology.

So, you may recall that in the earlier three lectures, we had considered the use of bootstrap methodology in estimating the value of the estimator, its bias as well as the standard error and we also considered that how the boot function works in R software. So, we have now covered the theory part as well as the computation part related to the bootstrap estimator, its bias as well as standard error.

Now, the next aspect is confidence interval. So, you can see that up to now, whatever I have done, we have used the point estimation, we have estimated the value of the statistics at a point so, that was point; point estimate, that was obtained as a average of all the bootstrap value of the statistic.

Now, we would like to estimate the confidence interval using the bootstrap methodology. The difference between estimating the bias standard error and confidence interval is that you have seen we have used the result like $\frac{\overline{x}-\mu}{\sigma/\sqrt{n}}$ follows normal distribution N(0, 1) if σ^2 is known and population with mean μ and variance σ^2.

So, my data is from the normal μ σ^2, but their statistics $\frac{\overline{x}-\mu}{\sigma/\sqrt{n}}$, the sampling distribution of this statistics is N(0, 1) and using the distribution of the statistics, we try to find out the cutoff values and based on that, we try to compute the confidence interval.

Problem is this, the computation of sampling distribution or the determination of sampling distribution is not so easy and without that, we cannot estimate the confidence interval. So, the bootstrap methodology helps once again, and the concept of the confidence interval and the estimation procedure remains the same whatever is in the usual statistical inference, whatever are the basic fundamentals related to the theory of confidence interval, they remain the same.

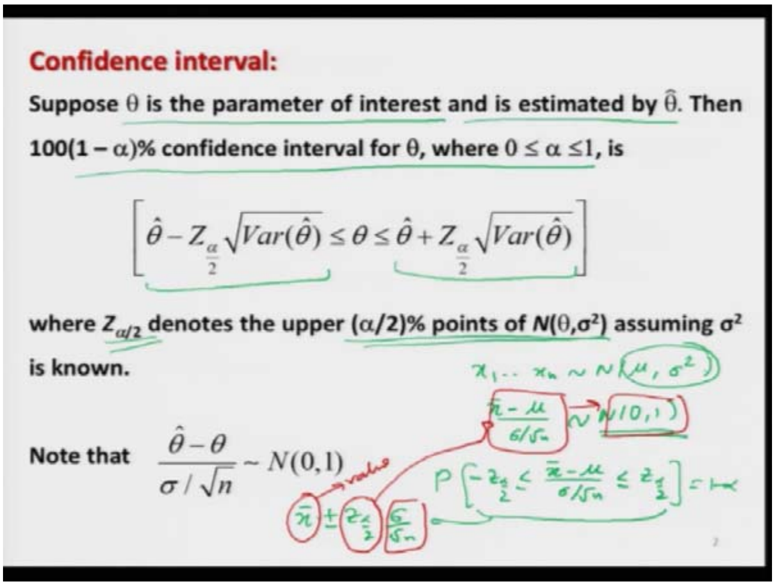
The only challenge before us that how to compute the values of the parameter, standard error and distribution of the statistic. So, we; so, out of these three components, we already have learnt how to find out the value of the parameter that is the bootstrap estimator of the parameter. We also have computed the standard error of the bootstrap statistics.

The only challenge which is left for the computation of confidence interval, how to find out the critical values which are dependent on the sampling distribution of the statistics? So, for that people have proposed different types of ways. So, in this lecture, we are going to consider this and based on that, we will try to construct the confidence interval.

Well, let me make it clear in the beginning that it is very difficult for me to give you the entire details behind the theory and fundamentals of the construction of those five types of confidence interval, but still I will try my best to give you a brief overview, a brief detail and particularly, in one type of confidence interval, we also need the concept of jackknife.

So, jackknife we are not covering here in this course. So, I will not be able to give you the more details about those things, but I will just try to show you that how the things work and ultimately, we are going to use the R software. So, R software will take care of all the theory behind the construction of the confidence interval. So, let us now begin our lecture with this note, ok.

(Refer Slide Time: 05:16)



Confidence interval:

Suppose θ is the parameter of interest and is estimated by $\hat{\theta}$. Then 100(1 – α)% confidence interval for θ, where 0 ≤ α ≤1, is

$$\left[\underbrace{\hat{\theta}-Z_{\frac{\alpha}{2}}\sqrt{Var(\hat{\theta})}} \le \theta \le \hat{\theta}+\underbrace{Z_{\frac{\alpha}{2}}\sqrt{Var(\hat{\theta})}}\right]$$

where $Z_{\alpha/2}$ denotes the upper (α/2)% points of $N(\theta,\sigma^2)$ assuming σ^2 is known.

$x_1 .. x_n \sim N(\mu, \sigma^2)$

$\boxed{\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}} \sim \boxed{N(0,1)}$

Note that $\frac{\hat{\theta}-\theta}{\sigma/\sqrt{n}} \sim N(0,1)$

value

$P\left[-z_{\frac{\alpha}{2}} \le \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \le z_{\frac{\alpha}{2}}\right] = 1-\alpha$

$\bar{x} \pm z_{\frac{\alpha}{2}} \boxed{\frac{\sigma}{\sqrt{n}}}$

So, first I try to give you a brief overview that how do we compute the confidence interval in a usual case. Suppose if I say that I have got here a sample $x_1, x_2,\ldots, x_n$ from some normal population with mean μ and variance σ^2, right. In this case, we know that $\frac{\overline{x}-\mu}{\sigma/\sqrt{n}}$, this follows a N(0, 1) distribution and this is the sampling distribution of $\frac{\overline{x}-\mu}{\sigma/\sqrt{n}}$, right.

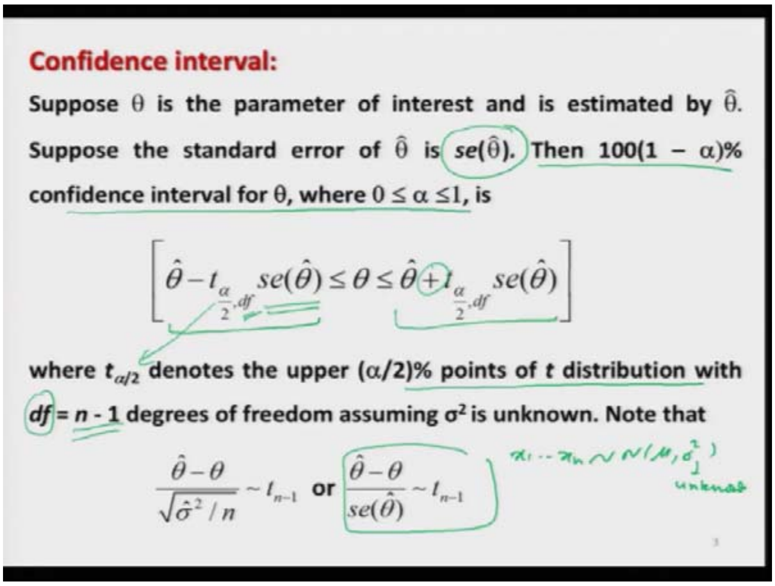
And based on that, what we try to do? We try to write down this $\frac{\overline{x}-\mu}{\sigma/\sqrt{n}}$ that will lie between $-Z_{\alpha/2}$ and $Z_{\alpha/2}$ and the probability of such an event has to be (1 – α), where α is lying between 0 and 1 and if you try to solve this inequality, you will get the, the lower and upper confidence limits, right. So, this is the basic theory what we have done.

So, essentially, if you try to obtain here the confidence interval that will look like $\frac{\overline{x}-\mu}{\sigma/\sqrt{n}}$. So, now, if you try to identify the components of this confidence limits, so, I have here the value of the statistic that is the estimator of the population mean μ, the critical value which is obtained from the sampling distribution, this value is going to be known only when we know that the sampling distribution of this statistics is N(0, 1).

And then, we have the standard error of $\overline{x}$, right, here we are assuming that σ is known, ok. So, these are the three components. So, based on these three components, I can write a general structure of the confidence interval. So, this is what is my starting point. So, suppose if I say the suppose that θ is the parameter of interest and this θ is being estimated by $\hat{\theta}$ for example, in the N(μ σ^2) means I have estimated μ by $\overline{x}$ sample mean.

Then, in this case the 100(1 − α)% confidence interval for θ for α lying between 0 and 1 is given by $\hat{\theta} - Z_{a/2}\sqrt{\text{var}(\hat{\theta})}$ that is the standard deviation of $\hat{\theta}$ or standard error of $\hat{\theta}$ and the upper limit of the confidence interval is $\hat{\theta} + Z_{a/2}\sqrt{\text{var}(\hat{\theta})}$, right. Where they say $Z_{\alpha/2}$ denotes the upper α/2% points on the distribution of N(θ, σ^2) and we are assuming here σ^2 is known.

(Refer Slide Time: 08:19)



Confidence interval:

Suppose θ is the parameter of interest and is estimated by $\hat{\theta}$. Suppose the standard error of $\hat{\theta}$ is $se(\hat{\theta})$. Then 100(1 − α)% confidence interval for θ, where 0 ≤ α ≤1, is

$$\left[\hat{\theta} - t_{\frac{\alpha}{2},df}\, se(\hat{\theta}) \le \theta \le \hat{\theta} + t_{\frac{\alpha}{2},df}\, se(\hat{\theta})\right]$$

where $t_{\alpha/2}$ denotes the upper (α/2)% points of t distribution with $df = n - 1$ degrees of freedom assuming σ^2 is unknown. Note that

$$\frac{\hat{\theta} - \theta}{\sqrt{\hat{\sigma}^2 / n}} \sim t_{n-1} \text{ or } \boxed{\frac{\hat{\theta} - \theta}{se(\hat{\theta})} \sim t_{n-1}}$$

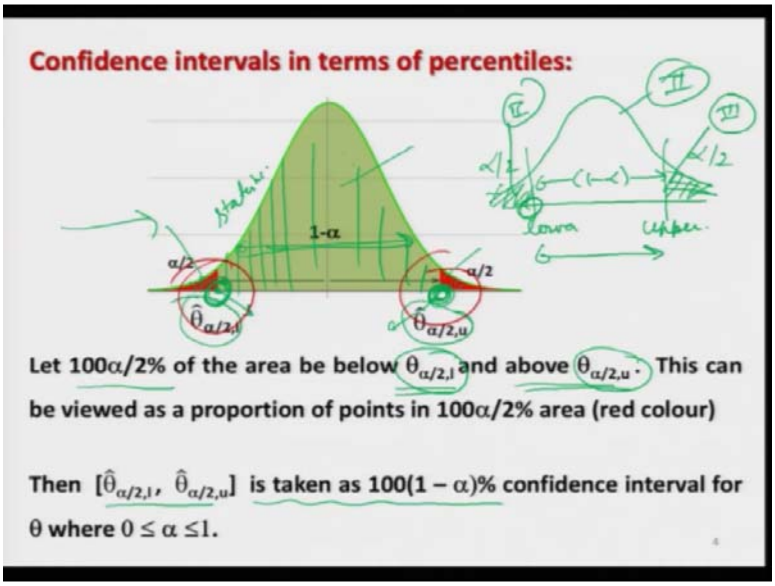
$x_1 \cdots x_n \sim N(\mu, \underset{\text{unknown}}{\sigma^2})$

And if σ^2 is unknown, then what happened that we can we try to estimate the σ^2 and instead of using the standard deviation of the statistics, we try to use the standard error of $\hat{\theta}$, right. So, in this case, we know that the distribution of the statistics in case of N(μ, σ^2), if I say that $x_1, x_2, \ldots, x_n$ are coming from N(μ, σ^2).

And if σ^2 is not known to us, then ($\hat{\theta}$ - θ)/standard error ($\hat{\theta}$) follows the t distribution with (n – 1) degrees of freedom, right. So, that is what I am trying to write down here that if σ^2 is unknown to us, then the $100(1 - \alpha)\%$ confidence interval for theta for α lying between 0 and 1 is given by this interval, where the lower limit is $\hat{\theta}$ minus the critical value $t_{\alpha/2}$ at the given degrees of freedom say df in general into standard error of $\hat{\theta}$ and the upper limit is given by the same quantity with a positive sign.

And here, this critical value $t_{\alpha/2}$ this denotes the upper α/2% points of the t distribution with say degree of freedom it denoted as df which in general which in this particular case is small n minus 1, right. So, this is my basic idea behind this confidence interval.

(Refer Slide Time: 09:51)



Confidence intervals in terms of percentiles:

Let 100α/2% of the area be below $\theta_{\alpha/2,l}$ and above $\theta_{\alpha/2,u}$. This can be viewed as a proportion of points in 100α/2% area (red colour)

Then $[\hat{\theta}_{\alpha/2,l}, \hat{\theta}_{\alpha/2,u}]$ is taken as $100(1 - \alpha)\%$ confidence interval for θ where $0 \leq \alpha \leq 1$.

And now, if you try to see, this confidence interval has got another thing. If you try to see here, if I try to draw the sampling distribution of something, then you see here on the left and, right hand side, we are trying to take the value, I am trying to divide the α into say two equal parts well; two equal parts are needed when we have a symmetric distribution. If the distribution of statistics is not symmetric, then it is going to beα/2 and 1 minusα/2.

Well, I am trying to give you the basic fundamentals. So, I will try to remain as simple as possible. So, we are interested in finding out the confidence interval. So, I want to find out the confidence interval; such that this area under the sampling distribution of the corresponding statistics is $(1 - \alpha)$ and the area, the shaded area on the left-hand side and the shaded area on the, right-hand side is $\alpha/2$.

And corresponding to which I have to find out here the lower limit of the confidence interval and upper limit of the confidence level; confidence interval and these limits have been defined in such a way such that they are the values on the x-axis here, which are trying to divide the entire area into three parts; part number I, part number II and part number here III and this area has been divided in such a way; such that the area in the mid it is $(1 - \alpha)$ and the sum of the areas on the two extremes it is α.

In this case, this is $\alpha/2 + \alpha/2$. So, essentially, if you remember what are we trying to do? This is the sampling distribution of my statistics whatever I want to compute. So, if I try to obtain various values of statistics and if I try to plot a density curve or a histogram, then essentially this limit, this lower limit of the confidence interval and this upper limit of the confidence interval here, these are actually the value of the percentiles.

And if you remember, we had discussed this thing in case of simple random sampling also. So, here in this figure, I have denoted the $(1 - \alpha)\%$ area in the green shade and the$\alpha/2$ area in the red shade. So, essentially, I have to divide this entire area into say 100 equal parts.

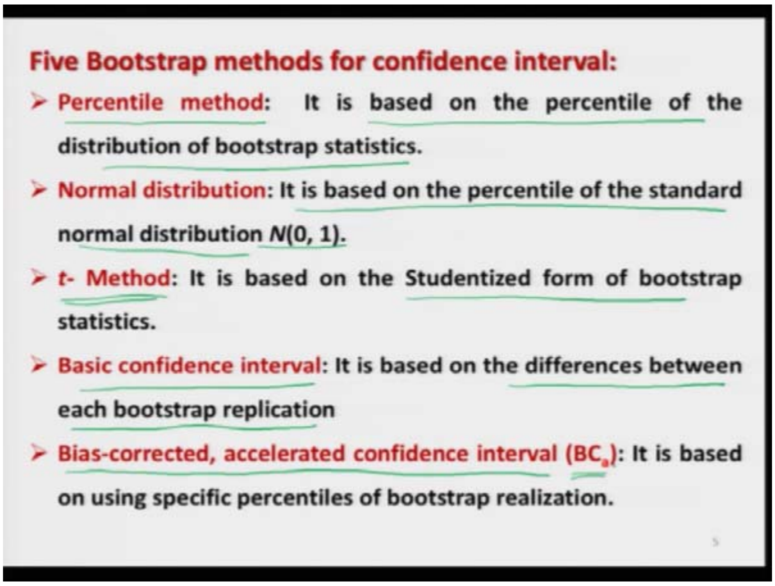
And then, I have to find out here this value here that what is the value of the percentile, which is trying to divide the total area into three equal I mean into three parts, right, such that the middle area is $(1 - \alpha)$ and the remaining two areas on left-hand side and, right-hand side in this case area$\alpha/2$ and $\alpha/2$ each, right.

And if I can find out the value of the percentile, which is indicated by here $\hat{\theta}_{(\alpha/2,\ l)}$ and $\hat{\theta}_{(\alpha/2,\ u)}$ that is lower and upper limits, then this is actually going to give us the $100(1 - \alpha)\%$ confidence interval. So, this is the basic idea behind the confidence interval that the confidence interval can be constructed or viewed what we have seen in this slide under normal or t distribution and in principle, the confidence interval can also be seen from the point of view of percentiles, right.

So, what are we going to do? We have to essentially find out the sampling distribution and then, I have to find out the cut off values which are here, right. So, I am trying to say here that there are $100\alpha/2\%$ values which are below $\theta_{\alpha/2,\text{ lower}}$ and above $\theta_{\alpha/2,\text{ upper}}$. And so, that means, essentially the proportion of the values in the here red zone is$\alpha/2$;$\alpha/2$ and the area of and the proportion of the points in the green zone here is $(1 - \alpha)$, right.

So, if I can obtain such a partition, then the estimates of this $\theta_{\alpha/2,\text{ l}}$ and $\theta_{\alpha/2,\text{ u}}$ they can be considered as $\hat{\theta}_{\alpha/2,\text{ l}}$ and $\hat{\theta}_{\alpha/2,\text{ u}}$ and which can be considered as $100(1 - \alpha)\%$ confidence interval for the parameter θa. So, these are the basic fundamentals which have been used in the construction of bootstrap confidence interval, right.

(Refer Slide Time: 14:44)



Five Bootstrap methods for confidence interval:

- **Percentile method:** It is based on the percentile of the distribution of bootstrap statistics.
- **Normal distribution:** It is based on the percentile of the standard normal distribution *N*(0, 1).
- ***t*- Method:** It is based on the Studentized form of bootstrap statistics.
- **Basic confidence interval:** It is based on the differences between each bootstrap replication
- **Bias-corrected, accelerated confidence interval (BC_a):** It is based on using specific percentiles of bootstrap realization.

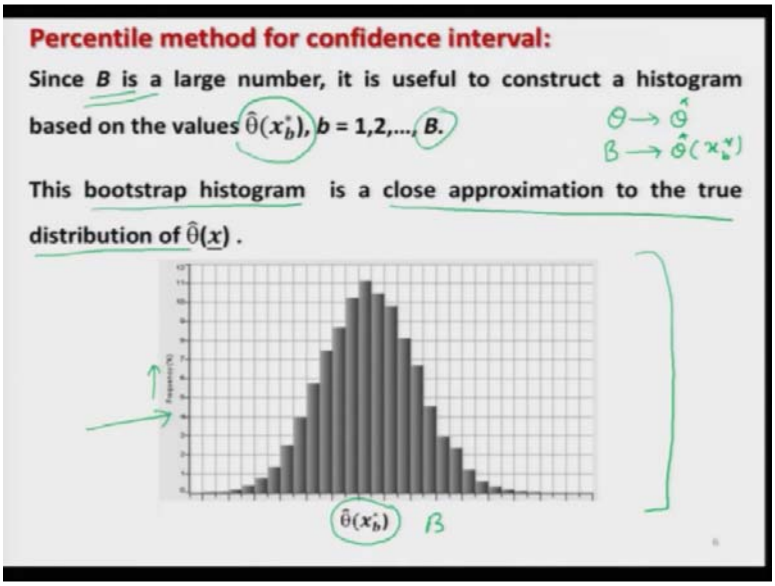
So, in bootstrap methodology, there are five popular ways of finding out the confidence interval. The first methodology is without any order of preference I am saying that in the order in which I am going to consider. The first method we are going to consider here is percentile method. So, this is based on the percentile of the distribution of the bootstrap statistics for example, I have just explained it.

Second thing is this we assume that the distribution of the statistics is N(0, 1) and then, we try to find out the percentile of the standard N(0, 1) distribution and we try to construct the confidence interval. And similarly, as I explained you the normal distribution and t distribution can be used for the confidence interval estimation so, based on that, we have third method which is t method: t-method is I mean t is the t distribution.

So, this is this method is based on the t distribution and we try to and this is based essentially on the studentized form of the bootstrap statistics. And then, the fourth one is basic confidence interval and this is essentially based on the differences between each bootstrap replication.

And finally, we will consider the bias corrected, accelerated confidence interval on which says indicate them by here BC a and it is based on using the specific percentile of a bootstrap realization, which is a different way to compute the percentile.

(Refer Slide Time: 16:19)



Percentile method for confidence interval:

Since B is a large number, it is useful to construct a histogram based on the values $\hat{\theta}(x_b^*)$, $b = 1,2,\ldots, B$.

This bootstrap histogram is a close approximation to the true distribution of $\hat{\theta}(\underline{x})$.

$\theta \to \hat{\theta}$

$B \to \hat{\theta}(x_b^*)$

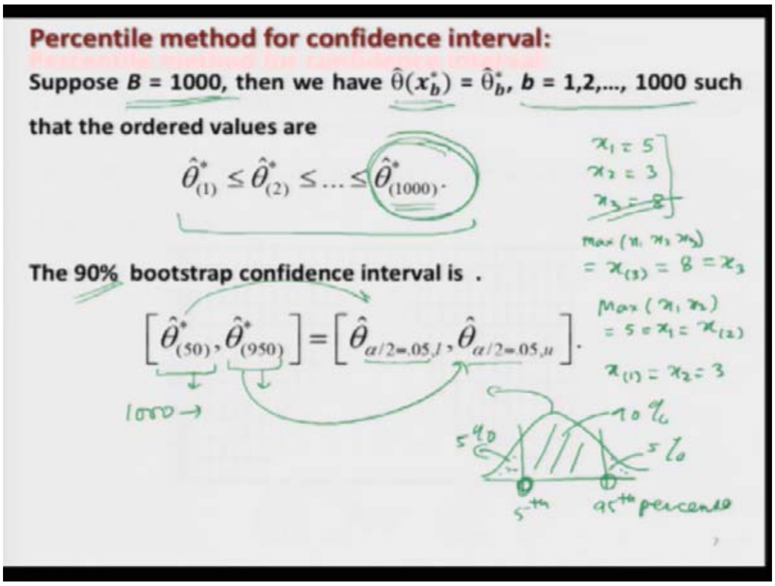
So, first let me try to explain you the basic behind the percentile method of confidence interval. So, we know that we have a parameter of interest θ and we want to estimate it by estimator $\hat{\theta}$, but it is difficult to obtain the finite sample properties of the or the sampling distribution of $\hat{\theta}$ so, we try to use the bootstrap methodology.

And we try to draw capital B number of bootstrap samples and in each samples, we try to compute $\hat{\theta}$ and this is denoted by here $\hat{\theta}(x_b^*)$ star means if you remember $\hat{\theta}_b^*$ was the say bth bootstrap sample. So, in every sample, you have to compute the statistics, and which is denoted by $\hat{\theta}(x_b^*)$.

So, since B usually we try to take a large number so that we can have a good histogram means if the B is very small, then the histogram will not actually look nice, right and if B is large, then the histogram which is constructed using the bootstrap values and that is called a bootstrap histogram is a good approximation is a close approximation to the true distribution of $\hat{\theta}(\underline{x})$, right. So, for example, just to illustrate you and just to convey my idea what I am trying to say, I have just taken here a possible histogram.

So, the values of $\hat{\theta}$ have been computed for say B bootstrap samples and they have been and based on that, this histogram has been computed. So, you can see here on the y axis, we have the frequency, right. So, , this can be anything actually, this is only a representative representation of the histogram that we can obtain.

(Refer Slide Time: 18:20)



Percentile method for confidence interval:

Suppose B = 1000, then we have $\hat{\theta}(x_b^*) = \hat{\theta}_b^*$, b = 1,2,..., 1000 such that the ordered values are

$$\hat{\theta}_{(1)}^* \le \hat{\theta}_{(2)}^* \le \ldots \le \hat{\theta}_{(1000)}^*.$$

The 90% bootstrap confidence interval is .

$$\left[\hat{\theta}_{(50)}^*, \hat{\theta}_{(950)}^*\right] = \left[\hat{\theta}_{\alpha/2=.05,l}, \hat{\theta}_{\alpha/2=.05,u}\right].$$

So, what we can obtain now here? That we have obtained here all the values of the theta hat x b and we can just arrange them in increasing or decreasing order some and then, we have to obtain basically the ordered value. So, suppose if I say suppose I have obtained 1000 bootstrap sample so, B is here 1000.

And I try to rearrange them in the increasing order so, whatever is the maximum value, which is denoted by here $\hat{\theta}_{(1000)}^*$ and I am writing this 1000 inside the parenthesis so that is indicating the ordered value, which is the standard notation. In say statistics for example, if I say here I have two values; three values here, x_1 equal to say 5 and x_2 is equal to say here 3 and x_3 is equal to suppose here 8.

So, out of this, I try to find out the maximum value and the maximum value of among x_1, x_2 and x_3, it is here this is denoted by here x order 3 which is equal to here see here 8 and this is same as here x_3, then now, x_3 is out. So, now, I try to find out the maximum value between x_1 and x_2.

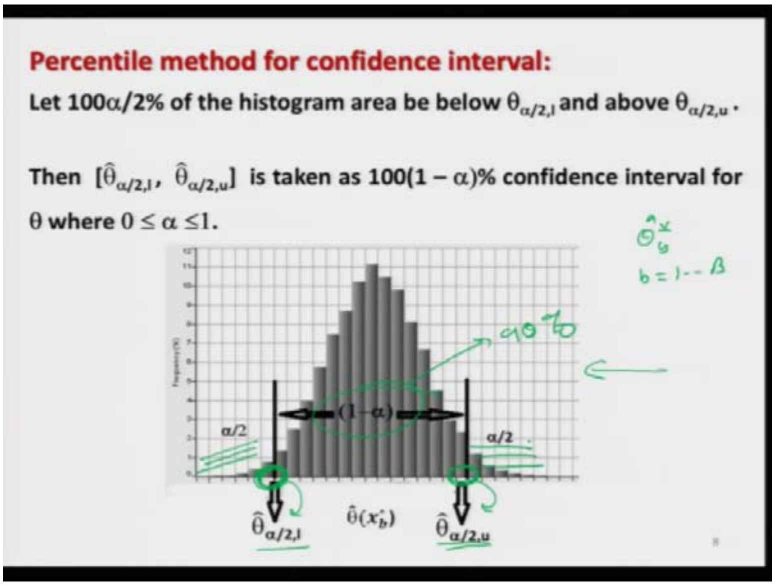
And this comes out to be here say here 5 and this is actually the first value x_1, but as an ordered value, this is the second largest value and obviously, then the maximum value x_1 is denoted is the same as here x_2 and which is here 3. So, this is the meaning of this notation.

So, I try to arrange the values in the increasing order for $\hat{\theta}_b^*$ and then, suppose I want to construct the 90% bootstrap confidence interval. So, what I try to do here? I simply try to; I simply have to obtain the 50th quantile and 950th quantile why?

Because we are taking here 1000 value so, I want here 90% so, what I want here the area in the mid, this is 90% and the area on the two sides is 5% each. So, if I try to construct here the histogram, then this is going to be the say the 5th percentile and this is going to be the 95th percentile.

So, essentially, if I try to consider this curve as a sampling distribution of $\hat{\theta}$, then essentially we are interested in finding out the 5th and 95th percentile, which are in this ordered value they are the 50th and 950th value of $\hat{\theta}^*$, right. So, at α is equal to 10 percent,$\alpha/2$ will become 0.05 and then, I try to find out the value here and here and these are your confidence interval. So, that is what I am trying to give you an idea that how the things are calculated.

(Refer Slide Time: 22:04)



Percentile method for confidence interval:

Let 100α/2% of the histogram area be below $\theta_{\alpha/2,l}$ and above $\theta_{\alpha/2,u}$.

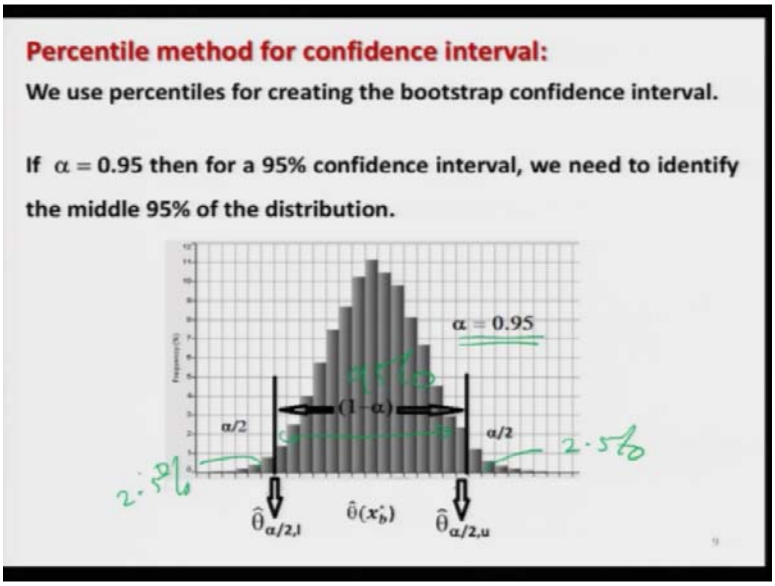
Then $[\hat{\theta}_{\alpha/2,l}, \hat{\theta}_{\alpha/2,u}]$ is taken as 100(1 – α)% confidence interval for θ where $0 \le \alpha \le 1$.

So, now, if I try to do it on the on a histogram so, I assume that we have obtained here suppose large number of values of $\hat{\theta}_b^*$ say b goes from 1 to here B and these values have been plotted here in this histogram and this is the histogram that we have obtained.

So, we try to find out here in on the x axis, we I try to find out the value such that the area on the left-hand side is$\alpha/2$ and the area on the, right-hand side of this point is$\alpha/2$ and whatever are the values that we have obtained, these are the percentiles which are trying to indicate the say lower value of the confidence interval and upper value of the confidence interval.

This is what I mean when I say that I want to construct the histogram of the bootstrap value. So, this area in between is $(1 - \alpha)$ so that so, if I take α is equal to 10% so, this area is going to be 90% and this area on the left-hand side and, right-hand side will be 5 percent, right.

(Refer Slide Time: 23:14)



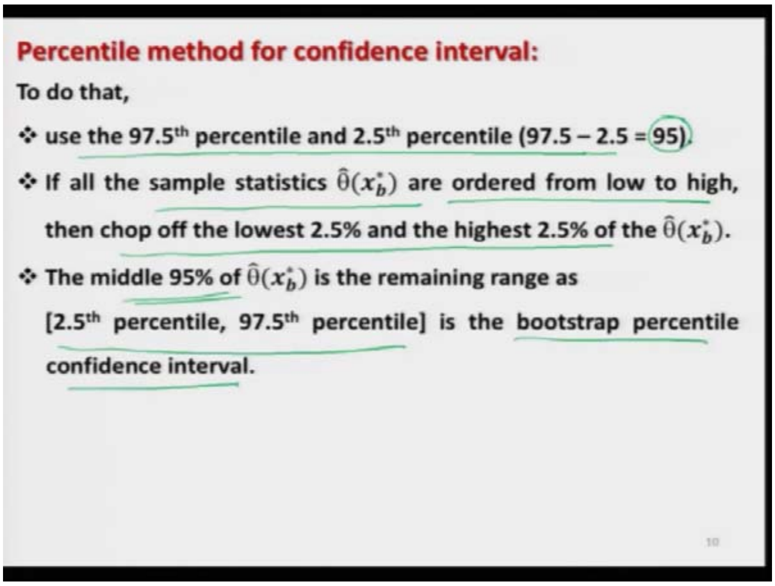
Percentile method for confidence interval:

We use percentiles for creating the bootstrap confidence interval.

If $\alpha = 0.95$ then for a 95% confidence interval, we need to identify the middle 95% of the distribution.

For example, so, I try to show you here the same thing and that if I try to take here for example, α is equal to 95 percent, which is the one of the popular values for the construction of confidence interval, then in this case, you have to take this area to be 95% and this area, and this area will be 2.5% each, right.

(Refer Slide Time: 23:45)



Percentile method for confidence interval:

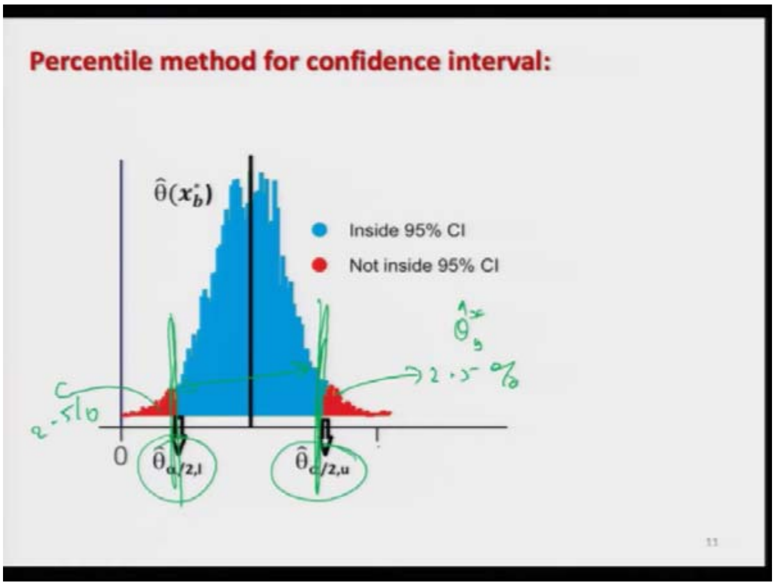
To do that,

- use the 97.5th percentile and 2.5th percentile (97.5 – 2.5 = 95).
- If all the sample statistics $\hat{\theta}(x_b^*)$ are ordered from low to high, then chop off the lowest 2.5% and the highest 2.5% of the $\hat{\theta}(x_b^*)$.
- The middle 95% of $\hat{\theta}(x_b^*)$ is the remaining range as [2.5th percentile, 97.5th percentile] is the bootstrap percentile confidence interval.

So, now if I try to formalize this procedure, what I am trying to do? I am trying to create the histogram and then, I am trying to use the 97.5th percentile and 2.5th percentile and the area between the between these two percentile will be 95 percent, right. So, in order to achieve it, what I can do?

All the sample statistics which I denoted here by $\hat{\theta}(x_b^*)$ are ordered from low to high and then, we try to chop off the lowest 2.5% and highest 2.5% of this distribution, right. And the and whatever is the middle 95% values that will be the 95% of the total area and the 2.5th percentile and 97.5th percentile will give us the lower and upper limit of the bootstrap percentile confidence interval.

(Refer Slide Time: 24:44)

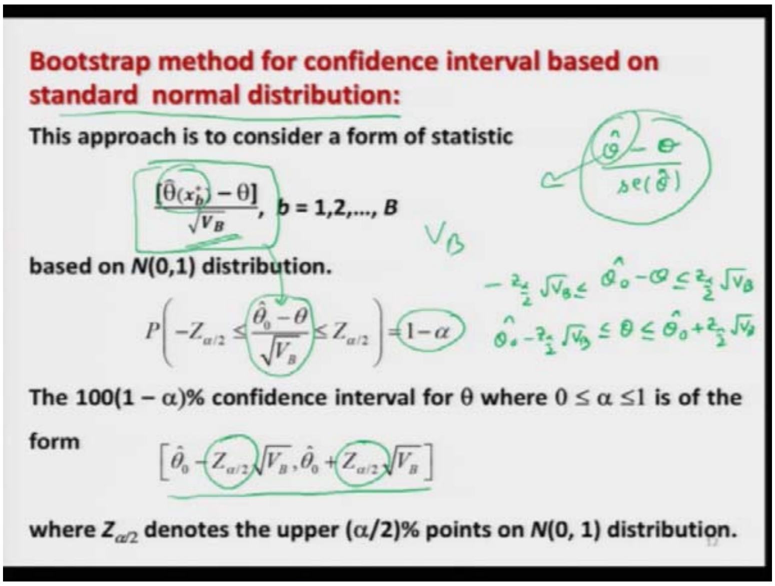


Percentile method for confidence interval:

For example, if you try to see here, I try to make one more this type of plot so, all the values of say this $\hat{\theta}_b^*$, they have been plotted here and then, I try to choose here a point where I can chop off the 2.5% on the left-hand side and 2.5% on the, right-hand side. So, this area has been denoted in color red and that area, which is between the two limits which is 95% it is indicated by blue color, right.

So, this $\hat{\theta}_{\alpha/2,\ \text{lower}}$ is the lower limit of the confidence interval and $\hat{\theta}_{\alpha/2,\ u}$ is the upper limit of the confidence interval and this by this approach, you will get the confidence interval through percentile method. This is one of the most simple method to obtain, ok.

(Refer Slide Time: 25:36)



Bootstrap method for confidence interval based on standard normal distribution:

This approach is to consider a form of statistic

$$\boxed{\frac{[\hat{\theta}(x_b^*) - \theta]}{\sqrt{V_B}}}, \quad b = 1,2,\ldots, B$$

$$\frac{\hat{\theta} - \theta}{se(\hat{\theta})}$$

V_B

based on *N*(0,1) distribution.

$$P\left(-Z_{\alpha/2} \le \frac{\hat{\theta}_0 - \theta}{\sqrt{V_B}} \le Z_{\alpha/2}\right) = 1-\alpha$$

$$-\frac{z_\alpha}{2}\sqrt{V_B} \le \hat{\theta}_0 - \theta \le \frac{z_\alpha}{2}\sqrt{V_B}$$

$$\hat{\theta}_0 - \frac{z_\alpha}{2}\sqrt{V_B} \le \theta \le \hat{\theta}_0 + \frac{z_\alpha}{2}\sqrt{V_B}$$

The 100(1 – α)% confidence interval for θ where $0 \le \alpha \le 1$ is of the form

$$\left[\hat{\theta}_0 - Z_{\alpha/2}\sqrt{V_B}, \hat{\theta}_0 + Z_{\alpha/2}\sqrt{V_B}\right]$$

where $Z_{\alpha/2}$ denotes the upper (α/2)% points on *N*(0, 1) distribution.

So, now I come to the next approach which is based on the use of standard normal distribution N(0, 1). So, in this approach, we simply try to use the basic concept that was used in the classical statistical inference for the construction of confidence interval, right.

So, if you remember, if I try to take any value here $\hat{\theta}$ and if I try to find here find the difference with the θ which is actually unknown and if I try to divide it by a standard error of theta hat usually, it will follow a N(0, 1) particularly, when your number of observations are large.

So, that is why we try to take here capital B to be large enough so that this result holds true. So, now, this theta is θ and θ is means given to us, but $\hat{\theta}$ we do not know how to; estimate, what is the form of $\hat{\theta}$ exact form and so, finding out the exact sampling distribution of this quantity will be difficult.

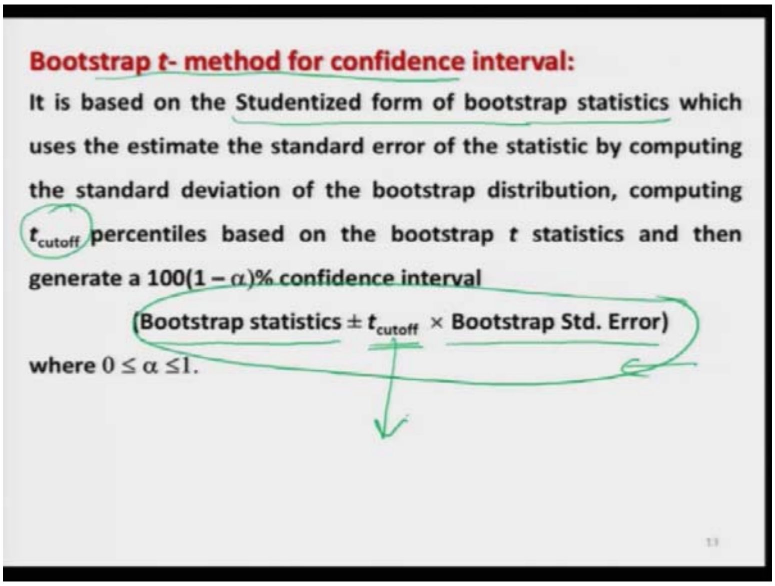
So, what I try to do here? Instead of this $\hat{\theta}$, we try to take here the bootstrap estimator and I try to consider here a statistics of this form, where theta hat is replaced by the bootstrap value of the statistics minus theta and the standard error of $\hat{\theta}$ is difficult to find so, we try to use the bootstrap estimate of the standard error of $\hat{\theta}$. So, that we have denoted by V_B if you remember.

So, now, this statistics can be used to construct the confidence interval. So, in case if I take the confidence coefficient to be (1 − α), then using this result that this statistic lies between -$Z_{\alpha/2}$ and $Z_{\alpha/2}$ where $Z_{\alpha/2}$ is the upper α/2% points on the N(0, 1) distribution.

The using this thing I can simply solve it and you can simply see here that $\hat{\theta}_0 - \theta$ is less than $-Z_{\alpha/2}\sqrt{V_B}$ and this is $Z_{\alpha/2}\sqrt{V_B}$. And if you try to take a θ, this will be here $\hat{\theta}_0 - Z_{\alpha/2}\sqrt{V_B}$ and on the, right-hand side, this will be $\hat{\theta}_0 + Z_{\alpha/2}\sqrt{V_B}$.

So, $\hat{\theta}_0$ is essentially your bootstrap statistics which you have obtained, the bootstrap estimator of the parameter theta. So, this is the confidence interval which you will obtain. So, you can see here, in this method, you are assuming that this distribution is normal, right, so and that is why you are using here the cut off values as here $Z_{\alpha/2}$, right.

(Refer Slide Time: 28:59)



Bootstrap *t*- method for confidence interval:

It is based on the Studentized form of bootstrap statistics which uses the estimate the standard error of the statistic by computing the standard deviation of the bootstrap distribution, computing t_{cutoff} percentiles based on the bootstrap *t* statistics and then generate a 100(1 − α)% confidence interval

(Bootstrap statistics ± t_{cutoff} × Bootstrap Std. Error)

where $0 \le \alpha \le 1$.

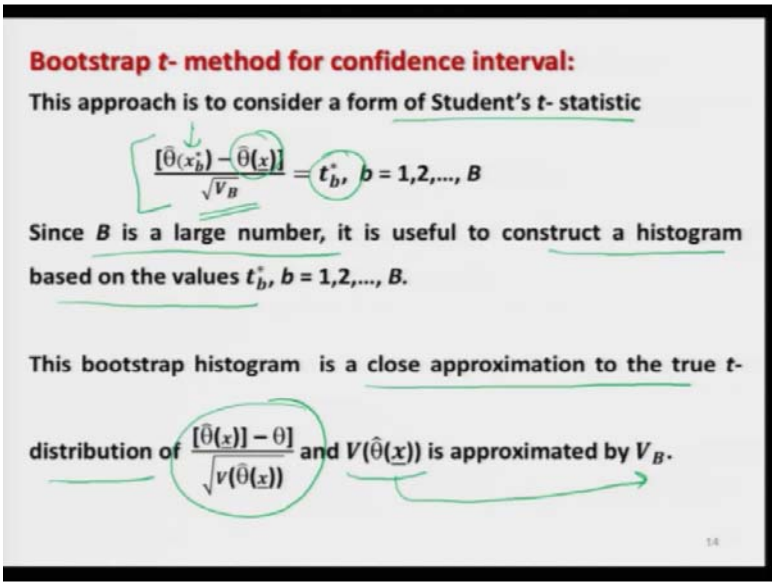
But now, the second approach is suppose I do not know and I can use the concept of t distribution and I would like to do the same thing with the concept of t distribution, right. So, in the case of bootstrap t method for the construction of confidence interval, what we try to do? Instead of taking this type of statistics which we have used here, right, we try to take the statistics which is of the form of which has got the form of a studentized t statistics and we try to create a studentized form of the bootstrap statistics.

And in this thing, the standard error of the statistics can be computed by the bootstrap methodology. Now, there can be two methodology that you try to compute the bootstrap and that you try to use the bootstrap estimator of every replicate or you try to use the bootstrap estimate of the variance.

So, based on that, we have got two different methodologies, but in this case, we are simply trying to say use the t statistics, which has been constructed on the basis of bootstrap distribution and try to compute the cutoff values which are denoted by in general here as a t_{cutoff}.

And then, try to use the general structure of the confidence interval like as bootstrap statistics plus minus the critical value which is denoted as t cutoff and then, the bootstrap standard error, right where this α is between 0 and 1. So, this will also give you another type of bootstrap confidence interval. So, in this slide, I have given you the basic idea behind now, how to do it, let me try to show you here.

(Refer Slide Time: 30:39)



Bootstrap *t*- method for confidence interval:

This approach is to consider a form of Student's *t*- statistic

$$\frac{[\hat{\theta}(x_b^*) - \hat{\theta}(\underline{x})]}{\sqrt{V_B}} = t_b^*, \quad b = 1,2,\dots, B$$

Since B is a large number, it is useful to construct a histogram based on the values t_b^*, $b = 1,2,\dots, B$.

This bootstrap histogram is a close approximation to the true *t*-distribution of $\frac{[\hat{\theta}(\underline{x})] - \theta}{\sqrt{V(\hat{\theta}(\underline{x}))}}$ and $V(\hat{\theta}(\underline{x}))$ is approximated by V_B.

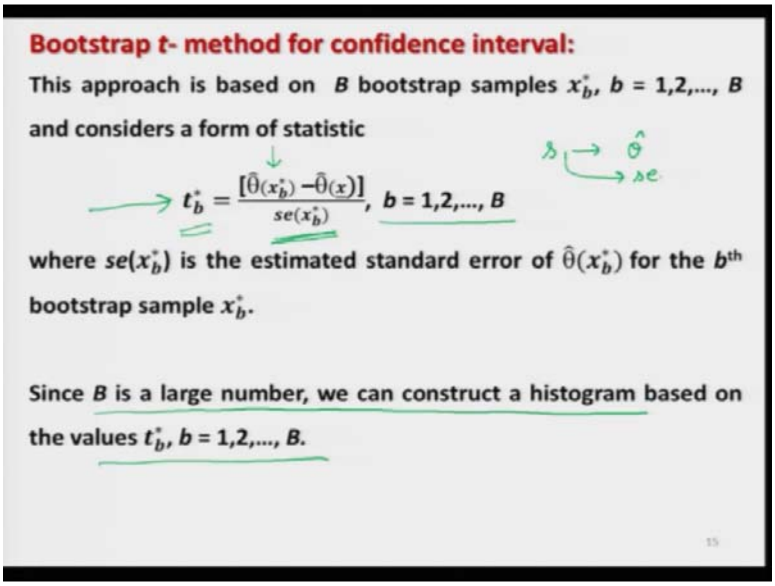
14

So, what we try to do here? In this approach, we try to consider the form of student t statistics. So, this is the value of the bootstrap estimator which we try to obtain in every bootstrap sample and then, I try to construct here the statistics $\hat{\theta}(x_b^*)$ which is from the bth sample minus $\hat{\theta}(\underline{x})$ and divided by the bootstrap standard error of the statistics, right.

And then, we try to take large number of bootstrap sample; such that the approximation holds good and since we have got the large number of observations so, we can construct the histogram of this value which are obtained here we which are denoted here as a t_b^*. So, t_b^* is the value of the studentized t statistics which is based on the bootstrap sample number b, a small b, right.

And in case if you try to construct this type of histogram of this of the values of t_b^*, this will be a close approximation to the t distribution of $\frac{[\hat{\theta}(x_b^*) - \hat{\theta}(\underline{x})]}{se(x_b^*)}$, where the standard error of $\hat{\theta}(\underline{x})$ is closely and well approximated by the quantity V_B. So, that is the basic idea. So, if you try to see, we have used the classical concept of construction of the t statistics and then, we have used it in the construction of a similar statistics based on the bootstrap samples, right.

(Refer Slide Time: 32:31)



Bootstrap *t*- method for confidence interval:

This approach is based on B bootstrap samples x_b^*, $b = 1,2,\ldots,B$ and considers a form of statistic

$$\longrightarrow \underset{=}{t_b^*} = \frac{\overset{\downarrow}{[\hat{\theta}(x_b^*) - \hat{\theta}(\underline{x})]}}{\underline{se(x_b^*)}},\ \underline{b = 1,2,\ldots,B}$$

$s \to \hat{\theta}$

$\quad \hookrightarrow se$

where $se(x_b^*)$ is the estimated standard error of $\hat{\theta}(x_b^*)$ for the b^{th} bootstrap sample x_b^*.

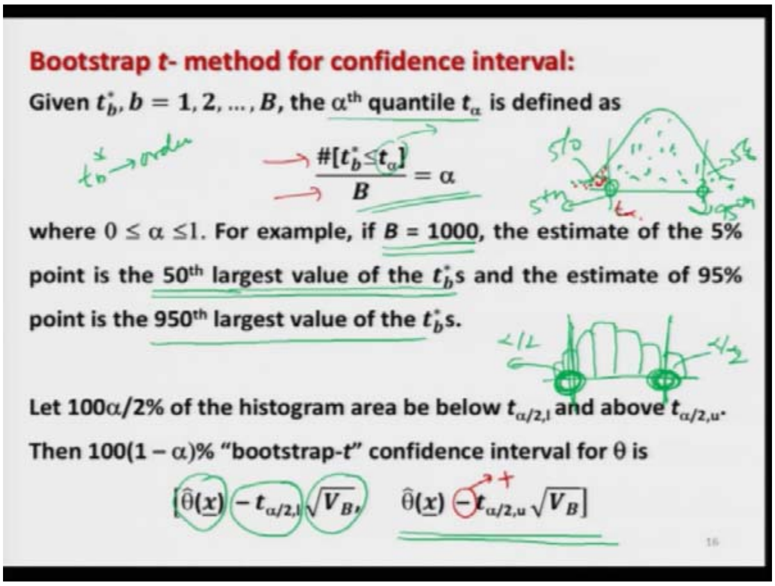
Since B is a large number, we can construct a histogram based on the values t_b^*, $b = 1,2,\ldots,B$.

So, now, the final form of the t_b^* that will emerge is that you that for every bootstrap sample, try to create this t_b^* where the value of the parameter will be obtained for each and every bootstrap sample minus $\hat{\theta}(\underline{x})$ and try to use the bootstrap sample and compute its standard error.

So, you have so, you try to suppose I try to take here the sample number 1, I try to obtain here $\hat{\theta}$ say based on the bootstrap and then, based on this sample s 1 itself I try to find out the standard error because you will have a sample of size small n so, from that, you can compute the simply the variance and then, take its standard, take its say positive square root.

So, that will give you the value of the estimator as well as the value of the standard error for each and every bootstrap sample, right. So, now, you try to repeat this for a large number of bootstrap sample and since the B is a large so, we can construct here a histogram based on this t_b^* values, ok.

(Refer Slide Time: 33:55)



Bootstrap *t*- method for confidence interval:

Given $t_b^*, b = 1, 2, \ldots, B$, the α^{th} quantile t_α is defined as

$$\frac{\#[t_b^* \le t_\alpha]}{B} = \alpha$$

where $0 \le \alpha \le 1$. For example, if $B = 1000$, the estimate of the 5% point is the 50th largest value of the t_b^*s and the estimate of 95% point is the 950th largest value of the t_b^*s.

Let $100\alpha/2\%$ of the histogram area be below $t_{\alpha/2,l}$ and above $t_{\alpha/2,u}$.

Then $100(1-\alpha)\%$ "bootstrap-*t*" confidence interval for θ is

$$[\hat{\theta}(\underline{x}) - t_{\alpha/2,l}\sqrt{V_B}, \quad \hat{\theta}(\underline{x}) - t_{\alpha/2,u}\sqrt{V_B}]$$

Now, once you create this type of histogram, then again you have to means define the critical values. So, as we have discussed, these critical values are essentially the percentiles. So, we try to compute this these percentiles like this means we have here all sorts of values of t_b^* something like this, which are scattered all over the this distribution and we try to find out the proportion of the t_b^* value which are smaller than say here t_α, right. And what is your t_α? t_α is the α^{th} quantile.

So, I am simply trying to say that how so, first you try to find out the α^{th} quantile say 5th quantile, 95th quantile etc. and then, try to see that how many values of the t_b^*, which you have estimated on the basis of bootstrap sample are smaller than t_α. So, you can see here all these values I will try to now mark them in red color, right.

Suppose these are the values which are say smaller than say here t_α, right. So, I try to compute that number and I try to divide it by that total number of values so, this will give me the proportion. So, from there, I can compute this α where α will be between 0 and 1.

So, now, for example, if I say suppose I choose B is equal to 1000, then I need suppose 5% and 95% quantiles on the left-hand side. Suppose this is 5% and this is here 5% so, this value will be 5th quantile and this value on the, right-hand side will be 95th quantile suppose I want to have this thing.

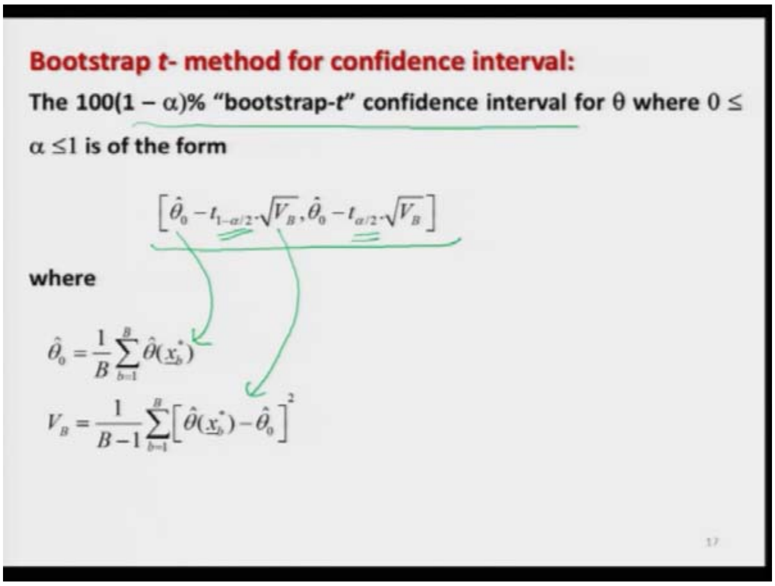
So, what I can do? That I have got here the value of t_b^*, I try to order them as I shown you earlier and then, I just try to choose the largest value among this the values of t_b^* and the 950th largest value of the values of t_b^* and then, these are going to give us the critical values.

Then what will happen? Then, now in case if I try to recreate this thing in the form of some histogram like over here, then what will happen? If I have partitioned this into two parts such that this part is$\alpha/2$% and this part is$\alpha/2$ percent, then whatever are the values on the x axis, they are going to give us an estimate of the t value I mean the critical value, which looks like as if this is based on the t distribution.

And then, I can use the classical approach of the confidence interval estimation something like $\hat{\theta} \pm$ standard error of $\hat{\theta}$ into the critical value, which I have explained you here, you can see here, this approach. So, now, I have estimated this t cutoff using this histogram and then, I can use here $\hat{\theta}(\underline{x})$ and then, the bootstrap, a standard error of $\hat{\theta}$ and then, this critical value is obtained from the, from this constructed histogram, right.

So, $t_{\alpha/2,\ 1}$ is going to denote the say the value on the x axis corresponding to width say$\alpha/2$% points are less than t_α and similarly, I can compute here the upper limit like as here, upper limit, like as here, right. So, this is how we can obtain the confidence limit using the t method.

(Refer Slide Time: 39:05)



Bootstrap *t*- method for confidence interval:

The 100(1 – α)% "bootstrap-*t*" confidence interval for θ where 0 ≤ α ≤1 is of the form

$$\left[\hat{\theta}_0 - t_{1-\alpha/2}\cdot\sqrt{V_B}, \hat{\theta}_0 - t_{\alpha/2}\cdot\sqrt{V_B}\right]$$

where

$$\hat{\theta}_0 = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}(\underline{x}_b^*)$$

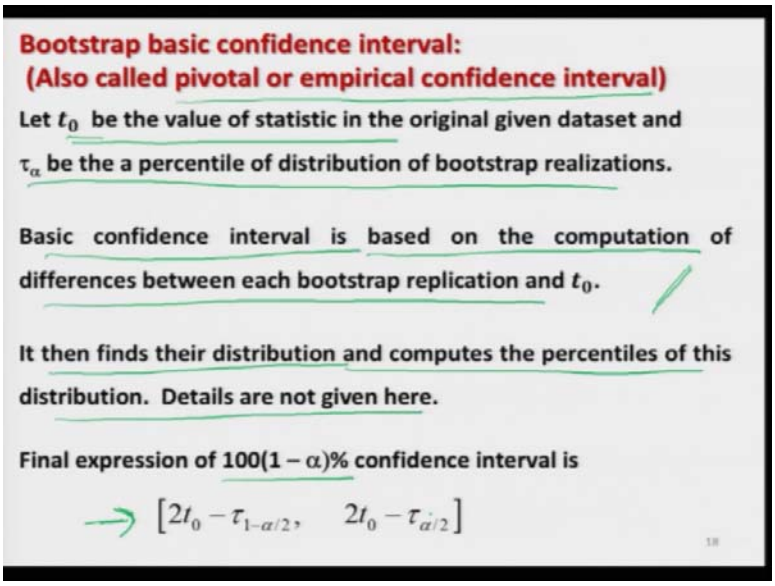
$$V_B = \frac{1}{B-1}\sum_{b=1}^{B}\left[\hat{\theta}(\underline{x}_b^*) - \hat{\theta}_0\right]^2$$

17

And in brief, if I try to comprehend all the results whatever I have used here. In general, I can say the 100(1 – α)% confidence interval based on the bootstrap t methodology for the perimeter theta for α between 0 and 1 will look like 0 hat if you remember, you had defined it to be the bootstrap estimator of theta.

And V_B here is defined as the bootstrap variance, right. And here, I have used here 1 - α/2 and α/2 just to make it here more general that if the distribution is not exactly symmetric, then if you and if you want to use the bootstrap t method over here, then at least you should be able to use it, ok.

(Refer Slide Time: 40:00)



Bootstrap basic confidence interval:
(Also called pivotal or empirical confidence interval)

Let t_0 be the value of statistic in the original given dataset and τ_α be the a percentile of distribution of bootstrap realizations.

Basic confidence interval is based on the computation of differences between each bootstrap replication and t_0.

It then finds their distribution and computes the percentiles of this distribution. Details are not given here.

Final expression of $100(1-\alpha)\%$ confidence interval is

$$[2t_0 - \tau_{1-\alpha/2}, \quad 2t_0 - \tau_{\alpha/2}]$$

18

Now, I give you a very quick review of the bootstrap basic confidence interval means this is another type of approach to construct the confidence interval using the bootstrap methodology, right. It is difficult for me to give you the complete details as we have not covered the entire statistics and you need something more, but I will just try to give you an idea.

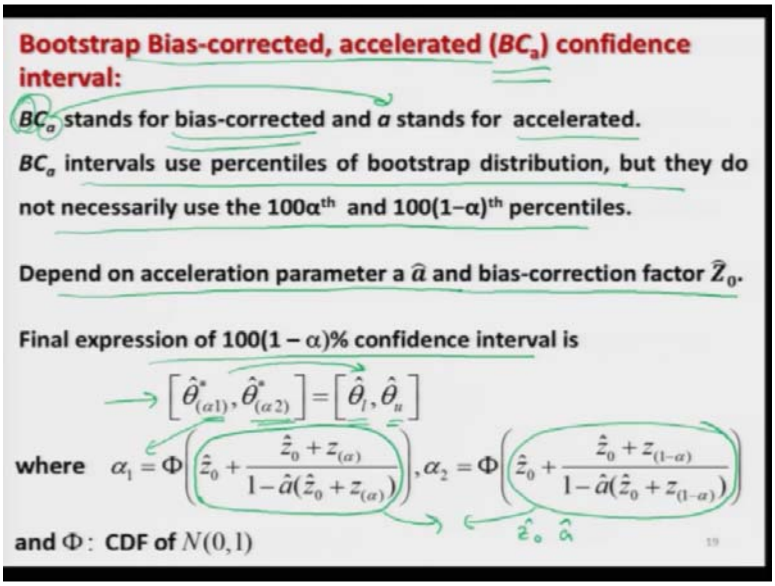
This confidence interval is also called as pivotal or empirical confidence interval and in this case. Suppose t_0 be the value of the statistics in the original sample, right. If you remember in the example of correlation coefficient, I had computed the correlation coefficient hours so; this is the value of t_0 statistics or statistics in the original sample.

And let tau α be the percentile of the distribution of bootstrap realization. So, you try so you try to estimate your parameter theta by $\hat{\theta}$ using the bootstrap estimators and you have large number of values of this estimator and then, you try to construct the bootstrap histogram and from there, you try to compute the percentile.

So, this basic confidence interval is actually based on the computation of differences between each bootstrap replication and t_0 and after this, it tries to find out the distribution and compute the percentile of this distribution, details we are not giving here.

So, we use this approach to construct the, to find out the percentile and then, we try to construct the $100(1-\alpha)\%$ confidence interval using this expression. So, after doing algebra, you will obtain this expression twice of t_0 minus $\tau_{1-\alpha/2}$, twice of t_0 minus $\tau_{\alpha/2}$, right.

(Refer Slide Time: 42:01)



Bootstrap Bias-corrected, accelerated (BC_a) confidence interval:

BC_a stands for bias-corrected and a stands for accelerated.

BC_a intervals use percentiles of bootstrap distribution, but they do not necessarily use the $100\alpha^{th}$ and $100(1-\alpha)^{th}$ percentiles.

Depend on acceleration parameter a $\hat{a}$ and bias-correction factor $\hat{Z}_0$.

Final expression of $100(1-\alpha)\%$ confidence interval is

$$\longrightarrow \left[\hat{\theta}^*_{(\alpha 1)}, \hat{\theta}^*_{(\alpha 2)}\right] = \left[\hat{\theta}_l, \hat{\theta}_u\right]$$

where $\alpha_1 = \Phi\left(\hat{z}_0 + \dfrac{\hat{z}_0 + z_{(\alpha)}}{1-\hat{a}(\hat{z}_0 + z_{(\alpha)})}\right), \alpha_2 = \Phi\left(\hat{z}_0 + \dfrac{\hat{z}_0 + z_{(1-\alpha)}}{1-\hat{a}(\hat{z}_0 + z_{(1-\alpha)})}\right)$

and Φ: CDF of $N(0,1)$

$\hat{z}_0$ $\hat{a}$

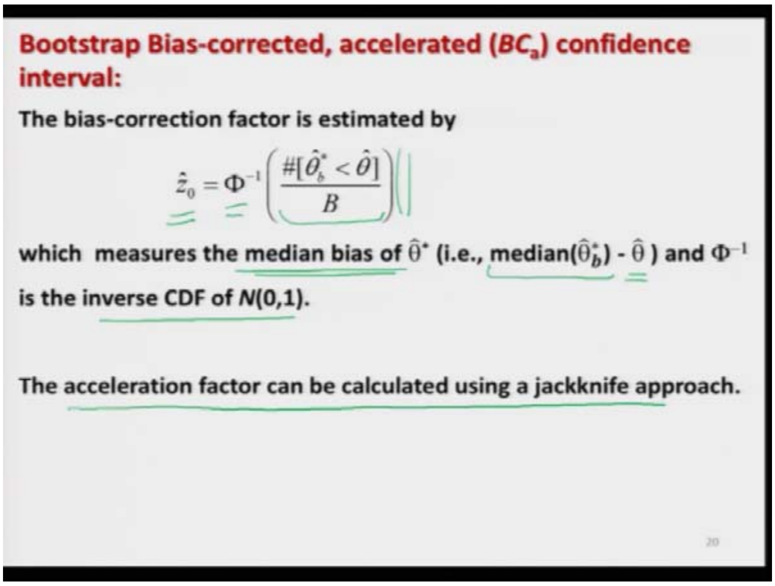
19

So, similarly I will try to give you a quick review of another approach to construct the bootstrap confidence interval, this is bootstrap bias corrected accelerated confidence interval, which is shortly called as BC a, a is in the subscript. And BC a stands for B stand for bias-corrected, this BC stands for bias corrected and this here a, this stands for accelerated.

So, actually this BC_a intervals use percentile of the bootstrap distribution, but they do not necessarily use the 100 alpha'th or $100(1-\alpha)$'th percentile. Actually, they are computed in a different way and it depends on the something called acceleration parameter and bias-correction factor $\hat{z}_0$. So, once you try to do all the algebra and all the statistical tool, then the final expression of $100(1-\alpha)\%$ confidence interval comes out to be here like this, which the lower limit is converted by $\hat{\theta}^*_{(\alpha 1)}$ and upper limit is denoted by $\hat{\theta}^*_{(\alpha 2)}$ inside the parenthesis.

And they are in briefly, I am trying to denote the confidence limit which we are going to obtain on the basis of data as $\hat{\theta}_l$ and $\hat{\theta}_u$, right, where this α1 is actually the value from the CDF of N(0, 1) and this α1 is the value of the CDF at this point, right. And similarly, this α 2 is the value of the CDF of N(0, 1) at this point and you can see here that these two values, which I have marked here, they depend on say factor like $\hat{z}_0\hat{a}$.

(Refer Slide Time: 44:40)



Bootstrap Bias-corrected, accelerated (BC_a) confidence interval:

The bias-correction factor is estimated by

$$\hat{z}_0 = \Phi^{-1}\left(\frac{\#[\hat{\theta}_b^* < \hat{\theta}]}{B}\right)$$

which measures the median bias of $\hat{\theta}^*$ (i.e., median($\hat{\theta}_b^*$) - $\hat{\theta}$) and Φ^{-1} is the inverse CDF of N(0,1).

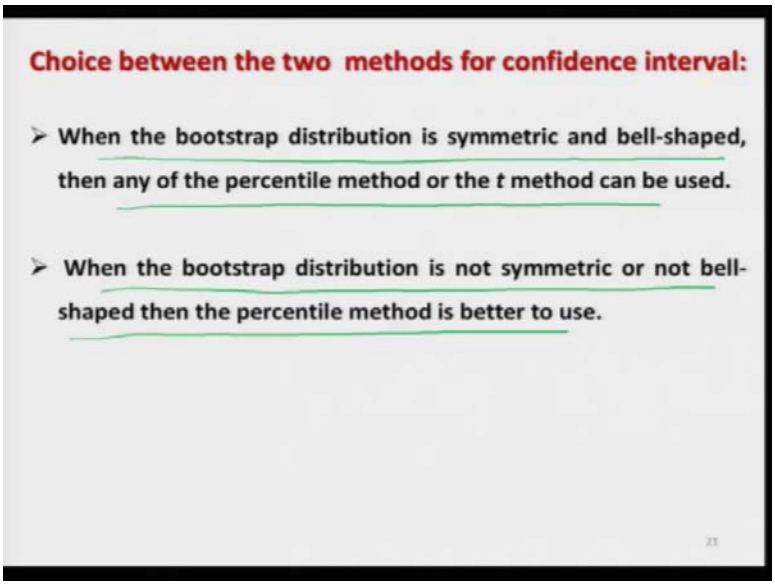
The acceleration factor can be calculated using a jackknife approach.

20

Now, how to compute those these two values? So, this bias-correction factor in this case is estimated by $\hat{z}_0$ and if you try to see, it is based on the inverse of the CDF of N(0, 1). So, Φ^{-1} is indicating the inverse CDF of inverse cumulative distribution function of N(0, 1).

And this is based on the value of the proportion of the values of bootstrap statistics, which are smaller than $\hat{\theta}$, right. And this actually bias correction factor essentially measures the median bias of theta hat star that means, the median value of $\hat{\theta}_b^*$ - $\hat{\theta}$, right. And second factor which is the acceleration factor actually, this is calculated using the jackknife approach, which we are not using here so, I am not giving you here the details, right.

(Refer Slide Time: 45:38)



Choice between the two methods for confidence interval:

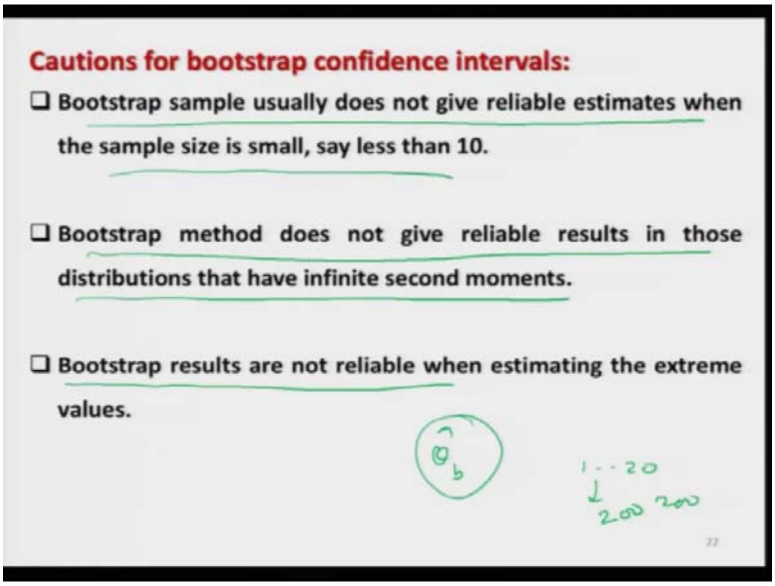
- When the bootstrap distribution is symmetric and bell-shaped, then any of the percentile method or the *t* method can be used.
- When the bootstrap distribution is not symmetric or not bell-shaped then the percentile method is better to use.

21

Now, just to give you an idea that between these two types of methods which are based on percentile and on the; and on the concept of using the sampling distribution, how one can choose? Well, these are only the guidelines, but they have to be followed in a proper way.

So, what I can say in a very simple language that when the bootstrap distribution is symmetric and bell-shaped, then any of the percentile method or the t method can be used. And when the bootstrap distribution is not symmetric or not bell-shaped, then the percentile method is better to use, right. And similarly, there are other type of constraints for other and other types of situation where these different types of confidence intervals can be used.

(Refer Slide Time: 46:29)



Cautions for bootstrap confidence intervals:

- ❑ Bootstrap sample usually does not give reliable estimates when the sample size is small, say less than 10.
- ❑ Bootstrap method does not give reliable results in those distributions that have infinite second moments.
- ❑ Bootstrap results are not reliable when estimating the extreme values.

$\hat{\theta}_b$

1 - - 20

200 200

Now, before I conclude the lecture, let me try to give you some cautions, which has to be kept in mind when we are trying to use the bootstrap confidence interval. Definitely, you know that bootstrap confidence interval are really trying to approximate the true confidence interval. So, definitely when you are trying to approximate, you have to be careful, and the approximation has to be used at a, right place and in the correct situation.

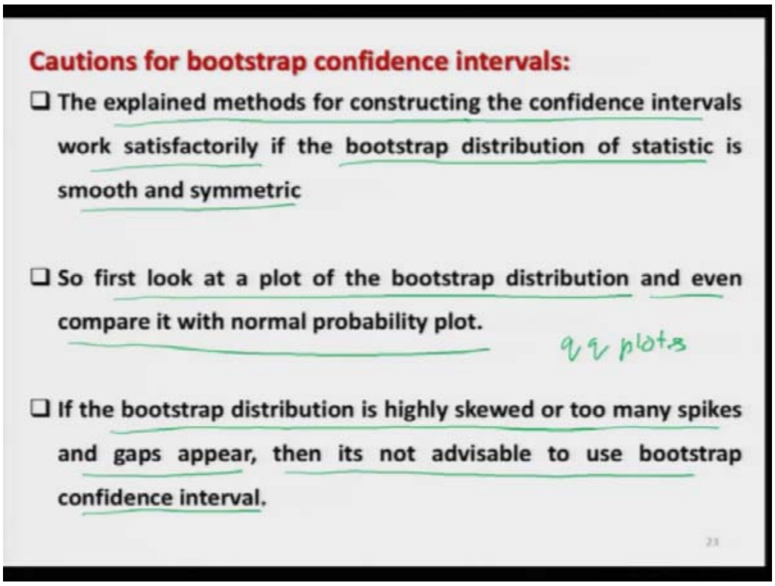
So, first caution is that the bootstrap sample usually does not give reliable estimate when the sample size is small, say a very small. Say as a rule of thumb, I can say if the sample size is smaller than 2, then it is difficult to believe on the accuracy of the, this bootstrap confidence intervals.

And the bootstrap method does not give reliable results in those distribution that have infinite second moments. There are some situations where the second moment does not exist. So, when the moment does not exist, you cannot estimate the variability using the bootstrap variance and so, in that case, it is not advisable to use the bootstrap confidence interval.

And when you see that the values are in such a way such that the sample values have got some extreme values means you try to estimate $\hat{\theta}^{2*}_{b}$ and if you try to that see from the histogram that some of the values are extreme, they are quite away from the usual values for example, all the values are between say 1 to say here 20 and suddenly there is some value which is 200 or 2000 or something like this.

Then in those situation when we have extreme values in the data in the bootstrap values, then this bootstrap confidence interval are not reliable. Actually, in this case, even the classical statistics has different types of tools to handle the situation so, the same story continues in the case of bootstrap samples also.

(Refer Slide Time: 48:38)



Cautions for bootstrap confidence intervals:

- □ The explained methods for constructing the confidence intervals work satisfactorily if the bootstrap distribution of statistic is smooth and symmetric
- □ So first look at a plot of the bootstrap distribution and even compare it with normal probability plot.

q q plots

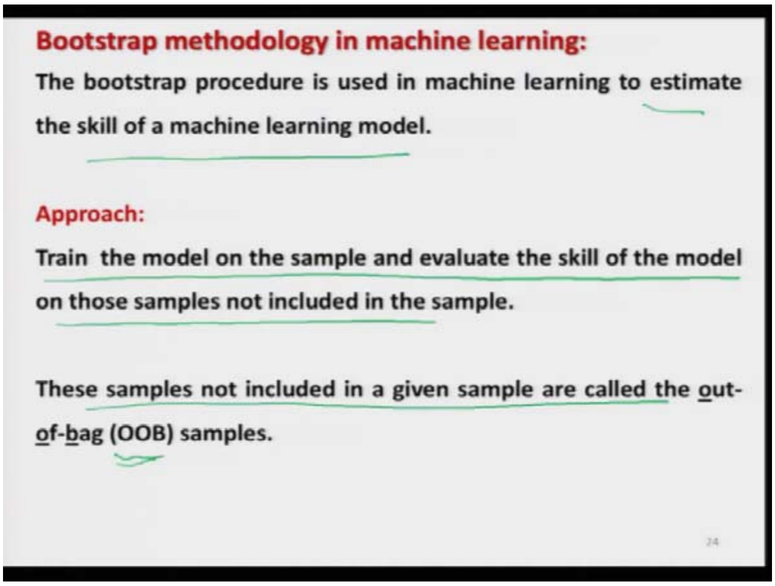
- □ If the bootstrap distribution is highly skewed or too many spikes and gaps appear, then its not advisable to use bootstrap confidence interval.

And whatever the explained methods we have considered here for the construction of confidence interval, they were quite satisfactorily if the bootstrap distribution of the statistics is smooth and symmetric. This idea you can get it get from the by plotting the bootstrapped values using histogram or plotted density and there are several types of option in R to plot such values and from there, you can have a fair idea whether you have to use the bootstrap confidence interval or not.

So, what is the rule? That first we try to look at a plot of the bootstrap distribution and many times, the statistical tools are based on the normal distribution. So, in that case, we would also like to compare whether the bootstrap distribution is close enough to the normal distribution or not. So, this can be achieved by normal probability plot.

In R, there is an option what is called as q-q plots, they can help you in plotting the normal probability plot and I will show you in the R that it is automatically obtained when you are trying to do the bootstrap confidence interval. And if the bootstrap distribution is highly skewed or there are too many spikes and gaps which are means clearly appearing in the histogram or the density curve, then in those situations it is not advisable to use the bootstrap confidence interval. And but that is also true for the usual tools of statistical inference, ok.

(Refer Slide Time: 50:17)



Bootstrap methodology in machine learning:

The bootstrap procedure is used in machine learning to estimate the skill of a machine learning model.

Approach:

Train the model on the sample and evaluate the skill of the model on those samples not included in the sample.

These samples not included in a given sample are called the out-of-bag (OOB) samples.

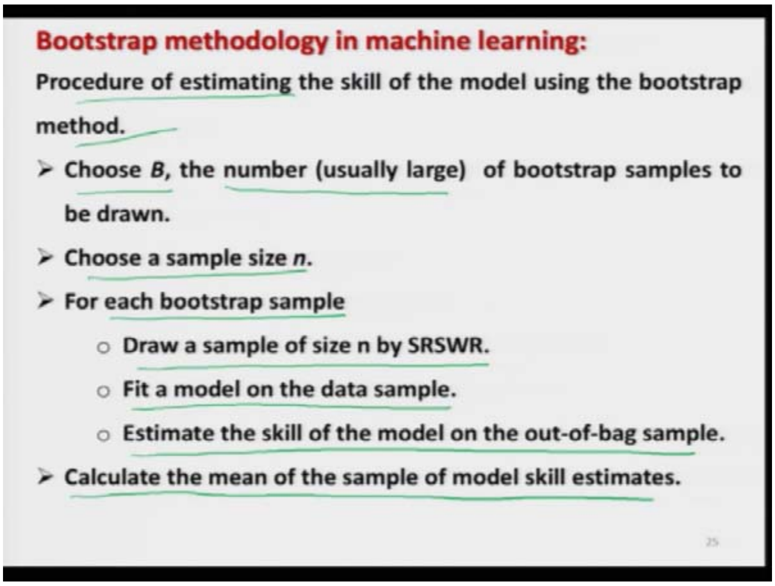
24

So, after giving you this limitation, now just to make you convinced, I try to give you an example that this that how this bootstrap and methodology is being used in data sciences. So, machine learning is a part of the data sciences which we all know. So, this bootstrap methodology is used in machine learning in a with a different name and this is actually, this procedure is actually used to estimate the skill of a machine learning model.

But, before I start, let me make it very clear that I am not really an expert in machine learning; I know little bit to survive, right. So, do not expect too µch and do not ask many questions, right.

So, the approach here is very simple that first we try to train the model on the sample and evaluate the skills of the model on those samples not included in the sample. And these samples not include in the given sample are called as out-of-bag sample and they are briefly denoted by OOB sample, that is the popular terminology which is used in machine learning, right.

(Refer Slide Time: 51:40)



Bootstrap methodology in machine learning:

Procedure of estimating the skill of the model using the bootstrap method.

- Choose *B*, the number (usually large) of bootstrap samples to be drawn.
- Choose a sample size *n*.
- For each bootstrap sample
  - Draw a sample of size n by SRSWR.
  - Fit a model on the data sample.
  - Estimate the skill of the model on the out-of-bag sample.
- Calculate the mean of the sample of model skill estimates.

25

And how do we estimate the skill of the model using the bootstrap method? So, this is here the procedure. The first step is this try to choose B, which is your number of bootstrap samples and usually, it is taken to be a large number and then, you try to draw a sample of size n and then from there, you try to draw B number of bootstrap sample.

and for each of the bootstrap sample, draw a sample of size n by SRSWR usual approach, then you try to fit a model on the data sample, estimate the skill of the model on the out-of-bag sample and after that once you obtain this thing, calculate the mean of the sample of model skill estimates. So, you can see here, if you try to translate this terminology or if you ask me to translate this terminology in the classical statistic this is nothing, but what we have done in the bootstrap methodology, right.

So, now, this is the time to stop after having a long lecture. Well, that was important and but before I go further, let me assure you whatever I have taught you in this long lecture, in R they can be done in a fraction of second, you simply have to use one command and it will give you all sorts of confidence interval, but as I said my objective is not to produce compounders, I want to produce good qualified data scientist. So, that is why I have taken this topic over here.

Now, once I try to use the R software, you will see it will give you all the four or five outcomes of based on this confidence interval, but your trouble is this which one to choose and what are the, what these things are giving you, all the confidence limits you will see they are going to be quite different from each other. So, this type of knowledge, this type of basic fundamental will help you in choosing the correct confidence interval.

And when you are using the this bootstrap methodology in a bigger data set having millions and billions of observation, you will not haven't have any opportunity to look into the data, you will simply be doing first the data cleaning and during data cleaning, you have to take care whether your observations have spikes or say gaps or extreme value and then, based on that you have to choose the correct confidence interval so, that was my objective to choose these topics over here before I go to R software.

So, you try to please revise this thing, try to settle down these things and try to read from the book also means I have followed this lecture from the book of Bradley Efron, which I told you earlier. So, you can have a look in that book, you will get more knowledge, more clarity and I will see you in the next lecture, till then, goodbye.