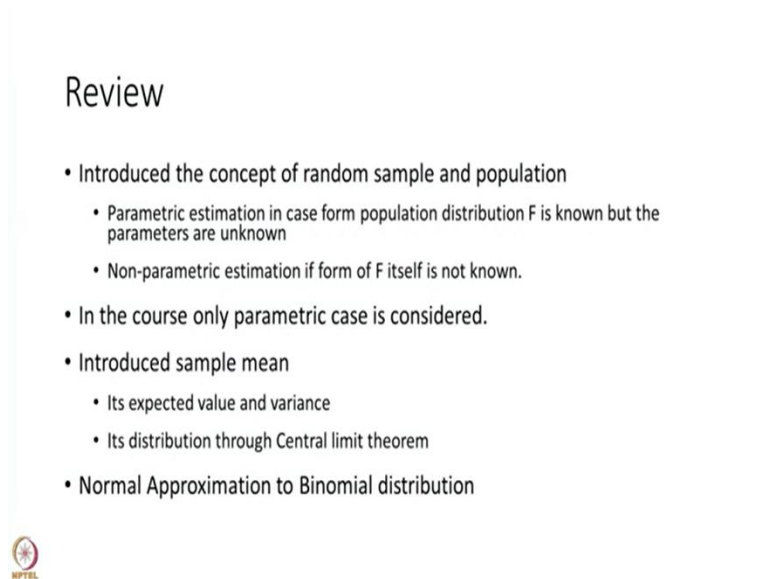


Dealing with Materials Data: Collection, Analysis and Interpretation
Professor. Hina A. Gokhale,
Department of Metallurgical Engineering and Materials Science
Indian Institute of Technology, Bombay
Lecture 48
Sampling Distribution 2


Hello and welcome to the course on Dealing with Materials Data. From the previous session, we have been dealing with area we are leaning about the area on sampling distributions.

(Refer Slide Time: 00:42)

A slide titled "Review" with a list of bullet points. The slide is white with a vertical line on the right side. At the bottom left, there is a small logo for NPTEL.

Review

- Introduced the concept of random sample and population
 - Parametric estimation in case form population distribution F is known but the parameters are unknown
 - Non-parametric estimation if form of F itself is not known.
- In the course only parametric case is considered.
- Introduced sample mean
 - Its expected value and variance
 - Its distribution through Central limit theorem
- Normal Approximation to Binomial distribution



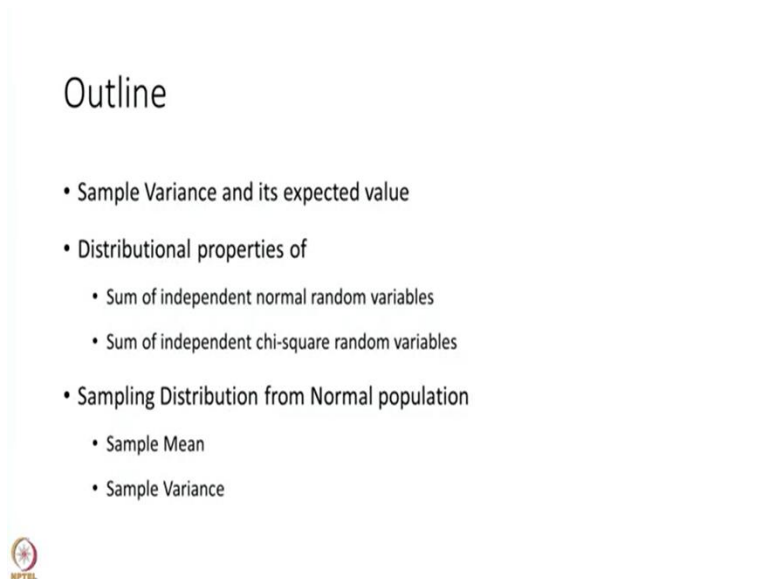
In the previous session, first thing we did was we introduced or sometimes reintroduced the concept of random sample and a population. What is a population and what is a random sample vis a vis a population and we also said that the whole purpose of doing statistics, the reason for following so much of science of statistics is to understand the population through a random sample. We said that if the population distribution function is known to us in the, up to a level of it is form then we say that it a case of parametric estimation.

Where we have to estimate the parameters of the distribution, but if the form of the distribution is not known then we call it a non-parametric case. In the present course, we are going to consider only the parametric case. Then we introduced what is known as sample mean, basically, we would like to understand the parameters of the distribution.

So, we assumed that the population distribution F has a mean μ and a standard deviation σ and then, we introduced what is a quantity statistic called sample mean we found it expected value

and its variance and we found its distribution through central limit theorem. We basically, say that it is expected value is same as the population mean and its variance is the population variance divided by its size of sample and the central limit theorem we said that as if when n is very large the population mean will tend to a normal distribution with as population mean. What I want to say is that the sample mean when n is large the sample size is very large will follow a normal distribution with mean as the population mean and the variance as the population variance divided by the size of the sample. Then we discussed a special case or an example in which we approximated the binomial distribution by normal distribution using central limit theorem. When the n the number of Bernoulli trials in the binomial distribution is very large.

(Refer Slide Time: 03:36)



In the present case, we are going to talk about the sample variance its expected value. Then we will also discuss certain distributional properties of sum of independent normal variables and sum of independent chi square random variables. We will use these properties to derive the sampling distributions form a normal population and we will see that if sample if the sample is drawn from the normal population what is the distribution of sample mean, and what is the distribution sample variance.

(Refer Slide Time: 04:12)

Sample Variance

- Let X_1, X_2, \dots, X_n be independent random variables from a common distribution function F with mean μ and variance σ^2 .
- Sample Variance is defined as

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

$$= \frac{1}{n-1} \left[\sum X_i^2 - n\bar{X}^2 \right]$$

$$(n-1)S^2 = \sum X_i^2 - n\bar{X}^2$$



So let us begin we define sample variance please recall the first few lectures in descriptive statistics it is exactly the same definition. $X_1, X_2, X_3, \dots, X_n$ be an independent random sample from a common distribution F with a mean value μ and a variance σ^2 . Then the sample variance is defined as shown here,

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

$$= \frac{1}{n-1} \left[\sum X_i^2 - n\bar{X}^2 \right]$$

$$(n-1)S^2 = \sum X_i^2 - n\bar{X}^2$$

(Refer Slide Time: 05:02)

Expected value of S^2

$$\begin{aligned}(n-1)E(S^2) &= E\left[\sum X_i^2 - n\bar{X}^2\right] \\ &= E\left(\sum X_i^2\right) - nE(\bar{X}^2) \\ &= nE(X_1^2) - nE(\bar{X}^2)\end{aligned}$$

• Note that for any random variable W : $E(W^2) = Var(W) + (E(W))^2$

$$\begin{aligned}(n-1)E(S^2) &= n\left[Var(X_1) + (E(X_1))^2\right] - n\left[Var(\bar{X}) + (E(\bar{X}))^2\right] \\ &= n\sigma^2 + n\mu^2 - n\left(\frac{\sigma^2}{n}\right) - n\mu^2 = (n-1)\sigma^2 \\ &\quad E(S^2) = \sigma^2\end{aligned}$$



Let us try to find its expected value to with this we get

$$\begin{aligned}(n-1)E(S^2) &= E\left[\sum X_i^2 - n\bar{X}^2\right] \\ &= E\left(\sum X_i^2\right) - nE(\bar{X}^2) \\ &= nE(X_1^2) - nE(\bar{X}^2)\end{aligned}$$

Because X_1, X_2, X_3, X_n all are distributed identically as F with a mean value μ n variance σ^2 . So they are identical, so instead of taking summation of n of them, I can as well take n times expected value of X_1 any one of. This is X_1 I can even take X_2 , so X_1 is not important but what it says is, that the common expected value from the distributed F is being taken. Now we apply the general rule of random variable that expected value of any random variable

$$E(W^2) = Var(W) + (E(W))^2$$

and applying these to each one of these component of expected value of S square we find this and it finally simplifies which you can verify very easily through simple algebra.

$$(n-1)E(S^2) = n\left[Var(X_1) + (E(X_1))^2\right] - n\left[Var(\bar{X}) + (E(\bar{X}))^2\right]$$

$$= n\sigma^2 + n\mu^2 - n\left(\frac{\sigma^2}{n}\right) - n\mu^2 = (n-1)\sigma^2$$

$$E(S^2) = \sigma^2$$

It is expected value of S square that is expected value of sample variance is the population variance.

We will learn in future this called in unbiased estimator of sample variance. When a statistic like S square its expected value is exactly the population variance and S square is the sample variance then it is called the when expected value of S square is equal to the population variance. It is called unbiased estimator S square is an unbiased estimator of sigma square. We will learn about it in future. Now, let us recall few things which we have mentioned in the past and in case it has not been it is the first time let us start it.

(Refer Slide Time: 07:44)

Sum of independent Normal RV's

- Let X_1, X_2, \dots, X_n independent normal random variables with mean $\mu_1, \mu_2, \dots, \mu_n$ and variances $\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2$ respectively. Then

$$\sum_{i=1}^n X_i \sim N\left(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2\right)$$



Let X_1, X_2, \dots, X_n be independent normal random variables with a mean $\mu_1, \mu_2, \mu_3, \dots, \mu_n$ and variance is $\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2$. One way I have to put a comma, so that it is $\mu_1, \mu_2, \dots, \mu_n$ it is $\mu_1, \mu_2, \dots, \mu_n$ and variances is $\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2$ etc. Then summation of these random variable is also distributed as normal with mean as summation of means and variance is summation of σ^2 .

Please remember when the 2 random variables are, they are independent then the sum of variance sum of the random variable, variance of sum of the random variables is sum of the variances. So, this is what we have used here.

(Refer Slide Time: 08:50)

Sum of Independent Chi-square RV's

- Let X_1, X_2, \dots, X_n independent chi-square random variables with degrees of freedom k_1, k_2, \dots, k_n , then

$$\sum_{i=1}^n X_i \sim \chi^2 \left(\sum_{i=1}^n k_i \right)$$



Next we would like to see in the case where $X_1, X_2, X_3, \dots, X_n$ are independent chi square random variables with degrees of freedom $k_1, k_2, k_3, \dots, k_n$.

$$\sum_{i=1}^n X_i \sim \chi^2 \left(\sum_{i=1}^n k_i \right)$$

This should be very obvious because the chi square with k degrees of freedom itself has been defined as a sum of squares of standard normal random variable.

So, the it just the additive nature of the independent chi square random variable comes very naturally. Now let us consider so far what we have been doing we said that f is some distribution with mean μ and variance σ^2 . Now I am defining the form of f and I am saying that it is a normal distribution, so I am saying that now I am taking sampling from the normal population.

(Refer Slide Time: 10:00)

Sampling from Normal Population

- Let X_1, X_2, \dots, X_n be iid $N(\mu, \sigma^2)$

Distribution of Sample Mean

- Sum of independent normal variables is distributed as normal thus, sample mean is distributed as normal with mean and variance as

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

$$E(ax) = aE(x)$$

- and therefore it follows that

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

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} : \text{normalizing r.v. } \bar{X}$$
$$\text{Standardizing r.v. } \bar{X}$$
$$\frac{W - E(W)}{Var(W)} = \text{Normalization}$$

So let X_1, X_2, X_n be independent identical distributed normal random variables with mean μ and variance σ^2 . Then what is the distribution of sample mean? Well sum of independent normal variable is distributed as normal. Therefore, sample mean is also distributed as normal with mean and variance as this, because remember here I think am telling something very obvious expected value of ax is a times expected value of x .

So, this gives you that expected value of

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

Which we have already proved and therefore if the sum of the independent normal variable is distributed as normal. Therefore, sample mean is distributed this and further normalizing the random variable \bar{X} with respect to its mean and variance. We get $\bar{X} - \mu$ divided by $\sigma \sqrt{n}$ is distributed as normal 01. Please remember $\bar{X} - \mu$ divided $\sigma \sqrt{n}$ is called normalizing random variable \bar{X} . It is also known as standardizing, standardizing random variable \bar{X} .

So, for any random variable if you do any random variable W minus expected value of W divided by variance of W is the normalization refers to normalization. So, it says this is of course this is not just normalization it is actually a normal distribution, so this defines the distribution of a sample

mean when the population itself is a normal population. Then the sample mean is distributed normally as mean as a population mean and variance as a population variance divided by size of the sample and if you take $\bar{X} - \mu$ divided by σ / \sqrt{n} it is distributed as normal 0 1.

(Refer Slide Time: 12:53)

Distribution of Sample Variance

$\sum (X_i - \bar{X}) = 0$

• Now, Consider,

$$\sum_{i=1}^n (X_i - \mu)^2 = \sum_{i=1}^n (X_i - \bar{X} + \bar{X} - \mu)^2 = \sum_{i=1}^n (X_i - \bar{X})^2 + n(\bar{X} - \mu)^2$$

$$\therefore \sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma} \right)^2 = \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{\sigma} \right)^2 + \left(\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \right)^2$$

$\underbrace{\left(\frac{X_i - \mu}{\sigma} \right)^2}_{\sim \chi^2(1)} = \underbrace{\left(\frac{X_i - \bar{X}}{\sigma} \right)^2}_{\sim \chi^2(n-1)} + \underbrace{\left(\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \right)^2}_{\sim \chi^2(1)}$

Now if you take a distribution of sample variance, we have to do some calculation in order to understand it, so let us start.

$$\sum (X_i - \mu)^2 = \sum (X_i - \bar{X} + \bar{X} - \mu)^2 = \sum (X_i - \bar{X})^2 + n(\bar{X} - \mu)^2$$

$$\sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma} \right)^2 = \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{\sigma} \right)^2 + \left(\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \right)^2$$

So I have basically, divided this by sigma square and this by sigma square and I get this identity. You see this is very beautiful can you see that because X_i is distributed as normally with mean μ and sigma square $X_i - \mu$ over sigma whole square is a chi-square variate because this itself is a standard normal variate.

You see this, this is distributed as normal 0 1, agreed? This we have already shown that it is distributed as normal 0 1 and these each individual 1 are distributed as normal 01 and therefore we are taking summation and then you are squaring it , you are squaring it.

So, the whole item will become a chi-square and it is only one normal standard normal variate, so it will be chi-square with 1 degrees of freedom. While here you are summing up chi-squares each individual if you look at this whole individual it is distributed as chi-square as 1. They are all independent because Xi's are independent and therefore the summation of n chi-squares will give you chi-squares with n degrees of freedom.

Shall I explain it again? Let us start from

$$\sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma} \right)^2 = \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{\sigma} \right)^2 + \left(\frac{\sqrt{n} (\bar{X} - \mu)}{\sigma} \right)^2$$

$$\sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma} \right)^2 \sim \chi^2(n), \text{ and } \left(\frac{\sqrt{n} (\bar{X} - \mu)}{\sigma} \right)^2 \sim \chi^2(1)$$

Now what am I saying is that this inner part Xi minus mu over sigma, because we have said that Xi is distributed as a normal distribution it is coming from a normal population with mean mu and variance sigma square, we get a Xi minus mu divided by sigma as a standard normal random variable. So, Xi minus mu over sigma is distributed as normal 01. Similarly here we know that X bar is distributed as a normal random variable with mean mu and variance sigma square divided by n therefore X bar minus mu divided by sigma divided by square root n is also distributed as normal 01, in other words standard normal random variate.

Now, we take a square of it, so this is where we take a square of it. So, if you take a square of a one single standard normal variate then it is distributed as chi-square. Here we take n of these standard normal variate and we take square of it and we sum it up. Now remember, each Xi is independent, so Xi minus mu divided as sigma is also independent and therefore Xi minus mu divided by sigma whole square are independent for i is equal to 1, 2, 3, 4, n and therefore you are

summing up n independent chi-square random variates and therefore it becomes chi-square with n decrease of freedom.

The question is, what is the distribution of this? Now if we use the fact that sum of two independent chi-square random variable with decrease of freedom n and m is a chi-square random variable with a degree of freedom n plus m, if we use that very reasonably we can say that this should be distributed as chi-square n minus 1.

I repeat, we know that if the two independent chi-square random variables are distributed with degrees of freedom respectively n and m then the sum of the two random chi-square random variables independent chi-square random variables will be a chi-square random variable with degrees of freedom as sum of the degrees of freedom, so it will be n plus m.

So, if you consider that this is one degree of freedom chi-square random variable which is added into something which gives you a n degree of freedom chi-square random variable in that case we can reasonably understand that this has to be chi-square random variable with n minus 1 degrees of freedom and this is what is our argument.

(Refer Slide Time: 20:33)

Distribution of Sample Variance

$\sum (X_i - \bar{X}) = 0$

• Now, Consider,

$$\sum_{i=1}^n \frac{(X_i - \mu)^2}{\sigma^2} = \sum_{i=1}^n (X_i - \bar{X} + \bar{X} - \mu)^2 = \sum_{i=1}^n (X_i - \bar{X})^2 + n(\bar{X} - \mu)^2$$

$$\therefore \sum_{i=1}^n \underbrace{\left(\frac{X_i - \mu}{\sigma}\right)^2}_{\sim N(0,1)} = \sum_{i=1}^n \underbrace{\left(\frac{X_i - \bar{X}}{\sigma}\right)^2}_{?} + \underbrace{\left(\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma}\right)^2}_{\sim N(0,1)}$$

$\chi^2(n)$ $\chi^2(n-1)$ $\chi^2(1)$

I have written it down again, this is the identity that we have got, is this the same as the previous one. We see that this as we argued before is chi-square n minus chi-square 1 degree of freedom, this is chi-square with n degrees of freedom, sum of two independent chi-square random variable is also chi-square with degrees of freedom as sum of their degrees of freedom. It is reasonably,

reasonable to conclude that the center one is also chi-square with n minus 1 degrees of freedom. So, what it says if you look at this carefully, this says that, this says that the S, let us go up, I again go to arrow we go back and then we use the pen.

So, now what we have is, remember that this quantity is

$$\sum_{i=1}^n \left(\frac{X_i - \bar{X}}{\sigma} \right)^2 \sim \chi^2(n-1)$$

$$\left(\frac{S}{\sigma} \right)^2 \sim \chi^2(n-1)$$

S square divided by sigma square and we are saying that this is distributed as chi-square n minus 1 degrees of freedom. Here we are saying that S square over sigma square is distributed as chi-square n minus 1 degrees of freedom. So, quickly if we see, we saw that if you assume that the population distribution is normal distribution with mean mu and variance sigma square then the sample mean is distributed, also as a normal distribution with a mean mu and variance as sigma square divided by n, the size of the sample.

And the sample variance in that case is distributed as a chi-square distribution with n minus 1 degrees of freedom when it is divided by sigma square. In other words this also shows that expected value of S square, here there is n minus 1, now it makes sense expected value of S square becomes sigma square, which is what we had shown earlier also. Here there should be n minus 1 because this divided by n minus 1 is S square. So, S this is n minus 1 time S square divided by sigma square.

(Refer Slide Time: 23:47)

t distribution

- Recall that t distribution with n degrees of freedom defined as

$$t = \frac{Z}{\sqrt{Y/n}}$$

where, $Z \sim N(0,1)$ and $Y \sim \chi^2(n)$ and that they are independent.

Hence, we have

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

Normalized sample mean

$$Y = (n-1) \frac{S^2}{\sigma^2} \sim \chi^2(n-1)$$

Therefore,

$$t = \sqrt{n} \frac{(\bar{X} - \mu)}{S} = \frac{(\bar{X} - \mu)}{S/\sqrt{n}} \sim t(n-1)$$



This brings us to another distribution, you remember we introduced a t distribution. In t distribution we said that if you if there is if z is a standard normal variate and y is a chi-square random variable with n degrees of freedom.

$$t = \frac{Z}{\sqrt{Y/n}}$$

where, $Z \sim N(0,1)$ and $Y \sim \chi^2(n)$ and that they are independent.

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

$$Y = (n-1) \frac{S^2}{\sigma^2} \sim \chi^2(n-1)$$

Therefore,

$$t = \sqrt{n} \frac{(\bar{X} - \mu)}{S} = \frac{(\bar{X} - \mu)}{S/\sqrt{n}} \sim t(n-1)$$

Please recall, I mean look at a certain similarity, similarities with this definition of random variable and this definition of random variable.

You see that when sigma is unknown in future we are going to do that if the population variance is not known, population standard deviation is not known. Then if you replace it by its estimated value which is sample variance or sample standard deviation then instead of a normal distribution the standardized or a normalized random variable having a standard normal distribution, it will have a t distribution with n minus 1 degrees of freedom.

Again, this is what we are going to use in future with respect to interval estimation as well as we are going to use it with respect to hypothesis testing. So, please remember what I have I said is that if you have a sample n sample of size n from a normal distribution with mean mu and standard deviation sigma or variance sigma square then the sample mean minus mu divided by sigma divided by square root n that is the standardized or a normalized value of X bar.

This is a normalized sample mean that is distributed as normal 0,1 but in case sigma is not known and you replace sigma by the sample standard deviation then the same normalized sample mean with a estimate of or the estimate of a population standard deviation it becomes a t distribution with n minus 1 degrees of freedom.

(Refer Slide Time: 27:45)

Summary

- Introduced sample variance, its expected value
- Under assumption of Normal Population
 - Sample mean is distributed as Normal with mean μ and variance $\frac{\sigma^2}{n}$ (where $\frac{\sigma^2}{\sqrt{n}}$ is crossed out)
 - Sample variance is distributed as chi-square distribution with $n-1$ degrees of freedom ($\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$)
- The ratio $\frac{(\bar{X}-\mu)}{S/\sqrt{n}}$ is distributed as t distribution with $(n-1)$ degrees of freedom

So now let us summarize, we first introduced here the standard the sample variance and its expected value by not assuming any form of the distribution, we only said that the population distribution is F with a common mean μ and a standard deviation σ .

Then, we made an assumption that the population is a normal population with mean μ and a variance σ^2 . Then we said that the sample mean is distributed also as a normal distribution with mean μ and variance σ^2/n , this should be σ^2/n , there is a mistake here, please correct it. It should be σ^2/n then, we found that sample variance is distributed as a chi-square distribution with $n-1$ degrees of freedom with certain multiplication please remember.

And I think I should make correction here also because it gives a wrong impression and this should not happen. What we really mean to say is that $(n-1)S^2/\sigma^2$ is distributed as chi-square, chi-square $n-1$ degrees of freedom.

Please make this correction, sorry for this mistake and then we revisited the t distribution by stating that the ratio of sample difference between sample mean and the population mean to the sample variance by square root n is distributed as a t distribution with $n-1$ degrees of freedom. Next we will consider further values on sampling distribution, thank you.