Essentials of Data Science with R Software – 1

Professor. Shalabh
Department of Mathematics & Statistics
Indian Institute of Technology Kanpur

Lecture No. 59

Consistency and Sufficiency of Estimators

Hello friends welcome to the course Essentials of Data Science with R Software – I, in which we are trying to understand the basic concepts related to the probability theory and statistical inference. So, if you recall that in the last two lectures, we had talked about two properties of the estimator, they are unbiasedness of an estimator and efficiency of estimator. Now, in this lecture, we are going to talk about two more properties. One is the consistency and another is the sufficiency.

So, what is the meaning of consistency and sufficiency? For example, if you say what do you really mean by consistent. Consistent means, if you try to repeat something, then the same thing is happening again and again, and the number of times that things are not happening that is very, very less. So, you can say that the person is consistent.

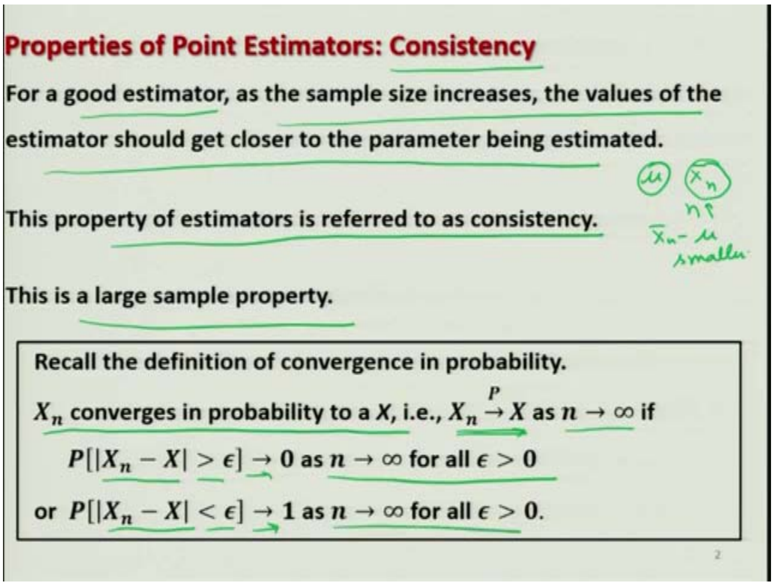
For example, if you take a help from your from one of your friend and every time the friend is helping you will say that okay the friend is dependable and different is consistent in helping the people, and similarly, if suppose you have a couple of friends and there is one more friend who is very dependable, and sometimes you make a statement that okay with having a friend like you is more than sufficient in my life. What does this mean?

That means, you can always depend on him. And whenever in your life whatever that needs, you will have which had to be fulfilled by your friend, you do not need many friends, but only one friend is sufficient. So, the similar is the story with the estimation also that whenever we are trying to estimate an unknown parameter by statistics which is a function of random variables that will be dependent on sample, so, every time you generate a sample the values are going to be changed.

So, now, what is really happening with the estimator when the sample size becomes larger? That is the property, which is related to the consistency. And sufficiency is the property of the estimator, which is trying to explain that how much the estimator is capable of explaining the information which is contained inside the statistic.

So, let us try to understand now this concept in more detail. But before that, do you recall or do you remember the definition of convergence in probability? I'm sure, yes. And if you do not remember it, I will say pause the video, first revise the concept of convergence in probability because we are going to use here, and consistency is simply that convergence in probability. So, with this aim, we try to begin our lecture now.

(Refer Slide Time: 03:09)



Properties of Point Estimators: Consistency

For a good estimator, as the sample size increases, the values of the estimator should get closer to the parameter being estimated.

μ X_n

$n \uparrow$

$\bar{X}_n - \mu$ smaller

This property of estimators is referred to as consistency.

This is a large sample property.

Recall the definition of convergence in probability.

X_n converges in probability to a X, i.e., $X_n \xrightarrow{P} X$ as $n \to \infty$ if

$P[|X_n - X| > \epsilon] \to 0$ as $n \to \infty$ for all $\epsilon > 0$

or $P[|X_n - X| < \epsilon] \to 1$ as $n \to \infty$ for all $\epsilon > 0$.

2

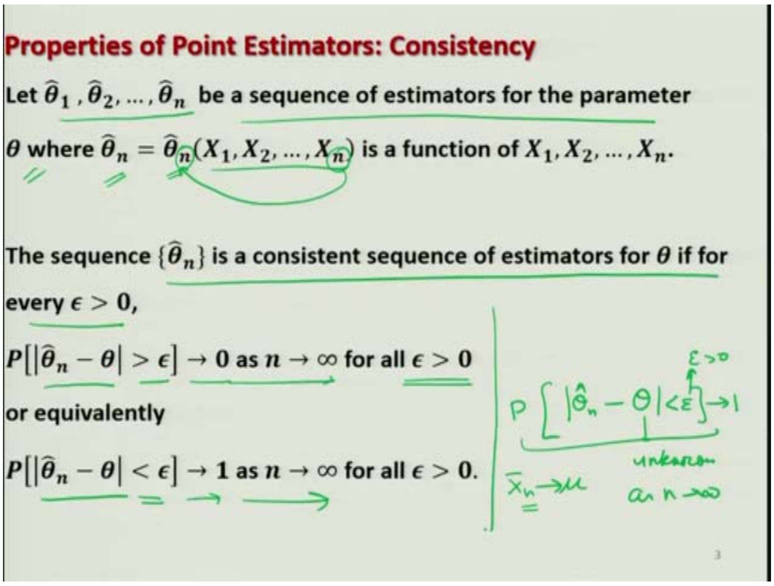
So, now, you see. First, we try to understand what is the property of consistency of an estimator. So, we expect that in case an estimator is good, then as the sample size increases, the values of the estimator should get closer to the parameter being estimated. That means if you try to estimate say for example, the population mean μ by say sample means say $\bar{X}$, which is dependent on the sample size n. So, as n is increasing, then this value will become more dependable, more dependable in the sense, the difference between say this $\bar{X}_n$ and μ will become say smaller.

Now, if you try to recall, have you done this type of property in the past? Think about it. This property that as the sample size is increasing, and the values of the estimator are getting closer to the parameter being estimated is referred to as consistency, and this is a large sample properties, it means, you cannot establish or you cannot talk about the consistency in case if your sample size is small.

And meaning of the consistency of an estimator is valid only when the sample size is large. So, do you remember the definition of convergence in probability? This was like X_n convergence in probability to a X that is X_n is tending toward X as n goes to infinity, if $P[|X_n - X| > \epsilon] \to 0$ as $n \to \infty$ for all $\epsilon > 0$

or $P[|X_n - X| < \epsilon] \to 1$ as $n \to \infty$ for all $\epsilon > 0$. I will explain you the meaning of this expression earlier.

(Refer Slide Time: 05:03)



Properties of Point Estimators: Consistency

Let $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_n$ be a sequence of estimators for the parameter θ where $\hat{\theta}_n = \hat{\theta}_n(X_1, X_2, \dots, X_n)$ is a function of $X_1, X_2, \dots, X_n$.

The sequence $\{\hat{\theta}_n\}$ is a consistent sequence of estimators for θ if for every $\epsilon > 0$,

$P[|\hat{\theta}_n - \theta| > \epsilon] \to 0$ as $n \to \infty$ for all $\epsilon > 0$

or equivalently

$P[|\hat{\theta}_n - \theta| < \epsilon] \to 1$ as $n \to \infty$ for all $\epsilon > 0$.

$\varepsilon > 0$

$\underbrace{P\left[|\hat{\theta}_n - \theta| < \varepsilon\right] \to 1}_{\text{unknown}}$

$\bar{X}_n \to \mu$

as $n \to \infty$

Now, you try to see. Now we try to define that how can you define the consistency of an estimator. So, let this $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_n$ hat be a sequence of estimator for the parameter θ, where they say $\hat{\theta}_n$ mean $\hat{\theta}$ is a function of X_1, X2,…, X_n where this sample size is indicated here as n in the subscript. So that is indicating that $\hat{\theta}$ is a function of n.

So, now, we say that the sequence $\hat{\theta}_n$ is a consistent sequence of estimator for θ if

$P[|\hat{\theta}_n - \theta| > \epsilon] \to 0$ as $n \to \infty$ for all $\epsilon > 0$, and this is very valid for all values of

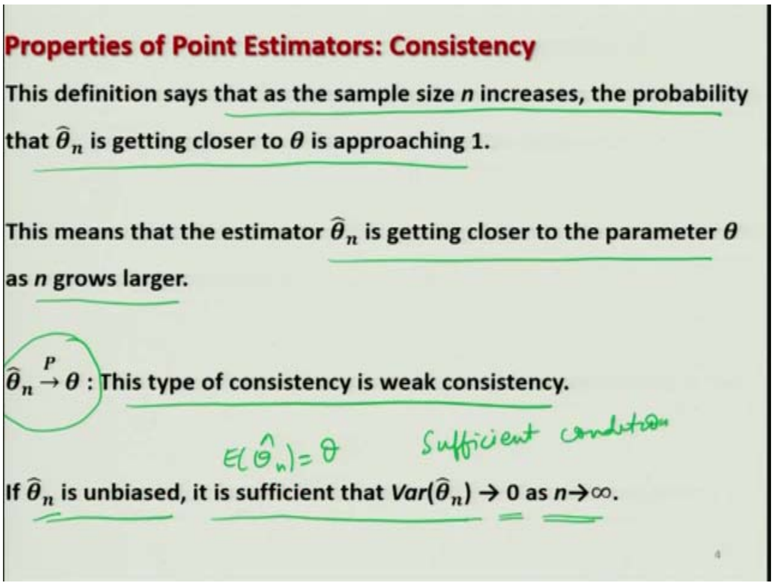
$P[|\hat{\theta}_n - \theta| < \epsilon] \to 1$ as $n \to \infty$ for all $\epsilon > 0$.

And you see that is expected. You are trying to estimate an unknown parameter, θ here is unknown, and you are trying to say that you are trying to estimate it by here estimator $\hat{\theta}_n$. So, now, what are you trying to say this estimator is going to be good if the difference between $\hat{\theta}_n$ minus θ is as small as possible. So, you are trying to say that in case if I try to take the deviations this can occur in the positive and negative direction both.

So, we try to consider here the absolute value and probability that the difference between $\hat{\theta}_n$ and θ is less than epsilon. The probability of such an event is tending to 1 as n goes to infinity where this epsilon is greater than 0. So, what are you trying to say that as we are trying to increase the sample size that means, in case if I take an example of sample mean $\bar{X}_n$ to estimate μ, so, as you are trying to take the sample size from say 5 to 10, to 100, 200, 5000 etc. this value of $\bar{X}_n$ is increasing then the value of $\bar{X}_n$ is getting close to μ. This is what we mean.

Now, if you try to compare this definition of consistency with this definition of probability, that is that convergence in probability, then, what do you see that they are the same thing actually.

(Refer Slide Time: 07:42)



Properties of Point Estimators: Consistency

This definition says that as the sample size n increases, the probability that $\hat{\theta}_n$ is getting closer to θ is approaching 1.

This means that the estimator $\hat{\theta}_n$ is getting closer to the parameter θ as n grows larger.

$\hat{\theta}_n \xrightarrow{P} \theta$: This type of consistency is weak consistency.

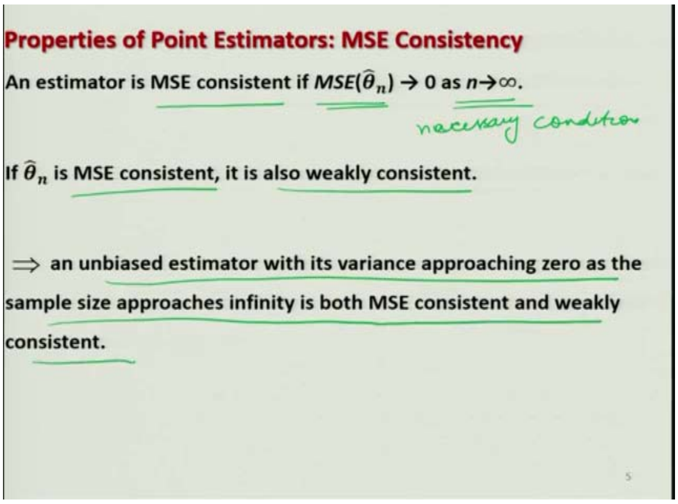
$E(\hat{\theta}_n) = \theta$ Sufficient condition

If $\hat{\theta}_n$ is unbiased, it is sufficient that $Var(\hat{\theta}_n) \to 0$ as $n \to \infty$.

And this definition is simply saying that as the sample size increases, the probability that $\hat{\theta}_n$ is getting closer to θ is approaching 1. So, this means that $\hat{\theta}_n$ is getting closer to the parameter θ as the sample size n grows larger. And we can indicate the same thing by saying that $\hat{\theta}_n$ is converging to θ in probability. If you try to compare this statement and here this statement, you simply have to replace here X and Y $\hat{\theta}_n$ and X by θ and you get here the conversion in probability.

So, I can say here that $\hat{\theta}_n$ is converging to θ in probability and this type of consistency is called as weak consistency. Now, in case if you try to assume that the $\hat{\theta}_n$ is an unbiased estimator of θ that is expected value of $\hat{\theta}_n$ is equal to θ then it is sufficient to show that if the variance of $\hat{\theta}_n$ is tending to 0 as n goes to infinity, then the estimator is consistent, but remember one thing this is only a sufficient condition.

(Refer Slide Time: 09:03)



Properties of Point Estimators: MSE Consistency

An estimator is MSE consistent if $MSE(\hat{\theta}_n) \to 0$ as $n \to \infty$.

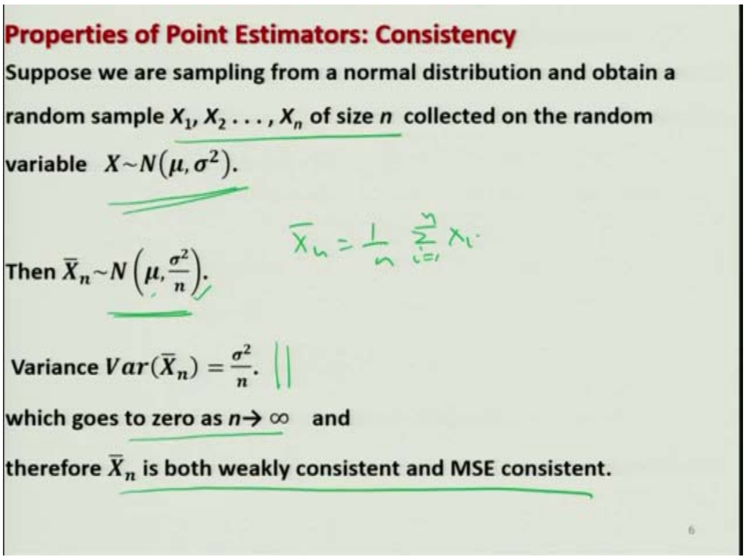
necessary condition

If $\hat{\theta}_n$ is MSE consistent, it is also weakly consistent.

$\Rightarrow$ an unbiased estimator with its variance approaching zero as the sample size approaches infinity is both MSE consistent and weakly consistent.

So, exactly on the same lines, as we have defined the consistency condition in terms of variance of the estimator. Similarly, we can define the MSE consistency, that means, an estimator is MSE consistent if MSE of $\hat{\theta}_n$ goes to 0 as n goes to infinity. And this is also a nice hurry condition for the consistency. So, in case $\hat{\theta}_n$ is MSE consistent, then it is also weakly consistent. So this implies that an unbiased estimator with its variance approaching 0 as the sample size approaches infinity is both MSE consistent and weakly consistent. That is what you have to just keep in mind.

(Refer Slide Time: 09:49)



Properties of Point Estimators: Consistency

Suppose we are sampling from a normal distribution and obtain a random sample $X_1, X_2 \ldots, X_n$ of size n collected on the random variable $X \sim N(\mu, \sigma^2)$.

Then $\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$. ✓

$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} x_i$.

Variance $Var(\bar{X}_n) = \frac{\sigma^2}{n}$.

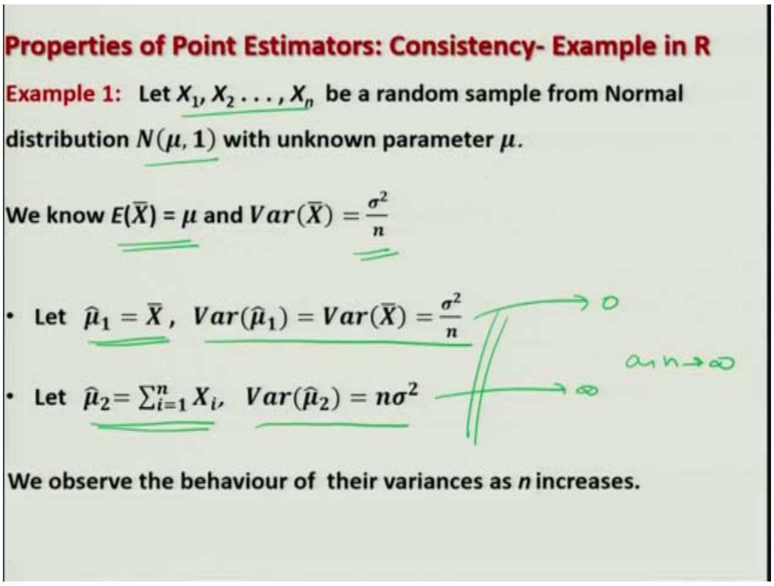
which goes to zero as $n \to \infty$ and

therefore $\bar{X}_n$ is both weakly consistent and MSE consistent.

Now let me try to take a very simple example to show you that how the sufficient condition helps us in finding or determining whether the estimator is consistent or not. So, suppose we have a sample $X_1, X_2, \ldots, X_n$ of size small n from N(μ , σ^2), then we know that the sample mean $\overline{X}_n$ which is like $\sum_{i=1}^{n} X_i$/n, this follows a normal distribution with mean μ and variance σ^2 /n.

So if you try to see here that variance of $\overline{X}_n$ is σ^2 /n which goes to 0 as n goes to infinity, and therefore, this $\overline{X}_n$ is both weakly consistent and MSE consistent. And similarly, you can now find the variance of any statistics and can check this condition and can find. But ideally you must take the convergence in probability of that estimator.

(Refer Slide Time: 10:47)



Properties of Point Estimators: Consistency- Example in R

Example 1: Let $X_1, X_2 \ldots, X_n$ be a random sample from Normal distribution $N(\mu, 1)$ with unknown parameter μ.

We know $E(\bar{X}) = \mu$ and $Var(\bar{X}) = \frac{\sigma^2}{n}$

- Let $\hat{\mu}_1 = \bar{X}$, $Var(\hat{\mu}_1) = Var(\bar{X}) = \frac{\sigma^2}{n} \rightarrow 0$
- Let $\hat{\mu}_2 = \sum_{i=1}^{n} X_i$, $Var(\hat{\mu}_2) = n\sigma^2 \rightarrow \infty$

as $n \rightarrow \infty$

We observe the behaviour of their variances as n increases.

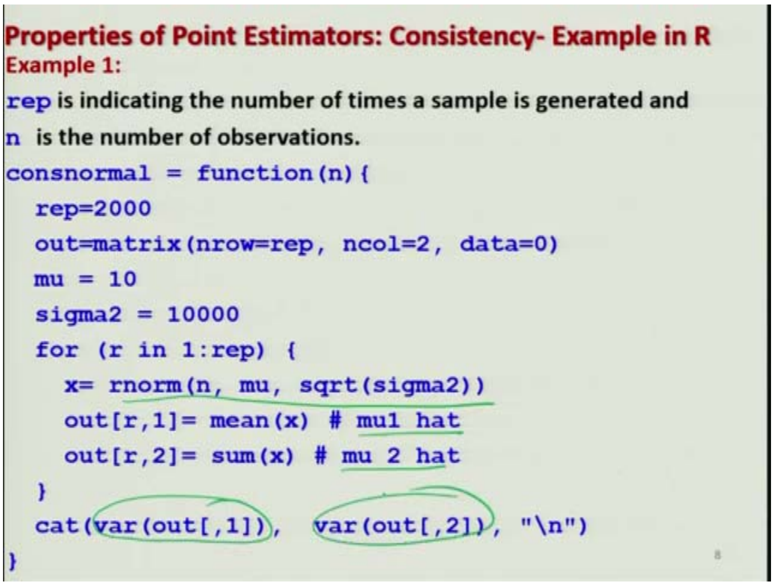
Now, let me try to take an example here to illustrate that how this consistency condition will appear when you try to look it from the data science point of view. And if you try to execute it on the R console or if you want to get remind whether a particular estimator is consistent or not and suppose it is quite complicated to get it from the theory. So, one can also look into the simulated results and possibly can get an indication whether the estimator is consistent or not.

So, let me try to take an example where we have observed a sample from a normal population with mean μ and variance 1 and here we try to see whether what is happening to the consistency property. So, suppose, we decide that we want to estimate μ and we also know that E($\bar{X}$)= μ and its variance is σ^2/n.

So, suppose we propose to estimate μ by two possible estimators $\hat{\mu}_1$ which is equal to sample mean $\bar{X}$ and $\hat{\mu}_2$ which is simply the summation of all the observations in the sample. Now in case if you will try to find out variants of $\hat{\mu}_1$, this is σ^2/n and variance of $\hat{\mu}_2$ that will be n times σ^2 because all the observations are IID identically and independently distributed.

Although, we know from here that what is really going to happen that as n goes to infinity, this value will go to 0 and this value will go to infinity as n goes to infinity, but we want to see how this behavior will look like when we are trying to execute it on the R console.

(Refer Slide Time: 12:33)



Properties of Point Estimators: Consistency- Example in R

Example 1:

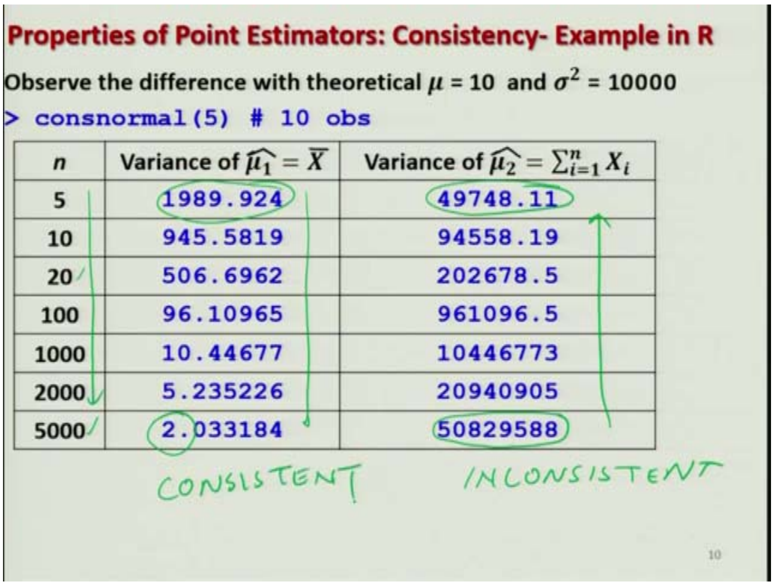
`rep` is indicating the number of times a sample is generated and

`n` is the number of observations.

```
consnormal = function(n){
  rep=2000
  out=matrix(nrow=rep, ncol=2, data=0)
  mu = 10
  sigma2 = 10000
  for (r in 1:rep) {
    x= rnorm(n, mu, sqrt(sigma2))
    out[r,1]= mean(x) # mu1 hat
    out[r,2]= sum(x) # mu 2 hat
  }
  cat(var(out[,1]), var(out[,2]), "\n")
}
```

So now, we have written here program in which I am trying to generate the observation from the normal distribution. And then I am trying to compute the value of $\hat{\mu}_1$ and $\hat{\mu}_2$ and then I am trying to find out the variance of $\hat{\mu}_1$ and $\hat{\mu}_2$.

(Refer Slide Time: 12:52)



Properties of Point Estimators: Consistency- Example in R

Observe the difference with theoretical $\mu = 10$ and $\sigma^2 = 10000$

```
> consnormal(5) # 10 obs
```

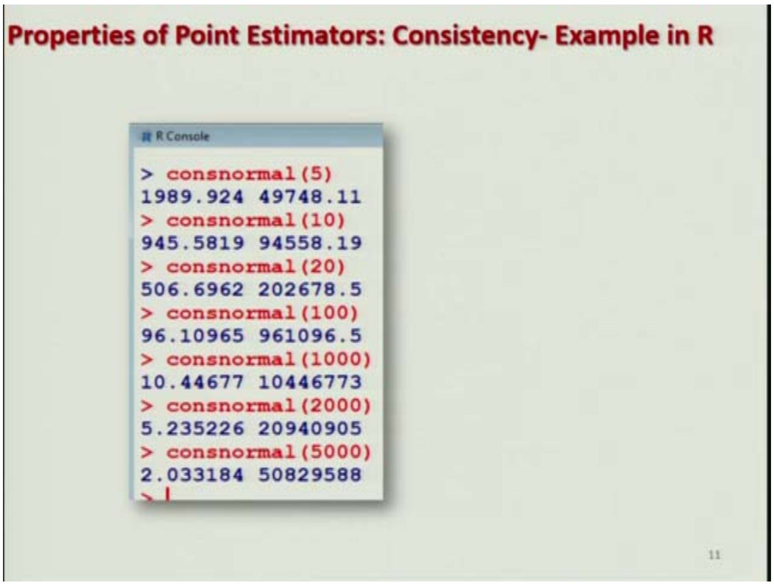
n	Variance of $\widehat{\mu_1} = \overline{X}$	Variance of $\widehat{\mu_2} = \sum_{i=1}^{n} X_i$
5	1989.924	49748.11
10	945.5819	94558.19
20	506.6962	202678.5
100	96.10965	961096.5
1000	10.44677	10446773
2000	5.235226	20940905
5000	2.033184	50829588

CONSISTENT INCONSISTENT

So, if you try to see this screenshot here. And now, if you try to see these are the values of the variances of $\hat{\mu}_1$ and $\hat{\mu}_2$, and you can see here that as the values of n is increasing from 5, 10, 20 to 5000 the value of the variance is decreasing here. That starts from 1989 and it is decreasing to 2, but, in the case of $\hat{\mu}_2$ the variance start from 49748 and it increases and it is increasing to a very large value. So, this is actually not increasing, but this is decreasing.

So, you can see here that from the siμlation results, it is clear that the variance of $\hat{\mu}_1$ is decreasing as n is increasing and the variance of $\hat{\mu}_2$ is increasing as n is increasing. So, I can be confident that at least this is not consistent. So, this is inconsistent and this is consistent. At least this is giving us a good confidence. Well, I will simply say that if you want to be 100 percent sure, you have to go through with the result from that theory.

(Refer Slide Time: 14:02)

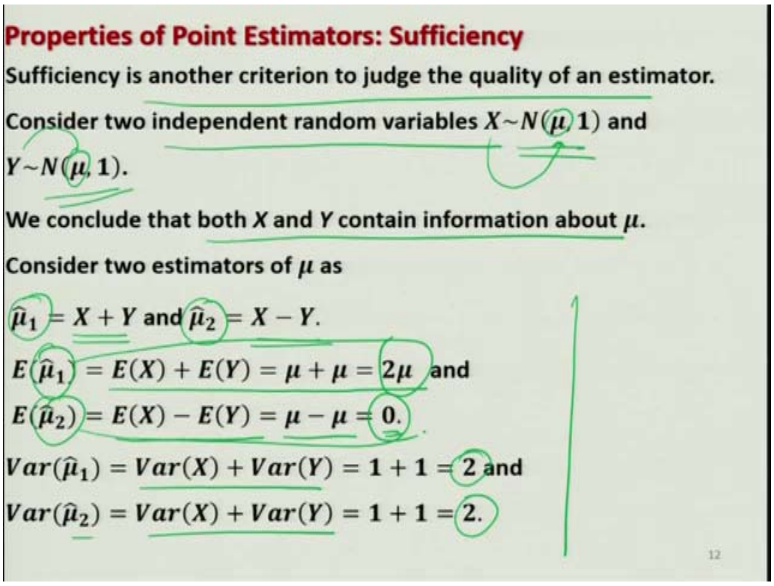


Properties of Point Estimators: Consistency- Example in R

```
> consnormal(5)
1989.924 49748.11
> consnormal(10)
945.5819 94558.19
> consnormal(20)
506.6962 202678.5
> consnormal(100)
96.10965 961096.5
> consnormal(1000)
10.44677 10446773
> consnormal(2000)
5.235226 20940905
> consnormal(5000)
2.033184 50829588
> |
```

So, this is how you can look for the consistency property on the basis of simulated values in data science.

(Refer Slide Time: 14:11)



Properties of Point Estimators: Sufficiency

Sufficiency is another criterion to judge the quality of an estimator.

Consider two independent random variables $X \sim N(\mu, 1)$ and $Y \sim N(\mu, 1)$.

We conclude that both X and Y contain information about μ.

Consider two estimators of μ as

$\hat{\mu}_1 = X + Y$ and $\hat{\mu}_2 = X - Y$.

$E(\hat{\mu}_1) = E(X) + E(Y) = \mu + \mu = 2\mu$ and

$E(\hat{\mu}_2) = E(X) - E(Y) = \mu - \mu = 0$.

$Var(\hat{\mu}_1) = Var(X) + Var(Y) = 1 + 1 = 2$ and

$Var(\hat{\mu}_2) = Var(X) + Var(Y) = 1 + 1 = 2$.

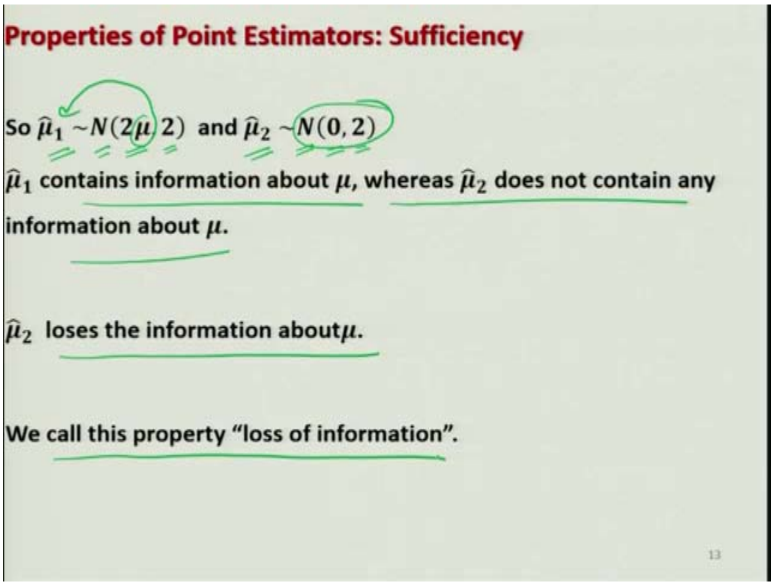
So, now, I come to another topic, which is about the sufficiency of the estimator. So, sufficiency is another criterion to judge the quality of an estimator. Now, in order to understand what is the sufficiency let me try to take a very simple example, and you please try to concentrate on the values, which I am going to obtain. So, we consider here two independent random variables X and Y, what are following the same distribution normal μ, 1.

So, you can see here that the random variable X has information about μ, because μ is there and Y also has information about μ, both variances they are actually known as 1 in X and Y. So, we can conclude by looking at the structure of action why that both X and Y contain information about μ. Now we want to estimate μ, so we try to consider here two possible estimators and I try to compute say X plus Y as one possible estimator indicator as $\hat{\mu}_1$ and X minus Y as another estimator to estimate μ, and it is indicated by $\hat{\mu}_2$.

Now, if you try to see, the expected value of $\hat{\mu}_1$ and $\hat{\mu}_2$, it comes out to behave like this. For $\hat{\mu}_1$ the expected value is twice of μ and for the $\hat{\mu}_2$ this is expected value of X minus expected value of Y which is μ minus μ is equal to 0. And if you try to find out their variances, so variances of $\hat{\mu}_1$ and $\hat{\mu}_2$ they are the same that will become variance of X plus variance of Y in both the cases and the covariance terms will become 0 because, because X and Y are independent and the variances of $\hat{\mu}_1$ and $\hat{\mu}_2$ comes out to be the same as 2.

Now you try to observe what is happening in these values. Here this $\hat{\mu}_1$ contains the value of μ, but now, this $\hat{\mu}_2$ is not containing the value 0 when you are trying to take the expectation.

(Refer Slide Time: 16:16)



Properties of Point Estimators: Sufficiency

So $\hat{\mu}_1 \sim N(2\mu, 2)$ and $\hat{\mu}_2 \sim N(0, 2)$

$\hat{\mu}_1$ contains information about μ, whereas $\hat{\mu}_2$ does not contain any information about μ.

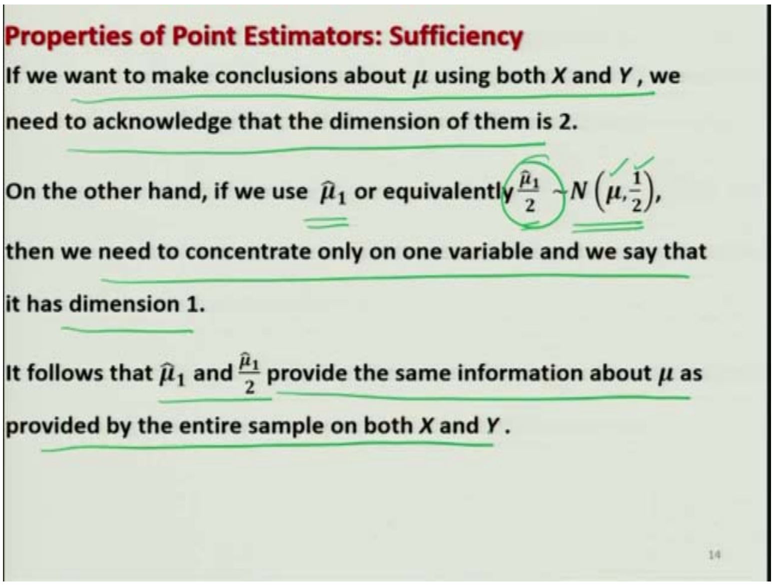
$\hat{\mu}_2$ loses the information about μ.

We call this property "loss of information".

So, now, I can say here that $\hat{\mu}_1$ is following a normal distribution with the mean twice of μ and variance 2 whereas, $\hat{\mu}_2$ is following a normal distribution with mean 0 and variance 2. So, now, you can see here you want to estimate the value of μ. And this $\hat{\mu}_1$ contains the value of μ but whereas, there is no μ involved in normal 0, 2. So, $\hat{\mu}_1$ contained information about μ, whereas, $\hat{\mu}_2$ does not contain any information about μ.

So, we can see here that $\hat{\mu}_2$ loses the information about μ. And we call this property as loss of information. So, you can see here that this x and y both had the information about μ, but with this $\hat{\mu}_2$ there is no information in the distribution, which is related to μ.

(Refer Slide Time: 17:13)



Properties of Point Estimators: Sufficiency

If we want to make conclusions about μ using both X and Y, we need to acknowledge that the dimension of them is 2.

On the other hand, if we use $\hat{\mu}_1$ or equivalently $\frac{\hat{\mu}_1}{2} \sim N\left(\mu, \frac{1}{2}\right)$,

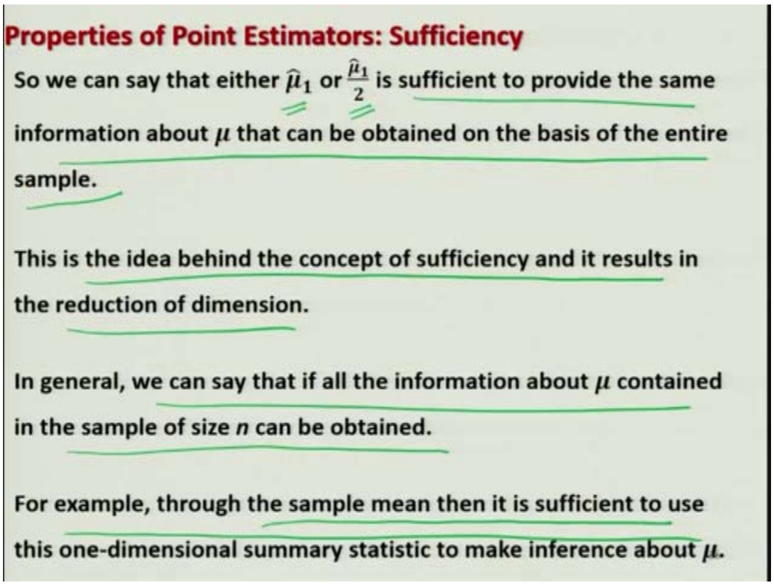
then we need to concentrate only on one variable and we say that it has dimension 1.

It follows that $\hat{\mu}_1$ and $\frac{\hat{\mu}_1}{2}$ provide the same information about μ as provided by the entire sample on both X and Y.

So, now, what we need to do? In case if you try to look at this structure, we can conclude that if we want to make conclusion about μ is using both X and Y, we need to acknowledge that the dimension of them is 2 because there are two variables. On the other hand, in case if you try to use $\hat{\mu}_1$ or say equivalently if you try to see $\hat{\mu}_1/2$ will also have a normal distribution with mean μ and variance and n/2. In this case, we need to concentrate only on one variable and we say that it has dimension 1.

We are considering only on $\hat{\mu}_1$ upon two. So, now, looking at both the results, we can conclude that $\hat{\mu}_1$ an $\hat{\mu}_1$ divided by 2 provide the same information about μ as provided by the entire sample on both X and Y.

(Refer Slide Time: 18:19)



Properties of Point Estimators: Sufficiency

So we can say that either $\hat{\mu}_1$ or $\frac{\hat{\mu}_1}{2}$ is sufficient to provide the same information about μ that can be obtained on the basis of the entire sample.

This is the idea behind the concept of sufficiency and it results in the reduction of dimension.

In general, we can say that if all the information about μ contained in the sample of size n can be obtained.

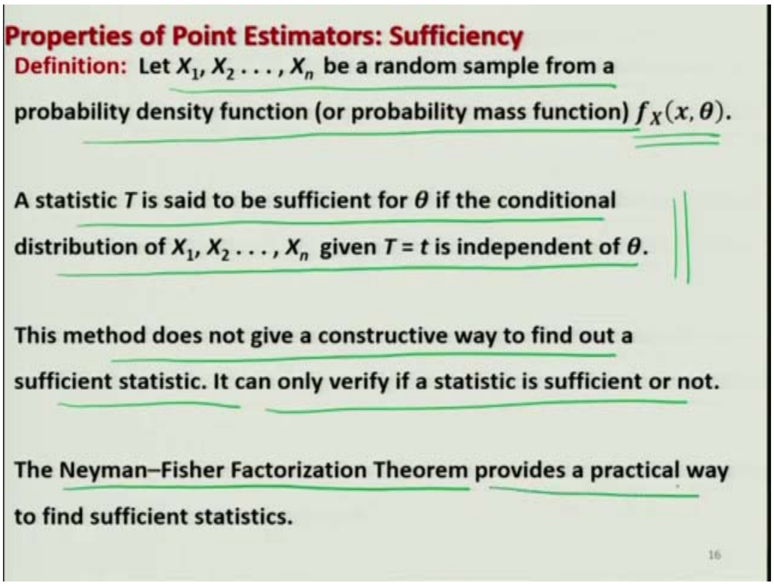
For example, through the sample mean then it is sufficient to use this one-dimensional summary statistic to make inference about μ.

So, we can say that either $\hat{\mu}_1$ or $\hat{\mu}_1$ by 2 is sufficient to provide the same information about μ that can be obtained on the basis of entire sample. You do not need the entire sample, but you simply can use here these estimator and they will give you the same information about μ that you can obtain if you try to use the entire sample. This is the idea behind the concept of sufficiency, and it helps us in reducing the dimensions.

So, in general, we can say that if all the information about μ contained in a sample of size n can be can be obtained directly by using this estimator that means this estimator is sufficient. For example, if you try to take a sample say $X_1, X_2, \ldots, X_n$ and you try to compute the sample mean, then what will happen, the dimension of the sample is n because there are n variables, which have to be considered and each of them is consisting or having the information about μ.

But when you are trying to estimate the parameter μ by the sample mean then also expected value of $\bar{X}$ is equal to μ. And through the sample mean, it is sufficient to use only the one dimensional summary statistics to make inference about the μ. Because sample mean is only unidimensional $X_1, X_2, \ldots, X_n$ that will be n dimensional. So, you can see here you have reduced the demonstration n to 1 without losing any information.

(Refer Slide Time: 19:53)



Properties of Point Estimators: Sufficiency

Definition: Let $X_1, X_2 \ldots, X_n$ be a random sample from a probability density function (or probability mass function) $f_X(x, \theta)$.

A statistic T is said to be sufficient for θ if the conditional distribution of $X_1, X_2 \ldots, X_n$ given $T = t$ is independent of θ.

This method does not give a constructive way to find out a sufficient statistic. It can only verify if a statistic is sufficient or not.

The Neyman–Fisher Factorization Theorem provides a practical way to find sufficient statistics.

16

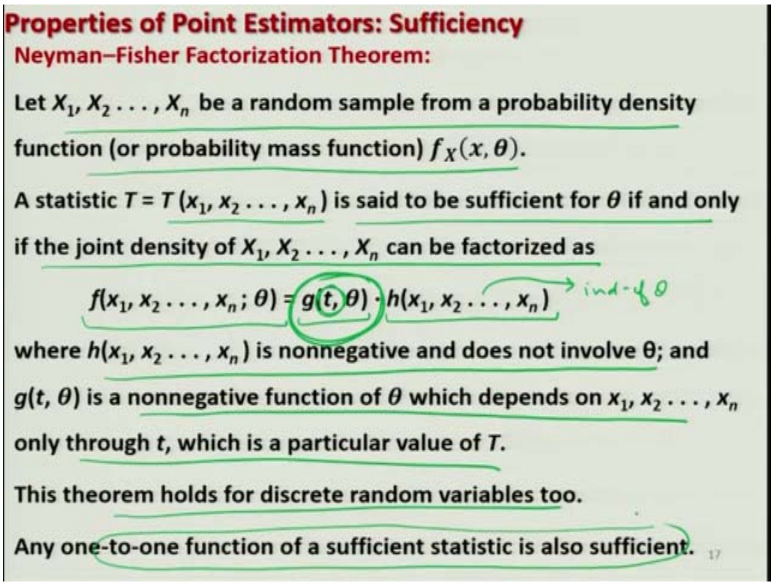
And that is why we call it that this is a statistic which is sufficient. Sufficient in the sense, we do not have to look into the entire sample, but we have to look only into the value of this statistics. This is going to give us the same information, which the complete sample can give us. Now, the question is how to find out such a statistic.

So, first I will try to give you here the definition, the formal definition of sufficient statistics and then I will try to show you that how you can find it out. The finding of sufficient statistics is not difficult at all. So, the formal definition is let $X_1, X_2, \ldots, X_n$ be a random sample from a probability density function or a probability mass function which is given by like the f (x, θ) now we try to define the sufficiency of the estimator.

A statistics T is said to be sufficient for θ if the conditional distribution of $X_1, X_2, \ldots, X_n$ given T is equal to small t is independent of the θ. Now, you understand all the terminologies here. You know the statistics, you know the conditional distribution you know the meaning of independent of θ. So, this definition is extremely simple, but finding out their statistics is more simpler.

Because this method, this definition does not give us a constructive way to find out a sufficient statistics. It can only verify that if a statistics is sufficient or not. That means, if you say that okay $\overline{X}$ is a sufficient statistics for the μ or not that can be verified using this definition. We have a theorem, which is called as Neyman-Fisher Factorization Theorem, this helps us, and this provides us a practical way to find sufficient statistics.

(Refer Slide Time: 21:41)



Properties of Point Estimators: Sufficiency

Neyman–Fisher Factorization Theorem:

Let $X_1, X_2 \ldots, X_n$ be a random sample from a probability density function (or probability mass function) $f_X(x, \theta)$.

A statistic $T = T(x_1, x_2 \ldots, x_n)$ is said to be sufficient for θ if and only if the joint density of $X_1, X_2 \ldots, X_n$ can be factorized as

$$f(x_1, x_2 \ldots, x_n; \theta) = g(t, \theta) \cdot h(x_1, x_2 \ldots, x_n)$$

ind. of θ

where $h(x_1, x_2 \ldots, x_n)$ is nonnegative and does not involve θ; and $g(t, \theta)$ is a nonnegative function of θ which depends on $x_1, x_2 \ldots, x_n$ only through t, which is a particular value of T.

This theorem holds for discrete random variables too.

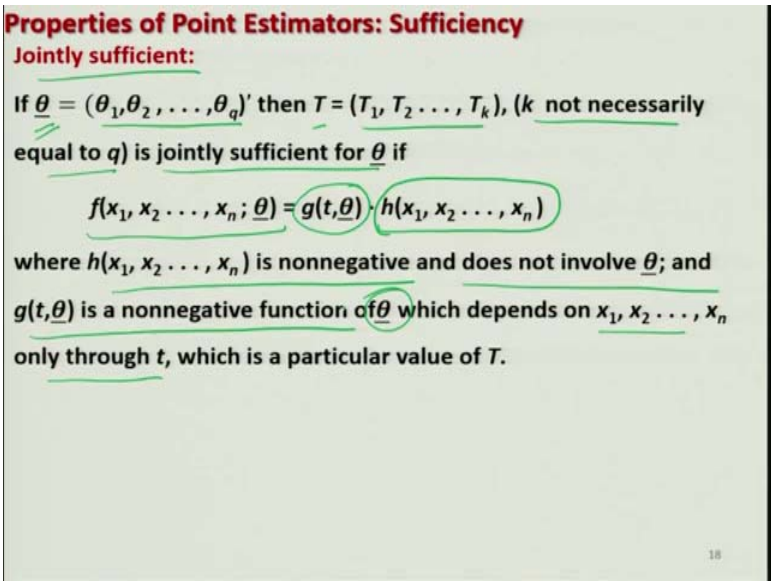
Any one-to-one function of a sufficient statistic is also sufficient.

What is this let us try to understand the Neyman-Fisher Factorization Theorem. So, the same assumption that let $X_1, X_2,\ldots, X_n$ be a random sample from a PDF or probability mass function f x θ now, a statistics T which is a function of $X_1, X_2,\ldots, X_n$ is said to be sufficient for θ if and only if the joint density of $X_1, X_2,\ldots, X_n$ can be factorized like this. This is here, the joint density function of $X_1, X_2,\ldots, X_n$ and θ now, this is being factorized into two parts. Say $g(t, \theta)$ and h of $X_1, X_2,\ldots, X_n$.

So, you can see here that this $X_1, X_2,\ldots, X_n$ this is independent of θ. So, we say that where this h $X_1, X_2,\ldots, X_n$ is a function which is non-negative and it does not involve the θ. And this quantity here $g(t, \theta)$, this is also a non-negative function of θ, which depends on the $X_1, X_2,\ldots, X_n$, but that is depending through only T, and a small t is a value of capital T.

So, in a very simple language you have to partition the joint density function into two parts such that one part is a function of only the observations and another part is a function of statistics and the parameter. And this theorem holds for the discrete random variables also. And after this, you have to just remember one result that any one-to-one function of a sufficient statistic is also sufficient. So, if you can find a statistic, which is sufficient and if you try to consider any other function, which is a one-to-one function of the statistics that will also be sufficient.

(Refer Slide Time: 23:26)



Properties of Point Estimators: Sufficiency

Jointly sufficient:

If $\underline{\theta} = (\theta_1, \theta_2, \ldots, \theta_q)'$ then $T = (T_1, T_2 \ldots, T_k)$, ($k$ not necessarily equal to q) is jointly sufficient for $\underline{\theta}$ if

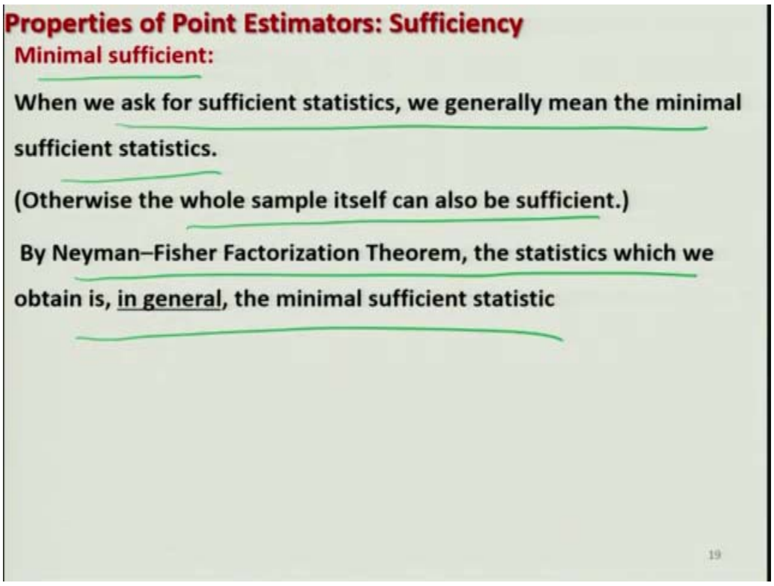
$$f(x_1, x_2 \ldots, x_n; \underline{\theta}) = g(t, \underline{\theta}) \cdot h(x_1, x_2 \ldots, x_n)$$

where $h(x_1, x_2 \ldots, x_n)$ is nonnegative and does not involve $\underline{\theta}$; and $g(t, \underline{\theta})$ is a nonnegative function of $\underline{\theta}$ which depends on $x_1, x_2 \ldots, x_n$ only through t, which is a particular value of T.

And actually this definition can be extended to a case when my parameters are in the form of a vector. So, and then in that case, we have a concept of jointly sufficient. So, for example, if I have the parameter θ_1, θ_2,…, θ_q, which are indicated by here underscore θ then the statistics T is a function of T_1, T_2,…, T_k well if k is not necessarily equal to Q, then this T is jointly sufficient for this θ if the joint density function of X_1, X_2,…, X_n and $\underline{\theta}$ can be expressed into two parts where one function or one part is only a function of X_1, X_2,…, X_n and it is independent of $\underline{\theta}$, whereas, the second part is a function of T and $\underline{\theta}$.

So, the same definition extends over here also that h X_1, X_2,…, X_n is non-negative it does not involve the parameter $\underline{\theta}$ and $g(t, \underline{\theta})$ is also non-negative and it depends on X_1, X_2,…, X_n only through T and this is also a function of $\underline{\theta}$. So, the same definition continues here.

(Refer Slide Time: 24:48)



Properties of Point Estimators: Sufficiency

Minimal sufficient:

When we ask for sufficient statistics, we generally mean the minimal sufficient statistics.

(Otherwise the whole sample itself can also be sufficient.)

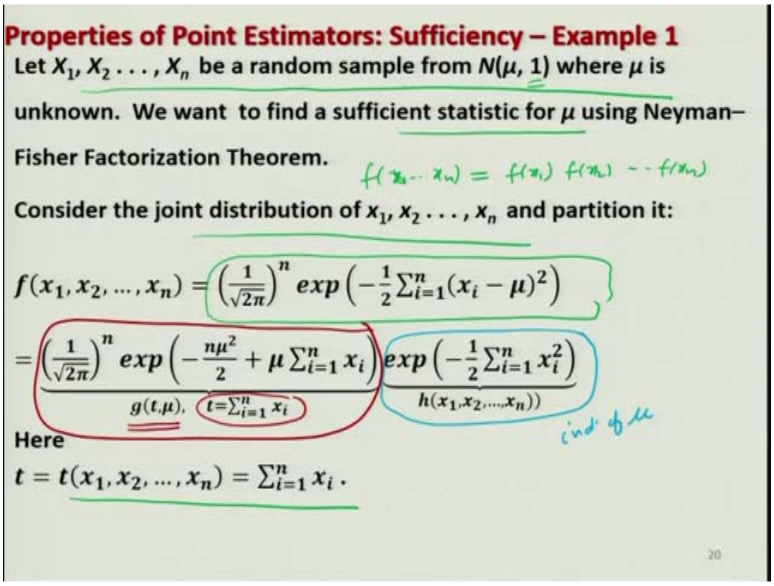
By Neyman–Fisher Factorization Theorem, the statistics which we obtain is, in general, the minimal sufficient statistic

So, and one thing there is another concept of minimal sufficient? I am not going into detail but let me just give you this idea that if suppose for a given parameter, if there more than one statistics which are sufficient, then we try to see that which of the statistics is better. So, for that we try to employ here the concept of minimal sufficiency, and we try to see that which of the statistics is trying to reduce the dimension more.

And that is statistic, which is trying to reduce the dimension to a minimum value that is called as a minimal sufficient. Well, I am not going into the details. Those who are interested they can look into the book, and this will try to give you an idea. But in general, when we ask for the sufficient statistics then we are generally trying to find out the minimal sufficient statistics.

Because, otherwise if you do not use it, then the whole sample itself will be sufficient. So, when you are trying to use the Neyman-Fisher Factorization Theorem, the statistic that you obtain is in general the minimal sufficient statistic. That is the advantage. So, you can see here this is the advantage of using Neyman-Fisher Factorization Theorem for finding outer sufficient statistics.

(Refer Slide Time: 26:06)



Properties of Point Estimators: Sufficiency – Example 1

Let $X_1, X_2, \ldots, X_n$ be a random sample from $N(\mu, 1)$ where μ is unknown. We want to find a sufficient statistic for μ using Neyman–Fisher Factorization Theorem.

$f(x_1 \cdots x_n) = f(x_1)\, f(x_2) \cdots f(x_n)$

Consider the joint distribution of $x_1, x_2, \ldots, x_n$ and partition it:

$$f(x_1, x_2, \ldots, x_n) = \left(\frac{1}{\sqrt{2\pi}}\right)^n exp\left(-\frac{1}{2}\sum_{i=1}^n (x_i - \mu)^2\right)$$

$$= \underbrace{\left(\frac{1}{\sqrt{2\pi}}\right)^n exp\left(-\frac{n\mu^2}{2} + \mu \sum_{i=1}^n x_i\right)}_{g(t,\mu),\ t=\sum_{i=1}^n x_i} \underbrace{exp\left(-\frac{1}{2}\sum_{i=1}^n x_i^2\right)}_{h(x_1, x_2, \ldots, x_n))}$$

ind of μ

Here

$t = t(x_1, x_2, \ldots, x_n) = \sum_{i=1}^n x_i$.

Now, let me try to take an example to show how you can employ this theorem to find out a sufficient statistics. So, let X1, X2,…, Xn be a random sample from a normal population with mean μ and variance one, μ is unknown and we need to find out a sufficient statistics for μ using the Neyman-Fisher Factorization Theorem.

So, for that first we need to understand that how we are going to find out the joint distribution function of X1, X2,…, Xn which has to be partitioned into two parts. So, you can see here in this case, if you try to find out here X1, X2,…, Xn. So, since you are assuming that the sample is random, so, all X1, X2,…, Xn and are independent. So, they can be written as the product of f of X1, f of X2 up to here of Xn.

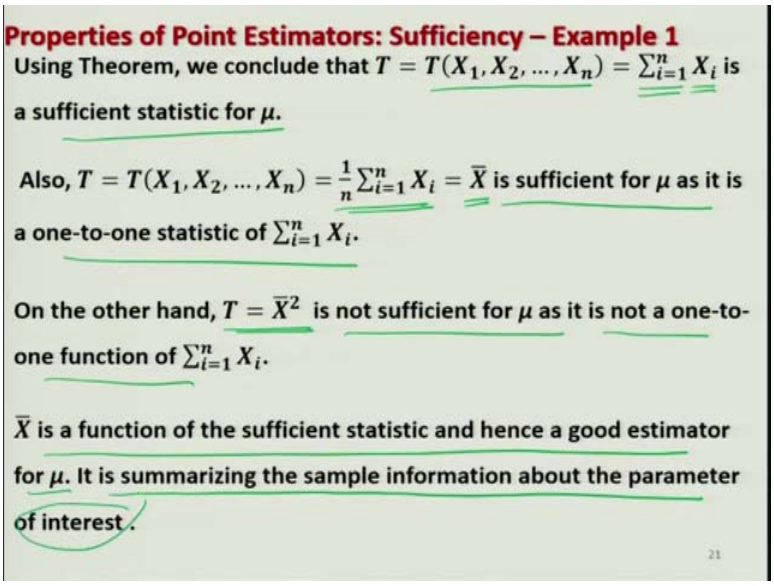
So, if you try to write down here the PDF of normal distribution then their product can be written here like this $\left(\frac{1}{\sqrt{2\pi}}\right)^n exp\left(-\frac{1}{2}\sum_{i=1}^n (x_i - \mu)^2\right)$. Now, now, after this if you try to partition it into two parts, we can operate here the following two partitions. You can see here first try to look here that you can see here this is a part which is depending only on X1, X2,…, Xn and this is independent of any parameter say here μ .

And similarly, if you try to look here in the second part, which is here like this one this is actually here trying to indicate the $g(t, \theta)$ of the Neyman-Fisher Factorization Theorem. So, in this case your θ is μ so this is actually a function which is g(t, μ) and you can see here this is depending only on the value of X1, X2,…, Xn and μ .

And you can see here, here the value of t is $\sum_{i=1}^{n} x_i$ that is the sum of all the sample values.

So, here now, I can say that in this case, we have found that T is equal to $\sum_{i=1}^{n} x_i$. So, this is only the sum of sample observation.

(Refer Slide Time: 28:23)



Properties of Point Estimators: Sufficiency – Example 1

Using Theorem, we conclude that $T = T(X_1, X_2, \dots, X_n) = \sum_{i=1}^{n} X_i$ is a sufficient statistic for μ.

Also, $T = T(X_1, X_2, \dots, X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$ is sufficient for μ as it is a one-to-one statistic of $\sum_{i=1}^{n} X_i$.

On the other hand, $T = \bar{X}^2$ is not sufficient for μ as it is not a one-to-one function of $\sum_{i=1}^{n} X_i$.

$\bar{X}$ is a function of the sufficient statistic and hence a good estimator for μ. It is summarizing the sample information about the parameter of interest.

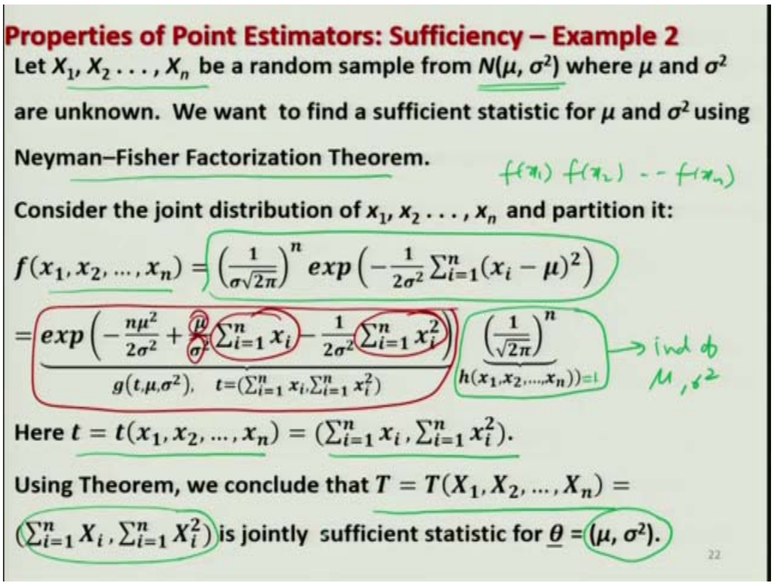
21

And based on this now, using the Neyman-Fisher Factorization Theorem, we can conclude that T is equal to t(X_1, X_2,…, X_n) is $\sum_{i=1}^{n} X_i$, this is a sufficient statistics for μ . And you see, since you have understood that any one-to-one function of a sufficient statistic is also sufficient. So, we can also consider the function summation Xi upon n which is simply your sample mean.

So, instead of writing the sample observe, the some of the sample observation, we can also say that sample mean is sufficient statistics for μ because this is a one-to-one function of summation Xi. On the other hand, if you try to see means another statistic T equal to square of that sample mean $\bar{X}^2$ is not sufficient for μ, because it is not a one-to-one function of summation Xi.

Hence, now, we can conclude that $\bar{X}$ is a function of the sufficient statistics and hence a good estimated for μ. And it is summarizing the sample information about the parameter of interest, which is here μ .

(Refer Slide Time: 29:35)



Properties of Point Estimators: Sufficiency – Example 2

Let $X_1, X_2 \dots, X_n$ be a random sample from $N(\mu, \sigma^2)$ where μ and σ^2 are unknown. We want to find a sufficient statistic for μ and σ^2 using Neyman–Fisher Factorization Theorem.

$f(x_1) f(x_2) -- f(x_n)$

Consider the joint distribution of $x_1, x_2 \dots, x_n$ and partition it:

$$f(x_1, x_2, \dots, x_n) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right)$$

$$= \underbrace{exp\left(-\frac{n\mu^2}{2\sigma^2} + \frac{\mu}{\sigma^2}\sum_{i=1}^{n} x_i - \frac{1}{2\sigma^2}\sum_{i=1}^{n} x_i^2\right)}_{g(t,\mu,\sigma^2),\quad t=(\sum_{i=1}^{n} x_i, \sum_{i=1}^{n} x_i^2)} \underbrace{\left(\frac{1}{\sqrt{2\pi}}\right)^n}_{h(x_1,x_2,\dots,x_n))=1} \rightarrow \text{ind of } \mu, \sigma^2$$

Here $t = t(x_1, x_2, \dots, x_n) = (\sum_{i=1}^{n} x_i, \sum_{i=1}^{n} x_i^2)$.

Using Theorem, we conclude that $T = T(X_1, X_2, \dots, X_n) = (\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i^2)$ is jointly sufficient statistic for $\underline{\theta} = (\mu, \sigma^2)$.

22

And now, he consider one more example, where we tried to get a scent sample from a normal population with mean μ and variance σ^2 and both μ and σ^2 are unknown to us, and we try to find out their sufficient statistics using the Neyman-Fisher Factorization Theorem.

So, we consider here the joint distribution of X_1, X_2,…, X_n, which can once again be obtained as a product of f of X_1 into f of X_2 into f of X_n because the observations are independent and that we can operate exactly in the same way as we have done in the earlier example. So, if you try to write down the joint PDF of f of X_1 is 2x and this can be written here like this one, simply multiply all f x size from the N(μ, σ^2).

And now, this part can be partitioned into two, two parts. You can see here this part, this is the part number one where you can see here this part is independent of μ and σ^2. And this is only here a function of here h of X_1, X_2,…, X_n where h of X_1, X_2,…, X_n is equal to here 1. Now, the second part here you can see here this is here like this. So, you can see here now, this is a function g k μ is σ^2 and where this T is a function of X_1, X_2,…, X_n which is here like the $\sum_{i=1}^{n} X_i$and summation Xi square, and summation is over the sample values.

So, now, you can see here that here this T is equal to function of X_1, X_2,…, X_n which is $\sum_{i=1}^{n} X_i$and $\sum_{i=1}^{n} X_i^2$, and here the parameter here is $\underline{\theta}$ which is μ and σ^2. So, I can see here that the $\sum_{i=1}^{n} X_i$and $\sum_{i=1}^{n} X_i^2$this statistics is jointly sufficient for μ and σ^2.

So, now, we come to an end to this lecture. And now, you can see here that I have given you a fair idea about the statistical properties of this estimator. Now, I think that time has come where you have to just practice these things in a separate way. For example, how to show unbiasedness, how to show consistency, how to find a sufficient estimator etcetera, you have to take one example and then you have to look for these properties that how you can verify them in a theoretical way as well as in the numerical way.

Once you can do this thing, then all these properties will get settled down in your mind. And obviously, when we are looking for a good estimator, we would always like to have all the properties in that estimator, but these are the different properties that we expect to be in a good estimator.

Now, the next question is this. Many times we are continuously, we have talked that this estimator how to get this estimator, so in the next lecture, I will try to give you two methods method of movements and maximum likelihood method, which helps us in finding out a good value of the parameter, as a function of sample values that is a statistics. So, but before that it is very important that all these properties should be very clear in your mind. So, you try to practice it, and I will see you in the next lecture. Till then, goodbye.