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 - 33
Bootstrap Methodology
What is Bootstrap and Methodology

Hello friends, welcome to the course Essentials of Data Science with R Software- 2 where we are trying to learn the topics of Sampling Theory and Linear Regression Analysis, in this module on the Sampling Theory with R Software we are going to begin with a new chapter on Bootstrap Methodology.

So, the 1st question comes what is this bootstrap? What is this bootstrapping methodology? Well, in the first line I can assure you that is a very useful thing and once you are aiming for the data science this is going to be a tool which you just cannot ignore, now the question is why? You see whenever you are trying to work in a statistics; you are always trying to work in a finite sample.

And as a trained mathematical statistician I can share with you my experience that, when you try to do some algebra, do some mathematics with the statistical tool many a time it becomes too complicated to handle; in those situations we have an option that ok we say ok we are unable to find out the properties of the estimator say in finite sample we are unable to find it out.

So, we try to see what happens if the sample size become large, and we try to employ different types of statistical approximation. One of the popular approximation theory is large sample approximation theory this means, what happens to the estimator or it is properties, when the sample size grow large but, obviously those properties are going to be valid only when the sample size is extremely large in practice it may not really happen.

That is the first thing. So, the question is what will happen to the properties of the estimator in a finite sample? This has to be known to us. And if not exactly means I will be happy if I can have a good approximation. Second thing is this you may recall that when I discuss the estimation of population mean during simple random sampling then, we discuss that why do we prefer to use sample mean as an estimator for the population mean.

We can also use something else- sample median, sample mode, sample geometric mean, sample harmonic mean, and also we have considered only the estimation of population mean and variance, but there can be other things also, there can be correlation coefficient, there can be coefficient of variation or there can be any complicated structure that you would like to estimate. Why you would like to estimate? Because, those things are making sense in your experimental setup.

So, at that moment saying that ok as a statistician since I cannot find out the finite sample properties of this estimator means you should not use it. That is a wrong sentence, that is not really acceptable in many situation under these type of circumstances this bootstrap methodology comes as a savior for us and it try to help us.

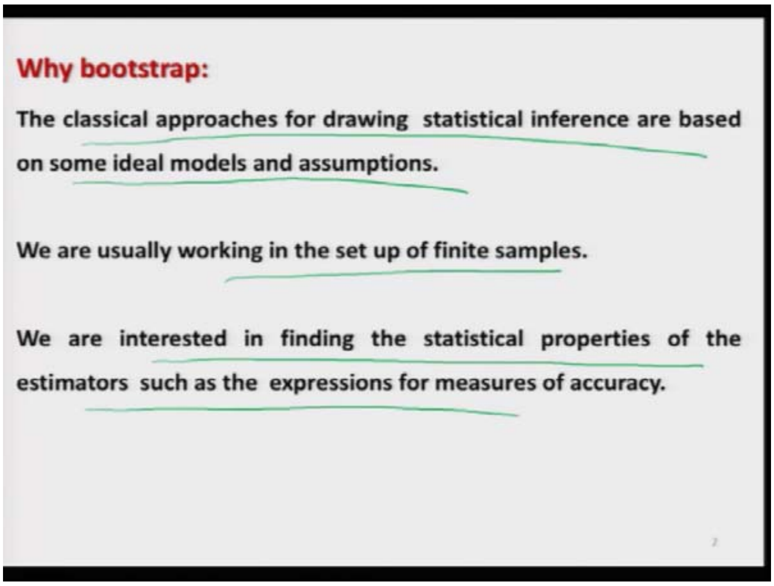
It will give us as a reasonably good solution at least something is better than nothing, but whatever is something that is also good that is not bad. So, that is the basic idea of the bootstrap methodology and it is a sort of computational technique. So, whenever you have got a sample just using that sample you can estimate anything you can estimate bias, you can estimate a standard error, you can estimate variance, you can conduct test of hypothesis, you can construct confidence interval or whatever you want to do those things are possible.

Yes, the success of the approximation or the goodness of the approximation depends on the structure of the estimator which you are trying to handle. Means, if you have a complicated structure possibly the results may not be that good as expected, but they will be not bad also.

So, this bootstrap is a bootstrapping is a computational procedure which is based on simple random sampling with replacement and it gives us a very simple way out simple formulation of the problem so, that we can execute it and we can find out the answers of the question like, finding out bias, standard error, confidence interval, etc, from the given set of data only you do not need any other information from outside ok.

So, this is the basic idea behind the bootstrapping methodology. So, let us start our discussion my objective in this chapter is this I will try to give you a reasonable introduction from the statistical point of view, but my focus will be more on the computation side. I will not be giving you the detail proofs etc, of the results, but I will simply state the result and I will show you how they can be used to obtain the properties of the estimator, right.

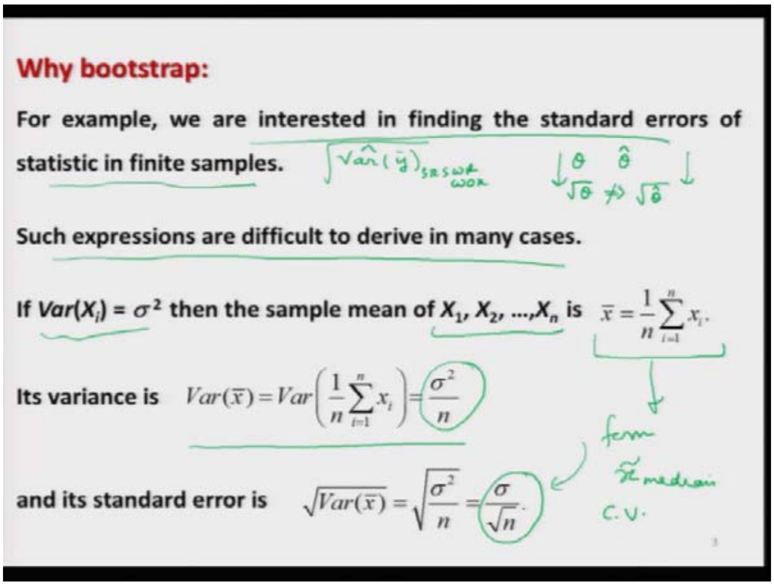
(Refer Slide Time: 06:19)



So, let us begin our lecture. So, whenever we are trying to do the classical inference then what we try to do? That we try to draw some sample and those samples are based on some ideal assumptions and ideal models. And such type of assumptions may or may not really work in finite sample and in real life experiment, they are difficult to meet.

So, in these type of conditions under these type of situation we are more interested in finding the statistical properties of the estimator such as for example, standard error or a absolute deviation or any measure of accuracy, confidence interval or anything else.

(Refer Slide Time: 07:02)



Now, let me take a very simple example to explain you what do we really mean by the finite sample properties and what is our interest and how do we fulfill it. Suppose we are interested in finding out the standard errors of any statistics in finite sample, right. And those expressions are difficult to derive in many many cases. You may recall that when we discussed the simple random sampling at that time we had discussed the we had obtained the variance of $\bar{y}$ under SRSWR and WOR.

And then we had estimated it and after that we had found its standard error. At that time I had explained you in detail that the properties of the estimator of variance and the properties of the standard errors are not the same, and at that time I had discussed a result that if I have a parameter θ and suppose if the estimator of θ is $\hat{\theta}$; and now suppose we are interested in estimating the $\sqrt{\theta}$ then, it does not imply that the estimator of $\hat{\theta}^2$ of $\hat{\theta}$ will be square root of $\hat{\theta}$, right.

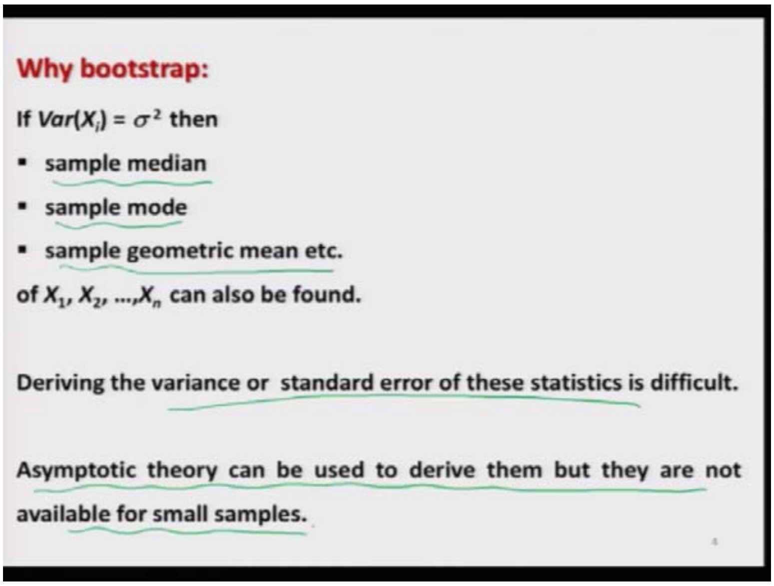
So, it does not mean that if you try to change the form of the estimator like as from θ to $\sqrt{\theta}$ then, you can also say that the form of the estimator will also change from $\hat{\theta}$ to $\sqrt{\hat{\theta}}$. , that is possible under certain type of condition, but not always. Those conditions are discussed in this statistical inference and then this is not the platform where I would be talking about those conditions, right.

Suppose I take a very simple example, suppose I have got a sample say $X_1, X_2, \ldots, X_n$ and where I am assuming that variance of each of X each of the X_i is constant σ^2. And suppose I consider sample mean. So, sample mean is defined at $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ now, if I try to find out the variance of $\bar{x}$.

So, will this will come out to be σ^2/n and a if I try to find out this the standard error that is going to be $\sqrt{\text{var}(\bar{x})}$ which will be here $\sigma / \sqrt{n}$. So, in this case you can see that this form $\bar{x}$ is pretty simple and we are able to find out the exact form of the standard error. What if I try to take any other estimator? Suppose now let me take here simply here sample median.

Then do you think that can you really obtain a neat and clean clear exact expression for the standard error and suppose if I take it take here coefficient of variation then if and then, if I ask you can you find out this the standard error of the coefficient of variation that might be very very difficult.

(Refer Slide Time: 10:02)

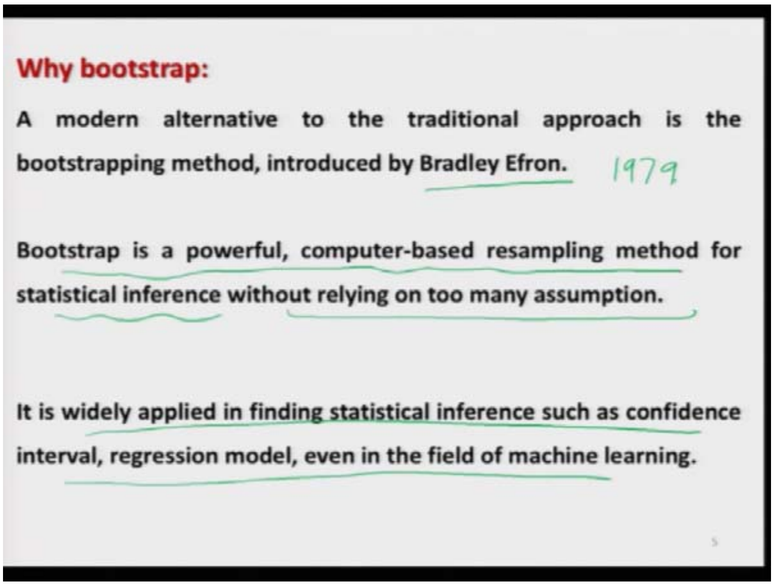


So, under the same situation instead of sample mean if I try to consider the sample median, sample mode, or say sample geometric mean or any other estimator any other statistics, then those statistics can be found, but finding out their variance or standard error is difficult.

And I am not saying impossible it is possible that somebody who is good in mathematics can find it out, but then the, but then there is a doubt that even if those expressions can be found sometime it may be so complicated that it is difficult to understand what they are trying to convey. And in such cases we try to approximate them, and one of the popular approximation is achieved by the asymptotic theory.

So, for example, the last sample asymptotic theory can be used to derive the approximations for the variance of standard errors, but definitely as the sample as the name suggest that asymptotic theory means when the sample size is large. So, when you are trying to find out the expressions which are valid for large sample size their validity in the small sample become questionable, and which may not really be a good option ok.

(Refer Slide Time: 11:24)

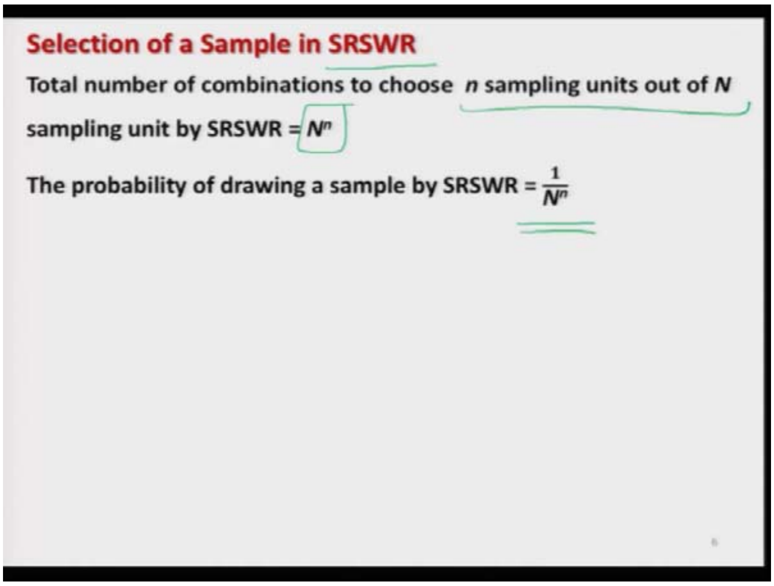


So, a modern alternative traditional approach is the bootstrapping method and this approach was introduced by Bradley Efron in 1979, right. Bradley Efron is a Professor at Stanford University in the Department of Statistics and now I think he is about 80 or say 81 years old, if you want to find out his detailed dig you can find it on the webpage of Stanford University, right.

So, he introduced the bootstrapping method. This bootstrapping method is a very powerful computer based resampling method for statistical inference and it does not relying on too many statistical assumption. So, now, you can see although this technique was proposed in 1979, right; this is almost 30 years 40 years back, right.

But since development in the technique gained importance and it became a very popular tool among the data scientist. And this technique is widely applied in finding statistical inference such as confidence interval, regression model, field of machine learning, whatever you want means there is no end.

(Refer Slide Time: 12:51)



And it is pretty simple to apply. How it can be used? First let me try to create here a base and I would like to give you a confidence that those who are thinking at this moment that they do not know bootstrap. I will try to prove them they are wrong, but they can be partially correct, right.

Because, now they know all the ingredients of the bootstrapping methodology, the only thing is this that they have to just combine those techniques together. Now if I ask you do you know simple random sampling? Answer is yes. Do you know the simple random sampling with replacement? Answer is yes. Do you know how to take the samples by simple random sampling with replacement? Answer is yes.

Now, if I say do you know how to draw the sample by SRSWR through a software? Answer is yes. Now next I ask you do you know how to compute the variance of an estimator? I am not saying sample mean I am saying any estimator. The answer is yes.

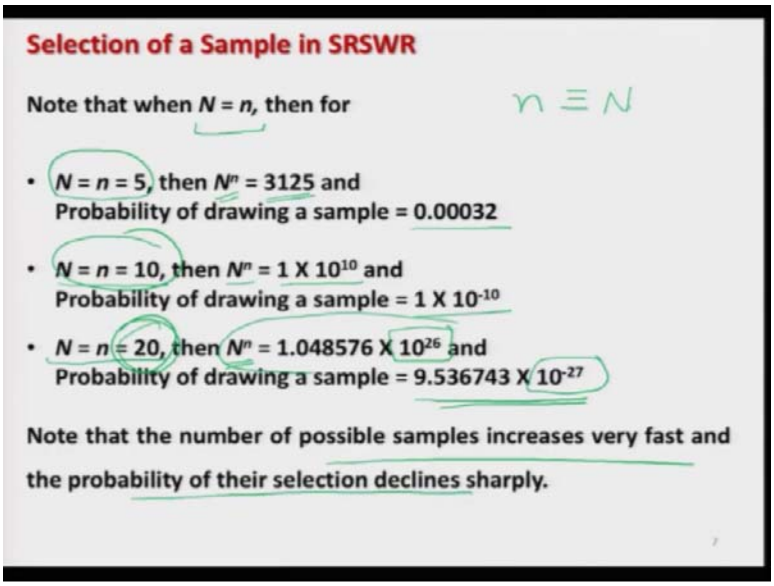
That is simply expected value of x_i minus or say expected value of the statistics minus it is mean whole square that you know for example, if I ask you how to find out the variance of sample mean you know how to find it out.

And if I ask you can you find out the variance of the statistics or for example, the sample mean for a given sample? Answer is yes. Now, these are the ingredients of bootstrap sampling. So, you know each and everything. The only thing which I need to tell you here how to combine those ingredients together to give you a final result, ok. So, let us begin. Now try to observe ok.

You have considered the simple random sampling with replacement in which you have a population of size capital N and from that population you try to draw a sample of size small n, that we know how to get it done. And the total number of samples which can be drawn from the population of size capital N and the samples are of size small n.

This is capital N^n and the probability of drawing a sample by SRSWR is $1/N^n$. Now, if you try to look at this probability. What you try to do in simple random sampling? You say well all the samples are drawn independently.

(Refer Slide Time: 15:37)



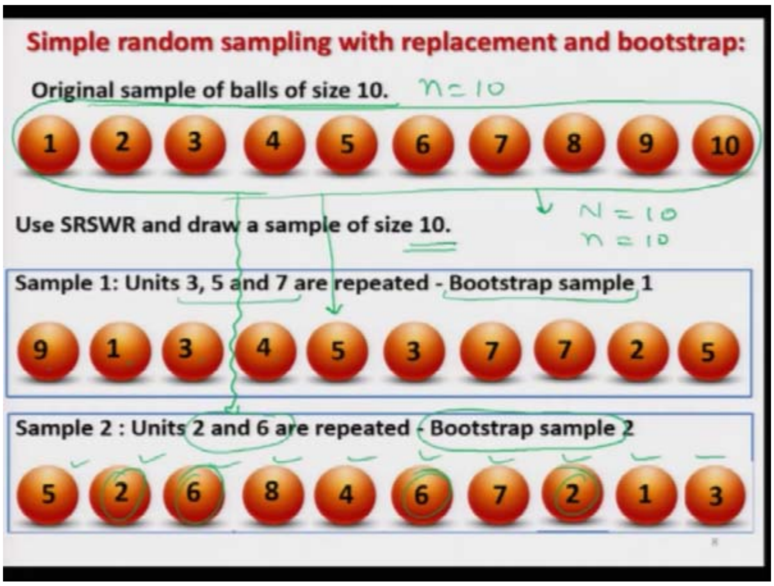
So, you can see here suppose if I take capital N and small n to be the same. That mean the you have drawn here a sample of size small n and now, let us take that this is also a population. So, if this happens then what will be the total number of samples total some number of possible sample?

So, you can see here I have counted it here if I try to take a sample of size 5 and the population is also of size 5, then the value of N^n is 3125 and the probability of drawing a sample is 1 upon 3125 which is 0.00032. And if you try to take capital N and small n to be 10, then N^n will become 10 power of 10 and the probability will become simply 1 upon 10 raise to the power of 10, right.

You can say this is 0.0000 up to and then 1, this is a very small quantity and remember this is only a sample of size 10. Now, in case if you make a population or sample of size 20 just equate them then the value of N^n becomes something of this magnitude 10^{26} and the probability of selection become something like this with the magnitude of 10 to the power of minus 27 this is very small.

And if you try to see here you have taken here only the value 20. So, with a value of 20 where small n and capital N are 20 the number of samples becomes too large. Now, you can imagine once you have a sample of size 100, 200 or 500 what will happen? This number this total number of possible sample, samples increases very fast and the probability of their selection declines very sharply.

(Refer Slide Time: 17:33)



So, I can use this property here and but a first let me try to take a sample and try to explain you the basic concepts and then I will come to the more formal definition. So, here I am trying to give you a sort of connect between the simple random sampling with replacement and bootstrap methodology. What are the basic concepts related to the bootstrap?

So, suppose I have a population and from there we have a sample of size a small say n is equal to 10. So, we have drawn a sample of size 10 and these are my 10 balls which I have numbered as 1, 2 up to here 10 you can see so, this is my original sample. Now, I ask you can you please draw simple random sample with replacement of size 10? That means, you had drawn a sample of size 10.

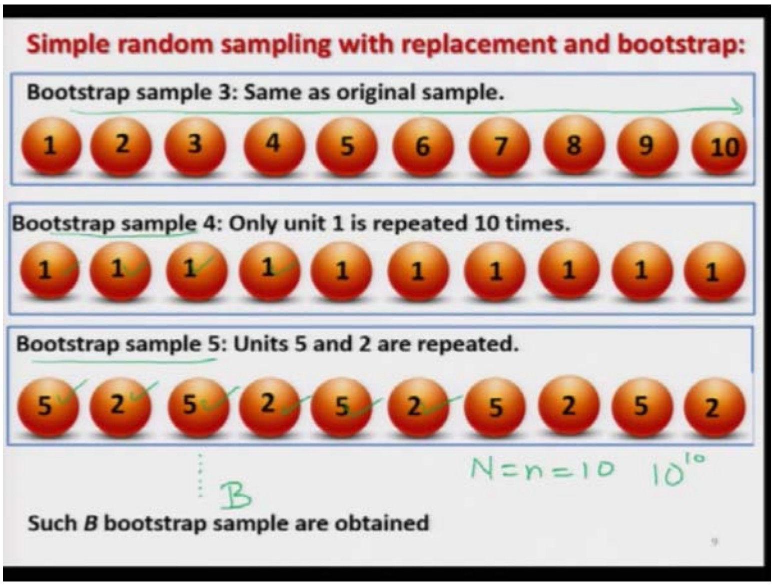
Now, you are trying to treat this sample as a population and now I am asking you from this population can you draw here a sample of size 10, right? So, now I am trying to treat the sample that was originally drawn from a population just like a new population. So, treating this sample as a population we are trying to draw the samples of the size equal to the size of the population.

So, if you try to draw here a sample of size 10, you can see here one possible sample can be 9 1 3 4 5 3 7 2 and 5 so, you can see here the ball numbers or the unit number 3, 5 and 7 are repeated. So, this is my one sample. Now, I give it a new name. I give it a name bootstrap sample and I am calling this sample number 1 as bootstrap sample 1.

So, now what do you understand? Whatever is your originally drawn sample from the same sample I am trying to draw here another sample of the as of the size of the original sample using SRSWR and whatever is your sample that I am calling as bootstrap sample. Now, in case if I try to repeat this process suppose from this population I try to draw here another sample by SRSWR of size 10.

So, you can see here that we will get here this sample of 10 balls yes, this sample is expected to be different from the first sample and here you can see that units 2 and 6 or the ball number 2 and 6 are repeated 2 is here, 2 is here, 6 is here, and 6 is here. So, this is also a simple random sampling sample which is drawn by with replacement. So, this sample I am calling here as a bootstrap sample and this is my second bootstrap sample which I have drawn from the same population.

(Refer Slide Time: 20:56)

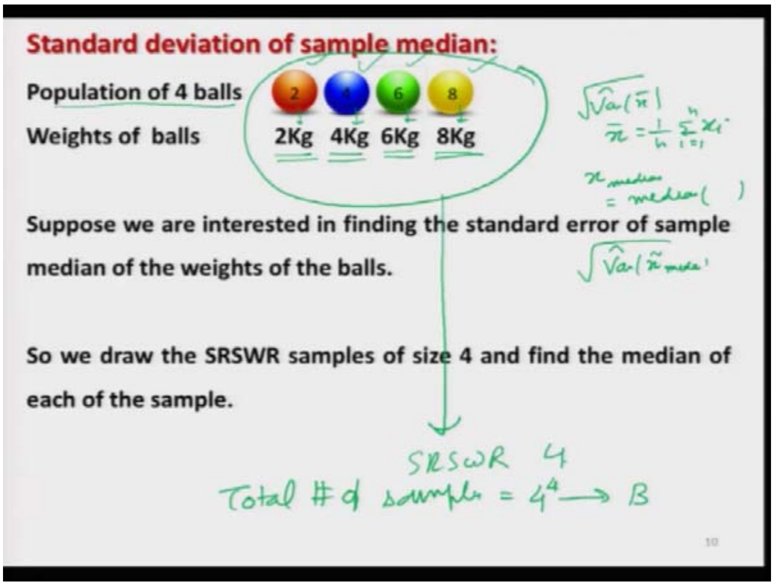


And now, you can see I can continue with this process means I can take here one more bootstrap sample. Suppose I take here bootstrap sample here like this and it is coming out to be here just like a the original sample this is possible. And similarly I try to take here one more sample and in this sample all the units are the same 1, 1, 1, 1, 1 that is possible.

So, that is another my this bootstrap sample and if you try to see here there is another possibility that I have taken here one more bootstrap sample in which box number 5 and 2 are repeated 5, 2, 5, 2, 5, 2 and so on. This is possible. Because, you have taken here N is equal to n small n is equal to 10 so the so there will be 10 to the power of here 10 possible combinations and I have taken here only 5 such possible combination.

So, these samples which you have drawn from the original sample using SRSWR they are called as bootstrap samples. And as we have drawn here only 5 sample suppose we continue further and say we have obtained capital B number of such B bootstrap samples, right.

(Refer Slide Time: 22:20)



Now, let me take here one more example. So, through this example I have given you an idea what is called a bootstrap sample. Now, I try to give you one more example and in this example I will try to connect the bootstrap methodology. Suppose I have a population of 4 balls. I have here red ball, blue ball, green ball, and yellow ball and we have obtained their weights.

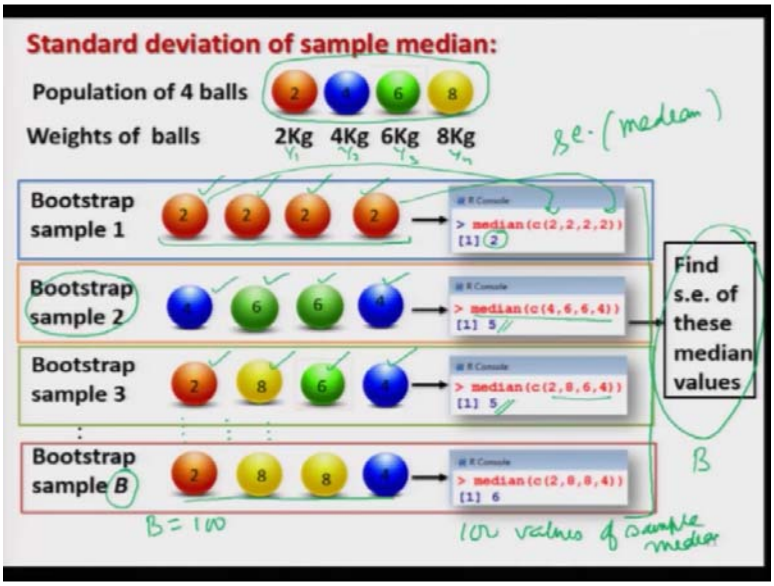
So, ball this red ball has got weight 2 kg, blue ball has got weight 4 kg, green ball has got weight 6 kg, and yellow ball has got a weight 8 kg. And suppose, we are interested in finding out the standard error of the sample median I mean just like you have computed the sample mean of the observation $x_1, x_2, \ldots, x_n$ by $\frac{1}{n}\sum_{i=1}^{n} x_i$. Similarly I can obtain the x median also which is using some software say R in you have a command median $x_1, x_2, \ldots, x_n$; so, and from there this will be another measure of central tendency.

So, I want to so just like we have found the variance of $\bar{x}$ earlier and its square root to give the standard error. Similarly, I want to find out the variance of x median and I want to find out its estimate I mean the standard error and then the square root. So, that means the standard error of sample median.

So, now how to get it done? Now, in this case we know that finite sample expressions are not available one can obtain the sample median but, obtaining the standard error is pretty difficult and we do not have any clear cut expression just like the expression in case of sample mean. So, what we try to draw here? What we try to do here? We try to use the bootstrap methodology we have got here a sample of size 4, I try to treat it here as a population and using this population I try to draw simple random sample with replacement of size 4.

And we can draw here more than one sample for example, total number of possible samples are going to be 4^4, right. So, I try to draw here suppose out of 4 power of here 4 suppose I try to draw here B such sample, right, B is a number that is a reasonably large number we will try to discuss about it when we discuss more details of the bootstrapping formally, ok.

(Refer Slide Time: 25:29)



So, I try to take here now such a bootstrap samples and then we try to find out their sample median. So, I try to consider this sample now as my population and these weights they are something like my y_1, y_2, y_3 and here y_4 these are the values of the variable. Now, from this population which was actually originally a sample I try to draw here bootstrap samples.

Suppose the first bootstrap sample comes out to be just like here the only the red ball having a weight 2 kg so all the balls are the same. And then, I try to find out the median of the values of weight of thus selected balls. So, in order to compute the median in R there is a command median. So, I have computed it and I have pasted here the screenshot so, you can see that here the median comes out to be 2, right.

You can see here this 2 is here, this 2 is here and so on, right. So, this is the median of 2, 2, 2, 2 ok. So, I have found the sample median of sample number 1. Similarly, I try to take here another sample another bootstrap sample and suppose I get here 2 green balls and 2 blue balls whose weights are 4 kg and 6 kgs. So, I try to find out the sample median of these 4 balls and this median comes out to be 5.

Now, I try to obtain here the 3rd sample. The 3rd sample comes out to be a red ball, yellow ball, green ball, blue ball and whose weights are say 2 kg, 8 kg, 6 kg, 4 kg and I try to find out their median this median comes out to be here 5. And then I can continue and then means suppose I try to draw here say capital B number of balls.

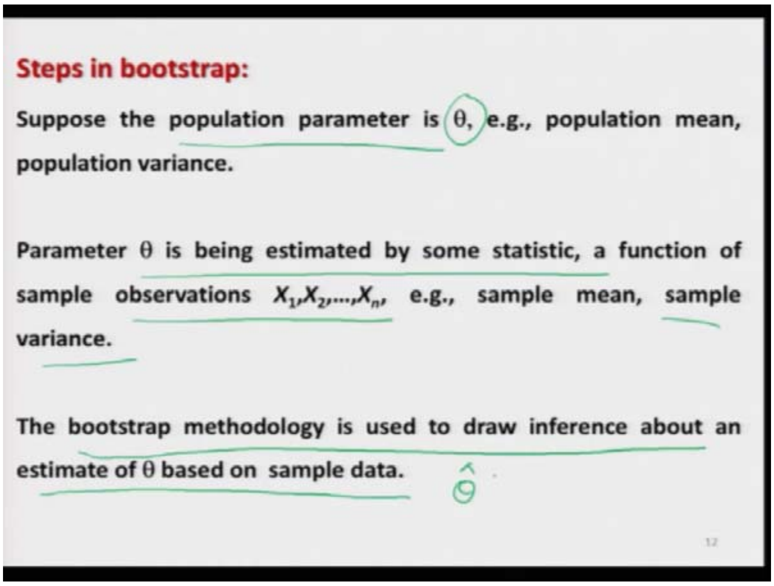
So, B can be 100, B can be 200 whatever you want, right. So, suppose if I say that suppose B is here 100 so that means, you have obtained such 100 samples then you have here 100 values of the samples of the weights and then you have found the found your median so, finally you have here capital B number of sample medians.

So, if I say B is equal to 100 then that means you have here 100 values of sample median. Now, what I am asking you suppose you want to find out the standard error of; standard error of median. Now, just do one thing. Try to find out the standard error of this 100 median values or in general try to find out the standard error of this B median values.

Now, if I say that the value of the standard error which you are trying to obtain here using the usual expression that is the positive square root of the variance of these median values, then this is going to give you a well approximated value of the standard error of sample median. So, this is actually what you have obtained here the standard error this is called as the bootstrap standard error of the sample median.

And this bootstrap estimate of the sample median the standard error of bootstrap estimate of the sample median is going to approximate the true standard error mean, as if you had knew that what is the correct expression to find out the standard error of sample median both the expressions are going to be pretty close under certain condition and those conditions are very simple.

(Refer Slide Time: 29:53)



So, now you I have given you good idea that what is this bootstrap methodology and if I ask you that this is a general methodology means, I have taken here the example of finding out the standard error of sample median similarly you can take any standard error you can take the standard error of sample mode, you can take the sample say this coefficient of variation, you can take the sample geometric mean, you can take the sample harmonic mean, whatever you want and you can find out the standard error number 1.

Number 2, I have given you the I have given actually an idea that if you want to obtain the standard error then you try to obtain the bootstrap sample try to find out the sample values compute the same statistics for example, in this case the sample median and then try to find out the standard error by the usual formula.

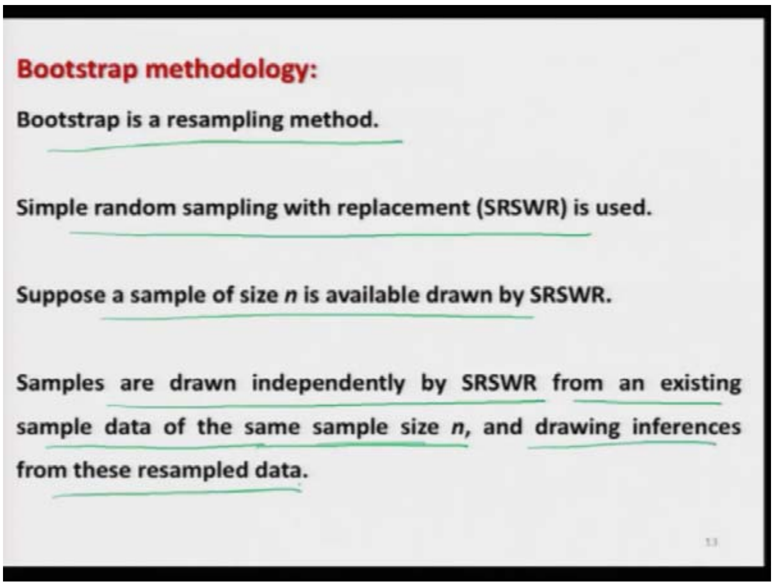
Similarly in case if you want to find out any other thing else like as you are not interested in say standard error, but you want to compute correlation or something else just try to use the same standard formula and try to use it on the bootstrap samples. And try to compute the value you will get the value. Now, can you believe this? Answer is yes.

You have to believe, this bootstrap methodology is as simple as this. Well, from the statistics point of view yes, this is the job of a mathematical statistician to ensure that the approximated value which are being obtained are they really good value or not. So, it has been proved mathematically that under certain type of condition which I will detail you this methodology will give you a good estimator, right.

So, now let us try to formalize whatever we have done through this example ok. So now, suppose we are interested in a population parameter which is denoted as θ. And this θ can be population mean, population variance, its skewness or correlation coefficient anything in the population. And definitely this θ is going to be estimated by some statistic, which is a function of sample observation $X_1, X_2, \ldots, X_n$.

For example, we you had estimated the population mean by $X_1, X_2, \ldots, X_n$ using the function sample mean you had obtained the population variance using $X_1, X_2, \ldots, X_n$ using the function sample variance and similarly you can take any other thing also. So, now I try to give you the steps that how would you try to do the same thing using the bootstrap methodology. So, this bootstrap methodology is used to draw inference about an estimate of θ that is $\hat{\theta}$ based on the sample data.

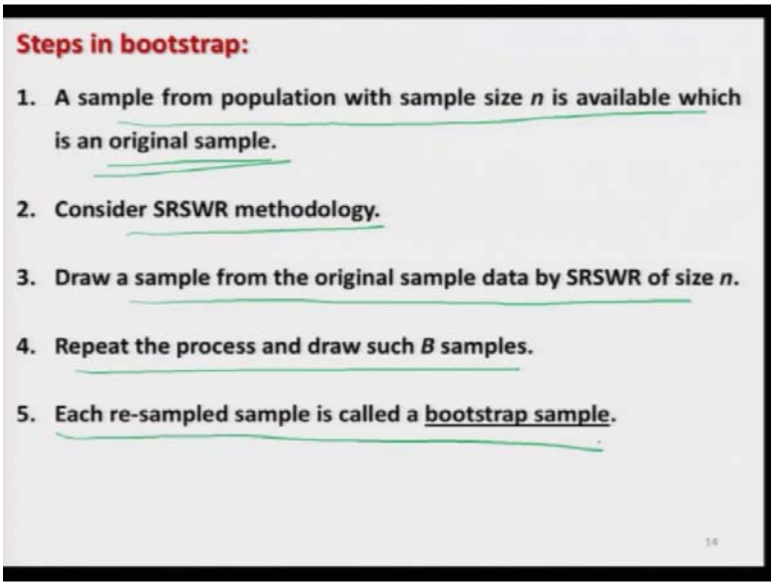
(Refer Slide Time: 33:09)



How? Ok. Before that, now we have understood that bootstrap is a resampling method, we are just trying to draw the samples by SRSWR again and again and then we try to move forward. And in bootstrap we are using simple random sampling with replacement and in practice we will always have a sample available so, we assume that a sample of size n is available to us which has been drawn by SRSWR.

And then after that what we do? We try to draw the samples independently by SRSWR that means all the b bootstrap sample have been drawn independently. They have been drawn from where from the existing sample data of the same sample size n, and then we try to draw the statistical inferences from these resample data.

(Refer Slide Time: 34:11)



And if you try to see here, let me try to give you this idea. What we try to do? That suppose there is some population and we would like to draw our inferences from that population, but suppose this population is not known to us. Then the question is who is going to tell us about the population.

Asking something to an unknown person about the population do not you think that this is a better option to ask the sample it is self what is your population? Yes we can do, but the but my problem is this the sample is data and I discuss earlier data is deaf and dumb that data cannot listen to our instructions that data cannot tell you what is my parent population. So, what we try to do? We try to take the help of data, we try to retrieve the information from the data, and we try to construct the population.

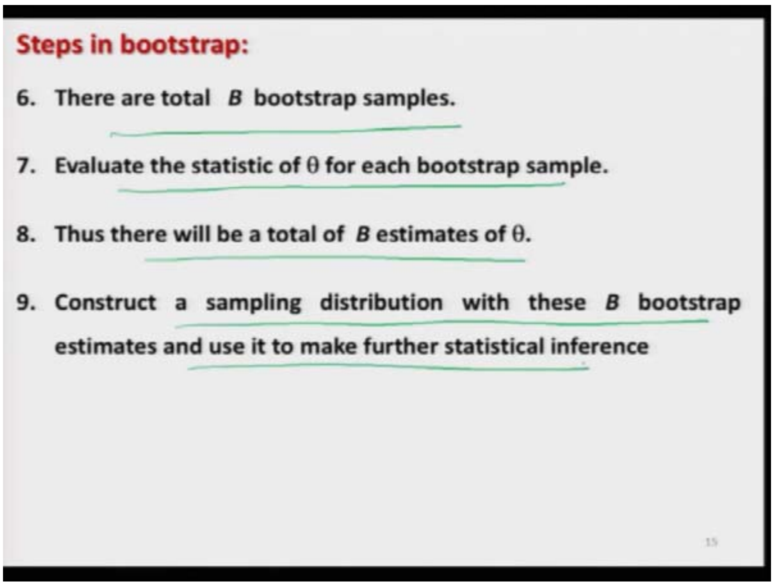
Their population may not be 100 percent same as of the original one, but since it is being constructed on the basis of the sample. So, we expect that the constructed population will be a good approximation to the true population. And this is the basic idea behind bootstrapping methodology.

So, we try to take b bootstrap samples and then we try to create our population and then we try to treat this bootstrap population which we have constructed ourselves as if this is the true population and we try to draw our statistical inferences from there. And how it can be done? This can be done by using the following steps these are the steps.

The 1st step is this we assume that a sample of size n is available and which has been drawn from a population and which is our original sample. Now, we consider the simple random sampling with replacement methodology and we draw a sample from the original sample by SRSWR of the size n.

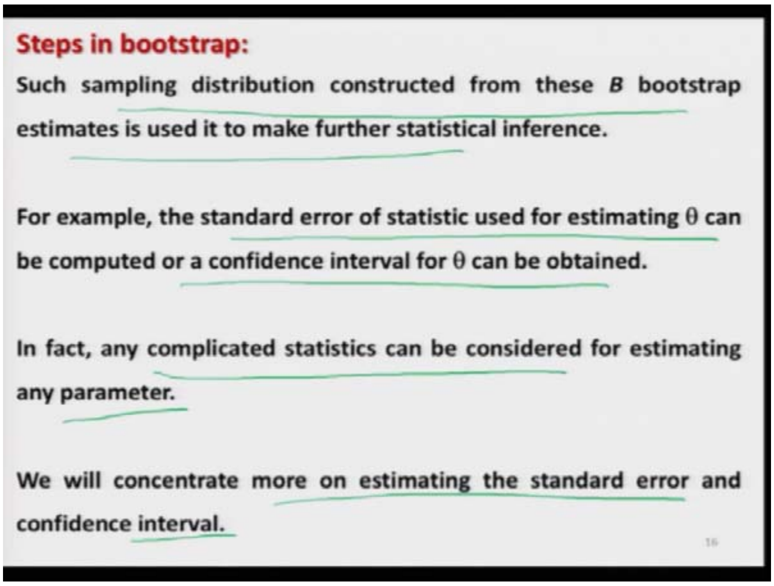
This is what you have to keep in mind that you are trying to draw the samples again and again and this and the sample size is the same as of the size of the original sample. And now, repeat this process and suppose we get here capital B number of sample. And each of this sample what we have drawn through re sampling this is called as bootstrap sample.

(Refer Slide Time: 36:50)



So, there are now a total of B bootstrap sample. Now, we would try to evaluate the statistics of θ for each of the bootstrap sample. For example, we estimated we found the sample median for each of the bootstrap sample. So, there will be all together capital B estimates of θ based on the bootstrap samples. Now, using this B bootstrap sample we can construct the sampling distribution and then we can use it to make further statistical inferences,, right.

(Refer Slide Time: 37:29)



So, this is what exactly I have done here,, right. For example, you have seen here that we have constructed the sampling distribution from these B bootstrap estimates and suppose we want to obtain the standard error of the statistics, which is used for estimating θ or we want to estimate the confidence interval estimation or any other things.

Means, any complicated statistics can be considered here for estimating any of the parameter. Now, you do not have any problem that you have to define a statistics which is mathematically tractable or it is easy to handle the mathematics or algebra of the statistics. You can think of any statistics, you can work with any statistics, and you can understand or you can study its properties.

In this chapter, we are going to concentrate more on estimating the bias, standard error, and confidence interval estimation. Bias is pretty simple and standard error I already have explained you and the confidence interval will also be on the similar lines as we construct the classical confidence interval, ok. So, now this is the time to stop in this lecture, so I have given you a fair idea what is bootstrap methodology.

And now, you should be confident that it is not difficult and it is basically a re sampling method, computational method, and since we have now good computational techniques so, it is not difficult to implement bootstrap in through computers. So, that is why this is a computer intensive technique.

So, now in the next lecture I will try to first understand that how we can estimate the bias of an estimator, Bias means bootstrap bias. And similarly how we can compute the standard error of an estimator a standard error means the bootstrap standard error and after that we will consider the confidence interval estimation. So, you try to study here.

And if you want to read it more my suggestion is that you can look into the book by Bradley Efron the Inventor of Bootstrap, right. That is a very nice book written in a very simple language and it gives you a fair idea. Yes, after that many developments have been made, but at this moment we are trying to understand the basic principles only. So, you study you revise and I will see you in the next lecture, till then good bye.