

Essentials of Data Science with R Software – 1
Professor Shalabh
Department of Mathematics and Statistics
Indian Institute of Engineering Kanpur
Lecture 62
Basic Concepts of Confidence Interval Estimation

Hello friends, welcome to the course, Essentials of Data Science with R Software 1 in which we are trying to understand the basic concepts of probability theory and statistical inference. So, from this lecture, we are going to start our new topic on the Confidence Interval Estimation. So, you can recall that in the last couple of lectures, we had talked about the point estimation that we are trying to estimate the value of a parameter at a given point.

Now, we had briefly discussed that this quantity can also be estimated this parameter can also be estimated in the form of an interval. So, let me try to take the same example that I took earlier suppose you want to know that how much time you will take from your home to your college and suppose, you conduct an experiment and you go to your college and try to note down the time and finally, you try to take the arithmetic mean of those values and give me the answer that that it will take around 25 minutes.

So, that means, you are trying to tell me the value of your time at a point as 20 minutes or 25 minutes. Now, can you also make a statement like I will take say between 20 to 30 minutes? So, that means, this value of the parameter can be any value in the interval 20 to 30. So, that means you are trying to estimate the parameter in the form of an interval.

So, this is actually related to Confidence Interval Estimation that on the basis of the given sample of data, you are trying to compute or estimate two values which are the lower and upper limits of an interval. So, this is actually related to interval estimation of the parameters and in this lecture, we are going to talk about that how one can estimate the parameters in the form of an interval.

So, in this lecture, I will try to introduce you with the concept and I will try to give you the basic fundamentals, terminologies etc. And in the next lecture, I will try to construct the confidence interval for the normal mean when variance is known, and then we will try to attempt to find out the confidence interval when variances are known. So, let us begin our lecture.

(Refer Slide Time: 03:00)

Interval Estimation of Parameters:
Consider an example to understand what we mean by interval estimation.

Consider a situation in which a student wants to know the time taken to travel from the home to the college. Suppose the student makes 30 trips and notes down the time taken.

To get an estimate of the expected time, one can use the arithmetic mean. Let us say $\bar{x} = 30$ minutes. This is the point estimate for the expected travelling time.

It may not be appropriate to say that the student will always take exactly 30 minutes to reach the college.

2

So, let us try to first consider a very simple example to understand what do we really mean by interval estimation. So, consider a situation in which the student wants to know the time taken to travel from home to the college and suppose the student makes 30 trips and notes down the time taken. So, now, to get an estimate of the expected time the one can use the arithmetic mean and suppose, the arithmetic mean of these 30 values comes out to be 30 minutes.

So, this is the value of the time at a given point. So, this is the point estimate for the expected traveling time. But if you try to see, when is one where you are trying to say that, that the student will take say 30 minutes, it does not mean or it may not be appropriate to say that a student will always take exactly 30 minutes to reach the college. And variability that is there in the values which have helped in obtaining the value \bar{x} equal to 30 is not represented here.

(Refer Slide Time: 04:11)

Interval Estimation of Parameters:
Rather the time may vary by a few minutes each time.

To incorporate this feature, the time can be estimated in the form of an interval.

A statement like the time varies mostly between 25 and 35 minutes is more informative.

Both – mean and variation of the data are taken into account.

The interval (25 minutes, 35 minutes) provides a range in which most of the values are expected to lie.

We call this concept interval estimation.

3

So, now, what will really be happening that that time of traveling every time maybe varying by a couple of minutes by a few minutes. So, to incorporate this feature, the time can be estimated in the form of an interval and a statement like the time varies mostly between 25 and 35 minutes is more informative and it and it is more meaningful. In such a case the both the aspects that is the mean and variation of the data are taken into account.

So, for example, I can write this information in the form of an interval like this 25 to 35. So, this interval is giving us a range in which most of the values are expected to lie. And this concept is called as Interval Estimation of Parameters.

(Refer Slide Time: 05:07)

Interval Estimation of Parameters:

A point estimate can not take into account the variation in the values of the estimate.

The deviation between the point estimate and the true parameter (e.g. $|\bar{x} - \mu|$) can be substantial in many situations, especially when the sample size is small.

To incorporate the information about the precision of an estimate in the estimated value, a confidence interval can be constructed for which we have a certain degree of confidence that the parameter, e.g., μ lies within.

So, a point estimate cannot take into account the variation in the values of the estimate, but what is really happening? There is always some difference between the point estimate and the observed value and this type of deviation between the point estimate and the true parameter can be substantial in many situations and particularly when that sample size is small. So, to incorporate the information about the precision of an estimate in the estimated value what we do?

We can construct a confidence interval. For which we have a certain degree of confidence that the parameter for example, say here μ lies inside this interval. So, this is how we try to proceed.

(Refer Slide Time: 06:02)

Interval Estimation of Parameters:

- An interval estimate for a population parameter is called a confidence interval.
- The length of the interval reflects the uncertainty about μ .
- The wider the interval is, the higher is our uncertainty about the location of μ .
- Information about the precision of estimation is conveyed by the length of the interval.
- A short interval implies precise estimation.

So, an interval estimate for a population parameter is called as confidence interval and what about this interval? The length of the interval reflect the uncertainty about μ . That means, whether you are trying to say the interval is 20 to 25 minutes or 20 to 35 minutes. So, the length of the interval in this case is the upper limit – lower limit which is 5 minutes and length of the interval in this case is upper limit 35 – lower limit 20 which is 15 minutes.

So, this is the length of the interval, the wider the interval is the higher is our uncertainty about the location of μ . That is obvious means, if you try to say that μ is lying suppose between see here 10 and 100. So, that means, the value of μ can be anywhere in this interval between 10 to 100. But, in case if you try to take the interval like as here 10 to see here 15.

So, that is indicating the value of μ is going to be most likely between 10 and 15. So, what do you think which is more precise? Differently a shorter interval is more precise. So, a shorter interval implies precise estimation and the information about the precision of estimation is conveyed by the length of the interval. So, this is the criteria that we try to impose when we are trying to construct the confidence interval.

(Refer Slide Time: 07:42)

Interval Estimation of Parameters:

- We cannot be certain that the interval contains the true but unknown population parameter.
- The confidence interval is constructed so that we have high confidence that it does contain the unknown population parameter.

In that sense, the location of the interval will give us some idea about where the true but unknown population parameter μ lies.

6

And we cannot be certain that the interval contained the true but unknown population parameter always. So, in order to take care of this, we try to associate a concept of confidence coefficient and we say that confidence interval is constructed so that we have high confidence that it does contain the unknown population parameter. Because you see, the population parameter that you are trying to simulate that is practically unknown to us.

So, we need to find out the confidence intervals or the lower and upper limit of the confidence interval in such a way says that this unknown parameter value is expected to lie inside this interval. So, in that sense, the location of the interval will give us some idea about where the true but unknown population parameter μ lies inside the interval.

(Refer Slide Time: 08:41)

Interval Estimation of Parameters:

Confidence interval of a parameter θ is a random interval with lower bound as $\hat{\theta}_L$ and upper bound $\hat{\theta}_U$, such that the unknown parameter θ is covered by a prespecified probability of at least $1 - \alpha$:

$$P_{\theta}[\hat{\theta}_L(X) \leq \theta \leq \hat{\theta}_U(X)] \geq 1 - \alpha$$

where $\underline{X} = (X_1, X_2, \dots, X_n)$.

$1 - \alpha$: probability, confidence level or confidence coefficient,
 $\hat{\theta}_L(X)$: lower confidence bound or lower confidence limit and
 $\hat{\theta}_U(X)$: upper confidence bound or upper confidence limit.

So, now, how to formally define the confidence interval? So, suppose, we have a parameter θ and θ lies in some random interval and upper and lower limits of this random interval they have to be found, they have to be estimated. So, suppose the lower bound of this interval is indicated by $\hat{\theta}_L$, L means lower and the upper bound is indicated by $\hat{\theta}_U$.

U means upper bound and when we are trying to know the value of unknown parameter θ , then we say that these two bound, $\hat{\theta}_L$ and θ U are defined such that the unknown parameter θ is covered by a prespecified probability of at least $1 - \alpha$ and through this statement, we try to impose our confidence on this statement. And we try to state it like this.

The $P_{\theta}[\hat{\theta}_L(X) \leq \theta \leq \hat{\theta}_U(X)] \geq 1 - \alpha$ and $\hat{\theta}_L$ and $\hat{\theta}_U$ they are dependent on the random sample X_1, X_2, \dots, X_n . So, I write it here as a $\hat{\theta}_L$ and inside the parentheses I write X so, that it is indicating that this is a function of X_1, X_2, \dots, X_n and same is true for the $\hat{\theta}_U$ also.

So, in this statement this $1 - \alpha$ is the probability here and this is called as confidence level or confidence coefficient and this limit $\hat{\theta}_L$, this is the lower confidence bound or the lower

confidence limit and this is based on X_1, X_2, \dots, X_n . So, we are writing here So, $\hat{\theta}_L(X)$ and similarly, this quantity $\hat{\theta}_U(X)$ this is the upper confidence bound or say upper confidence limit based on the X_1, X_2, \dots, X_n .

(Refer Slide Time: 11:04)

Interval Estimation of Parameters:

It is important to note that the bounds $\hat{\theta}_L(X)$ and $\hat{\theta}_U(X)$ are random and the parameter θ is a fixed value.

This is the reason why we say that the true parameter is covered by the interval with probability $1 - \alpha$ and not that the probability that the interval contains the parameter is $1 - \alpha$.

Some software packages use the term "error bar" when referring to confidence intervals.

8

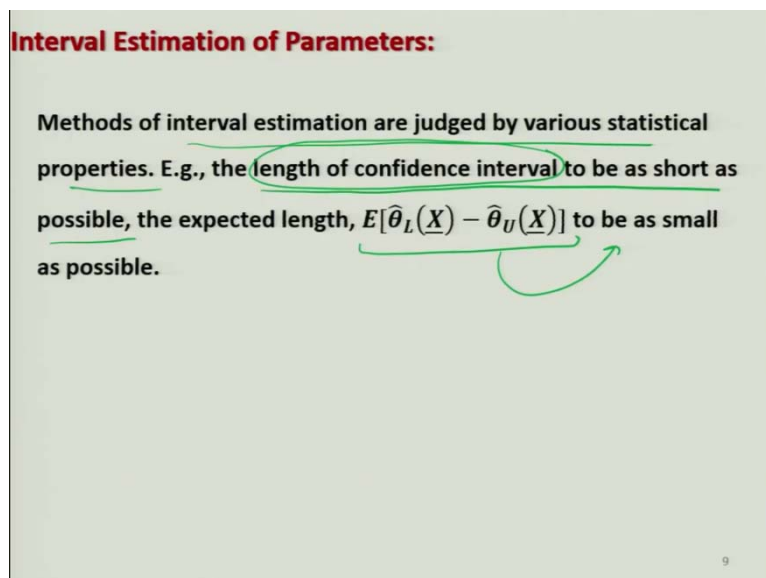
One thing you have to keep in mind here that both these value $\hat{\theta}_L(X)$ and $\hat{\theta}_U(X)$ they are statistic they are random and the parameter θ is expected to lie between these two limits is a fixed value and this is the reason that why we say that the true parameter is covered by the interval with probability $1 - \alpha$ and not that the probability that the interval contains the parameter is $1 - \alpha$.

This is very important to understand. Because many times people make this type of mistake in giving a proper interpretation to the confidence interval and when you are trying to compute this interval in some software package, sometimes you will see that they are using a term like error bar when they are referring to the confidence interval. Anyway, we are going to find it on the basis of the software only.

(Refer Slide Time: 12:09)

Interval Estimation of Parameters:

Methods of interval estimation are judged by various statistical properties. E.g., the length of confidence interval to be as short as possible, the expected length, $E[\hat{\theta}_L(X) - \hat{\theta}_U(X)]$ to be as small as possible.



Now, there comes once you can find out the value of $\hat{\theta}_L$ and $\hat{\theta}_U$ which are the lower and upper limits of the parameter θ , we also need to judge whether these values are good or bad. So, there are different methods to judge the interval estimation that whether the values of $\hat{\theta}_L$ and $\hat{\theta}_U$ are good or not. So, one criterion is that the length of the confidence interval should be as small as possible, as short as possible.

And we try to compute the expected length. And this expected length is expected to be as small as possible. Because the length of the confidence interval that is in the common language, but the length of the confidence interval depends on two random quantities $\hat{\theta}_L$ and $\hat{\theta}_U$. So, that is why we try to base our criteria on its expected value.

(Refer Slide Time: 13:25)

Frequency Interpretation of the Confidence Interval:

Suppose N independent samples $\underline{X}^{(j)}, j = 1, 2, \dots, N$ of size n are sampled from the same population and N confidence intervals of the form $[\hat{\theta}_L(\underline{X}^{(j)}), \hat{\theta}_U(\underline{X}^{(j)})]$ are calculated.

If N is large enough, then on an average $N(1 - \alpha)$ of the intervals

$$P_{\theta}[\hat{\theta}_L(\underline{X}) \leq \theta \leq \hat{\theta}_U(\underline{X})] \geq 1 - \alpha$$

cover the true parameter θ .

N ind. samples

$\underline{X}^{(1)} = (X_1^{(1)} \dots X_n^{(1)})$
 $\underline{X}^{(2)} = (X_1^{(2)} \dots X_n^{(2)})$
 \dots
 $\underline{X}^{(N)}$

10

Now, the next question comes that when you are trying to conduct an experiment, then what will be the interpretation of the confidence interval from the frequency point of view. Because, if you remember when we did the probability, we had defined the probability and we also had defined the probability in terms of relative frequency. So, similarly, there is a frequency interpretation of this confidence interval also, and this is as like this.

Suppose, we tried to draw capital N number of independent samples. And those samples are indicated by \underline{X} and in their superscript, this is here j . So, what is really happening this X here is say, X_1, X_2, \dots, X_n , and when I am trying to say here one, so, this is the first sample. So, there are going to be $X^{(1)}_1, X^{(1)}_2$, up to here $X^{(1)}_n$ number of values and similarly, if you try to obtain here the second sample, this will look like X_1, X_2, \dots, X_n .

But now, I am trying to indicate that this is the second sample and similarly, you can find out here X_n . So, we have such capital N independent samples and each of the sample. And each of samples and you can see here there are N observation, there are here N observations and they are sampled from this same population. And based on that, we tried to construct the confidence interval for each of the sample.

So, there is going to be one confidence interval for this sample and there is going to be another confidence interval for this second sample and similarly, we have the capital N^{th}

confidence interval for the capital N^{th} sample. And we try to compute the lower and upper limits of each of the interval like as $\hat{\theta}_L(X_j)$ and $\hat{\theta}_U(X_j)$.

Now, if capital N is large enough, then on an average N into $1 - \alpha$ of the intervals like this one to cover the true parameter θ . That probability that θ is lying between $\hat{\theta}_L(X)$ and $\hat{\theta}_U(X)$ is greater than or equal to $1 - \alpha$.

(Refer Slide Time: 16:02)

Frequency Interpretation of the Confidence Interval:
Let a random variable follow a normal distribution $N(\mu, \sigma^2)$.
Suppose we draw a sample of n observations repeatedly, say N times.
Compute $\hat{\theta}_L(X)$ and $\hat{\theta}_U(X)$ for each sample. *N samples*
We have N values of $\hat{\theta}_L(X)$ and $\hat{\theta}_U(X)$.
The samples will differ in each draw, and hence, the mean and the confidence interval will also differ.
Most of the $\hat{\theta}_L(X)$ and $\hat{\theta}_U(X)$ vary with respect to the mean and the confidence interval width.
So most confidence intervals, but not all, include μ .

So, now, we try to take an example and try to explain you this statement. Suppose, there is a random variable which is following a normal distribution with mean μ and variance sigma square and from this population suppose we draw capital N samples and each of the sample is having a small n number of observations. Now, we compute the lower and upper bounds of the confidence interval for each of the samples.

So, we have $\hat{\theta}_L(X)$ and $\hat{\theta}_U(X)$ for each of the sample. So, we are going to have capital N values of $\hat{\theta}_L(X)$ and $\hat{\theta}_U(X)$ and the samples will differ from each other in every draw. And hence, the sample mean and the confidence interval for each of the sample will also differ that is obvious means, if you try to take two different sample, we expect that the sample means are going to be different and similarly, the confidence interval will also be different.

And most of the $\hat{\theta}_L(X)$ and $\hat{\theta}_U(X)$ will vary with respect to the mean and the confidence interval with because you are trying to compute them again and again based on a random sample. So, the width of the confidence interval for each of the sample as well as the sample

mean of each of the sample will also be vary. So, what will happen? That most of the confidence interval but surely not all will include μ .

(Refer Slide Time: 18:02)

Frequency Interpretation of the Confidence Interval:
Following is the idea of the frequency interpretation of the confidence interval:

- Different samples will yield different point and interval estimates.
- Most of the times the interval will cover μ , but not always.
- The coverage probability is specified by $1 - \alpha$. $0 \leq \alpha \leq 1$
- The frequency interpretation means that we expect that (approximately) $(1 - \alpha) \cdot 100\%$ of the intervals to cover the true parameter μ . $\alpha = 0.05$ 95%

12

This is the frequency interpretation of the confidence interval. Now, in case if I try to comprehend this discussion, I can say very simply that different samples will a different point and interval estimates and most of the times the interval will cover μ , but surely not always, and the coverage probability is specified by and the term $1 - \alpha$; α is a quantity which is lying between 0 and 1.

And frequency interpretation means that we expect that approximately $1 - \alpha$ into 100 percent of the intervals will cover the true parameter μ . So, in case if I try to take here α is equal to 0.05 that means 95 percent of the intervals will cover the value of the true μ . This is what we mean by the interpretation of this confidence interval.

(Refer Slide Time: 19:09)

Pivotal Quantity Method to Derive a Confidence Interval:

A general method for finding a confidence interval for an unknown parameter is as follows:

1. Let X_1, X_2, \dots, X_n be a random sample of n observations.
2. Suppose we can find a statistic $g(X_1, X_2, \dots, X_n; \theta)$ such that
 - $g(X_1, X_2, \dots, X_n; \theta)$ depends on both the sample and θ but
 - the probability distribution of $g(X_1, X_2, \dots, X_n; \theta)$ does not depend on θ or any other unknown parameter.

Such $g(X_1, X_2, \dots, X_n; \theta)$ is called as pivotal quantity .

13

Now, the question is how can you construct the confidence interval? So, first of all I am going to give here a general methodology and after this I will try to construct the confidence interval for the mean and variance from a normal population. So, a general method for finding a confidence interval for an unknown parameter is like this. The first step is that we need to draw a sample or rather a random sample of a small n observation.

And then, suppose, we can find a statistic g which is a function of X_1, X_2, \dots, X_n and g is defined such that or g is found such that g depends on both the sample X_1, X_2, \dots, X_n and the parameter data, but the probability distribution of the g does not depend on θ or any other unknown parameter.

So, these are the two points that you have to find here, a statistic g is a function of X_1, X_2, \dots, X_n and θ and you have to find it in such a way so, that the statistics depends on both the sample and the parameter, but the probability distribution of this g does not depend on the θ or any other unknown parameter and such a g is called as pivotal value or pivotal quantity. Once you can obtain such a pivotal quantity, then you can find out the confidence interval very easily.

(Refer Slide Time: 20:52)

General Method to Derive a Confidence Interval:

For example, if X_1, X_2, \dots, X_n is a random sample from normal distribution $N(\mu, \sigma_0^2)$ where σ_0^2 is known, then

$$g(X_1, X_2, \dots, X_n; \theta = \mu) = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma_0}$$

depends on both the sample and θ but the probability distribution of $g(X_1, X_2, \dots, X_n; \theta = \mu)$ is $N(0, 1)$ which does not depend on $\theta = \mu$ or any other unknown parameter.

Now one can compute $\hat{\theta}_L(X)$ and $\hat{\theta}_U(X)$ so that

$$P_{\theta}[\hat{\theta}_L(X) \leq \theta \leq \hat{\theta}_U(X)] \geq 1 - \alpha$$

Handwritten notes in the image:
 - A box around $\bar{X} - \mu$ with an arrow pointing to μ .
 - A box around (σ_0/\sqrt{n}) with the word "known" written below it.
 - The text "g: pivotal quantity" written in green next to the distribution description.
 - The final probability expression is underlined.

Say for example, suppose X_1, X_2, \dots, X_n is a random sample from normal μ , sigma square where the variance sigma squared is known at sigma 0 square. Now, you can see here that $g(X_1, X_2, \dots, X_n)$ and θ , θ is equal actually here μ is like this square root of n , $\bar{X} - \mu$ upon sigma 0. So, it is actually $\bar{X} - \mu$ upon the standard deviation like this one. So, now, you can see here, here this quantity is known.

So, this whole quantity this is depending on μ that is θ , but you also know what is the probability distribution of this statistics? This is normal 0 1. So, you can see here that this is test G is depending on μ , but the distribution of g is normal 0 1 which is non depending on μ or any other unknown parameter. So, now, this g is your pivotal quantity and this can be used to obtain the lower and upper limits of the confidence interval for a parameter θ .

So, finally, I can say that interval estimation depends upon the finding the value of $\hat{\theta}_L$ using the value of X_1, X_2, \dots, X_n and $\hat{\theta}_U$ using the values X_1, X_2, \dots, X_n in such a way such that θ is lying between these two bounds and the probability that θ is lying between these two bound is greater than or equal to $1 - \alpha$. And once you can find out, such forms of $\hat{\theta}_L$ and $\hat{\theta}_U$, you have obtained the confidence interval for the parameter θ .

So, now, we come to an end to this lecture. And in this lecture, that was a brief lecture and I have tried my best to give you some basic concepts about the Confidence Interval Estimation and I am stopping here intentionally so that you can think about these things. And first you

try to settle down these definitions like as pivotal quantity and other things, lower bound, upper bound, confidence coefficient.

And then in the next lecture, I am going to consider the case of construction of the confidence interval for the parameter μ from one normal population. But in case if you try to settle down these things inside your brain, try to clear this concept inside your mind then possibly finding out confidence interval is a very simple and you can understand it very easily. So, try to revise this lecture and I will see you in the next lecture with the confidence interval on μ . Till then, goodbye.