Applied Statistics and Econometrics
Professor Deep Mukherjee
Department of Economic Sciences
Indian Institute of Technology Kanpur
Lecture 8
Introduction to Statistical Inference

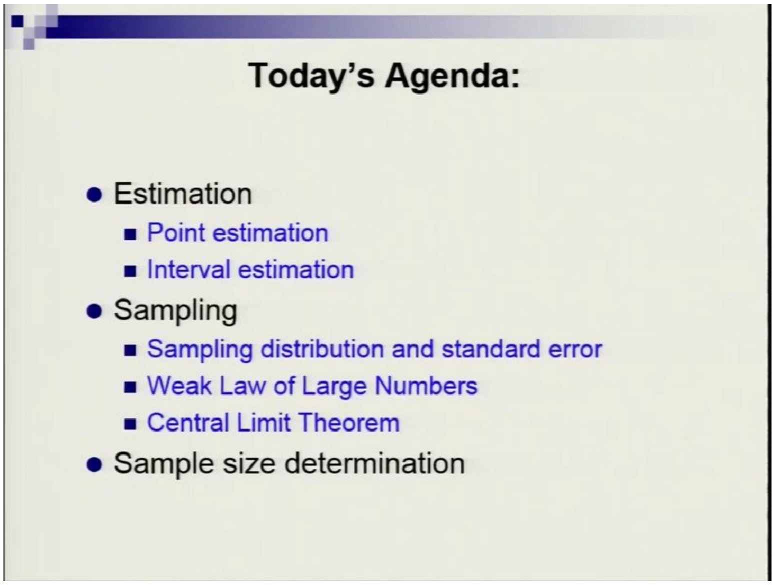
Hello, friends. Welcome back to the lecture series on applied statistics and econometrics. Before we start formal discussion involving normal distribution and other related topics, let us see what we are going to discuss today in terms of our agenda items. So before we discuss the agenda items one by one, let me give you a brief 30 seconds or 1 minute introduction. And actually from this lecture, we are going to start our discussion on a vast topic called statistical inference.

Now, statistical inference is basically the heart and soul of the subjects, statistics and econometrics. Why is this so important? Why is this heart and soul? Because as a statistician or an econometrician you have this challenge that you actually do not know the true values of the population parameters of your empirical model.

And you are out there to collect some data and draw some meaningful inferences from the data by following some robust, quantitative tools from statistics and econometrics. So, this entire journey, by looking at the data and by applying certain tools to draw some inference on the unknown population parameters is called statistical inference. And statistical inference has got 2 parts, one is called estimation theory and the other one is called hypothesis testing.

So, in today's lecture, we are going to cover a little bit of estimation theory and then, we will talk a lot about the sample collection and sampling distribution. Then, we are going to come back to the estimation theory in greater details in future. And then, of course, we will save a couple of lectures to discuss hypothesis testing. So, now, let us look at the agenda items.

(Refer Slide Time: 02:17)



Today's Agenda:

- Estimation
  - Point estimation
  - Interval estimation
- Sampling
  - Sampling distribution and standard error
  - Weak Law of Large Numbers
  - Central Limit Theorem
- Sample size determination

So, we are going to start our discussion with a very brief introduction to estimation theory, we are going to just define point estimation and interval estimation methods. We are going to discuss these things in detail later on. Then, we are going to discuss a little bit on sampling theory. We are going to talk about how samples are drawn in reality. Then, we are going to talk about what is the concept of sampling distribution and standard error. Then, we are going to talk about 2 statistical results, namely weak law of large numbers and central limit theorem. And finally, we are going to end today's discussion by having some formula for sample size determination.

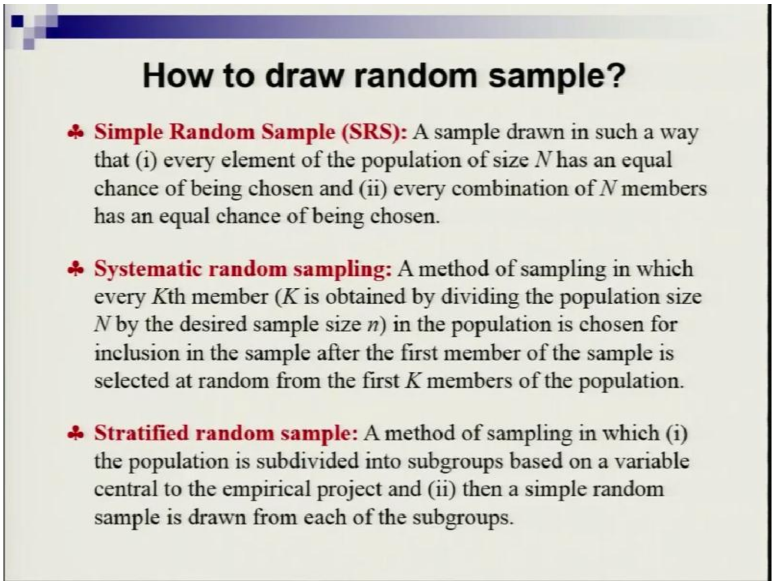
Note that, today's lecture is going to be mostly theoretical in nature. But, I believe that discussion of some theory is important, so that you understand why we are doing certain things. Other than that, it is going to remain like a cookbook, that I tell you do this, do that, follow steps 1, 2, 3, 4, and you do that blindly, get some results. But, you will never be able to answer the question to someone that why have you done so? So, why have you done certain things?

Why you have done a particular approach and why you have not followed the other approach? What is actually happening under curtain when you have adopted a particular statistical approach? To understand all these things, to understand the merits and demerits of a particular approach, it is important to understand or have some idea about the theory behind estimation theory and sampling theory. And that is what we are going to talk about in today's lecture.

I do not want to show you lots of proofs or lemma to bother you. But I will try to tell the story in simple layman's language. Occasionally, I will show you some symbols and some equations. But mostly, I will try to tell you what actually is there in the statistical theory textbooks in layman's language. So, we are going to start our discussion by telling you how to draw random samples in practice.

So, there are several methods and some of them are extremely complicated, but we are not going to cover all of them, because that is not the focus area for us. We just want to know that what are the simplest possible statistical sampling techniques which are available and you can adopt. And we are going to cover only 3 of them. But, I am telling you that there are more better or sophisticated sampling techniques available. We are going to only cover the tip of the iceberg.

(Refer Slide Time: 05:07)



How to draw random sample?

- ♣ **Simple Random Sample (SRS):** A sample drawn in such a way that (i) every element of the population of size *N* has an equal chance of being chosen and (ii) every combination of *N* members has an equal chance of being chosen.
- ♣ **Systematic random sampling:** A method of sampling in which every *K*th member (*K* is obtained by dividing the population size *N* by the desired sample size *n*) in the population is chosen for inclusion in the sample after the first member of the sample is selected at random from the first *K* members of the population.
- ♣ **Stratified random sample:** A method of sampling in which (i) the population is subdivided into subgroups based on a variable central to the empirical project and (ii) then a simple random sample is drawn from each of the subgroups.

So, we start with the first technique that is called simple random sample or it is abbreviated popularly as SRS. So, it is basically a sample drawn in such a way that it follows or satisfies 2 conditions. First of all, every element of the population of size N should have an equal chance of being chosen. And number 2, every combination of those N members of the population, should have an equal chance of being chosen.

The next in the list is systematic random sampling. So, it is a method of sampling in which every Kth member, where K is obtained by dividing the population size n by the desired sample size small n, in the population is chosen for inclusion in the sample. And there is a procedure, first, you have to figure out the first K members in the population and then you have to randomly draw 1 member from the first K members in the population and then, every Kth member from that initially chosen member has to be selected.

So, for an example, suppose this is the sampling strategy that you have devised for your men. So, you have decided that you want every tenth household to be surveyed. So, basically, what will you do? You have these voter list or address list of that locality and then you choose the first 10 and then, from these 10 you choose 1. And then, suppose that is the 7th element or the member in the first 10 household addresses, and then, from there every 10th house in the list, you will choose for your random sample. So that is called a systematic random sample.

Now, the last in the list is called a stratified random sample. This is a bit interesting. So, a stratified random sample satisfies 2 conditions; first, the population has to be divided or sub-divided into sub-groups based on our variable central to the empirical project. And then second, a simple random sample has to be drawn from each of these sub-groups. So, let me explain this idea of stratified random sampling through an example.

Suppose you are interested to statistically estimate a model or do some econometric research and you know a priori that wealth actually plays a big role and as per different wealth groups or socio-economic status groups, your population parameter values are actually different. So, when you want to test that hypothesis, that whether across the population groups or the socio-economic groups, the population parameter varies or not, you may want to estimate different models or estimate different values of sample statistics from various classes.

So, to give you a very simple example, suppose you are interested to figure out some population parameter value in a rural setup and you know that in a village you should have proportionate samples from the wealthy households and from the poor households and the middle class households as well. Because if you somehow draw a random sample where you have lot of representation from the poor households or the middle class households, you may not have sufficient number of households representing the wealthy group.

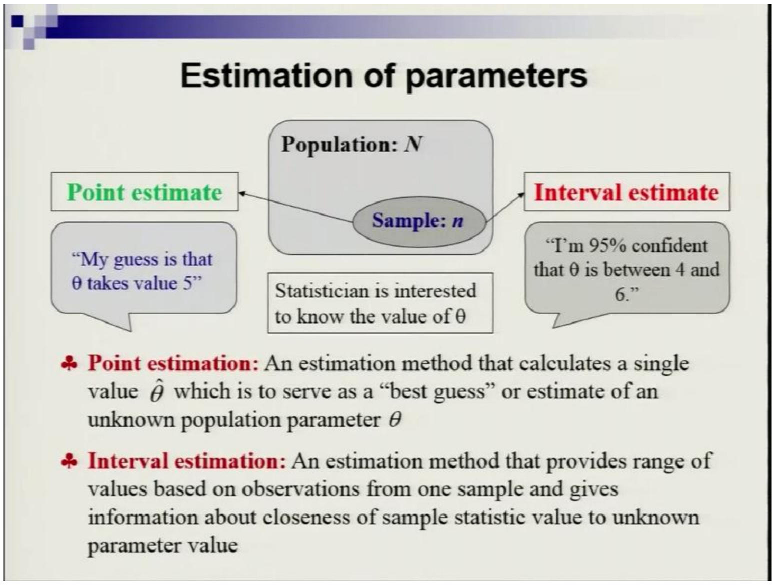
So, you can say that, I will try to make a balanced sample, I will try to get a balanced sample. So, you break the or divide the village population into 2 groups and you can call them that, one is wealthy group; and the other one is poor group. And how to decide? Because, of course, you do not have a priori information on income, so you can actually look at the quality of house that this household is residing at.

So, if it is a pakka house, it is a 2 storey building and they have garage, some car, something like that, then you may or may be a tractor standing, they own a tractor. So based on these visual observations, you can decide that, maybe this household is a rich one. And then of course, by looking at the standard of the house where a particular household is living, you can also decide whether this particular household is poor or not.

So if it is a kaccha house, with mud floors and all, then maybe bricks maximum, then, you can say that probably this is not such a well to do household. So, you actually generate 2 different strata; one is wealthy households and one is the poor or middle class households and then you draw random samples from each strata as per your need. So, that is basically the idea behind stratified random sampling.

So, now, we are going to look at the fundamentals of estimation theory in very simple manner. I do not want to spend more than 3 minutes time on these concepts, because, I am going to come back to statistical estimation theory in the next lectures to give you details. Because it is such a vast item that cannot be covered in 2 - 3 minutes. And as this is not the focal point for today’s lecture, I would just like to give you some idea about the estimation problem and then we will move on to the other things that are there in the agenda items.

(Refer Slide Time: 11:03)



Estimation of parameters

Population: N

Point estimate

Interval estimate

Sample: n

"My guess is that θ takes value 5"

Statistician is interested to know the value of θ

"I'm 95% confident that θ is between 4 and 6."

- ♣ **Point estimation:** An estimation method that calculates a single value $\hat{\theta}$ which is to serve as a "best guess" or estimate of an unknown population parameter θ
- ♣ **Interval estimation:** An estimation method that provides range of values based on observations from one sample and gives information about closeness of sample statistic value to unknown parameter value

So, we start with a population of size capital N and unfortunately, we do not have enough resources to collect data from all of these capital N members or elements of the population. So, we have to draw a simple sample of size small n, which is very small. Now, the question is that suppose there is a population parameter theta whose value is not known to the researcher and a statistician or an econometrician is interested to know what could be the value of theta. So, you have to make guesses.

Now, the question is basically how to make guesses? And estimation theory actually gives you or provides you certain tools to make good guesses. So, these are the 2 types of estimation methods available in the toolkit of a statistician or econometrician and they are called point estimate and interval estimate. First, let me talk about the point estimate. Suppose, the statistician is asking this question or maybe his or her boss is asking a question that what is the value of theta? And then, the statistician replies that my guess is that theta takes value 5.

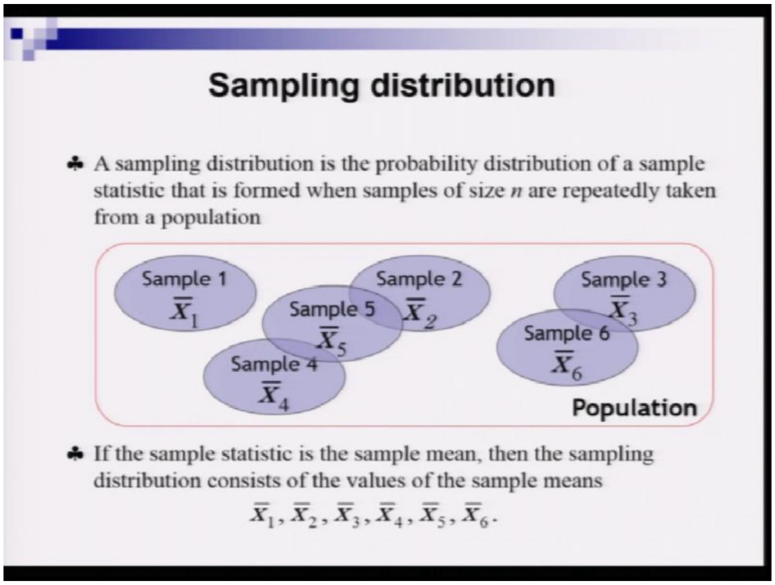
So, basically the statistician is reporting 1 number or 1 point from the set of all possible numbers for the population parameter and that is why it is called a point estimate. Now, the statistician can also reply in a bit probabilistic manner and the statistician can also reply to his or her boss by saying that, well, I am 95 percent confident that theta is going to be between 4 and 6. So, note that here the statistician is actually reporting a range and not only a range, he or she is also reporting a probability associated with that range. And this approach is called an interval estimation approach and the confidence interval is basically the interval estimate.

So, with this story, I think now you are ready to look at the formal definition. So, point estimation is an estimation method that calculates a single value theta hat, which is to serve as the best guess or estimate of an unknown population parameter theta. Now, an interval estimation is an estimation method that provides the range of values based on observations from one sample only and gives information about closeness of the sample statistic value to the unknown parameter value.

So, what is a sampling distribution? You may remember that we have already differentiated between population and sample. So population is the entire set of objects that you want to study or you want to draw some inferences about. But sample is basically a small section of that universe. So it is basically some chosen elements from that entire set or universe. So the issue emerges, how we are so sure that the sample statistic, like mean and median and variance, etc., that we calculate from a chosen sample is going to give us the true picture about the population.

The answer to this question is very difficult to address. In fact, I should tell you that the sample statistics like mean, median and variance that we have studied so far is going to change from 1 sample to the other. Because imagine, you have an universe or a population of people 5,000 objects and you just have drawn a simple random sample of 500 objects from that population.

(Refer Slide Time: 15:07)



Sampling distribution

- A sampling distribution is the probability distribution of a sample statistic that is formed when samples of size *n* are repeatedly taken from a population

- If the sample statistic is the sample mean, then the sampling distribution consists of the values of the sample means

$$\overline{X}_1, \overline{X}_2, \overline{X}_3, \overline{X}_4, \overline{X}_5, \overline{X}_6.$$

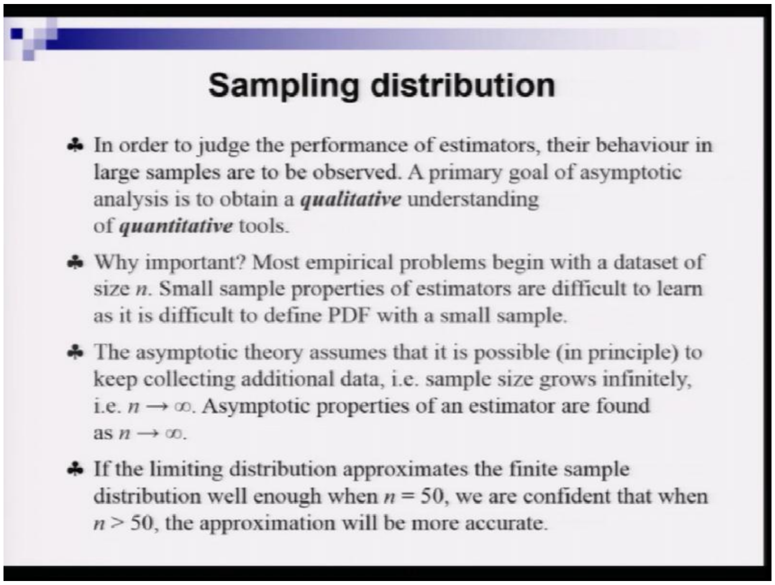
So, here in this slide, I am showing you that exact scenario where the big box is your population where you have some 5,000 or 6,000 objects and it could be any large number. And then, here, I have drawn several oval shaped objects, which are basically chosen sample. Now, how these sample is being chosen, that we forget at this moment. Suppose, some samples are chosen and we are calculating the sample statistic of our interest. Here, in this diagram, I am showing you, if we are interested in sample mean.

So, think about the sample 1, which has sample 1 written inside it and you see that suppose it is a sample of 100 objects from that population and based on these 100 objects, you get 1 X1 bar. Now, if you change your sample from the same population, then you may land up with getting a different sample, sample 2 and none of these objects may be part of the earlier chosen sample 1. Based on this sample 2, you can derive a measure of sample statistics a sample mean, which is denoted as X2 bar. And similarly, you can draw other samples.

Now, note that, we can draw infinite number of samples from the same population in this manner. And when I said that in the first illustration, I showed you that, I assume there that no two samples have the same objects. But it is an assumption, we can also relax this assumption. It may be the case that we have two samples where some elements are common. And if this happens, then of course, from one sample to the other, your measure of sample mean or sample variance is bound to differ. So, if that happens, then actually we can generate a distribution for the sample statistic that we are interested in. And that is basically the essence of sampling distribution.

So, in this case, again back to the diagram, so, here we have chosen six samples and we are interested in sample mean. So, we generated sample mean X1 bar, X2 bar, up to X6 bar, and these basically gives me the sampling distribution. Of course, these sample means will now be associated with some probability of actually being observed. So, that is basically the sampling distribution is all about.

(Refer Slide Time: 17:56)



Sampling distribution

- ♣ In order to judge the performance of estimators, their behaviour in large samples are to be observed. A primary goal of asymptotic analysis is to obtain a ***qualitative*** understanding of ***quantitative*** tools.
- ♣ Why important? Most empirical problems begin with a dataset of size *n*. Small sample properties of estimators are difficult to learn as it is difficult to define PDF with a small sample.
- ♣ The asymptotic theory assumes that it is possible (in principle) to keep collecting additional data, i.e. sample size grows infinitely, i.e. $n \rightarrow \infty$. Asymptotic properties of an estimator are found as $n \rightarrow \infty$.
- ♣ If the limiting distribution approximates the finite sample distribution well enough when $n = 50$, we are confident that when $n > 50$, the approximation will be more accurate.

So, let me continue a little bit with sampling distribution on theoretical side. In order to judge the performance of our estimators, their behavior in large samples are to be observed. A primary goal of asymptotic analysis is to obtain a qualitative understanding of quantitative tools. So, what do we mean by the quantitative tools? So, that is basically the same old thing sample mean or sample variance, etc., that we calculate from the sample. And what do we mean by the qualitative understanding of these tools?

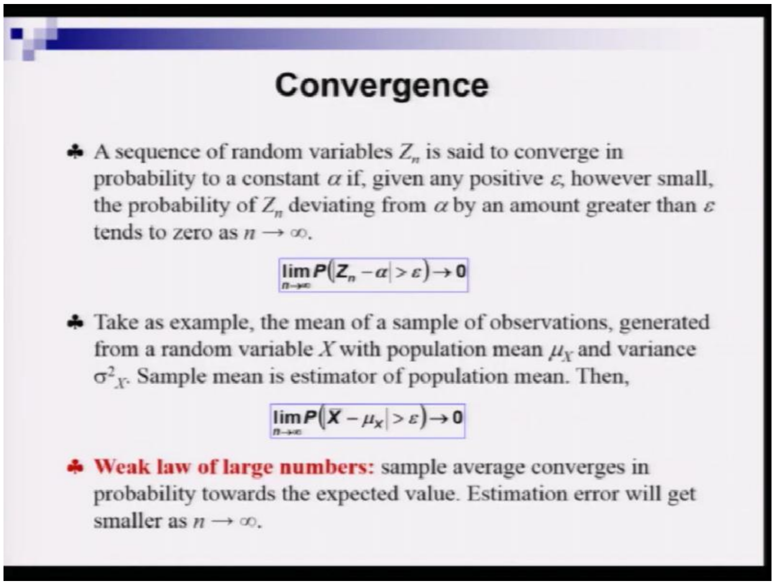
So, the qualitative understanding implies that we may be interested to see if we draw samples many, many times. Is there any pattern in the sample statistic that we calculate from these large number of samples? Do they converge to some common number? Something like that. Now, why is this important? Now, asymptotics are very important because most empirical problems begin with a data set of size n and small sample properties of estimators are difficult to learn as it is difficult to define probability density function with a small sample.

So, here let me explain it again. So, when we draw a small sample size say of 100 from a population of say 5,000 or 10,000, generally, we do not have enough money to draw samples repeatedly. We just stop after one sample in hand. And then, how can we draw inference based on that small sample or just one sample about the properties of the estimators? We want to know how these estimators behave if we take repeated samples? And asymptotic theory helps us in this matter.

So, asymptotic theory assumes that it is possible, at least in principle to keep collecting additional data. It implies sample size grows infinitely, it implies a n tends to infinity and asymptotic properties of an estimator are found as n tends to infinity. So, we can keep on drawing samples from the same population and as we keep on the number of samples will tend to infinity and then asymptotic properties can be derived about an estimator.

Now, what is an estimator? Well, estimator is basically the formula that we use to calculate some sample statistic like sample mean. Now, if the limiting distribution when n tends to infinity approximates the finite sample distribution well when n equal to 50, so suppose we assume that our asymptotic properties are holding well enough for n equal to 50 from some simulation study, in that case, we can confidently say, that when n is greater than 50, the approximation will be more accurate. So, you see, the, when we say asymptotics we actually mean the large sample case.

(Refer Slide Time: 21:10)



Convergence

- ♣ A sequence of random variables Z_n is said to converge in probability to a constant α if, given any positive ε, however small, the probability of Z_n deviating from α by an amount greater than ε tends to zero as $n \to \infty$.

$$\boxed{\lim_{n\to\infty} P(|Z_n - \alpha| > \varepsilon) \to 0}$$

- ♣ Take as example, the mean of a sample of observations, generated from a random variable X with population mean μ_X and variance σ^2_X. Sample mean is estimator of population mean. Then,

$$\boxed{\lim_{n\to\infty} P(|\bar{X} - \mu_X| > \varepsilon) \to 0}$$

- ♣ **Weak law of large numbers:** sample average converges in probability towards the expected value. Estimation error will get smaller as $n \to \infty$.

Now, we are going to talk about the concept of convergence and probability limit. So, a sequence of random variables Zn is said to converge in probability to a constant alpha, if given any positive epsilon, however small it is the probability of Zn deviating from that alpha by an amount greater than epsilon, tends towards 0, as n tends to infinity. Apparently this definition may look very hard. But let me try to explain in layman’s terms.

So here, what is the sequence of random variables Zn that is there in the definition, so, if we collect a lot of samples from the population at least theoretically, then suppose, you know we get 10 samples from the population. So, if we are interested in sample mean then from these 10 samples, we get to see 10 different values of the sample mean. And as realization of one particular sample is random, it is uncertain, there is a probability associated with that. So, the sample mean from each and every sample that we have collected from the data actually is a random variable.

So here, the Zn the sequence of random variables could be seen as a sequence of sample means, so if we are interested in sample mean to study. So, here you can ignore that complicated mathematical form, because I have already explained you the Zn variable, what is that. Alpha is basically a constant and what is that constant it will be clear to you soon. And this epsilon is just some arbitrary positive number, it could be very small number like 0.00001 or 0.0005, something like that.

So, what actually is the essence of this expression in the blue box, the first box in the slide that you are seeing. So, it basically shows that number of samples increases towards infinity, then, actually this Zn, this sequence of random variables, you can assume that sequence of sample means coming from different samples is going to converge toward some number, alpha and as it converges towards that constant alpha, the gap between that constant number and the sample mean in each case is going to decrease and decrease and decrease.

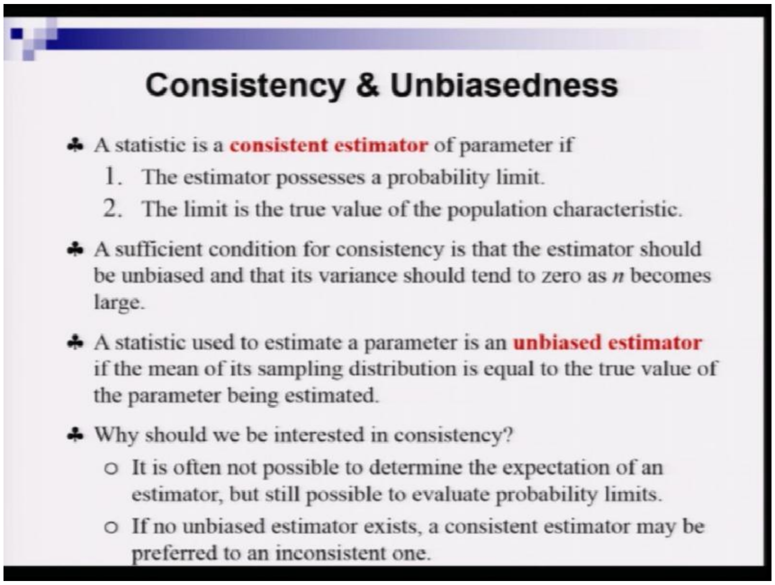
Ultimately, statistically, it is going to be 0, that is basically the essence of the expression. Now, let us take an example, the mean of a sample of observations generated from the random variable X with population mean mu x and variance sigma square x. So, sample mean is the estimator of the population mean, I talked about that thing. And now, you try to link the x mathematical expression that I have shown you in the first box to the second box.

Here, you see, as I was talking about, that Zn is replaced by x bar the sample mean, now it is the random variable, very specific random variable we are talking about. And this alpha, the constant I was talking about in the definition of convergence is replaced by a mu of x. So, that is basically the population mean that we want to actually infer about or we want to know. And then, this difference between the sample mean from different samples and the true population mean is getting smaller and smaller and smaller as n tends to infinity and when n is extremely large, then actually the difference boils down to almost 0.

Now, we are going to talk about an interesting property of sampling and that is called the weak law of large numbers. And why this law is very important, because this law actually is the fundamental basis for running of certain risky industries, like insurance. So, what does a weak law of large numbers say?

So, it says that sample average convergence in probability. So, what does weak law of large numbers say? It says that the sample average converges in probability towards the expected value of the true population parameter and estimation error will get smaller and smaller as n tends to infinity. Basically nothing but the statement of the previous mathematical expression that we have came across.

(Refer Slide Time: 26:36)



Consistency & Unbiasedness

- A statistic is a **consistent estimator** of parameter if
  1. The estimator possesses a probability limit.
  2. The limit is the true value of the population characteristic.
- A sufficient condition for consistency is that the estimator should be unbiased and that its variance should tend to zero as *n* becomes large.
- A statistic used to estimate a parameter is an **unbiased estimator** if the mean of its sampling distribution is equal to the true value of the parameter being estimated.
- Why should we be interested in consistency?
  - It is often not possible to determine the expectation of an estimator, but still possible to evaluate probability limits.
  - If no unbiased estimator exists, a consistent estimator may be preferred to an inconsistent one.

Now, we are going to talk about two interesting properties about the estimators and they are consistency and unbiasedness. Note that, we will not make use of consistency and unbiasedness properties here in the first part of the course. But, when we will be doing econometrics, the second part of the course, there this consistency and unbiasedness properties play a big role. And we will just inform you about these 2 properties in layman's language at this moment. Later, when we will do econometrics in the second part, you will find more meaningful use of these concepts in econometric theory.

So let me start by defining a statistic. Well, statistic we have all seen. So I do not think that I at this moment, I should define statistic. Let me start with the definition of a consistent estimator. So, a statistic is a consistent estimator of a population parameter, if 2 conditions are satisfied. Number 1, the estimator possesses a probability limit, I have already explained what is a probability limit. And number 2, the limit is the true value of the population characteristic.

Now, there is a sufficiency condition for consistency as well and it says that the sufficient condition for consistency is that the estimator should be unbiased and that its variance should tend to 0 as n becomes pretty large. So, in mathematical terms, n tends to infinity. Now, we are going to provide a simple definition for unbiased estimator. A statistic used to estimate a parameter is called an unbiased estimator, if the mean of its sampling distribution is equal to the true value of the, or population parameter being estimated.

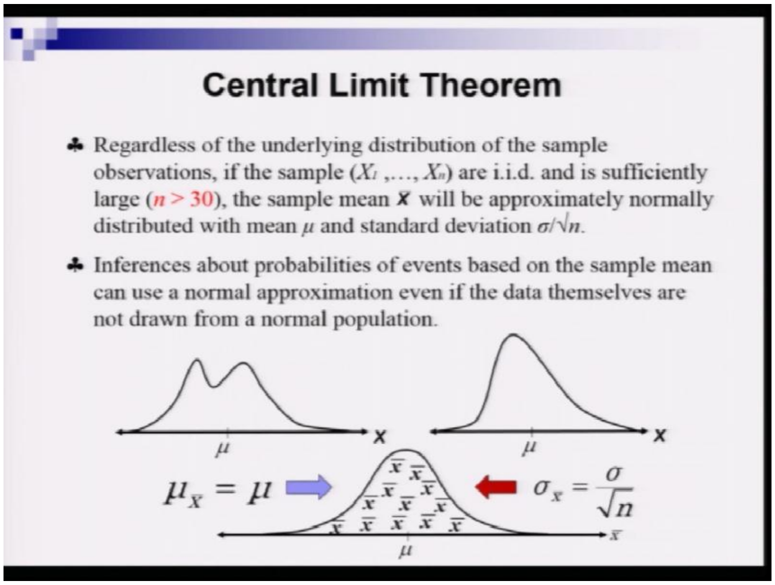
So here, in this case, the bias is 0. So bias actually is the difference between the true value of the population parameter that is unknown to us, and the estimated value of the population parameter by applying a particular formula of estimator. So why are we so worried about consistency? There are 2 reasons and both are theoretical reasons, but if you do econometrics at an advanced level or if you want to study statistics at a higher level, then you will see that these issues are very important issues. And that is why he actually I just have highlighted these 2 points here.

I do not want to get into deeper details of this, but at least I thought that it is good to inform you that there are certain problems. It is often not possible to determine the expectation of an estimator, but still, possible to evaluate the probability limits. So, why we are talking about this, because I just said that when you want to prove that a particular statistic or estimator formula is unbiased estimator or not, then basically we have to focus on the mean of the sampling distribution.

But mean, means what? It means the expected value. So that is what this first bullet point is saying, sometimes expected value of the sampling distribution may not be possible to obtain. And what is the second point? Why we should be interested at least in theoretical statistics? So, if no unbiased estimator exists, because of the first fact or even if he, you find that there are some estimators, but none of them are unbiased, you need to choose one estimator.

So that you can provide an estimate of the unknown population parameter. Then which one to pick? So in that case, you will go with a consistent estimator, because consistent estimator, at least you know has a probability limit and it is the true value of the population characteristic. So, an estimator, which is consistent may be preferred to an inconsistent estimate, even if all the available estimators are kind of biased or the mathematical form for estimated is not available.

(Refer Slide Time: 30:46)



Central Limit Theorem

- ♣ Regardless of the underlying distribution of the sample observations, if the sample (X_1 ,…, X_n) are i.i.d. and is sufficiently large ($n > 30$), the sample mean $\bar{X}$ will be approximately normally distributed with mean μ and standard deviation $\sigma/\sqrt{n}$.
- ♣ Inferences about probabilities of events based on the sample mean can use a normal approximation even if the data themselves are not drawn from a normal population.

$$\mu_{\bar{x}} = \mu \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

In the last lecture only we have spoken about central limit theorem, but very briefly. So, in today’s lecture, I want to actually link that central limit theorem to the sampling theory, so that, we find how central limit theorem plays a very big role in Applied Statistics and Applied Econometrics. So, let me formally state the central limit theorem, we have already defined the central limit theorem in last lecture. But let us provide a somewhat different looking definition or statement for the theorem.

So, here we go, regardless of the underlying distribution of the sample observations, if the sample X1 to Xn so, there are n samples we are talking about are iid, what do we mean by iid? Its, it means, the abbreviation means independently and identically distributed, and if the sample is sufficiently large, so, by large sample we mean that n is greater than 30, then the sample mean X bar will be approximately normally distributed with mean mu and standard deviation sigma divided by root n.

We can also say that inferences about probabilities of events based on the sample mean can use a normal approximation even if the data themselves are not drawn from a normal population. The second point will be very clear from the diagram below. Look at the distributions that I am showing here. So, the distribution that you are seeing at the right hand side of the slide you, that is basically a bimodal distribution. Now, we know that normal is an unimodal distribution, normal is never bimodal. So, this distribution is not normal distribution, although it has some mean mu.

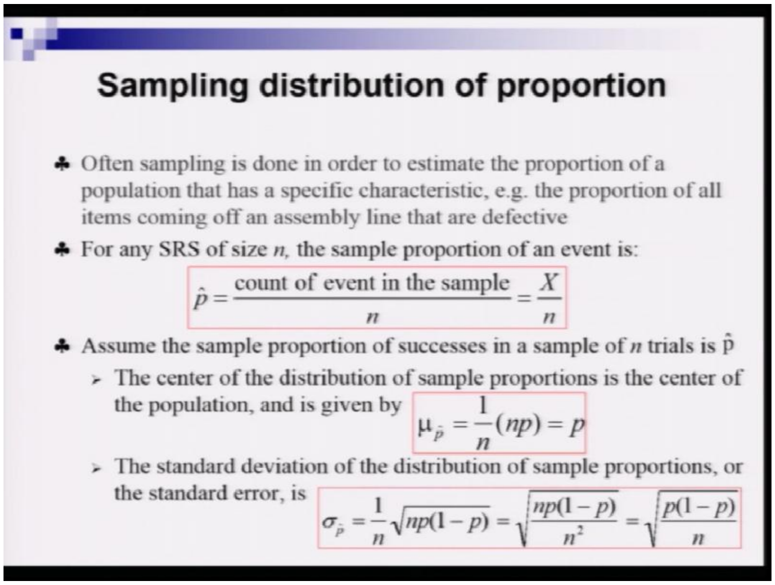
Now, the distribution that I am showing you towards the left hand side of the slide, there you see that we have a positively skewed distribution. We know that normal distribution is symmetric, but here, we are not showing an asymmetric or positively skewed right skewed distribution, although, it has mean mu. So, we are telling that even we have abnormal probability shapes, like no symmetry, multiple modes, even in that case the central limit theorem actually holds.

So, basically if we apply the central limit theorem in this case of sampling statistics distribution and all then what do we get? That is basically the last curve is talking about. So here, along the x axis I am measuring X bar, note the difference. In the previous 2 cases, I have measured a normal X not X bar. So here, I am interested about the estimator of the population mean, which is the sample mean denoted by X bar. So, I told you before that X bar is basically a random variable. So, I expect a sampling distribution for this random variable, because if I change my sample, I am expecting to see a different value of X bar.

Now, what I am saying is that the sampling distribution will look like more or less normal distribution. So, note that these values of these distribution are all x bars from different samples. But the mean of the sampling distribution will be mu of X bar and that is going to be equal to mu that is the true population parameter. And we have the variance of X bar or the sample mean or the estimator of population mean. And that is given by sigma divided by root of n.

So, you can now link the central limit theorem that I have stated at the very beginning and the way I ended this slide with showing you the sampling distribution for sample mean. Well note that, when we talk about sample statistic, it is not only sample mean or sample variance all the times, there could be sample proportion as well. So, if we are interested in sample proportion in some particular problems, then how do we go about the sample sampling distribution of sample proportions? That is the thing that we are going to study next.

(Refer Slide Time: 35:06)



Sampling distribution of proportion

- Often sampling is done in order to estimate the proportion of a population that has a specific characteristic, e.g. the proportion of all items coming off an assembly line that are defective
- For any SRS of size *n*, the sample proportion of an event is:

$$\boxed{\hat{p} = \frac{\text{count of event in the sample}}{n} = \frac{X}{n}}$$

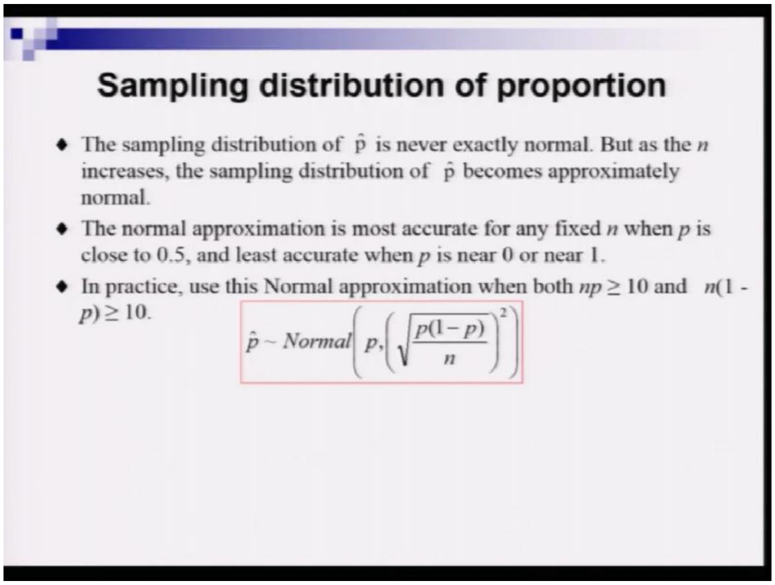
- Assume the sample proportion of successes in a sample of *n* trials is $\hat{p}$
  - The center of the distribution of sample proportions is the center of the population, and is given by $\boxed{\mu_{\hat{p}} = \frac{1}{n}(np) = p}$
  - The standard deviation of the distribution of sample proportions, or the standard error, is $\boxed{\sigma_{\hat{p}} = \frac{1}{n}\sqrt{np(1-p)} = \sqrt{\frac{np(1-p)}{n^2}} = \sqrt{\frac{p(1-p)}{n}}}$

So, often sampling is done in order to estimate the proportion of a population that has a specific characteristic. For example, you can say, proportion of all items coming off an assembly lines that are defective. So, here we are interested in the percentage defect being produced by an assembly line in the production process.

Now, for any SRS, here by SRS, I mean simple random sampling of size n, the sample proportion of an event can be defined as count of event in the sample divided by the sample size. And that can be denoted by the symbol p hat or p caret. Assume that the sample proportion of successes in a sample of n trial is p hat.

Now, the center of this distribution of sample proportions is the center of the population and is given by the mu of p hat equal to 1 over n times np. np is coming from the binomial distribution, it is the mean of binomial distribution, if you remember. And as we are taking the sample, then we are dividing this by 1 over n and that basically ultimately gives me p. Now, the standard deviation of the distribution of sample proportions or the standard error can be given as sigma of p hat and I am showing you how to get that expression in the box.

(Refer Slide Time: 36:52)



Sampling distribution of proportion

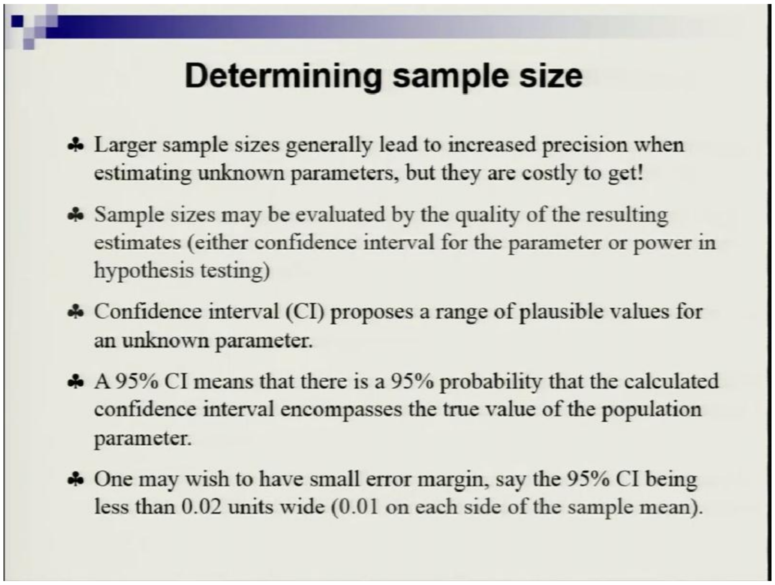
- The sampling distribution of $\hat{p}$ is never exactly normal. But as the n increases, the sampling distribution of $\hat{p}$ becomes approximately normal.
- The normal approximation is most accurate for any fixed n when p is close to 0.5, and least accurate when p is near 0 or near 1.
- In practice, use this Normal approximation when both $np \geq 10$ and $n(1 - p) \geq 10$.

$$\boxed{\hat{p} \sim Normal\left(p, \left(\sqrt{\frac{p(1-p)}{n}}\right)^2\right)}$$

So, we are not going to deeper in the sampling distribution of sample proportions, but we will end with 3 major points. So, the sampling distribution of p hat is never exactly normal. But as the n increases, the sampling distribution of p hat becomes approximately normal. So, you can see that, we are making use of these weak law of large numbers and central limit theorem. Now, the normal approximation is the most accurate for any fixed n when p is close to 0.5 and the least accurate when p is near 0 or near 1.

So, probably you can find similarity of the second point with the normal approximation to the binomial case, what we discussed earlier. Now, in practice use this normal approximation, when np is greater than 10 and n times 1 minus p is also greater than 10. So here, I am showing you the final form of normal distribution. So, p hat the sample proportion follows a normal distribution with mean p and the variance p times 1 minus p divided by n.

(Refer Slide Time: 39:09)



Determining sample size

- Larger sample sizes generally lead to increased precision when estimating unknown parameters, but they are costly to get!
- Sample sizes may be evaluated by the quality of the resulting estimates (either confidence interval for the parameter or power in hypothesis testing)
- Confidence interval (CI) proposes a range of plausible values for an unknown parameter.
- A 95% CI means that there is a 95% probability that the calculated confidence interval encompasses the true value of the population parameter.
- One may wish to have small error margin, say the 95% CI being less than 0.02 units wide (0.01 on each side of the sample mean).

Now, we come to the last topic of today’s discussion and that is sample size determination. Now, sample size determination is very important, because, of course, we do not have enough resources to collect data from the entire population or universe. So, we have to remain happy with a small sample.

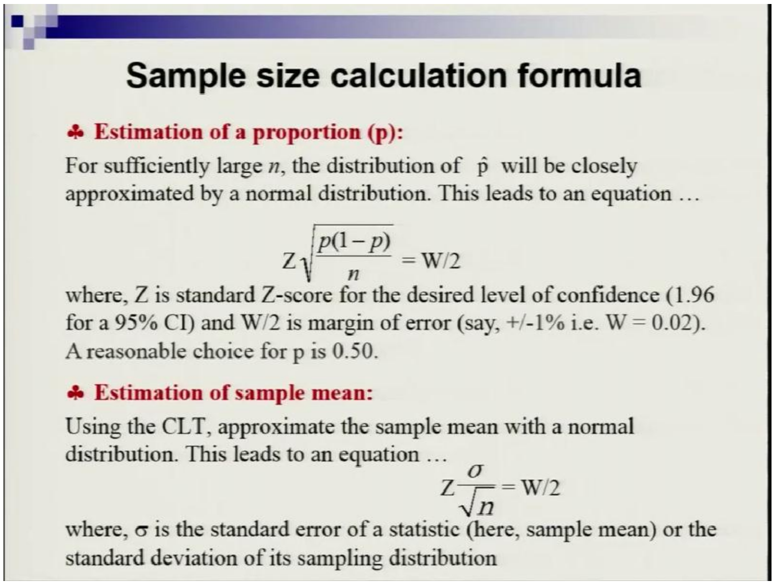
Now, how much small is going to do the work for you? That is the question. So, if you are if you know that the population size is say 10,000 is 50 data points or a sample size of 50 is good enough or it is better if you get 100. Of course, it is far better if you get say 1000, but, of course, you may not have enough resources to collect such a large sample.

So, where to stop? Because look that there is a trade off, the trade off is between the resources to collect a large sample size and it is basically the trade off between the money and the precision, because as you increase your sample size, you get higher precision from your estimates. And as you decrease the sample size, you save on resources, but then you also lose out on the accuracy of your estimate.

So here, a confidence interval is denoted by CI and that proposes a range of plausible values for the unknown parameter that we all know. So, in the previous slide, I say that the statistician is saying that, well, I am 95 percent confident about this, this. So what does you know this 95 percent confidence interval means? So it means that there is a 95 percent probability that the calculated confidence interval will encompass the true value of the population parameter say theta.

So, one may also wish to have a very small error margins, say the 95 percent confidence interval being less than 0.02 units wide. So 0.01 on each side of the sample mean. Why this is important? Because confidence interval ultimately it is a range. So, you want to have a thinner gap between these extreme points in the range or the limiting values of this range. So thinner or the narrower the interval it is, the more accurate your estimate is. So, of course, you want to have smaller error margin.

(Refer Slide Time: 40:39)



Sample size calculation formula

♣ **Estimation of a proportion (p):**

For sufficiently large *n*, the distribution of $\hat{p}$ will be closely approximated by a normal distribution. This leads to an equation …

$$Z\sqrt{\frac{p(1-p)}{n}} = W/2$$

where, Z is standard Z-score for the desired level of confidence (1.96 for a 95% CI) and W/2 is margin of error (say, +/-1% i.e. W = 0.02). A reasonable choice for p is 0.50.

♣ **Estimation of sample mean:**

Using the CLT, approximate the sample mean with a normal distribution. This leads to an equation …

$$Z\frac{\sigma}{\sqrt{n}} = W/2$$

where, σ is the standard error of a statistic (here, sample mean) or the standard deviation of its sampling distribution

So, now, we are going to discuss 2 different cases; one is going to be the case of sample proportion, because that is one kind of sample statistics. And next, we are going to talk about the case of sample mean that is another popular sample statistic. So, first we are going to talk about the estimation of proportion which is denoted by p.

So, for a sufficiently large small n or sample size, the distribution of p hat, the sample statistic will be closely approximated by a normal distribution. So, this comes from statistical theories that we have covered so far. This leads to an equation Z times square root of p times 1 minus p divided by n and that is equal to W over 2.

So here, the Z is the standardized Z score for the desired level of confidence and generally 1.96 is the Z score for a 95 percent confidence interval. Please remember this number, this is a magic figure in statistics. And you will see more use of this number 1.96 for 95 percent confidence interval later in this course as well. And W over 2 is the margin of error, say plus minus 1 percent.

So, if we assume that plus minus 1 percent, then W will take value 0.02. So then, we are talking about that around the sample mean there will be width of 0.01 units or 1 percentage point, that is plus minus 1 percent margin of error. And, of course, a priori we do not know the value of p, but a reasonable choice for p could be 0.25. And if is there some prior from previous research, then you can use that prior information to replace p as well.

Now, we are going to quickly look at the estimation of the sample mean case. So here, we use CLT, we approximate the sample mean with a normal distribution and that statistical steps lead us to an equation Z times sigma divided by root n and that is equal to W over 2. So here, the only new term well, it is not new, we all know that sigma is the standard error of statistic. So here, we are talking about sample mean as a statistic.

So, this is will be the standard deviation of the sampling distribution of sample mean. So, of course, sigma you have to first get an idea and I have shown how to get a proxy measure and estimate for sigma. So, you can use that and then from this formula the only unknown is n and that can be calculated. So, we are done with today’s lecture. In the next lecture, I am going to give you details about statistical estimation theory. Thank you.