

**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 - 34**  
**Bootstrap Methodology**  
**EDF, Bootstrap Bias and Bootstrap Standard Errors**

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 continue with our chapter Bootstrap Methodology. So, you may recall that in the earlier lecture we introduced the concept of bootstrap methodology and I have taken an example to show you that how it works.

My idea of taking the example was that I wanted to take out the fear from your heart, that ok, this is not something very complicated or difficult which you cannot understand. Now, you should be confident that, that was a very simple methodology and the example which I took for computation of the standard error that is pretty simple. You can always compute the statistics; you can always compute the standard error, variance, correlation coefficient whatever you want,, right.

So, now in this lecture I will try to concentrate on two aspects; whenever we are trying to find an estimator after finding out the estimator we are always interested in finding out its bias and its standard error. Why? You may recall that I already have explained it in couple of lectures several times, that why this bias is important and why do we consider the unbiased estimators, why standard error is important, why do we consider the concept of variability.

So, now how to estimate this bias and standard error using the bootstrap methodology? This means, now there are no barriers for you, whatever statistics you want to consider for the estimation of the parameters. Just consider it, forget about the theoretical properties, your interest lies in the amount of bias or standard error that will be available to you that is my promise. So, that is the advantage.

And that is why this, these are the topics which lay the foundations of data sciences, these are the computational intensive techniques. Once you are working in a more complicated environment, then you need better computing power, better computers, better computer structure, better database management system. But at least you know the way out now, that if I try to use this bootstrap I can handle the situation.

You will have no option now onwards in your life to say, sorry sir I cannot do, it is difficult, this should this sentence should be deleted from your life as far as you are working in the statistics. You can always say sir this may not be the 100 percent correct value, but this is a very good value. And this is our approach of life. In our life we always have two option either to leave the problem or we try to solve the problem.

Once I have taught you now you have no option in your life, you have only one option, you will never say no, but you will try to solve the problem as much as you can. After that you can improve the solution that is always welcome, but you are not allowed to say no.

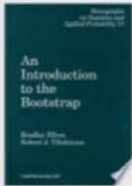
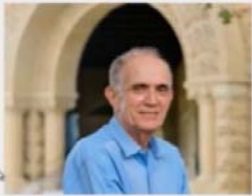
So, now let us start our this lecture, in this lecture I am going to first introduce the concept of empirical distribution function. And, then I will try to show you how you can estimate the bias and standard error, right. So, let us start, but before we go to our methodology let me try to show you something over here.

(Refer Slide Time: 04:16)

**Bootstrap:**

**Bradley Efron (Stanford University)**

Bootstrap introduced in 1979.

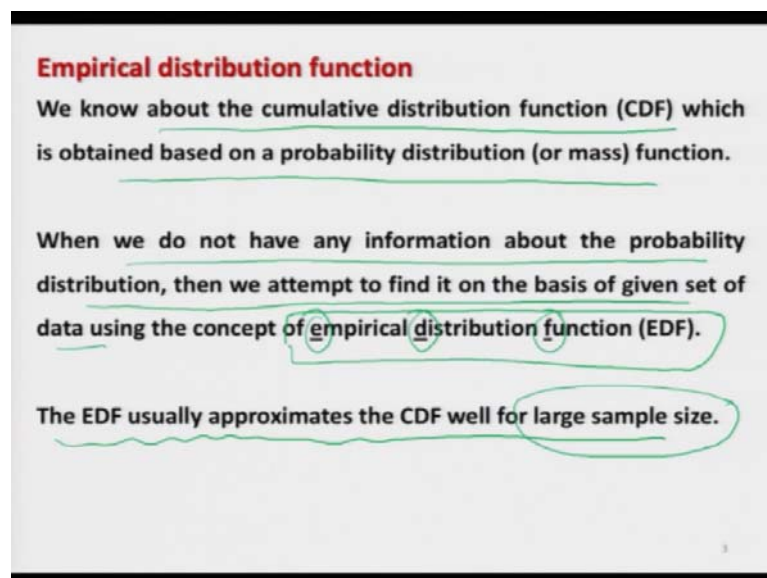


**An Introduction to the Bootstrap**  
**Bradley Efron, R.J. Tibshirani**  
Publisher: Chapman and Hall/CRC Press

So, he is Bradley Efron, he is working at Stanford University. And, he introduced the bootstrap methodology in 1979. And after that he and his colleague Robert J. Tibshirani, they had written a book this An Introduction to the Bootstrap. And if you try to read it from this book, this is a very simple and elementary book.

And so, we are considering the contribution of Bradley Efron, later on you will see that I will also consider the contribution of Robert Tibshirani, because he introduced the Lasso l a s s o. Lasso that we are going to discuss later on ok. So, now, we let us begin our lecture.

(Refer Slide Time: 05:07)



Now, you see in statistics whenever you are trying to do something, the first basic ingredient is distribution function, probability distribution function or cumulative distribution function. And that is our starting point and then we try to draw all statistical inferences based on that population, which is characterized by the corresponding distribution function.

Now, think about a situation that in practice you have no population you have no idea what is my population. What you have in your hand? It is only a sample of data, a small sample of data. Now, the question is how will you get the probability function or the distribution function? The data is not going to tell you well my parent distribution is this, but now you have to depend on your own logic and hands these hands only.

You have to do something so, that you can know the distribution function, otherwise you cannot move forward. So, now a very good option is that instead of hitting at random here and there to find out the parent distribution of the sample, one good option is this why not to ask the sample itself. But, when I ask the sample, the sample is different dumb, the data is defined dumb, it is not going to reply us.

So, now we have to develop some tools so, that we can retrieve the information about the parent distribution from where the sample has come. It is just like you find a small child somewhere and his parents are not around. And if you ask the child who are your parents, what is your address the child is so, small that it just cannot tell you what will you do, will you leave that child there itself or you will try to do something?

You will try to find out that who are the parents and after some time, you can reach close to the truth distribution, right. You can reach close to the family of that boy. And if you work more finally, you will reach to the true parent of that children; the same story goes here in the bootstrapping.

So, now we are going to first understand that once you have got a sample, how can you construct the distribution function just on the basis of sample, and I promising you this is very simple when you try to do it with bootstrap. So, let us try to do it ok. So, now, we know that the cumulative distribution function is obtained on the basis of a probability distribution function, or a probability mass function, or a probability density function.

And this is and this actually creates the foundation of all statistical inference, whatever we want to do in a statistics this is based on this cumulative distribution function this is called a CDF. Now, when we do not have any such information about the probability distribution, then we cannot obtain the CDF. And, then we try to attempt to find it on the basis of given set of data.

And, we use the concept of Empirical Distribution Function, which is called, denoted as EDF, E is coming from empirical, D is coming from the distribution, and F is coming from function. And well, I can tell you the final outcome, this EDF usually approximate the CDF well for last sample size.

Now, the next question is how large is the large and how small is the say small. So, I will say the rule of thumb is that if your statistics is very simple, then the sample size say I mean 30 and 40 will be sufficient, but if your statistics is complicated, then you have to take larger sample size.

(Refer Slide Time: 09:44)

**Empirical distribution function:**  
Suppose we have a sample of size  $n$  as  $x_1, x_2, \dots, x_n$ .  $P(x_i) = \frac{1}{n}$

The EDF is a discrete distribution that gives equal weight to each data point.

It assigns probability  $\frac{1}{n}$  on each  $x_i, i = 1, 2, \dots, n$ .

It forms a CDF that is a step function.

The step jumps up by  $\frac{1}{n}$  at each of the  $n$  data points.

So, now let us try to understand what is this EDF. Suppose we have a sample of size small  $n$  denoted as  $x_1, x_2, \dots, x_n$ . This EDF is a discrete distribution that gives equal weight to each data point, weight or equivalently I can say probability. So, I have got here a small  $n$  number of points.

So, if I say here probability of each  $x_i$  is simply  $\frac{1}{n}$  upon  $n$ , so, all  $x_1, x_2, \dots, x_n$  are equally probable. And based on this concept, now we conclude or we fix that we assign probability  $\frac{1}{n}$  upon  $n$  on each of this  $x_i$ . And based on this we try to create the CDF, then the CDF is going to be a step function. Step function mean, it will be something like this here, here, here and like this and these jumps are going to be here at  $\frac{1}{n}$  upon  $n$ , right.

(Refer Slide Time: 11:01)

**Empirical distribution function:**

Suppose a dice is rolled 100 times and the outcomes 1, 2, 3, 4, 5, and 6 are obtained as

Point on upper face ( $k$ )	1	2	3	4	5	6
$\#\{x_i = k\}$	12	20	10	18	24	16
Frequency ( $f_i$ )	0.12	0.20	0.10	0.18	0.24	0.16

Relative  $\frac{12}{100}$

$\hat{f}_k = \frac{\#\{x_i = k\}}{n}$ : Proportion of  $x_1, x_2, \dots, x_n$  in A

The EDF of  $x_1, x_2, \dots, x_{100}$  is defined as

$\hat{F} = (\hat{f}_1, \hat{f}_2, \hat{f}_3, \hat{f}_4, \hat{f}_5, \hat{f}_6)$

$\#(x_i = k)$

So, now based on this we have to construct our empirical distribution function. So, let me try to take a simple example to explain you how it is done. Once you understand the example, you will see that it is very simple. Suppose a dice is roll 100 times. What are the possible outcomes of the dice? They are 1, 2, 3, 4, 5 and 6 numbers on the upper face.

So, we roll the die 100 times and we try to count that how many times 1 has come, how many times 2 has come and similarly how many times 6 has come. And we record that the number of times this  $x_i$  equal to  $k$  that is  $k$  is the point on the upper face. So, once I take  $k$  equal to 1; that means, the number of points on the upper face of the dice which are obtained as 1 is 12.

So, 12 times number 1 is obtained, 20 times number 2 is obtained, 10 times number 3 is obtained, 18 time number 4 is obtained, 24 times number 5 is obtained and 16 time number 6 is obtained. So, what is the frequency? The frequency is defined by the total number of occurrence divided by the total number of possible number.

So, here the number of occurrences here is 12 out of 100, so, this becomes here 0.12. Similarly the frequency or I should actually call it relative frequency, because there are two types of frequencies; absolute frequency and relative frequency. So, here frequency means relative frequency. So, the relative frequency of point number 2 is 0.20, the

relative frequency of point number 3 is 10 upon 100 which is 0.1. And, similarly the relative frequencies of 4, 5 and 6 are also obtained, right.

So, this relative frequency is obtained here by this expression that the number of  $x_i$  equal to  $k$  divided by say small  $n$ . So, that is essentially the proportion of  $x$  in the sample  $x_1, x_2, \dots, x_n$ , right. And  $\Omega$  is here my sample space, right that is the set  $A$ , ok. So, now based on this I can define the empirical distribution function.

Remember when you study the probability theory in statistics and when we try to teach the probability distribution and cumulative distribution function PDF and CDF. Then, you try to obtain the CDF using the PDF, ok. The same concept is being used here, now we have obtained the distribution of the probabilities on the given data set on the basis of given data sets, right.

And similarly as you try to obtain the CDF from PDF in statistics that is the cumulative density function, cumulative distribution function from the probability density function. Similarly, we try to find out here empirical distribution function from this frequency distribution, which has been obtained by the relative frequencies, which is defined here by this  $f_k$  which is the proportion of  $x_1, x_2, \dots, x_n$  in  $A$ , where  $A$  is denoting the even that is number of say this  $x_i$  is equal to  $k$ , right.

So, this is the set of those values. So, now, this can be defined as simply here  $\hat{f}$ ;  $\hat{f}$  is a symbol for EDF, because CDF is usually defined by capital F. So, we are trying to estimate it, so, we define it by or indicated by  $\hat{f}$ . So, this is going to be simply collection of all this relative frequencies. So, now, I can see here whatever relative frequency I have obtained here, they are going to define our EDF, right. Do you think that is it very difficult, ok?

(Refer Slide Time: 15:39)

**Empirical distribution function:**

In other words,  $\hat{F}$  assigns to a set  $A$  in the sample space of  $x$  its empirical probability

$$Prob\{A\} = \frac{\#\{x_i \in A\}}{n}; \text{ Proportion of } x_1, x_2, \dots, x_n \text{ in } A$$

It can be proved that the vector of observed frequencies  $\hat{F} = (\hat{f}_1, \hat{f}_2, \dots)$  is a sufficient statistic for the true distribution  $F = (f_1, f_2, \dots)$ .

$\hat{F}$ : Obs values:  $\hat{\theta}$ : known  
 $F$ : True dist: Unknown:  $\theta$

This means that all the information about  $F$  contained in  $x_1, x_2, \dots$  is also contained in  $\hat{F}$ .

So, in other words if you want to make it a make a general, then I would say here that  $\hat{F}$  assigns to a set  $A$  in the sample space of  $x$  its empirical probability, which is computed by like this. The estimated probability of  $A$  is number of the data point, number of  $x_i$ 's which are belonging to the set  $A$  divided by the total number of points.

So, this is essentially the proportion of  $x_1, x_2, \dots, x_n$  in  $A$ . And yes, I am not giving you here the proof mathematical proof, but that proof is available in the books and from there, I can assure you and that we can prove that the vector of observed frequencies which is obtained here  $\hat{f}_1, \hat{f}_2, \dots$  etc., which is our empirical distribution function EDF is a sufficient statistic for the true distribution.

What does this mean? So, you can see here that your  $F$  is your here something like true distribution it is a true distribution, but it is unknown to us. So, it is just like our unknown parameter. Now, when I am trying to estimate it here by  $\hat{F}$  which is based on the observed values and this is something like  $\hat{\theta}$  which is known on the basis of given sample of data then we are trying to say this is a sufficient statistic.

So that means,  $\hat{F}$  is the sufficient statistics for  $F$ . What is these sufficient statistics? So, in statistics when we try to derive an estimator, then we want to judge whether the estimator is good or bad. So, for that several criteria have been proposed in statistical



inference, they are like unbiasedness, efficiency, consistency, sufficiency, completeness etcetera.

So, each of these criteria has got a different interpretation and sufficient statistics means, that if you have a sample then this sample has certain amount of information and that information is contained in the any small say  $n$  sample value like  $x_1, x_2, \dots, x_n$ . Suppose if I say  $n$  equal to 20.

So, this so the information contained in the; is contained in the 20 values. Now, suppose I obtain a statistic, suppose I say I obtain sample mean, and I want to estimate the population mean using sample mean. So, sample mean is going to be estimated on the basis of these 20 observations.

So, now I am saying whatever the information which is contained in those 20 observation, the same information is contained in sample mean also or the sample mean is conveying us the same information, which these 20 observation or a small  $n$  number of observations are trying to convey.

This is one of the most simple way to understand, the concept of sufficiency although this is a mathematical, there is a mathematical definition and there are some theorems and rules, which have to be satisfied before we can be confident that my statistics is a sufficient statistic.

Well I am not going into those details over here, but if you wish you can refer to any statistical inference book and you will find the concept of sufficiency. So, in a very layman and simple language I can say if as if an estimator is sufficient, we have a reason to be happy that it is going to give us a good value. So, this is exactly what I am trying to prove here that whatever EDF you have obtained on the basis of sample of data, this is the sufficient statistic for the true data true distribution function.

True distribution function is unknown to us. So, whatever value you are going to get from the EDF, you can depend upon it that is going to give you a good value ok. So, this means that all the information about capital  $F$  is contained in the sample  $x_1, x_2, \dots, x_n$  is also contained in  $\hat{F}$  that is the empirical distribution function.

(Refer Slide Time: 20:18)

**Plug-in principle:**

The plug-in principle is a simple method of estimation of parameters based on empirical distribution function.

The plug-in estimator of a parameter  $\theta = t(F)$  is defined to be  $\hat{\theta} = t(\hat{F})$ .

In other words, we estimate the function  $\theta = t(F)$  of the probability distribution  $F$  by the same function of the empirical distribution  $\hat{F}$ ,  $\hat{\theta} = t(\hat{F})$ .

$\hat{\theta}$  is called a plug-in estimate for the population parameter  $\theta$ .

The diagram illustrates the plug-in principle. It shows a flow from the true distribution  $F$  to the parameter  $\theta = t(F)$ . A handwritten note indicates that  $t$  is a 'same function'. Then, the empirical distribution  $\hat{F}$  is used to estimate  $\theta$  by  $\hat{\theta} = t(\hat{F})$ , where  $t$  is the same function as before.

So, now I come to another aspect. You may recall that if we have a true distribution function  $F$  and suppose we want to estimate a parameter  $\theta$ . Then, this  $\theta$  is actually a function of some function of here  $F$ ; I am not using here the symbol here  $F$ , because otherwise that will create confusion, so, I am calling it here is a  $\theta$  is equal to  $t(F)$ . So, this is some function.

And from this one we try to estimate  $\theta$ , but now my problem is this we do not know this  $F$  or  $\theta$ , rather I have here empirical distribution  $\hat{F}$ , but my objective is the same that I want to estimate the same parameter  $\theta$  by the same estimator  $\hat{\theta}$ . So, the question is how to estimate  $\theta$  on the basis of  $\hat{F}$ .

So, for that I use here the plug in principle, the literal meaning of plug-in is that you simply try to insert or replace the value. So, now, let us try to first understand what is this plug-in principle and what are the plug-in estimators. So, the plug-in principle is a simple method of estimation of parameters which is based on Empirical Distribution Function; EDF.

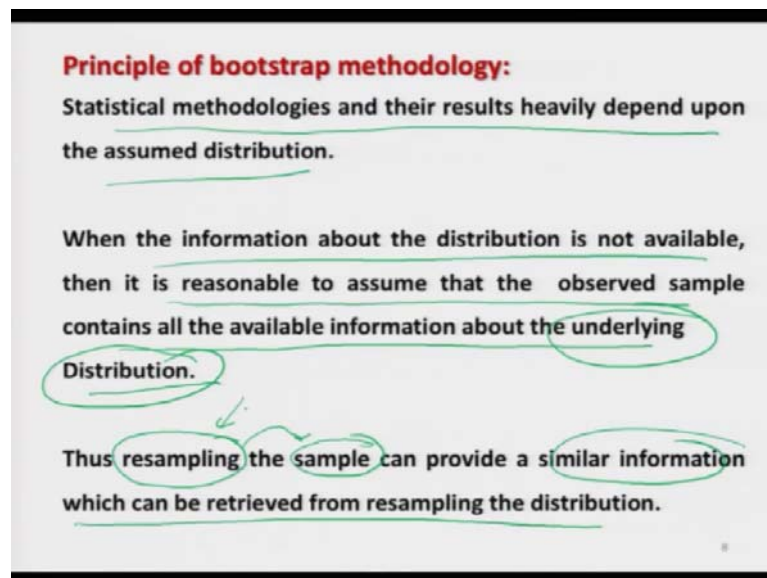
And suppose I have a parameter  $\theta$  which is a function of  $F$  and it is denoted by  $t(F)$ , then this parameter can be estimated by its plug-in estimator which is defined as  $\hat{\theta}$  is equal to

$t(\hat{F})$ . That means, what are you trying to do? That you wanted to estimate the parameter or the function  $\theta$  or  $t F$  based on the actual distribution function capital  $F$ .

And, but now I am trying to use it use the empirical distribution  $\hat{F}$  and we are trying to estimate  $\theta$  by  $\hat{\theta}$  and  $\hat{\theta}$  is obtained just by replacing by  $\hat{F}$  and note one thing that  $t$  remains the same,  $t$  here in  $\theta$  and  $t$  here in  $\hat{\theta}$  they are the same, right.

So, we are using the same function and then we are simply trying to replace the unknown frequencies by estimated frequencies or the unknown  $F$  by now empirical. So, this  $\hat{\theta}$  is called as say plug-in estimate of the population parameter  $\theta$ , right.

(Refer Slide Time: 22:56)



So, now this plug-in principle is going to play a very important role in estimating different types of quantities using the bootstrap methodology. So, now, I try to explain you what is the basic principle of bootstrap methodology. So, we know that all the statistical methodologies and their results they heavily depend upon the assumed this probability distribution.

If you assume that the sample is coming from normal distribution, then the properties will be entirely different, if the sample is coming from binomial. So obviously, when this information is not available, then one possible solution is that and it is reasonable to

assume that the observed sample contains all the available information about the underlying distribution, which is unknown to us.

So, now the question is how to know this underlying distribution. So, one simple solution is that we can use the resampling methodology. We can use the resampling technique and using the resampling technique on the available sample itself will provide many more samples. And we assume that every sample will provide a similar information which can be retrieved from the resampling of the distribution, right.

(Refer Slide Time: 24:32)

**Principle of bootstrap methodology:**

Let  $\theta$  be the unknown parameter.

Let  $\hat{\theta} = s(X_1, X_2, \dots, X_n)$  be the statistics based on the sample  $X_1, X_2, \dots, X_n$  to estimate the parameter  $\theta$ .

We need the sampling distribution of  $\hat{\theta}$  for drawing any statistical inference on  $\theta$ . Finding the exact sampling distribution of  $\hat{\theta}$  may be difficult or even too complicated to handle.

Such information can be used for finding the standard errors of any estimator or the confidence intervals for any estimate of  $\theta$ .

*Handwritten notes:*  
 $X_1, \dots, X_n \sim N(\mu, \sigma^2)$   
 $\sigma^2$  known  $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$   
 $\sigma^2$  unknown  $\frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t(n-1)$

So, when I say resampling it is something like suppose if I know the population and suppose I draw here thousands of samples, but which is not possible here. So, I am trying to draw suppose thousand sample from the same sample, and then we are trying to go back in the reverse direction. And we are trying to construct the probability distribution.

So, suppose  $\theta$  is the unknown parameter which we want to estimate. And suppose  $\hat{\theta}$  is a function of  $x_1, x_2, \dots, x_n$  this is a statistics which is based on the sample values and it is used to estimate the parameter  $\theta$ . So, in order to do anything we need the we first need the sampling distribution of  $\hat{\theta}$ , for drawing any type of statistical inferences on  $\theta$ . Means ideally we should have known the exact distribution of  $\hat{\theta}$ .

What does this mean? For example, you know in statistics that if  $x_1, x_2, \dots, x_n$  this is following a normal distribution with  $\mu$  and  $\sigma^2$ . And, suppose if you assume that  $\sigma^2$  is known, then you know that  $\frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$  this follows a normal distribution. And if  $\sigma^2$  is unknown, then you know that  $\frac{\bar{x} - \mu}{s / \sqrt{n}}$  which is the estimated standard error  $s$  this follows a t distribution with  $n$  minus 1 degrees of freedom.

So, ideally what I am trying to say that this is my statistics, these are my two statistics and these are the exact distribution normal or say here t under two different condition. So, now, what I am trying to say finding such exact sampling distribution of  $\hat{\theta}$  may always not be possible, and many time it is difficult or too complicated to handle, right.

This is what I mean by exact sampling distribution. So, now, means if I can find out such type of information, then such information can be used for finding the standard error of any estimator or confidence interval for any of the estimate of  $\theta$  including the bias, right.

(Refer Slide Time: 26:59)

**Principle of bootstrap methodology:**

Suppose the true distribution was known. Then different samples can be drawn from it and an empirical sampling distribution of  $\hat{\theta}$  can be constructed.

Now the true distribution is unknown. So samples can not be drawn from it.

Instead, the samples are resampled from the original sample and an empirical distribution can be constructed from them.

Bootstrapping method mimics the data-generating process as if the samples are drawn from the unknown true distribution.

10

So, now as I said that this principle is going to immediate the usual statistical procedure. For example, suppose you assume for a while that the true distribution is known to us. Then, you try to draw different samples from this population and you try to estimate the

parameters and based on those parameter values you try to create an empirical distribution of  $\hat{\theta}$ .

Now, this condition is that true distribution is unknown to us that is not known to us. So, sample cannot be drawn from there. One sample cannot be drawn from there how you can construct the empirical sampling distribution of  $\hat{\theta}$ ? So, what we try to do? Instead of doing this, we try to draw the samples through resampling technique from the original sample. And, we try to construct an empirical distribution from the values of  $\hat{\theta}$ .

So, this bootstrapping method mimics the data generating process as if the samples are drawn from the original true distribution. So, now, this bootstrapping methodology is simply trying to copy the true generating process, right and that is why whatever the results you are going to get from here you can depend upon them.

(Refer Slide Time: 28:42)

**Bootstrap methodology:**

Consider a population with  $N$  units :  $U_1, U_2, \dots, U_N$

The unknown  $N$  values on variable  $X$  :  $\underline{X} = (X_1, X_2, \dots, X_N)$  associated with  $1, 2, \dots, N$ .

Suppose a sample  $s = (i_1, i_2, \dots, i_n)$  is drawn by SRSWR.

Write for  $j = 1, 2, \dots, n$ ,

$$x_j = X_{i_j}$$

and define

$$\underline{x} = (x_1, x_2, \dots, x_n)'$$

Let  $\hat{\theta} = \hat{\theta}(\underline{x})$  be an estimator of  $\theta$ .

*Handwritten notes:*

$i = 3$   
3<sup>rd</sup> sample =  $(3, 3, \dots, 3)$

4<sup>th</sup> sample =  $(4, 4, \dots, 4)$   
 $(1, 2, 3, 4)$

$s_1 = (4, 1, 2, 3)$   
 $\rightarrow 1_1 \rightarrow 1_2 \rightarrow 1_3 \rightarrow 1_4$

$s_2 = (4, 3, 2, 1)$   
 $\rightarrow 2_1 \rightarrow 2_2 \rightarrow 2_3 \rightarrow 2_4$

So, now let me try to translate the bootstrap methodology, what we have understood through example in a more formal way. Suppose there is a population of capital  $N$  units and this population has got units  $U_1, U_2, \dots, U_N$ . Well once you try to associate a random variable with this  $U$ , then these values will become  $x_1, x_2, \dots, x_n$ .

For example,  $U$  is some human being and  $X$  is the height. So,  $X_1$  will become the height of that concerned human being, ok. So, now I can say that suppose there is a variable

capital X, and though and then these capital N number of values will also have the capital N number of values on this variable.

So, let all these values  $x_1, x_2, \dots, x_N$ , they are clubbed together and they are indicated by here capital X. Now, I am using here the symbol underscore, right. So, this X is  $\underline{X}$  is going to indicate all the values which are associated with 1<sup>st</sup>, 2<sup>nd</sup> up to nth unit. Now, suppose from this population I try to draw here a sample.

This sample is going to be some small n number of values out of this capital  $x_1, x_2, \dots, x_N$ , right. This  $i_1, i_2, \dots, i_n$  there i is going to indicate the particular sample. Suppose if I say here i equal to here say here 3, then this is my 3<sup>rd</sup> sample. And 3<sup>rd</sup> sample and then whatever the values I am getting here, they are indicated as follows.

Suppose we are getting a small n number of values. So, I will be getting here 1<sup>st</sup> value, 2<sup>nd</sup> value up to here nth value. And these values are obtained in the 3<sup>rd</sup> sample, so, I will try to write down here I mean 3 here. So, 3 subscript 1 is going to denote the 1<sup>st</sup> value in the 3<sup>rd</sup> sample. The 3 subscript 2 is going to denote the 2<sup>nd</sup> value in the 3<sup>rd</sup> sample. And similarly here 3 subscript n is going to denote the nth value in the 3<sup>rd</sup> sample.

Similarly, if you try to take here the 4th sample, then 4 sample will be denoted as a 4 1, 4 2 up to here 4 n, right. So, what does this mean? Suppose if I say that I have got here a sample here, say 1, 2, 3, 4. Now, I try to draw here sample say first sample  $s_1$  and I get here a value 4 1 2 3.

So, this is my here sample number 1; 1<sup>st</sup> value, this is my here sample number 1; 2<sup>nd</sup> value. This is my here sample number 1; 3<sup>rd</sup> value and this is my here sample number 1; 4<sup>th</sup> value. And similarly if I try to draw here the second sample, suppose this comes out to be 4 3 2 1.

So, this is 4 is now going to be 1st value in my 2<sup>nd</sup> sample, 3 is going to be the 2<sup>nd</sup> value in the 2<sup>nd</sup> 3 is going to be the 2<sup>nd</sup> value in the 2<sup>nd</sup> sample. And 2 is going to be the 3<sup>rd</sup> value in the 2<sup>nd</sup> sample and 1 is going to be the 4<sup>th</sup> value in the 2<sup>nd</sup> sample. So, this is what I mean here by this symbol, right.

So, we get here these numbers and then we try to define the sampling units, this is what I have explained you here in this example. And, then I try to define the values  $x_1, x_2, \dots, x_n$ ,

right. So, you can see here now this  $x_1, x_2, \dots, x_n$  is denoted by say  $\mathbf{x}$  symbol. And, then I try to define  $\hat{\theta}$ ,  $\hat{\theta}$  is going to be defined as  $\hat{\theta}(\mathbf{x})$ ; that means, this is a function of  $x_1, x_2, \dots, x_n$  and this is an estimator of  $\theta$ , right ok.

(Refer Slide Time: 33:08)

**Bootstrap methodology:**

For example,

$\theta = \bar{X} = \frac{1}{N} \sum_{i=1}^N X_i$  is the population mean. *unknown*

$\hat{\theta} = \hat{\theta}(\mathbf{x}) = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$  is the sample mean.

Choose a sample  $\mathbf{x} = (x_1, \dots, x_n)$  *n permutation*

$s^* = (i_{11}^*, i_{21}^*, \dots, i_{n1}^*)$  : Bootstrap sample

Define

$\mathbf{x}_1^* = (x_{11}^*, x_{21}^*, \dots, x_{n1}^*)$  : 1<sup>st</sup> bootstrap sample

For example, let  $s = (4, 2, 4, 5)$ , then one possible sample choice is

$s^* = (2, 5, 4, 2)$  and

$\mathbf{x}_1^* = (x_2, x_5, x_4, x_2)$

So, now what I try to do here? That suppose my parameter is suppose population mean. So, I am not trying to take a particular example. So, that I can really explain you what is happening. So, this population parameter is defined as here  $\frac{1}{N} \sum_{i=1}^N X_i$ , which is the population mean and this is unknown to us.

And, suppose we decide that we would like to estimate  $\theta$  by  $\hat{\theta}$  and  $\hat{\theta}$  is going to be a function of  $x_1, x_2, \dots, x_n$  say  $\mathbf{x}$ , which is suppose you try to choose here  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ , right. Now, I try to choose here a sample.

So, we already have drawn here a there a sample  $\mathbf{x}$ , now that is my original sample. And, now from this sample which was  $x_1, x_2, \dots, x_n$  we are trying to say choose here  $n$  possible values. So, basically I am trying to choose a permutation of  $n$  values. And whatever is my permutation that will give me the value of my 1st sample.

So, suppose this is a rearrangement of small  $n$  numbers. And then corresponding to these numbers I try to choose the value of corresponding  $\mathbf{x}$  and which is going to create my



bootstrap sample. So,  $x_1^*$  is going to give me the first bootstrap sample. For example, if I try to suppose my original sample is something like four values 4, 2, 4, 5, right.

And, now I try to draw here a sample random sample with replacement, which is going to be one of the possible permutations of these four values. Suppose this sample comes out to be here  $s^*$  which is here 2, 5, 4, 2, right. Now, corresponding to this 2 you try to choose thus the value of the corresponding value of  $x$ , which is your here  $x_2$ .

And, now it is indicated here as say  $x_2^*$  corresponding to this 5, you try to choose  $x_5$  which is going to now the 2<sup>nd</sup> unit in my sample 2<sup>nd</sup> drawn unit. And this is indicated by  $x_5^*$  here. Similarly, you try to take the choose the 3<sup>rd</sup> value in the sample. And this is going to be the 4<sup>th</sup> value in the original sample which is denoted here by  $x_4^*$ .

And, similarly you try to take the 2<sup>nd</sup> value from the original sample which is going to constitute the 4<sup>th</sup> value in the bootstrap sample that is the 4th value of  $x_4$ , right.

(Refer Slide Time: 36:09)

**Bootstrap methodology :**

- Repeat the process and obtain  $B$  bootstrap sample independently.
- Let these  $B$  bootstrap samples are  $x_1^*, x_2^*, \dots, x_B^*$ .
- Calculate bootstrap replicates  $\hat{\theta}(x_1^*), \hat{\theta}(x_2^*), \dots, \hat{\theta}(x_B^*)$ .
- Here  $B = 500, 1000$  or even larger. With the aid of computer, we can make  $B$  as large as we like to approximate to the sampling distribution of statistic
- Then calculate the mean and variance of  $\hat{\theta}(x_1^*), \hat{\theta}(x_2^*), \dots, \hat{\theta}(x_B^*)$  as

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

$$V_B = \frac{1}{B-1} \sum_{b=1}^B [\hat{\theta}(x_b^*) - \hat{\theta}_0]^2$$

*Handwritten notes:*  $S^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$  mean & variance of these B values

So, this is how one can create the bootstrap samples. And this process is repeated a large number of times. And suppose we repeated capital  $B$  number of times and we obtain  $B$  bootstrap samples independently. So, now, these bootstrap samples are denoted as  $x_1^*, x_2^*, \dots, x_B^*$ , you can see here I have denoted here the symbol here  $x_1^*$ .

So, this is actually indicating the first sample. So, whatever is the process which I have explained to you here this can be repeated and say capital B number of times, and we can obtain  $x_1^*, x_2^*, \dots, x_B^*$ . Now, based on this  $x_1^*, x_2^*, \dots, x_B^*$  simply trying to estimate your  $\theta$ .

So, you will have here first value of  $\hat{\theta}$  using the first sample, second value of  $\hat{\theta}$  using the second sample and Bth capital Bth value of  $\hat{\theta}$  using the capital Bth sample, right. So, now, I have here as  $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_B$ . Now, this B can be actually 500, 1000 or even larger.

Now, with the help of computer actually it does not make any difference, this is a matter of only just couple of seconds or couple of minutes. Means, if you try to draw the sample of 500 or 1000, I think usually this is a matter of only some seconds only. So, the rule is this if I try to make B as bigger as possible my approximation will become better and we will be getting good results, right.

So, with the help of computers we can make this B as large as we want to approximate the sampling distribution of the statistics. Now, so you have obtained such B samples and you have obtained the capital V values of  $\hat{\theta}$ . Now, I am asking please see what I am doing.

I am saying just find out the mean and variance of these B values that is all, I am not asking you a very big thing to do. So, the arithmetic mean of these B values will be

defined as  $\frac{1}{B} \sum_{b=1}^B \hat{\theta}(x_b^*)$ , I mean the that is the bth sample.

And the variance using the concept of capital  $S^2$  which was  $S^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2$ . This quantity can be the variance can be computed by this quantity  $V_B$ , right. So, this is simply the variance that can be obtained in r using the var command, right ok.

(Refer Slide Time: 39:24)

**Bootstrap methodology:**

For large  $n$  and  $B$ ,

- $V_B$  approximates the variance of  $\hat{\theta}(\underline{x})$  → True popn  $F$   $\text{Var}(\hat{\theta})$
- the empirical distribution of  $[\hat{\theta}(x_b^*) - \hat{\theta}(\underline{x})], b = 1, 2, \dots, B$  approximate closely the distribution of  $[\hat{\theta}(\underline{x}) - \theta(\underline{X})]$  where  $\underline{X} = (X_1, X_2, \dots, X_N)$  →  $\frac{n}{y}$   $\text{Var}(\bar{y})$

$(\hat{\theta}(\underline{x}) - \theta(\underline{x})) \rightarrow \text{dist}$

The bootstrap estimate of standard error of  $\hat{\theta}$  is  $\sqrt{V_B}$ .

Now, this is the statistical rule. Because, up to now we do not know what we have done is really good or bad. So, now, here comes the statistics and it has been proved using the statistical tool statistical methodology that for large small  $n$  and capital  $B$  if you means if your sample is reasonably large and your  $B$  is also large, then this quantity  $V_B$ . What is here  $V_B$ ? You can see here this is your here  $V_B$ , right.

- This  $V_B$  approximate the variance of  $\hat{\theta}(\underline{x})$

; that means, suppose you had known the true population. That means, you had knew the, suppose you knew the capital  $F$  from this capital  $F$  you try to find out the exact variance of  $\hat{\theta}$  means you try to compute  $\hat{\theta}$ .

And, then you try to find out its exact variant just like we have done in the case of simple random sampling, that we had a population from there we draw a sample of size  $n$ . Then we computed  $\bar{y}$  and then we computed variance of  $\bar{y}$ , right. So, what bootstrap is saying that  $V_B$  will approximately while approximate the value of variance of  $\hat{\theta}(\underline{x})$ . And, the empirical distribution of theta hat  $x_b^* - \hat{\theta}(\underline{x})$  will be approximately close to the true distribution of theta has  $\underline{x} - \underline{X}$  which is here the population.

This means here what? Suppose you knew the true population, then you would like to find out these value  $\hat{\theta}(\underline{x}) - \theta(\underline{X})$ , because everything is known to you and then you will try

to find out the distribution of this quantity. But, this is known to us sorry this is not known to us  $\theta$  population value is unknown to us. So, we cannot compute it. So, we are trying to estimate it on the basis of sample and those samples have been obtained on the basis of this bootstrapping.

And you can notice here that this  $\hat{\theta}$  is the bootstrap estimate of  $\theta$ , right. So, if you try to use it here, then both the distribution are going to be close enough. And in case if you want to find out the bootstrap estimate of standard error of  $\hat{\theta}$ , you simply have to take the positive square root of the variance that you have obtained that is square root of V B.

So, you can see here this is very simple and straight forward to obtain any distribution any standard error. And similarly this concept can be extended to any parameter.

(Refer Slide Time: 42:26)

**Bootstrap methodology of bias :**  
**The bootstrap estimate of bias of  $\hat{\theta}$  is**

$$\widehat{Bias}_B = \hat{\theta}_0 - t(\hat{F})$$

where

$$\hat{\theta}_0 = \frac{1}{B} \sum_{b=1}^B \hat{\theta}(x_b^*),$$

→  $\theta = t(F)$  is the parameter of interest,  
 →  $\hat{\theta} = t(\hat{F})$

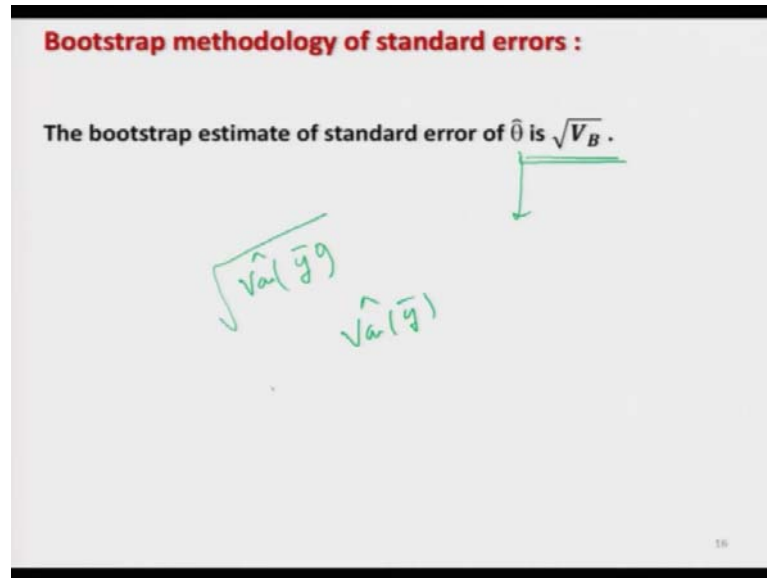
is the plug-in estimate of  $\theta$  based on the empirical distribution function, i.e., the vector of observed frequencies  $\hat{F} = (\hat{f}_1, \hat{f}_2, \dots)$  which is a sufficient statistic for the true distribution  $F = (f_1, f_2, \dots)$ .

This I will try to show you with some example. Now, in case if you want to find out the bias of this estimate, then now it is very simple. The estimate of bootstrap bias can be obtained by just using the bootstrap estimate of  $\hat{\theta}$  minus its empirical distribution function,, right.

So,  $\theta$  is the true parameter and  $\hat{\theta}$  is based on the empirical distribution  $\hat{F}$ . And which is simply the plug-in estimate of  $\theta$ . And, if you try to do it here it is very easy to compute

this bootstrap estimate. Well, you are not going to compute it manually because software will give you this output, but my idea of to, of explanation is you should know what are you getting.

(Refer Slide Time: 43:35)



And, similarly if you want to estimate or find the standard error of  $\hat{\theta}$ , earlier I said that if you have got the variance of  $\bar{y}$  high bar and if you take its positive square root, possibly the properties of square root of variance of  $\bar{y}$  or estimate are not going to be the same as the variance of  $\bar{y}$  estimator.


But, now I am saying that if you want to do it just take the positive square root of the variance of the bootstrapped values that you have obtained, through the resampling methods that is all. Now, what this  $\hat{\theta}$  can be your correlation coefficient  $\hat{\theta}$  can be your coefficient of variation or  $\hat{\theta}$  can be any function of  $x_1, x_2, \dots, x_n$ .

Now, you can see the importance. This can be any complicated function of  $x_1, x_2, \dots, x_n$ . And, this function can be the same function which earlier used to makes a make our life difficult from the algebra point of view. But, now at least we are unable to possibly solve the algebra, but at least numerically for a given sample, we can have a nice value, we can have a good value ok.

(Refer Slide Time: 44:45)

**Standard deviation of sample coefficient of variation:  
Example**

Population of 4 balls  
Weights of balls



Suppose we are interested in finding the standard error of coefficient of variation (CV) of the weights of the balls.

$$CV = \frac{\text{Standard deviation}}{\text{mean}} \quad \text{Popm} \quad \begin{matrix} \sigma \rightarrow \\ \mu \rightarrow \end{matrix} \quad \begin{matrix} \hat{\sigma} \\ \hat{\mu} \end{matrix}$$

So we draw the SRSWR samples of size 4 and find the CV of each of the sample.

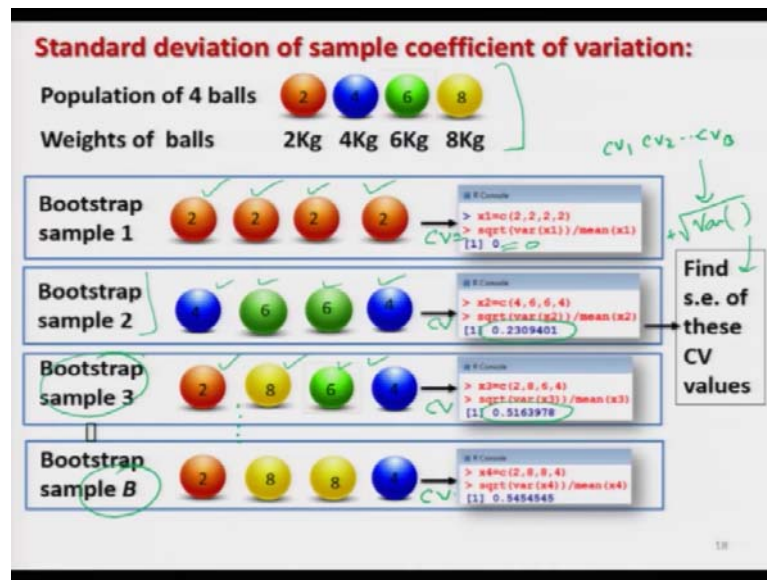
17

So, now let me try to take a very simple example to just illustrate what I have done. So, suppose I have the same example which I took earlier in the lecture that I have here 4 balls, which are of different colors and they have got different weights, 2 kg, 4 kg, 6 kg and 8 kg. And, suppose we are interested in finding the standard error of the coefficient of variation of the weights of these balls. The coefficient of variation is defined as the standard deviation divided by mean. So, these are your here population value.

So, the it is something like  $\sigma/\mu$  means  $\sigma$  is the, so  $\sigma$  is the standard deviation and  $\mu$  is the population mean and both are actually unknown to us. So, what we propose? That we replace  $\sigma$  by  $\hat{\sigma}$  and  $\mu$  by  $\hat{\mu}$ , where  $\hat{\sigma}$  and  $\hat{\mu}$  are based on some sample values.

So, what we try to do here now we use the bootstrap methodology. So, we have here a sample of size 4, I try to draw here bootstrap samples of size 4 by SRSWR. And, then I try to find out the standard error of this CV, how? Let me try to show you.

(Refer Slide Time: 46:03)



Suppose, this is my population or the original sample, I try to draw here 4 samples. Say first sample comes out to be 2, 2, 2, 2; that means, all are red colored ball, they have got the same weight. Now, I try to find out the CV of this sample in R I have given you here the screenshot of the outcome, I have just taken the square root of variance of these values 2, 2, 2, 2 divided by mean of these values. And this comes out to be 0 ; obviously, because all the values are the same. So, the variance is going to be 0.

But this is possible, so, I have taken this example to assure you that this can be there though do not get confused. Now, I try to take one more bootstrap sample, suppose this comes out to be 4, 6, 6, 4 and I try to find out its CV. So, now, the CV is coming out to be 0.2309401. And, similarly if I try to take here 3rd bootstrap sample over here, and whose values are coming out to be 2, 8, 6, 4 and I try to find out its CV which is coming out to be close to 0.51.

And, then I can continue here and then I can means, obtain here B such samples. And for each of the sample I try to compute the CV, now you can obtain you can see here you have got here the values  $CV_1, CV_2$  up to here  $CV_B$ . Now, I am saying just try to find out the variance of these values and take their positive square root.

So, simply trying to compute this, this is the variance of this take the positive square root of the variance this is going to give you the standard error of these CV values, right. So,

you can see here this was very simple and now you can imagine that instead of CV you can take any other thing. In some more example, I will try to show you that when we have more than one variable, say two variable then I can also compute the correlation coefficient ok.

So,, right, so now, this is the time to stop in this lecture. I have given you the basic methodology basic principle behind this bootstrapping and then I have taken an example to convince you that it is not difficult and how it can be executed. Well, on the next class I will try to do the same thing in the software also. And after that we will consider the confidence interval estimation.

First I want to make sure that you have understood these things, and you know how to compute at least these quantities or you know how to execute these things in R software. So, I would say try to think about it, try to settle down these concepts in your mind, try to convince yourself that the results are going to be good, try to take some artificial data sets where you know that what is the population values. From that population you try to draw.

Suppose a sample and from there you try to find out the bootstrap values, then you try to take couple of more samples. And you try to repeat the same bootstrap methodology. And, then try to see well you will see that the values which you are obtaining even by taking different sample from the same population, they are not differing much . They will be differing, but you always have to keep in mind that you do not know the true value; once you do not know the true value you have no other option.

So, try to make this approximation as good as possible, and now you can see that this will be depending on whether your sample is representative or not. So, that is why all the sampling techniques once again comes into picture, all those fundamentals whatever you have studied they come here in the picture. And, if you do not know and suppose somebody draws a wrong sample even the bootstrap sample bootstrap estimates will also be wrong.

Then, please do not blame the statistics. The problem lies with the candidate who do not know statistics. And the bigger problem is this without knowing the statistics he has jumped into the area of data science. So, once again I have proved that statistics is an



integral part of data science. So, if you want to become a data science you have no escape route now. You have to study only then you can be a successful data scientist. So, wish you all the best and I will see you in the next lecture, till then good bye.