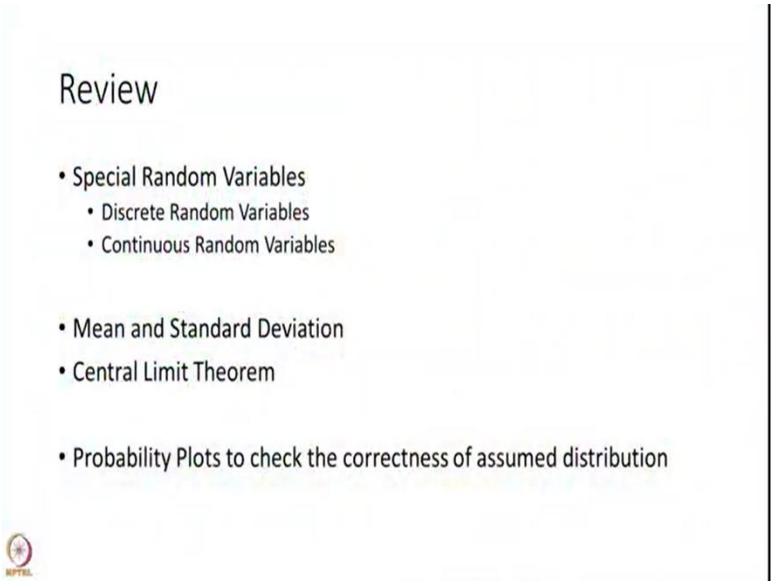
Dealing with Materials Data: Collection, Analysis and Interpretation
Professor. Hina A. Gokhale,
Department of Metallurgical Engineering and Materials Science
Indian Institute of Technology, Bombay
Lecture 47

Sampling Distribution

Hello and welcome to Dealing with Materials Data course. In this session, we would consider the case of Sampling Distributions.

(Refer Slide Time: 00:33)



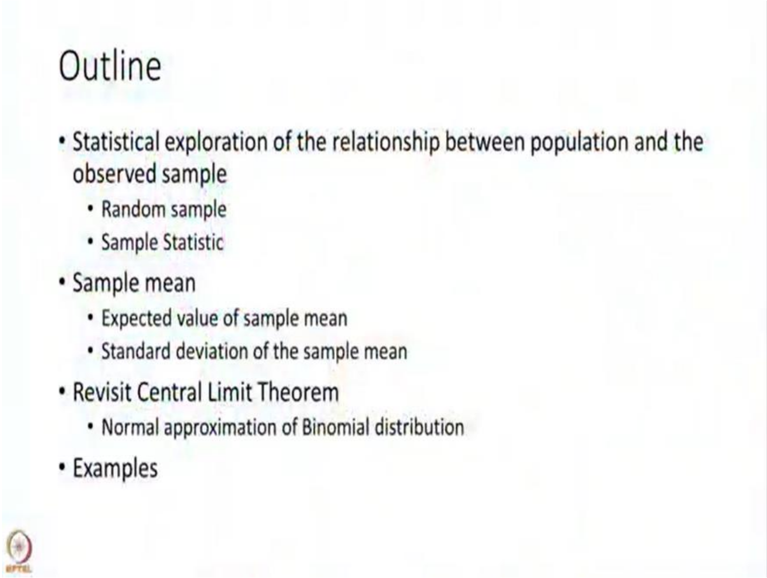
Review

- Special Random Variables
  - Discrete Random Variables
  - Continuous Random Variables
- Mean and Standard Deviation
- Central Limit Theorem
- Probability Plots to check the correctness of assumed distribution

Let us first review what we have done in the past, we introduced certain special random variables with some discrete random variables, we call them discrete distribution functions, continues random variables can also be called a continuous random, continues distribution functions.

With each one of them, we introduced what is their mean. What is their standard deviation? We once again introduced central limit theorem and in just the previous session, we looked into how to generally check the correctness of assumed distribution graphically, through probability plots.

(Refer Slide Time: 01:36)



Outline

- Statistical exploration of the relationship between population and the observed sample
  - Random sample
  - Sample Statistic
- Sample mean
  - Expected value of sample mean
  - Standard deviation of the sample mean
- Revisit Central Limit Theorem
  - Normal approximation of Binomial distribution
- Examples

What we want to do in this sessions or coming up all the sessions from now onwards is make a relationship between the observed sample and the population. Let us try to understand, why are we doing all this? The whole purpose of learning statistics is that we all are aware that there is a general population, there is larger information in the world, there is an entirety of certain kind of data in the world. But, we cannot look into each and every data points that is available in the population this entirety I call a population.

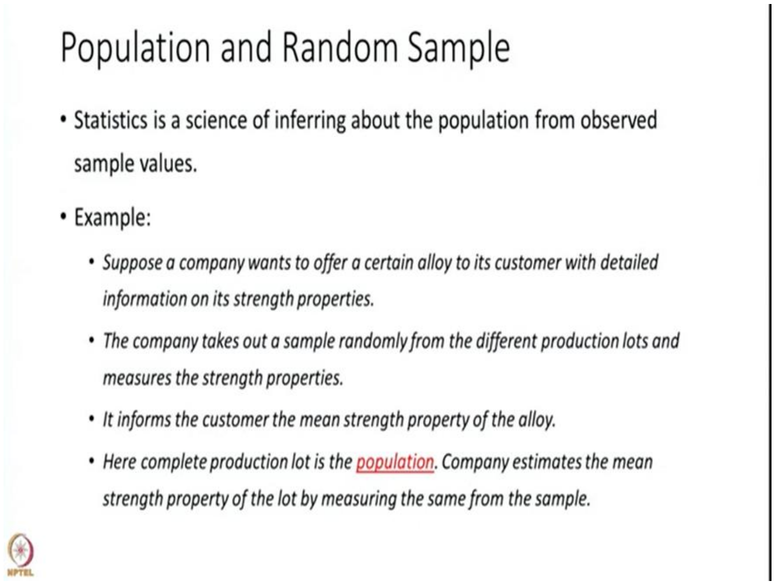
But what we do is we take a sample out of it we observe it, we do certain calculations on it and we try to make judgment on what the population would be like, this is the whole game of statistics. So, here now onwards in this course we trying to do something to setup our understanding of population through a random sample. So, we are going to introduce what is called a random sample once again, we have already done it once in the past we are going to do it again.

We will introduce something called a sample statistic, then we will talk about the most common way to understand any population or any sample for that matter. As we studied in the descriptive statistics, the most common numerical method is a central tendency or we can call here sample mean, and then dispersion of the data through sample, sample variance.

In this particular session we are going to concentrate only on the sample mean, we will find what is the expected value of sample mean. Because remember sample mean is also a random variable, so it has a expected value it has a distribution and therefore we would like to find out what is its expected value? What is its standard deviation of the sample mean?

We will revisit central limit theorem, because it tells us more about the nature of distribution of sample mean when the sample size is very large. We will also give an example we will also show that how a binomial distribution can be approximated as a normal distribution. Remember in the past we have approximated binomial by Poisson. So, we will clarify is to when it is to be distributed approximated as a normal distribution we will give some examples.

(Refer Slide Time: 04:35)



Population and Random Sample

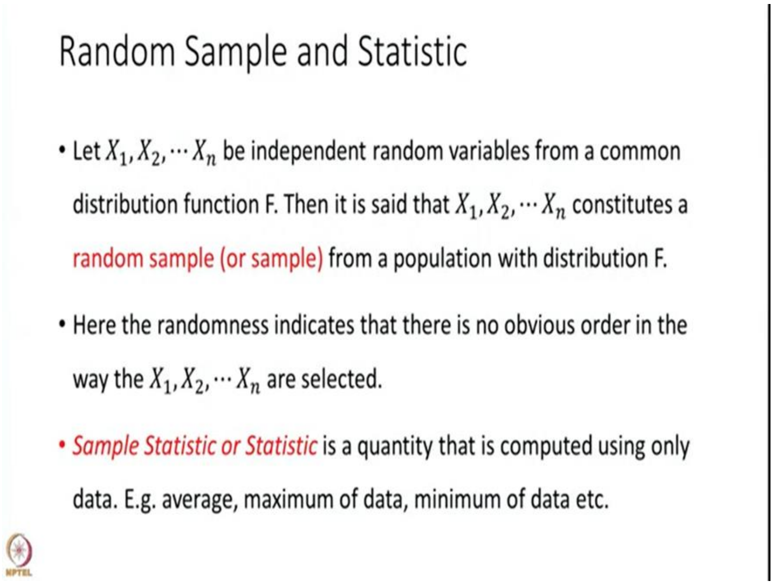
- Statistics is a science of inferring about the population from observed sample values.
- Example:
  - *Suppose a company wants to offer a certain alloy to its customer with detailed information on its strength properties.*
  - *The company takes out a sample randomly from the different production lots and measures the strength properties.*
  - *It informs the customer the mean strength property of the alloy.*
  - *Here complete production lot is the population. Company estimates the mean strength property of the lot by measuring the same from the sample.*

So, let us understand as I said the population and the random sample. As I said statistics is a science the whole purpose of statistics is to understand a population through a small observed sample value, small compare to the population itself. So, here I have given an example, this example concern that you imagine that in a production line in an industry certain alloy is being produced and the customer is requesting an alloy with a certain value of strength property of the alloy.

Now, it is the industry's responsibility marketing department responsibility to tell them that our alloy has a certain strength property. It has a mean strength property of this, how will it do it. Well it cannot start testing each and every alloys that are produced, instead what it may do it is, it may take a lot of it, the randomly chosen ingot.

And then from each ingot it will randomly choose a sample and it will put through a test and find it strength property and then it will say that our lot of production will have this much of strength property. So here, the complete production lot is what I call or what we call in statistics population and whatever that small sample you will take is called a random sample.

(Refer Slide Time: 06:26)



Random Sample and Statistic

- Let $X_1, X_2, \cdots X_n$ be independent random variables from a common distribution function F. Then it is said that $X_1, X_2, \cdots X_n$ constitutes a random sample (or sample) from a population with distribution F.
- Here the randomness indicates that there is no obvious order in the way the $X_1, X_2, \cdots X_n$ are selected.
- *Sample Statistic or Statistic* is a quantity that is computed using only data. E.g. average, maximum of data, minimum of data etc.

So, formally speaking X1, X2 up to Xn be an independent random sample from a common distribution function f. This common distribution function f refers to the population distribution function. Then it said that X1, X2, X3, Xn constitutes a random sample or just a sample from a population with a distribution F.

Here the randomness indicates that there is no obvious order in which X1, X2, Xn are selected there should not be it should not be it should not happen that you take always the first ingot from a lot or you take every tenth ingot from a lot. You must choose it in a random fashion, so that the sample becomes a representative of the population.

A sample statistic or simply a statistic is a quantity that is computed using only the data. You have come across such quantities a lot in this course for example average, standard deviation, maximum of a data, minimum of a data, all these quantities are calculated without making any assumptions on what the distribution of the population is, it is purely from the data. Such a quantity is called statistic, please remember it is singular it is statistic. Now as I said our whole interest is in finding out F, the population distribution F, capital F.

There are two cases here if you know the form of the distribution f as in the past I said that is you are dealing with the strength property of a material there is a possibility that it follows a log normal distribution. So, your f has a form of a log normal distribution, if you are looking at the differences suppose you are already knowing that this whole alloy is designed to have a particular value of strength properties say yield strength and in that case what you are going to do is you are going to actually measure the yield strength and take the difference.

If you take such difference between the yield strength you are sort of trying to calculate an error and therefore it may follow a normal distribution with mean 0 and some variance sigma square. So, it means that you are able to decide a priory, right from the beginning you are able to decide what is the form of the distribution F.

But what you do not know, you know the form of the distribution up to a point up to a level where you know you do not know the parametric values of that distribution. Say if have you assume normal distribution for an error, you say that the mean value is 0 or you can say that you do not know what is mean value the two unknown parameters are mu are sigma square.

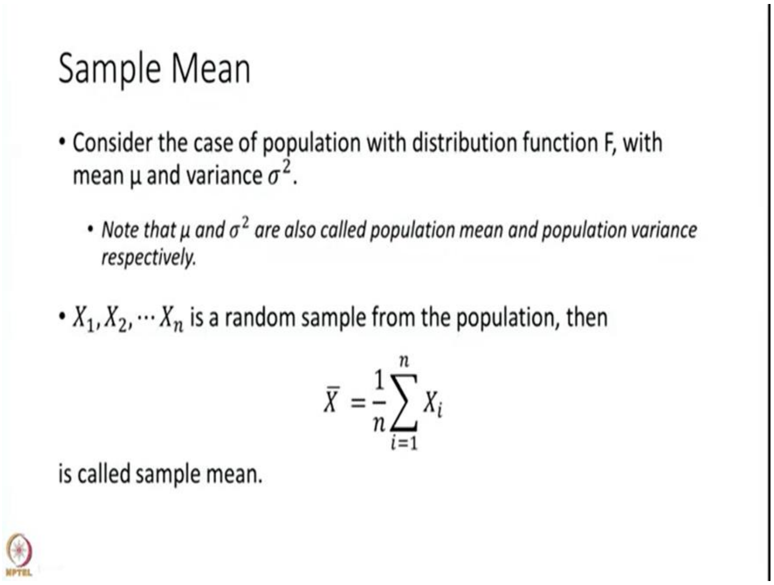
If you are looking at a weigh bull distribution and you want to look at only two parameter Weibull distribution, then you are worried about the scale and shape parameter. If you introduce you think that no no it may be a three parameter Weibull distribution, then you have the form of the distribution which is Weibull.

But what you do not have is the three parameters values. So, the psi which is location, scale and shape, so such situation it is called a process of parametric estimation. You know the form of the distribution but you do not know the parametric value, so then it falls in the first case which is called a parametric estimation and inference.

Suppose, you do not even know the form of the distribution that happens or you are not able to justify it properly that it should fall in this particular form of distribution, then it is called a non-parametric inference and estimation. In this present course, we are going to consider only the case of parametric estimation and therefore, the common distribution will be referred as a population distribution, we know its form but we do not know its parameter and we are going to go for parametric estimation case.

The first and foremost estimation that we would like to learn about is as I said in descriptive statistics also, mean place a central role and so does the standard deviation or a variance. So, we will talk about two sample statistics, one statistic is sample mean and another is sample variance which will follow in the next session. This session we will consider sample mean.

(Refer Slide Time: 11:48)



Sample Mean

- Consider the case of population with distribution function F, with mean μ and variance σ^2.
  - *Note that μ and σ^2 are also called population mean and population variance respectively.*
- $X_1, X_2, \cdots X_n$ is a random sample from the population, then

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

is called sample mean.

So, consider the case of a population with distribution function F with mean value mu and a variance sigma square. Remember we are not saying that this is a normal distribution, this is a Weibull distribution, we are just saying that it is some distribution which has a mean mu and variance sigma square. mu and sigma square are also called population mean and population variance respectively, which you do not know these are now your unknown parameters.

So, we are looking at a very generalized case right now. Let X1, X2, X3, Xn be a random sample from a same population with mean mu and variance sigma square and then we define sample mean as an average of the random sample.

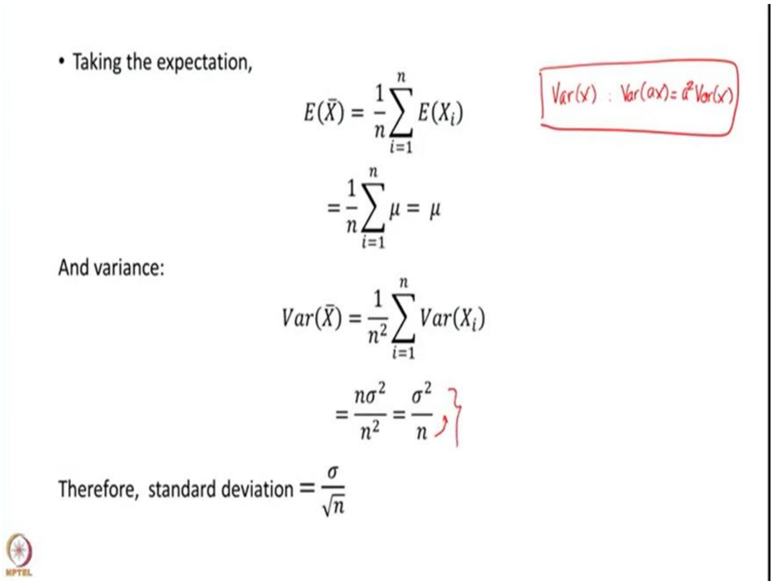
$X_1,\ X_2,\ \cdots X_n$ is a random sample from the population, then

$$\overline{X}\ = \frac{1}{n}\sum_{i=1}^{n} X_i$$

is called sample mean.

But here we would like to show that you see I take a sample and I find an X bar, someone else takes a sample and finds an X bar there are many values of X bar. Therefore, we are talking here random sample as a random sample as a you know random variables arising out of the same distribution F with mean mu and sigma square and therefore X bar is a function of random variable and therefore X bar is also a random variable.

(Refer Slide Time: 13:33)



- Taking the expectation,

$$E(\bar{X}) = \frac{1}{n}\sum_{i=1}^{n} E(X_i)$$

$$= \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$$

$\boxed{Var(X) \ : \ Var(aX) = a^2 Var(X)}$

And variance:

$$Var(\bar{X}) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i)$$

$$= \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Therefore, standard deviation $= \frac{\sigma}{\sqrt{n}}$

So then we can take an expectation of X bar, so if you take expectation of $\bar{X}$

follow the algebra of expectations it is

$$E(\bar{X}) = \frac{1}{n}\sum_{i=1}^{n} E(X_i)$$

$$= \frac{1}{n}\sum_{i=1}^{n} \mu$$

You can say that mu can be estimated using x bar. In the further session you will, we will call this a point estimation, what is variance of $\bar{X}$,

$$Var(\bar{X}) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i)$$

$$= \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

This n square comes in this form, it comes from this identity and therefore it is sigma square by n or the standard deviation is

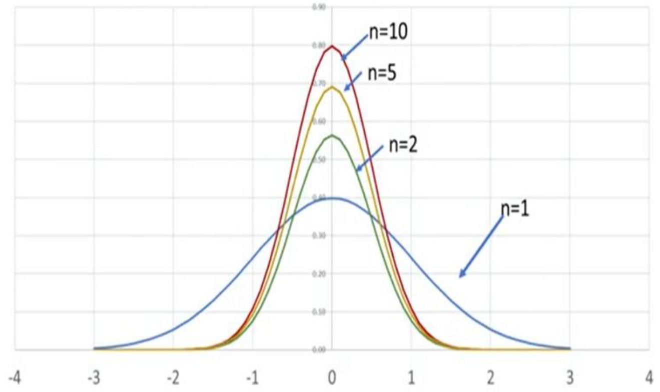
$$\text{standard deviation} = \frac{\sigma}{\sqrt{n}}$$

Let us try to understand this, this says that the expected value of $\bar{X}$ is mu it does not depend on the sample size it is always mu the population mean value.

But the variance of X bar is going to decrease as n will increase. So, if n increases then the whole value actually decreases it means that the spread of the distribution becomes smaller and smaller and therefore the distribution mean value X bar value comes closer and closer to the actual population mean. This is shown in the next slide.

(Refer Slide Time: 16:31)

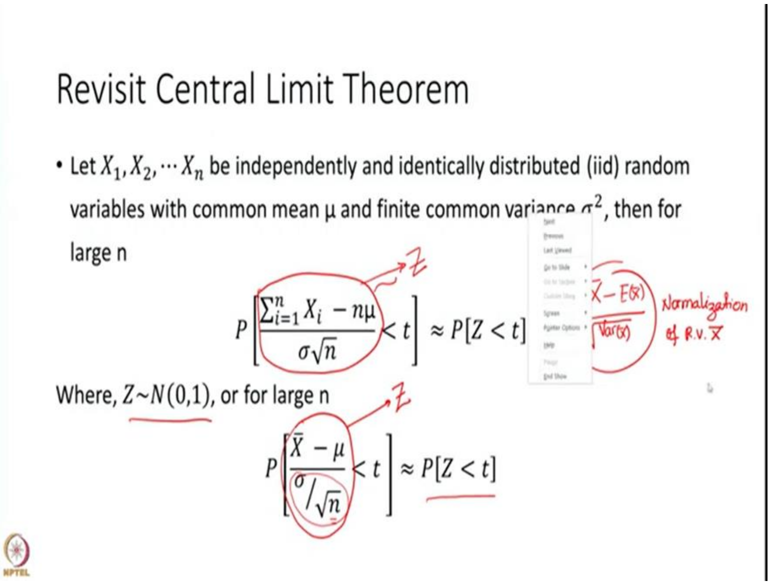
Effect of sample size on shape of the distn.



Here you see that we have plotted, I have plotted standard normal distribution function standard normal pdf, not distribution function probability density function. Blue line is when n is equal to 1, the green line is when n is equal to 2. You see that the spread is drastically decreased when you come to n is equal to 2.

If you take n is equal to 5 it is further decreased and if you go to n is equal to 10 it is even further decreased. It means that the S your n value becomes larger and larger the spread the distribution becomes closer and closer to the mean value population mean because we have taken standard normal distribution. It is a population mean value mu and it is coming closer and closer to that.

(Refer Slide Time: 17:38)



Revisit Central Limit Theorem

- Let $X_1, X_2, \cdots X_n$ be independently and identically distributed (iid) random variables with common mean μ and finite common variance σ^2, then for large n

$$P\left[\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} < t\right] \approx P[Z < t]$$

Where, $Z \sim N(0,1)$, or for large n

$$P\left[\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < t\right] \approx P[Z < t]$$

Z
Z
$\frac{\bar{X} - E(\bar{X})}{\sqrt{Var(\bar{X})}}$
Normalization of R.V. $\bar{X}$

Now let us revisit central limit theorem. What does central limit theorem say let us recall, let X1, X2, Xn be independently and identically distributed random variables, so it is a random sample with a common mean and a finite common variance sigma square. Then for large value of n

$$P\left[\frac{\sum_{i=1}^{n} X_i \ - n\mu}{\sigma\sqrt{n}} < t\right] \ \approx P[Z < t]$$

Or

$$P\left[\frac{\bar{X} \ - \mu}{\sigma/\sqrt{n}} < t\right] \approx P[Z < t]$$

where Z is a standard normal variate.

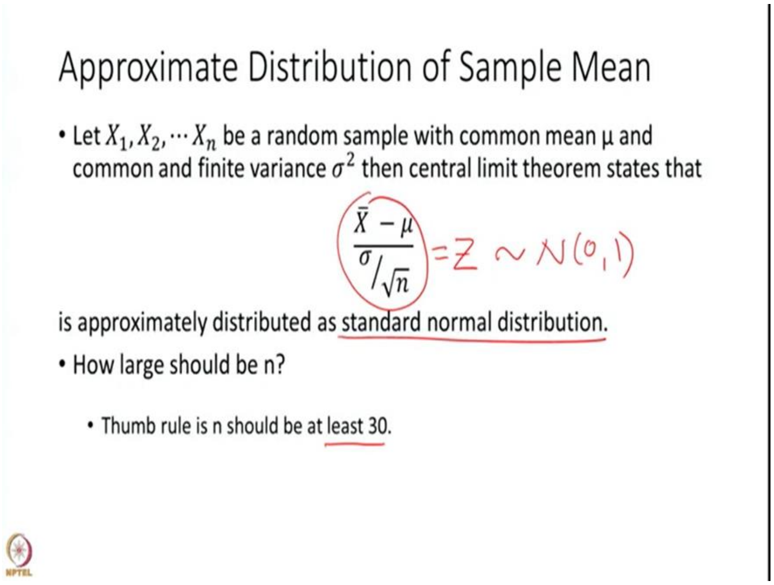
What it tries to say is that this quantity, what it tries to say is that this quantity comes closer and closer approximately to Z which is distributed as normal random variable with 0 mean and 1 standard deviation, it means that it is a standard normal variable. If you divide the numerator and denominator by n then you find that X bar is minus mu divided by sigma square root n is less than t also has behaves like a standard normal variate.

Now what let us look at this quantity very carefully, as n tends to infinity, the sigma over square root n becomes smaller and smaller and therefore it actually means that X bar comes closer and closer to mu.

So, this is now with reference to the present understanding of random variable n population we can say that a population mean comes closer and closer to the, the sample mean comes closer and closer to the population mean and if you take the standardized variable, this is another term that I would like to introduce, any random variable like here it is X bar, if you subtract it from its expected value X bar and divided by its variance square root of variance X bar then this is called normalization of random variable X bar.

Why it is called normalizing? Because it starts behaving like a standard normal variate, it behaves like a standard normal variable. So, this is called a normalization. So, now in future when you do when you hear anywhere when the values are normalized what is happening this is what it really means.

(Refer Slide Time: 21:09)



Approximate Distribution of Sample Mean

- Let $X_1, X_2, \cdots X_n$ be a random sample with common mean μ and common and finite variance σ^2 then central limit theorem states that

$$\boxed{\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}} = Z \sim N(0,1)$$

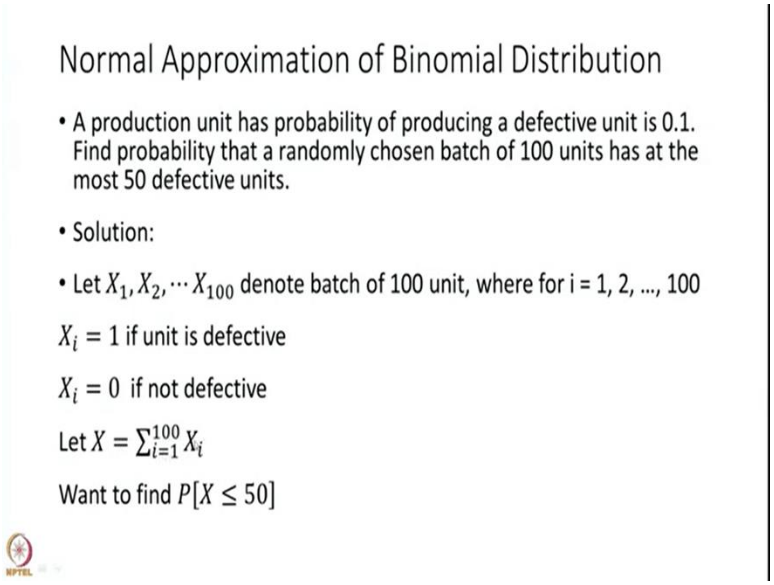
is approximately distributed as standard normal distribution.

- How large should be n?
  - Thumb rule is n should be at least 30.

So now let us worry about what is a approximate distribution of sample mean, again we say that X1, X2, Xn is a random sample with a common mean and a common finite variance sigma square. Then the central limit theorem, tells us that the normalized value of X bar follows a standard normal distribution, so then it follows a standard normal distribution.

This is, is equal to Z, where Z follows standard normal this is what exactly we talked about. The question again comes how large should be n? When you can consider that it is tending to infinity. Well the thumb rule is that it should be at least 30, please recall the discussions we had in the previous sessions when we discussed the central limit theorem, the same thing applies here also.

(Refer Slide Time: 22:46)



Normal Approximation of Binomial Distribution

- A production unit has probability of producing a defective unit is 0.1. Find probability that a randomly chosen batch of 100 units has at the most 50 defective units.
- Solution:
- Let $X_1, X_2, \cdots X_{100}$ denote batch of 100 unit, where for i = 1, 2, ..., 100

$X_i = 1$ if unit is defective

$X_i = 0$ if not defective

Let $X = \sum_{i=1}^{100} X_i$

Want to find $P[X \leq 50]$

So, let us now see another approximation, applying the normal that is the central limit theorem we would like to study the another approximation of binomial distribution when n tends to infinity, that is, when the number of independent Bernoulli trials tend to infinity or when the number of Bernoulli trials are very large. Let us consider the case, you remember the previous question we had a production unit has a probability of producing defecting unit as 0.1 and a randomly chosen batch of 100 units.

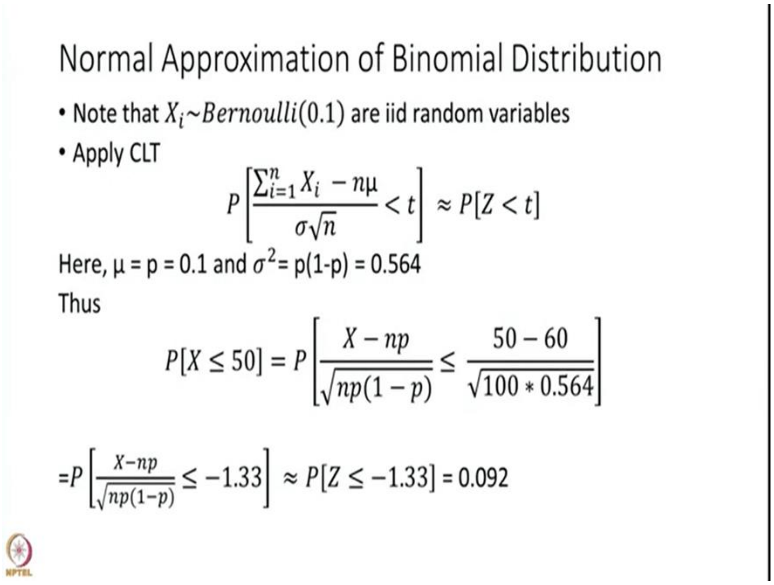
So, random this is, there is a random variable with 100 units, n is equal to 100. What is the probability that is has at the most 50 defective units? Maximum number of defective units it can have is 50. It does not have more than 50 units is what detective units is what it says. You recall we did the same thing by calculating the exact value in order to show the binomial approximation to the Poisson distribution.

Now we are going to the normal approximation, so X1, X2, X3, Xn 100 denote the 100 units and we say that Xi is 1 if the unit is defective and Xi is 0 if it is not defective and we defined X as

Let $X = \sum_{i=1}^{100} X_i$

Want to find $P[X \leq 50]$

(Refer Slide Time: 24:31)



Normal Approximation of Binomial Distribution

- Note that $X_i \sim Bernoulli(0.1)$ are iid random variables
- Apply CLT

$$P\left[\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} < t\right] \approx P[Z < t]$$

Here, μ = p = 0.1 and σ^2= p(1-p) = 0.564

Thus

$$P[X \leq 50] = P\left[\frac{X - np}{\sqrt{np(1-p)}} \leq \frac{50 - 60}{\sqrt{100 * 0.564}}\right]$$

$$= P\left[\frac{X-np}{\sqrt{np(1-p)}} \leq -1.33\right] \approx P[Z \leq -1.33] = 0.092$$

So, Xi is a Bernoulli trail with probability of defective as 0.1. So if you look at it, if you apply the central limit theorem, it says that

$$P\left[\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} < t\right] \approx P[Z < t]$$

μ = p = 0.1 and σ^2= p(1-p) = 0.09

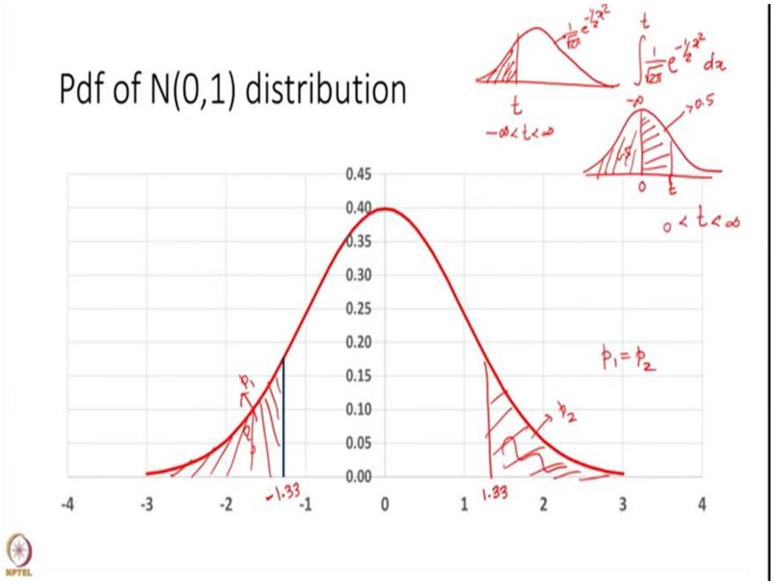
You can calculate it out and therefore probability that X is less than or equal to 50 is probability that

$$P[X \leq 50] = P\left[\frac{X - np}{\sqrt{np(1-p)}} \leq \frac{50 - 10}{\sqrt{100 * 0.09}}\right]$$

$$= P\left[\frac{X-np}{\sqrt{np(1-p)}} \leq -1.67\right] \approx P[Z \leq -1.67] = 0.05$$

This probability you can calculate either using normal tables, there are normal tables available, I will discuss about it in the next slide or you can use R and find out what is the probability value or you can use even excel table, excel spreadsheet there also they have a function to find out this particular value. But in all the cases I say that please remember to look the help.

(Refer Slide Time: 26:26)



Pdf of N(0,1) distribution

So, let us look at here, here I have a standard normal distribution plotted. This is a standard normal distribution red line and this is minus 1.33 value as the z is going to be less than or equal to that. So, we are looking for this probability, the question is how much is this? General normal tables that are available tend to give you the probability of any value t less than or equal to this.

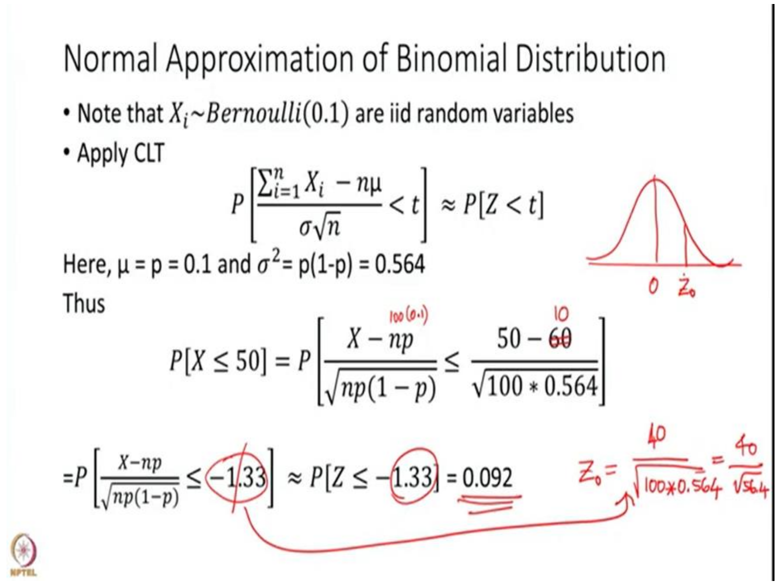
$$P(x < t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-\frac{1}{2}\{x^2\}}\, dx$$

This is the curve, so this is the value which is generally tabulated. But, remember that in the normal distribution it is a symmetric bell shaped curve, so if you come exactly at the 0 value this you already know is a 0.5, it is half and so is this is also, this is also 0.5. So, number of times here it takes a value of t between minus infinity and plus infinity. Sometimes it takes the value of t greater than 0 less than infinity, and it will give you values of this nature.

How do you tackle this? It says that the t is here. Well remember that, whether you take it here or you take 1.33 here, this area and this area are exactly the same. So, this probability and this probability if I call this probability p1 and this I call probability p2 then because of symmetry p1 is equal to p2 and therefore instead of finding the probability on this side you can find a probability here.

So what I am trying to say, which must have been said in your R session I just want to repeat it, that when you apply a normal distribution table or you use a standard routine to find out the normal probability, please check the help and make sure whether it considers this kind of probability or it considers this kind of probability.

(Refer Slide Time: 29:55)



Normal Approximation of Binomial Distribution

- Note that $X_i \sim Bernoulli(0.1)$ are iid random variables
- Apply CLT

$$P\left[\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} < t\right] \approx P[Z < t]$$

Here, μ = p = 0.1 and σ^2= p(1-p) = 0.564

Thus

$$P[X \leq 50] = P\left[\frac{X - \overset{100(0.1)}{np}}{\sqrt{np(1-p)}} \leq \frac{50 - \overset{10}{\cancel{60}}}{\sqrt{100 * 0.564}}\right]$$

$$=P\left[\frac{X-np}{\sqrt{np(1-p)}} \leq \cancel{-1.33}\right] \approx P[Z \leq -1.33] = \underline{\underline{0.092}}$$

$$Z_0 = \frac{40}{\sqrt{100 * 0.564}} = \frac{40}{\sqrt{56.4}}$$

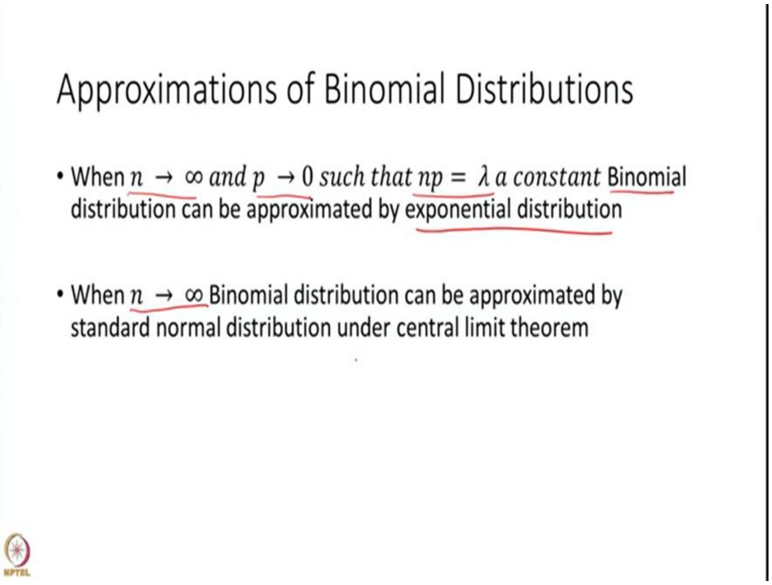
In our present case this value turns out to be 0.092, so here I have so if you calculate it out this turns out to be 0.092. If you look at this if you have to put this t value so when you put mu is equal to point 1 and sigma square is equal to 0.564 you want to have your X less than or equal to 50, so X minus np divided by square root np times 1 minus p turns out to be 50 minus 60 because it is 100 with possibility probability of failure probability of a defective is point 1.

So, probability of is point, where does this came from, wait a minute, wait a minute, I think I have made a mistake we must correct it. I think it is time I correct it. Probability so n is this so this will be actually 10. So, we are looking for this denominator is correct this value will be different we are looking for 10, 50 minus 10, so 40 divided by square root of 100 times 0.564 and therefore it may not be 1.33, you please check what is the value and what should be the distribution. Maybe in the further section I will clarify but please make sure that there is a mistake here.

Mu is when you find an np, n is 100 and p is 0.01, I do not know where did I get 60 from. So, actually it is 50 minus 10, so this is wrong, it should this value should be this much. 40 divided by square root of 100 times 0.564, you can workout this is this will come to 40 divided by square root of 56.4 and you can work out put that value n as shown in the previous case. Now it is a positive value 0 is here, so your t value will be somewhere here, your 1.33 will be some positive value.

If I call this value as Z0, if I call this value Z0 then your Z0 will be somewhere here and you are looking for this probability. So, it will be more than a half, please check it out there is some error here please correct it.

(Refer Slide Time: 32:54)



Approximations of Binomial Distributions

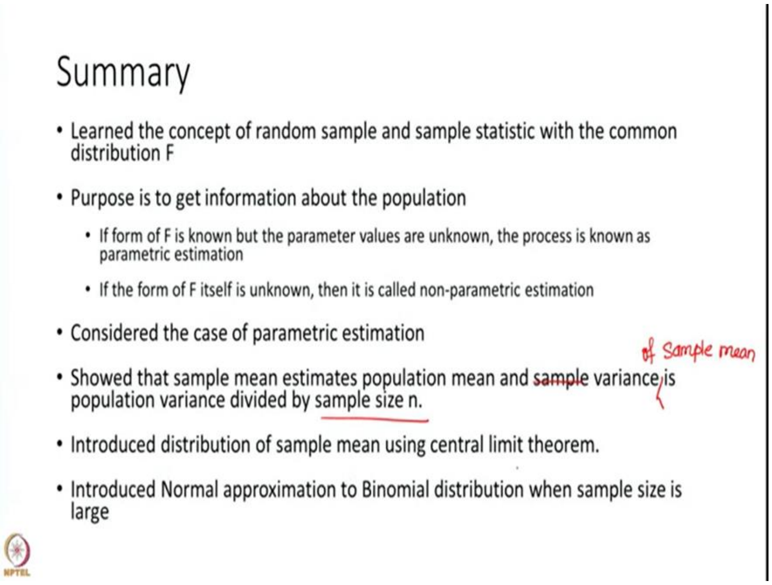
- When $n \to \infty$ *and* $p \to 0$ *such that* $np = \lambda$ *a constant* Binomial distribution can be approximated by exponential distribution

- When $n \to \infty$ Binomial distribution can be approximated by standard normal distribution under central limit theorem

So, what we have done? So, we have approximated binomial distribution by the normal distribution. Now there is a confusion once we approximate it by exponential and now we are approximating it by the normal distribution, so which one to use that is clarified in here.

By stating that if n tends to infinity and p tends to 0 such that np remains constant, then it binomial can be approximated by exponential distribution. But if only n is very large and you have no restriction over p then n can be approximated by a standard normal distribution under central limit theorem. So, with this let us summarize what we learnt today.

(Refer Slide Time: 33:43)



Summary

- Learned the concept of random sample and sample statistic with the common distribution F
- Purpose is to get information about the population
  - If form of F is known but the parameter values are unknown, the process is known as parametric estimation
  - If the form of F itself is unknown, then it is called non-parametric estimation
- Considered the case of parametric estimation
- Showed that sample mean estimates population mean and ~~sample~~ variance of Sample mean is population variance divided by sample size n.
- Introduced distribution of sample mean using central limit theorem.
- Introduced Normal approximation to Binomial distribution when sample size is large

We learned the concept of random sample and sample statistic with the common distribution function F, which is a population distribution function. The whole purpose of this session and the sessions coming up is to get information about the population through the random sample. If you know the form of the distribution function F population distribution function F in that case you know what is the it will be called I am sorry it will be called a parametric distribution function parametric estimation. If you do not know the form of F then it is called a non-parametric estimation.

In this course, we are considering only parametric estimation. We showed that the sample mean estimates the population mean and the variance of sample mean here there is also a correction to be added, let us do that and variance of sample mean is population variance divided by sample size n. we showed that a sample size n increases the variance of sample mean decreases.

We also introduced a distribution of sample mean as a using a central limit theorem and we showed using central limit theorem binomial distribution can be approximated as a normal distribution when the only sample size is large and you have to make a no assumption on the probability of success p being small, thank you.