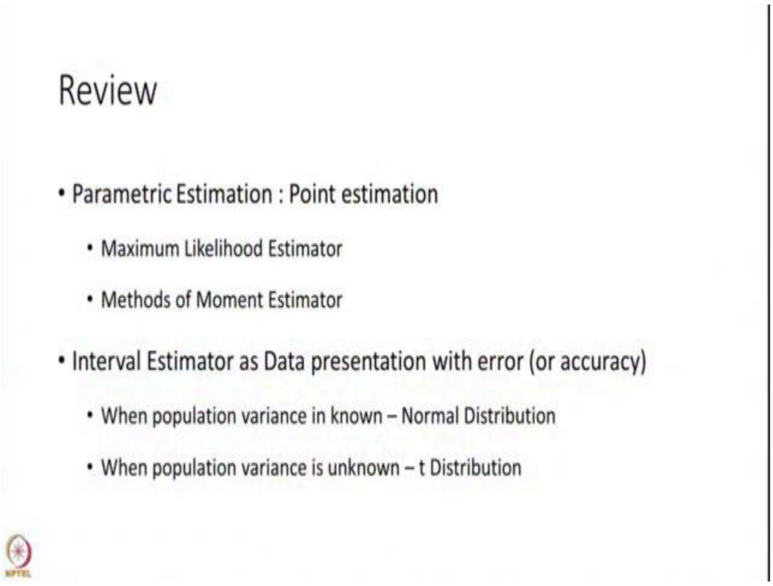
Dealing with Materials Data
Professor M P Gururanjan
Professor Hina A Gokhale
Department of Metallurgical Engineering and Materials Science
Indian Institute of Technology, Bombay
Lecture 53
Parameter Estimator 4

Hello and welcome to the course on Dealing with Materials Data. Presently we are going through the sessions on Parametric Estimation.

(Refer Slide Time: 00:29)



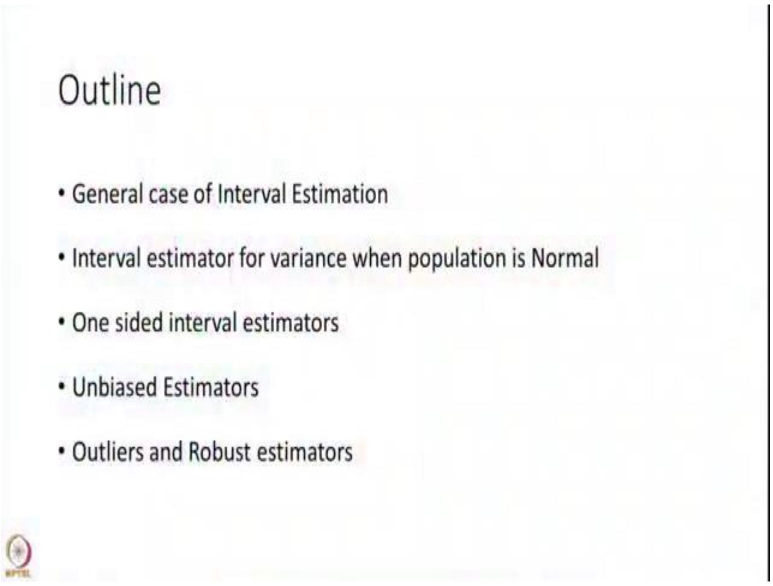
Review

- Parametric Estimation : Point estimation
  - Maximum Likelihood Estimator
  - Methods of Moment Estimator
- Interval Estimator as Data presentation with error (or accuracy)
  - When population variance in known – Normal Distribution
  - When population variance is unknown – t Distribution

We have gone through the session we have described point estimators such as, maximum likelihood estimator and methods of moment estimator. Then we discussed in the previous session the interval estimators as a data presentation with the error and we discussed two cases. One is a case of a normal distribution when the sigma is known, the standard deviation of the population is known.

Then we can use normal distribution to come up with the interval estimator of the parameter mean the population mean Mu. And when the variance is unknown, we can make use of t distribution to come up with the interval estimator for the population mean Mu.

(Refer Slide Time: 01:41)



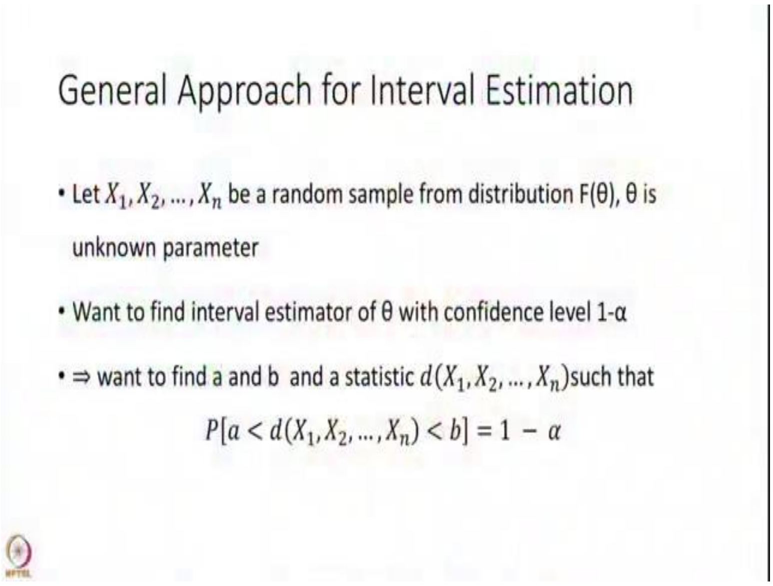
Outline

- General case of Interval Estimation
- Interval estimator for variance when population is Normal
- One sided interval estimators
- Unbiased Estimators
- Outliers and Robust estimators

In this present case very loosely I will try to explain, what is interval estimation in general. And in particular we will take a case of interval estimator of a variance when population is normal. And we will briefly discuss one sided interval estimator and then we will talk about the unbiased estimators, outliers and the robust estimators.

So, as I said this general approach I am just trying to give you a general feel is to, what is being done when you do the interval estimator, because we started with an example of a data representation. It is necessary to have a general understanding as to what is happening in this case.

(Refer Slide Time: 02:34)



General Approach for Interval Estimation

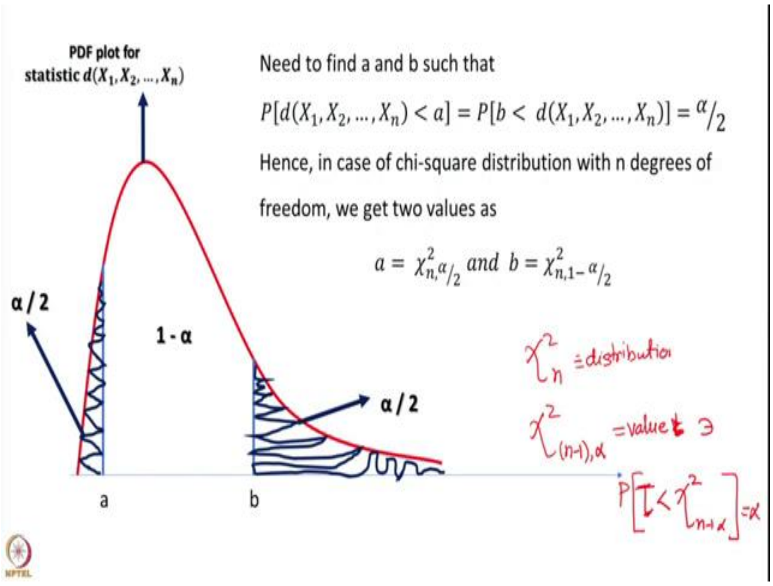
- Let $X_1, X_2, \dots, X_n$ be a random sample from distribution F(θ), θ is unknown parameter
- Want to find interval estimator of θ with confidence level 1-α
- ⇒ want to find a and b and a statistic $d(X_1, X_2, \dots, X_n)$such that

$$P[a < d(X_1, X_2, \dots, X_n) < b] = 1 - \alpha$$

So, let us consider that we have a sample of size n from a population with a distribution F theta and theta is an unknown parameter. And now we want to find an interval estimator of theta with some confidence level 1 minus alpha.

So, what we want to do is we want to find 3 things. We want to find the two numbers a and b and a statistic that is a function only data, d of X1, X2, X3, Xn such that probability of a less than d less than b is 1 minus alpha. And this statistic d will be some kind of an estimator or a function of the estimator of theta.

(Refer Slide Time: 03:43)



Need to find a and b such that

$$P[d(X_1, X_2, \dots, X_n) < a] = P[b < d(X_1, X_2, \dots, X_n)] = {}^{\alpha}/_{2}$$

Hence, in case of chi-square distribution with n degrees of freedom, we get two values as

$$a = \chi^2_{n, \alpha/2} \text{ and } b = \chi^2_{n, 1-\alpha/2}$$

χ^2_n = distribution

$\chi^2_{(n-1), \alpha}$ = value t ∋

$P[T < \chi^2_{n-1\,\alpha}] = \alpha$

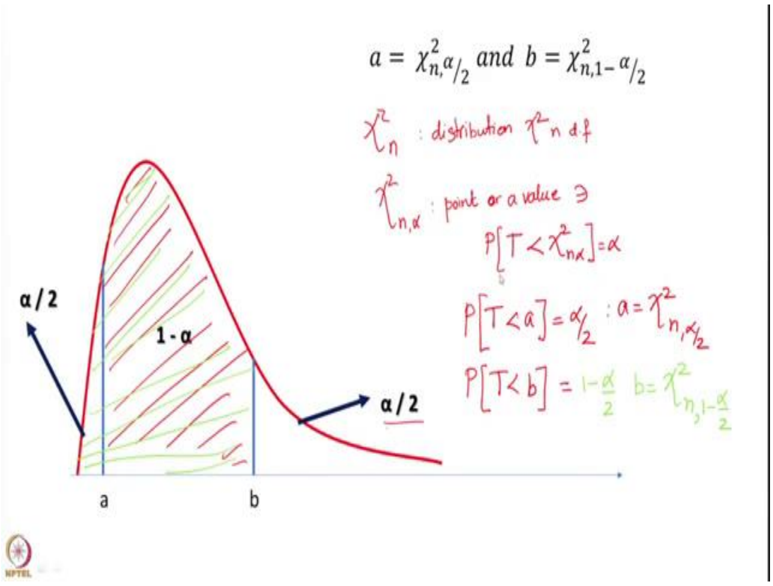
So, in this situation let us see that if we have a probability density plot for a statistic d and suppose it takes this form, I have very carefully shown and skewed density function. Because symmetric density function in terms of normal distribution and T distribution we have already discussed in the previous session.

So, I am very purposefully showing a very general density function which is a skewed density function. And now we are looking for a and b such that between a and b, the area under this density function curve is 1 minus alpha or we can say that area beyond b is alpha by 2 and area below a is alpha by 2. So, for example, chi square is one such skewed function and in that case we are looking for a chi square a value a which is chi square with n degrees of freedom with alpha by 2.

Because once again please remember we calculate the tabulate the value as probability of random variables smaller than the given value a. So, here the probability is alpha by 2, so I am considering it, a is equal to chi square with n degrees of freedom alpha by 2 as a value. And b is the chi square n, 1 minus alpha by 2 because this is alpha by 2, so this area is 1 minus alpha by 2.

I want to make this clear that this points are understood it correctly so let us see some things which I feel I should not clarify. See when I write chi square n degrees of freedom I mean a distribution. But, when I say chi square n minus 1 alpha what I mean is a value t such that probability of I would say a value such that probability of a random variable t less than chi square n minus 1 alpha is equal to alpha. This is what I mean, this is what I will explain in the next session.

(Refer Slide Time: 06:54)



$a = \chi^2_{n,\alpha/2}$ and $b = \chi^2_{n,1-\alpha/2}$

χ^2_n : distribution χ^2 n d f

$\chi^2_{n,\alpha}$: point or a value ∋ $P[T < \chi^2_{n\alpha}] = \alpha$

$P[T < a] = \alpha/2$: $a = \chi^2_{n,\alpha/2}$

$P[T < b] = 1-\frac{\alpha}{2}$ $b = \chi^2_{n,1-\frac{\alpha}{2}}$

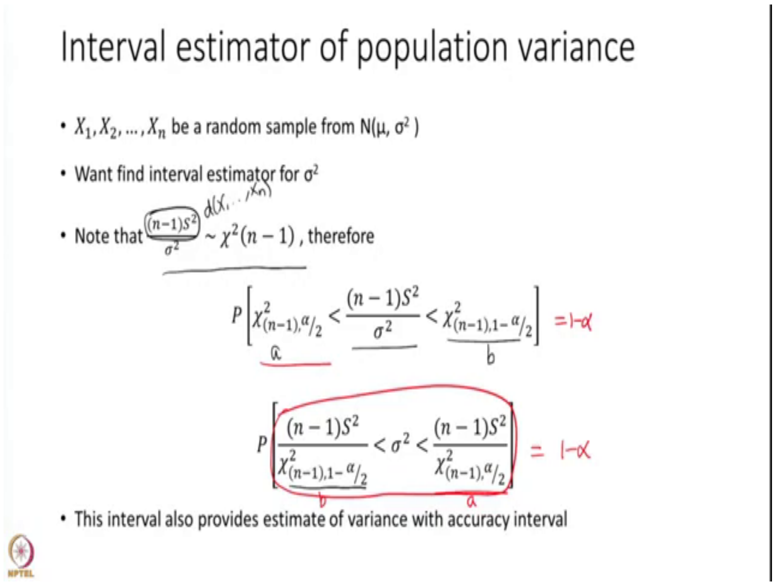
So, here what I want to say let us make it clear and this is true not just with chi square but the T distribution as well as others. Whenever, we say that chi square with n degrees of freedom I mean the distribution chi square with n degrees of freedom. But, when we write chi square n degrees of freedom alpha I mean a point or a value such that probability that T less than this value chi square n alpha is equal to alpha.

So please make sure, so in this case as I explained, this is 1 minus alpha, this area is 1 minus alpha so a, the area below a, we call it alpha by 2 and the area above a we call, above b we call alpha by 2. So, the first point a is very clear because you are taking probability of value less than a is equal to alpha by 2. So a has to be equal to in our notation chi square n alpha by 2, instead of alpha we have alpha by 2.

But, when you look at the value of b what happens is that, probability T less than b when you take that is not alpha by 2. It is probability T greater than b is alpha by 2, so you have to consider the complete probability which is here. Which is actually 1 minus alpha plus alpha by 2 which turns out to be 1 minus alpha by 2 and therefore b becomes chi square n, 1 minus alpha by 2 this is what I wanted to explain.

So, I hope now it is clear and this is how you look into the table because table always generally gives you these kind of, these values. But, as I and in this case there is no doubt because this is not a symmetric distribution so it always gives you this value and therefore you have to look up the value accordingly.

(Refer Slide Time: 10:06)



Interval estimator of population variance

- $X_1, X_2, \dots, X_n$ be a random sample from N(μ, σ²)
- Want find interval estimator for σ²
- Note that $\frac{\boxed{(n-1)S^2}}{\sigma^2}$ $d(X_1 \dots, X_n)$ $\sim \chi^2(n-1)$, therefore

$$P\left[\underbrace{\chi^2_{(n-1),\alpha/2}}_{a} < \underline{\frac{(n-1)S^2}{\sigma^2}} < \underbrace{\chi^2_{(n-1),1-\alpha/2}}_{b}\right] = 1-\alpha$$

$$P\left[\boxed{\frac{(n-1)S^2}{\underbrace{\chi^2_{(n-1),1-\alpha/2}}_{b}} < \sigma^2 < \frac{(n-1)S^2}{\underbrace{\chi^2_{(n-1),\alpha/2}}_{a}}}\right] = 1-\alpha$$

- This interval also provides estimate of variance with accuracy interval

Now, let us go to the interval estimator of population variance. So, we have an N size random sample from a normal distribution with mean Mu and variance sigma square. And we want to find an interval estimator of sigma square so we are assuming that sigma square is unknown. We have to find a statistic, you remember I said that you have to find a statistic which is in some kind of an estimator or you know a representative of an, a function of an estimator of the theta.

- $X_1, X_2, \dots, X_n$ be a random sample from N(μ, σ²)
- Want find interval estimator for σ²
- Note that $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$, therefore

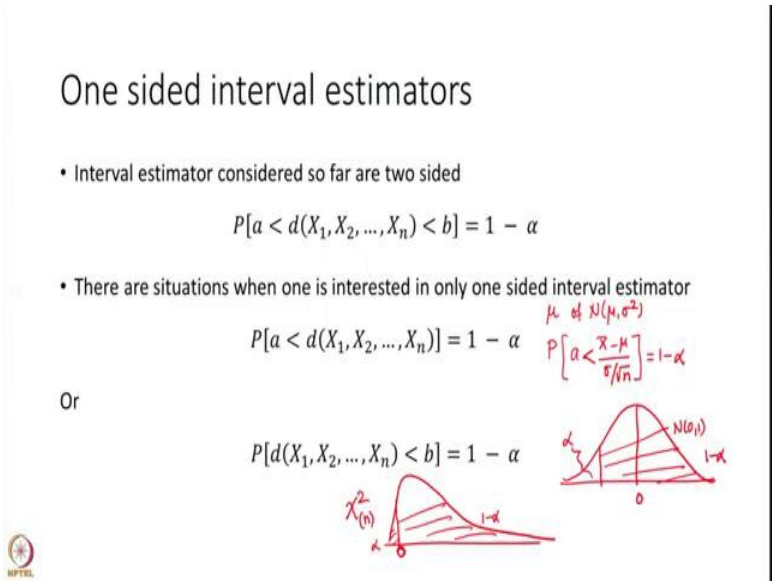
$$P\left[\chi^2_{(n-1),\,\alpha/2} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{(n-1),\,1-\alpha/2}\right]$$

$$P\left[\frac{(n-1)S^2}{\chi^2_{(n-1),\,1-\alpha/2}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{(n-1),\,\alpha/2}}\right]$$

This interval also provides estimate of variance with accuracy interval

So, remember that this is b and this is a, it is because we have taken the inverse function and then multiplied with then denominator this. So, this becomes an interval estimator and you know that this has a probability 1 minus alpha, this is 1 minus alpha. And therefore this provides the interval estimator of sigma square, the population variance of a normal population.

(Refer Slide Time: 13:10)



One sided interval estimators

- Interval estimator considered so far are two sided

$$P[a < d(X_1, X_2, \dots, X_n) < b] = 1 - \alpha$$

- There are situations when one is interested in only one sided interval estimator

$$P[a < d(X_1, X_2, \dots, X_n)] = 1 - \alpha$$

μ of $N(\mu, \sigma^2)$

$P\left[a < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right] = 1 - \alpha$

Or

$$P[d(X_1, X_2, \dots, X_n) < b] = 1 - \alpha$$

α
N(0,1)
$1-\alpha$
0
$\chi^2_{(n)}$
$1-\alpha$
α
0

Now, you see so far we have considered two sided estimators, it means that the estimator had the bound on two sides. Probability of a less than d less than b is equal to 1 minus alpha. Suppose we have only one sided estimator that is we take probability of a less than d is equal to 1 minus alpha or probability d less than b is 1 minus this.

Situation arises, this situation arises when you actually are, you know that a certain variable is always a greater than a certain number or you are not interested whether it is bounded by two sides. But, you just like to know if it is above one certain value or if it is below certain value.

In this case the one sided interval estimator are needed and I do not think we need to calculate. Because for every case that we calculated this interval estimator can also be derived using the same statistic d and the probability or the probability density function of that estimator.

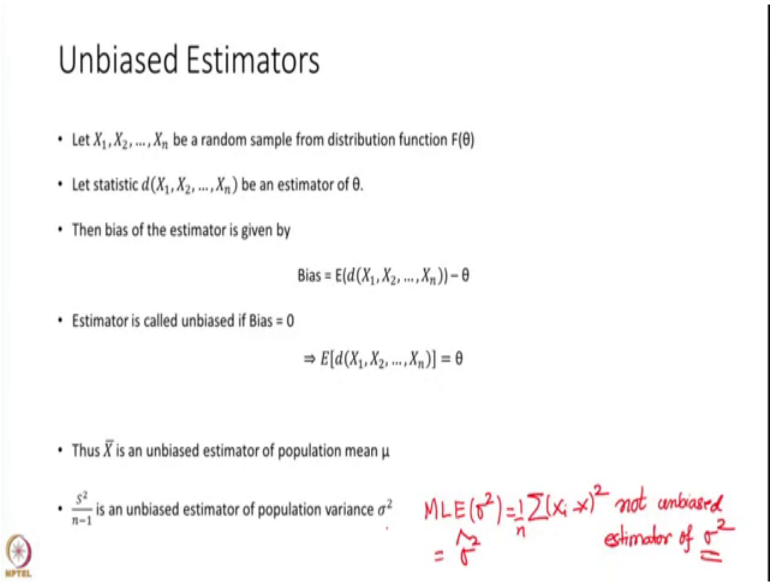
So for example, we can say that if you are talking about Mu of normal Mu sigma square and it is your sigma square is known. Then you are going to be talking about a less than x bar minus Mu over sigma square root n, which is going to be 1 minus alpha and so you are looking for in a normal distribution.

This is the normal 0, 1, this is 0. You are going to look for a value of a such that this probability is 1 minus alpha. You can look into the normal table and do it. Similarly, if you are looking for if you are into T distribution, you are going to follow the same method. In case it is a chi square distribution with n degrees of freedom then you have a skewed distribution.

And then you are looking for an a, so you are going to look for a which is like this. This is 1 minus alpha and therefore you are going to look at his table where this is alpha. Similarly, here also you look into the table where this is alpha. So, instead of alpha by 2, alpha by 2 on two sides, you will have alpha on one side. In the second case you will have alpha on the other side. So, I am not going into the details of this but it can easily be worked out in the same way as we have done in the two sided interval.

Now, we will move to what is called the unbiased estimators. This is a quality or evaluation of an estimator which is a for a point estimator.

(Refer Slide Time: 16:53)



Unbiased Estimators

- Let $X_1, X_2, \dots, X_n$ be a random sample from distribution function F(θ)
- Let statistic $d(X_1, X_2, \dots, X_n)$ be an estimator of θ.
- Then bias of the estimator is given by

$$\text{Bias} = \text{E}(d(X_1, X_2, \dots, X_n)) - \theta$$

- Estimator is called unbiased if Bias = 0

$$\Rightarrow E[d(X_1, X_2, \dots, X_n)] = \theta$$

- Thus $\bar{X}$ is an unbiased estimator of population mean μ
- $\frac{S^2}{n-1}$ is an unbiased estimator of population variance σ^2

$MLE(\sigma^2) = \frac{1}{n}\sum(x_i - x)^2 = \hat{\sigma}^2$ not unbiased estimator of $\underline{\underline{\sigma^2}}$

So, we are from interval estimators we are back to point estimator and we will first define, what is called unbiased estimators. So, let X1, X2, Xn be a random sample from a distribution F theta, theta is unknown and let statistic d X1, X2, Xn be an estimator of theta.

Then the biased of the estimator is defined by

$$\text{Bias} = \text{E}(d(X_1, X_2, \dots, X_n)) - \theta$$

Estimator is called unbiased if Bias = 0

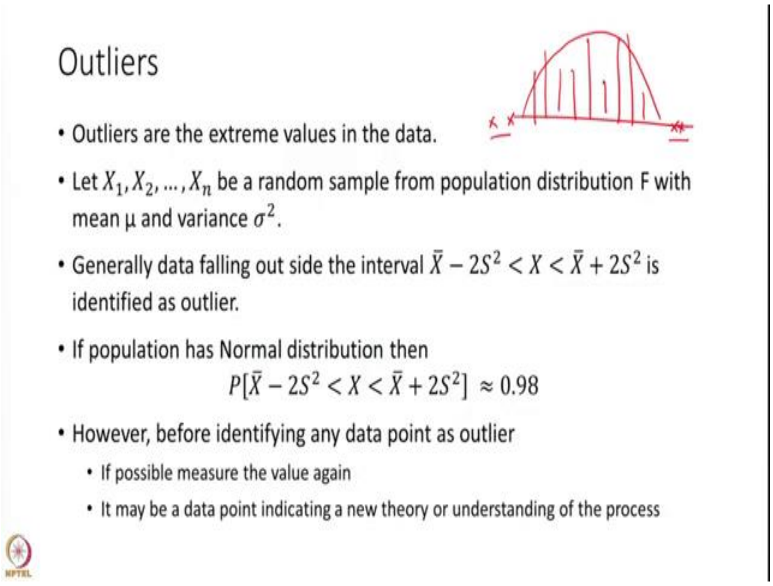
$$\Rightarrow E[d(X_1, X_2, \dots, X_n)] = \theta$$

And if the estimator if this biased is 0 then the estimator is called an unbiased estimator. So, if you recall from our past, we know that I mean our past lessons we know that sample mean is an unbiased estimator of a population mean Mu.

The sample variance is an unbiased estimator of a population variance sigma square. But remember that maximum likelihood estimator of sigma square which is 1 over n summation xi minus x bar whole square is not an unbiased estimator of sigma square. Please remember MLE of sigma square we generally call it sigma square hat is not an unbiased estimator. It is the sample variance which is an unbiased estimator of sigma square.

Sigma square hat maximum likelihood estimator of sigma is not an unbiased estimator of sigma. There is another we would like to talk about another kind of estimator which is called a robust estimator.

(Refer Slide Time: 19:18)



Outliers

- Outliers are the extreme values in the data.
- Let $X_1, X_2, \dots, X_n$ be a random sample from population distribution F with mean μ and variance σ^2.
- Generally data falling out side the interval $\bar{X} - 2S^2 < X < \bar{X} + 2S^2$ is identified as outlier.
- If population has Normal distribution then

$$P[\bar{X} - 2S^2 < X < \bar{X} + 2S^2] \approx 0.98$$

- However, before identifying any data point as outlier
  - If possible measure the value again
  - It may be a data point indicating a new theory or understanding of the process

But before that we would like to understand outliers. What is an outlier? Outlier are the extreme values in the data. So, if you look at any histograms you must have made some histogram plots. Suppose you have got some histograms like this for some data and you find that few points are just lying here.

Then these are called outliers in the data. It means that they do not fall into any shape that this histogram takes this it separately. So, let X1, X2, X3, Xn be a random sample from a population with distribution F with mean Mu and variance sigma square. I am not calling it a normal I am calling it a general distribution F, which has a population mean of Mu and population variance of sigma square.

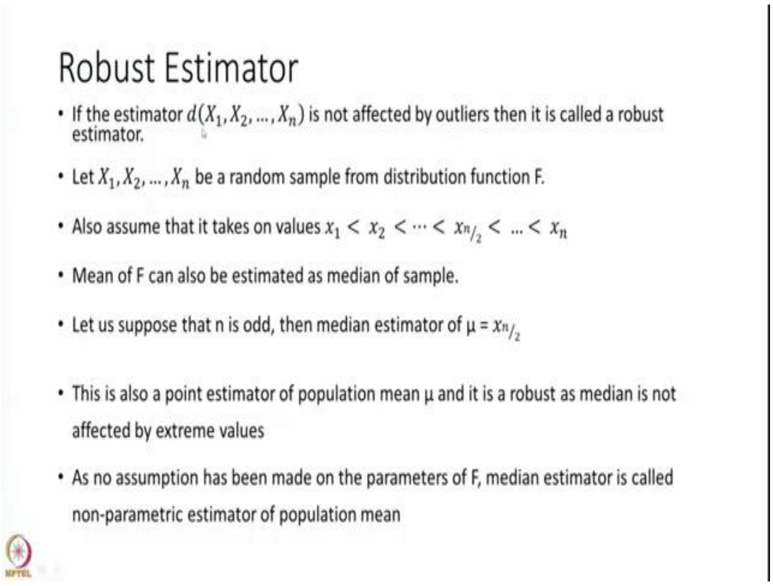
Generally, the data falling outside the interval of x bar minus 2, that is the sample mean minus 2 sample variance and sample mean plus 2 sample variance is identified as an outlier. This a general thumb rule, it is a thumb rule there is no proof to it but generally whatever falls outside X bar plus or minus 2 S square is identified as an outlier.

If you assume the normality of the F distribution in that case X bar minus 2, this interval which is X bar minus 2 S square less than X less than X bar plus 2 S square is approximately 98 percent. However, before identifying anything as an outlier, it is advised that you better measure the value again to make sure, whether it is an outlier or not? It may be a measurement error or we have to realise that when you are experimenting and truly a very different value pops up.

Very different value comes up may be you have to sit back and think whether it really indicate some new phenomena or some new theory which you have left out. So, it is not a good practice to just ignore or throw away the outlier data and consider only the good consistent data. It is the good idea to know that these are the outliers in the data and it is very important to report what you are going to do with the outlier when you do the analysis of the data.

Now, I come to the definition of robust estimator why I am talking about outliers.

(Refer Slide Time: 22:50)



Robust Estimator

- If the estimator $d(X_1, X_2, \ldots, X_n)$ is not affected by outliers then it is called a robust estimator.
- Let $X_1, X_2, \ldots, X_n$ be a random sample from distribution function F.
- Also assume that it takes on values $x_1 < x_2 < \cdots < x_{n/2} < \ldots < x_n$
- Mean of F can also be estimated as median of sample.
- Let us suppose that n is odd, then median estimator of μ = $x_{n/2}$
- This is also a point estimator of population mean μ and it is a robust as median is not affected by extreme values
- As no assumption has been made on the parameters of F, median estimator is called non-parametric estimator of population mean

An estimator d is called robust, if it is not affected by any outliers. That is the value of estimator d does not change significantly if there are extreme values in the data. You must be thinking we have heard this before we have when we were talking about measures of central tendency we said that the average, arithmetic average is sensitive to the extreme value, while the median of the data is not sensitive to the extreme value and this is exactly what we want to say here.

So, if we say that we have a sample then the mean of the mean value of the population can also be estimated by a median of the sample. You remember what is the median? If you have odd number of data points then it is exactly the middle value. It is xn by 2 which is the median value. If you have even number then you have to take the two middle values and the average of the two gives you the median value.

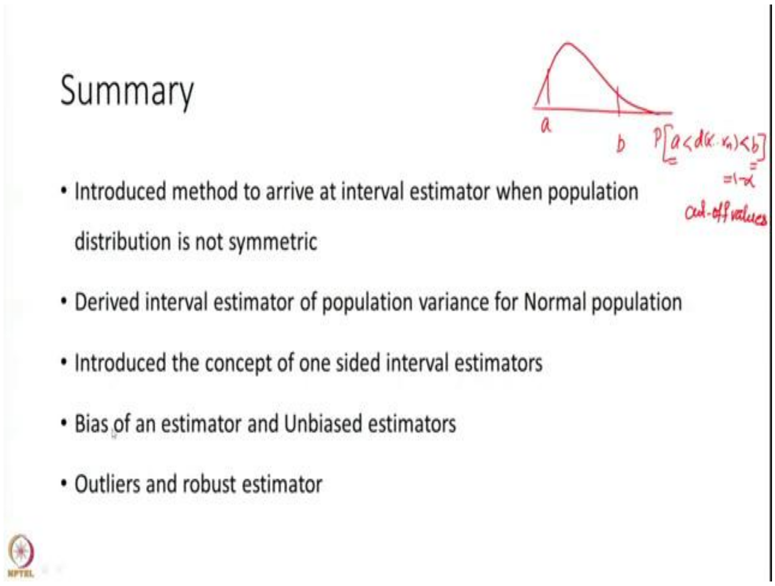
So, if x1, x2, x3 you have n data points and you order them from larger from smaller to the largest. In that case the middle value gives you median and that is also an estimator of mean value of the population. And as we know that median is robust against affected value and therefore median is a robust estimator of sample I am sorry population mean.

While average is not sample mean is not a robust estimator of the population mean. So, this is how the robust estimator is defined. Remember that when you find a median of the population, you are not going to make any assumption on the shape of the population distribution. And therefore median also represents what is known as a non-parametric estimator.

This is just for your information that non-parametric estimators are estimators which are based on the ordered value of sample they are based on the ordered values of sample. So, what we want to like to say here is that the mean of the population can also be estimated using the median of the sample. And median of the sample is a robust estimator because it is not affected by the extreme values or the outliers of the data.

While mean is so though the sample mean is an unbiased estimator of the population mean, it is not robust. Robust estimator of the population mean is median.

(Refer Slide Time: 26:45)



$P[\underline{a} < d(x_1 \ldots x_n) < \underline{\underline{b}}] = 1-\alpha$

cut-off values

Summary

- Introduced method to arrive at interval estimator when population distribution is not symmetric
- Derived interval estimator of population variance for Normal population
- Introduced the concept of one sided interval estimators
- Bias of an estimator and Unbiased estimators
- Outliers and robust estimator

So, let us summarize, we introduced methods to arrive it interval estimator when population distribution is not symmetric. We try to explain it in a very general terms but we immediately arrive and gave specific example of chi square distribution.

We also explain how to find the, what is also known as cut-off value, I did not use this term. But, whenever you find the values let me write down here. When you are trying to find a value a and b such that probability of a less than d , x1, x2, x3, xn is less than b is 1 minus alpha. Then a and b are called cut-off values. They are also called cut-off values and we look for it in the respective distribution tables.

So, we showed how to look for a and b and how to calculate this a and b in the table. So, we derive the interval estimator for a population variance for normal population which is derived from the chi square distribution. Because the sample variance is taken as a statistic and from there the interval estimator of population variance is derived. We introduced the concept of one sided interval estimator.

It would be smear repetition of what we have done for two sided intervals. So we have not repeated it here. We talked what is called bias of an estimator and what is known as an unbiased estimator, we found, we stated that sample mean is an unbiased estimator of a population mean. Sample variance is a unbiased estimator of a population variance. But, we said that maximum likelihood estimator of population variance under normality is not an unbiased estimator of normal population.

Unbiased estimator of variance of the normal population we discussed briefly about outliers, we emphasize that outliers are the extreme values in the data. They should be noted, they should be studied, they should not be unanimous decision every time remove the outliers, no. Because at times outliers represent the measurement errors, so we may have to conduct the experiment again to make sure that this is actually the value we are getting.

Or it may be an indicator of a new theory or a new phenomenon which we had never expected out of the experiment. After brief introduction to the outliers we introduced, what is known as robust estimator. Robust estimator are those which are not affected by outliers. And we showed that sample mean is not a robust estimator of a population mean though it is unbiased but it is not a robust estimator.

Median is a robust estimator of a population mean because it does not get affected by outlier as we discussed in the very beginning of descriptive statistics lectures. And we also said that median is a kind of a non-parametric estimator of population mean. Thank you.