


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 34
Special Random Variables 4


Hello and welcome to the course on dealing with materials data. We are going to continue our sessions from the previous few sessions on Special Random Variables. Let us review what we have done in the past.

(Refer Slide Time: 00:35)

Review

- Discrete Distributions
 - Uniform
 - ~~Brenoulli~~ Bernoulli
 - Binomial
 - Geometric
 - Negative Binomial
 - Poisson
 - Hypergeometric
- Continuous Distributions
 - Uniform
 - Normal
 - Chi square
 - t distribution
 - F distribution





We first considered the discrete random variables, which have a very special distribution. The first one we considered was the uniform, discrete uniform random variable, then we considered Bernoulli trial I see that there is a spelling mistake in Bernoulli trial it is should be this way, let me just correct it, it should be Bernoulli trials.

Then we found that there are three distributions, which come out of repeated Bernoulli trials, the first one is a binomial distribution, which has, over here. So, first is the binomial distribution, where you carry out n independent Bernoulli trials and look for the number of successes or number of trials, which has resulted in success.

Then comes the Geometric distribution, in which you wait till the first success occurs in your trial. So, you carry out number of trials still the first successes encountered. Negative

binomial is a rather generalization of Geometric distribution in which you try to get the X , you want to get a probability of conducting X successive trials to get exactly n number of successes. So, you have to conduct X independent Bernoulli trials until you get the n th successful trial.

Then we came to Poisson distribution and we showed that a Poisson distribution occurs in the case where the probability of any occurrence is very small, when the sample value is very large. So, when a Binomial distribution has a very large n , but a small probability of success, in such a way that n multiplied by P , that is the number of trials multiplied by the probability of success remains constant.

It tends to Poisson distribution. You will see or you might have already seen in the tutorials and R sessions that are being conducted that Poisson occurs also during the nucleation of atoms in the physics, in the field of physics, so there also the Poisson distribution is useful.

We also saw then the distribution which is called Hypergeometric distribution, in which we understand that there are m items of which n items have a special characteristic maybe they are defective or they are certain kind of atoms or they are certain kind of elements. So, it has a special characteristic.

So, there are m number of total items of which n number has a special characteristic, characteristic and you are drawing a sample of size X , and you want to estimate a probability that exactly k of them will have those characteristics. In that case, X follows Hypergeometric distribution.

And we gave an example of 3d atom probe filled ion microscopy in which most of this except for Poisson most of the distributions are covered, Poisson as I said before, it has been given separately in R session, giving you certain examples in material science and materials engineering for Poisson distribution. It is very interesting that even in that R session this the same 3d atom probe filled ion microscopy example is discussed in more details.

Then we moved on to Continuous distribution and we introduced first Continuous Uniform distribution. And we mentioned that it has a very special importance when it comes to

generating random numbers from different distributions using pseudo random number generator and pseudo random number generator actually tries to generate the uniform distribution, the variates of uniform distribution.

Then we introduced a Normal distribution and we gave some distributions which I derived from the normal distribution, which are Chi square, which are Chi square, t distribution and F distribution, when we move on and we look into the inference in statistics, these distributions are going to play a very significant role.

(Refer Slide Time: 06:11)

Outline



- Importance of Normal Distribution
- Central Limit Theorem
- Steps to demonstrate Central Limit Theorem



Importance of Normal Distribution

- Galileo observed that observations taken under the identical conditions tend to vary....200 years later Gauss established that the error in observation vary as normal distribution or Gaussian Distribution
- Central Limit Theorem (CLT) states that if independently distributed random variables X_1, X_2, \dots, X_n with finite mean μ_i and finite variance σ_i^2 for $i = 1, 2, \dots, n$, then as $n \rightarrow \infty$, $X = \sum_{i=1}^n X_i$ is distributed as normal distribution with

$$\text{Mean} = \underline{\mu} = \sum_{i=1}^n \underline{\mu}_i$$

$$\text{Variance} = \underline{\sigma^2} = \sum_{i=1}^n \underline{\sigma_i^2}$$



So we come to our present session, we want to cover the importance of Normal distribution, the Central Limit Theorem and steps to demonstrate central limit theorem. So, where is the importance of central limit theorem? Well, many years ago when Galileo was taking observations from the out in this space, and he was observing stars and constellation, under the exactly same circumstances, he found that the observations did not have, were not exactly identical.

He was actually unhappy with this fact. But it took 200 years for Gauss to come and establish that such an error, which can be attributed to human error or to machine error is a common factor and these errors are generally distributed as a normal distribution, we call this Normal distribution also as a Gaussian distribution to give a respect to Gauss who realize this particular distribution.

Now, what is central limit theorem? central limit theorem actually says that, when you have large number of random variables of observed independently from nearly identical distributions, nearly in the sense that the parameter value may not be the same, but the distributions are same, then the mean of these largely observed random variable also follows, no matter what is the original distribution of the random variable.

The mean of these observations follow Normal distribution.

So, here it is exactly stated,

- X_1, X_2, \dots, X_n with finite mean μ_i and finite variance σ_i^2 , then as $n \rightarrow \infty$,

$X = \sum_{i=1}^n X_i$ is distributed as normal distribution with

$$\text{Mean} = \mu = \sum_{i=1}^n \mu_i$$

$$\text{Variance} = \sigma^2 = \sum_{i=1}^n \sigma_i^2$$

Let us try to understand this. we are not saying that $X_1, X_2, X_3 \dots X_n$ follow normal distribution, we are saying that it may come from any distribution, only condition we are putting is that it has a finite mean and a finite variance, particularly finite variance. Yes,

there are distributions which have infinite means or infinite variance. So, we are ruling out all such distributions and then we are saying that if you take sum of all these independently distributed random variables X_i and if the sample size become extremely large, in that case, this summation of the variable, random variable, which is also a random variable let us call it X , then X follows a normal distribution, here is what we have to realize that it follows a normal distribution with a mean μ and variance σ^2 , μ is a summation of all the μ_i and σ^2 is summation of all the σ_i^2 .

(Refer Slide Time: 10:04)

Central Limit Theorem (CLT)

- Central Limit theorem can be proved using Characteristic Function and Taylor series expansion.
 - Here characteristic function of any random variable is defined as

$$E\left(e^{i\sqrt{-1}tX}\right) = E\left(e^{itX}\right)$$
- This shows that no matter what may be the distribution, sum of independent random variables, when n is large, is distributed as normal random variate with mean as sum of its means and variance as sum of its variances. (under the assumption that the underlying distribution has finite variance.)

Now, let us look into the proof, as to generally how it goes. The Central Limit Theorem is proved using characteristic function and Taylor series expansion, characteristic function so far in our course we have not introduced however, we have introduced what is called Moment Generating Function. Do you recall? The Moment Generating Function of any distribution we had introduced.

And this is a another version, sort of another version of a Moment Generating Function. And it is called Characteristic Function. It has the same property as Moment Generating Function that is its first derivative with the limiting value at t is equal to 0. Will give you the first moment, the second derivative will give you the second moment, these are all the raw moments, these are not central moments.

Only difference is that this particular distribution has an imaginary part in it, which is i , which is the square root of minus 1. If you use this expansion for any random,

$$E(e^{(\sqrt{-1})tX}) = E(e^{itX})$$

summation of any random variable and use, use this as expected value and use the Taylor series expansion and then let n go to infinity, you will find that all of them follow normal distribution.

I am going to skip the proof, proof is available in variety of textbooks, you can go through it. The people who would like to pursue this kind of Mathematical Statistics further they should go through the proof, because taking a Characteristic function and using a Taylor series expansion before letting n tend infinity is a very common technique used for proving many theorems in Mathematical Statistics.

So, here it is repeated again that no matter what may be the original distribution of the random variable, sum of the independent random variables, when n is large is distributed as normal random variate with mean as sum of its mean and variance as sum of its variance with only one assumption that underlying distribution has a finite variance.

(Refer Slide Time: 13:15)

Another version of CLT

- A more simple version of CLT states that if X_1, X_2, \dots, X_n are n independent and identically distributed random variables with mean μ and finite variance σ^2 , then as $n \rightarrow \infty$

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

\downarrow
 $\sim N(0,1)$

$E(\bar{X}) = \mu$
 $\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$
 $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$

where, $Z \sim N(0,1)$

- This can be easily demonstrated using R or even using Excel, as shown in the following steps.



So, this also has a very specific version, which is a simpler version, which you might have come across in your other courses of statistics and this is, it states that $X_1, X_2, X_3 \dots X_n$ are n independent and identically distributed random variables. Remember that time we did not say they are identically distributed, they were independent and they all had a mean μ_i and variance σ_i^2 .

Here, we are saying that they are identically distributed random variable with mean μ and finite variance σ^2 . Then as n tends to infinity, the standardized normal variate.

$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

Let us do some revision here as well before we say that it tends to normal distribution. Please recall, that expected value of \bar{x} is μ and the variance of \bar{x} in such situation is σ^2 / \sqrt{n} and therefore, a standard normal variate Z would be defined as

$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

So, this is what is written here that $Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$ the probability that this variate is less than t is approximately equal to probability that a Z which is no $N(0, 1)$ distributed random variable.

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

So, Z is less than t where Z is normal, standard normal variate, this can be easily demonstrated also by using R which will be done in your R sessions or even using Excel, which is what I would like to show you here that quickly without using R even by using Excel and generating random numbers, you can carry out the same, you can work out the sort of a demonstration proof of this particular version of Central Limit Theorem.

(Refer Slide Time: 15:48)

- Let $X_1, X_2, \dots, X_n \sim \text{Chisquare}(k) (= \chi^2(k))$, then as per CLT

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

as $n \rightarrow \infty$, where, $Z \sim N(0,1)$

- To demonstrate this let $n = 500$ and $k = 5$, and follow the steps below:

- Generate U_1, U_2, \dots, U_{500} using `Rand()` function
- Copy and Value paste U_1, U_2, \dots, U_{500} : Y
- Apply function `CHISQ.INV(U, 5)` to generate $\chi^2(5)$ random variable $X_i, i = 1, 2, \dots, 500$
- Calculate average \bar{X}
- Calculate $Z = \frac{\bar{X} - 5}{\sqrt{10}/\sqrt{500}}$ (note that mean and variance for $\chi^2(5)$ are 5 and $2*5 = 10$ respectively)
- Repeat steps 1 to 5 100 times to generate Z_1, Z_2, \dots, Z_{100}
- Plot histogram for Z_1, Z_2, \dots, Z_{100}

$X \sim \text{any r.v.} \sim \chi^2(5)$

$Y = F(X) \sim \text{Uni}(0,1)$

$X = F^{-1}(Y)$ $F = \chi^2$

$F^{-1} = \chi^{-2}$

$\mu = E(X) = k = 5$

$\sigma^2 = \text{Var}(X) = 2k = 10$ } when $X \sim \chi^2(5)$



So, let us see how that happens. So, here I have given the step by step method to follow. So, here what I am trying to show is that I am taking $X_1, X_2, X_3 \dots X_n$ as a Chi square random variable with k degrees of freedom. Why am I taking Chi square? Chi square is a skewed distribution. And remember Normal is a symmetric distribution.

So, we want to show that we take a skewed distribution and then the Central Limit Theorem applies how it comes very close to the Normal distribution, how it actually turns into a Symmetric distribution. So, we want to show that if $X_1, X_2 \dots X_n$ follow Chi square distribution with k degrees of freedom, then the probability of the standardized variable of \bar{X} . that is

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

As n tends to infinity, where Z is a standard normal variate. So, to demonstrate for time being we will take n to be 500 and k degrees of freedom of Chi square to be 5, and then we take the following steps, we first generate uniform random variables using Rand function, 500 of them.

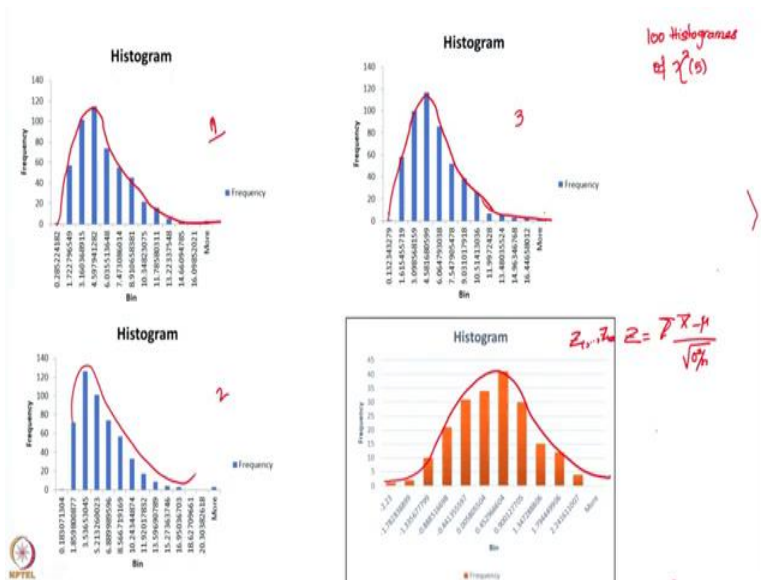
Then, we copy the value and value paste it here, because with every operation on the Excel sheet, these numbers tend to change, these numbers tend to change. If you do not continue with these function, these values tend to change. So we will copy and value paste it, then we apply the Chi squared inverse function. You please recall what we did, what we said about generating a random variate, random variate distributed by any distribution, so, if you recall, we said that if X is distributed as any random variable, any random variable, then $F(X)$ is distributed as $U(0, 1)$ random variate.

So, if you call this Y , then you can generate X by saying that it is F inverse of Y . So, if you have Y , you can find X and this is exactly what we are doing. Here we have generated sort of Y for $U_1, U_2, U_3 \dots U_{500}$. Then we apply Chi square inverse function. So, our F is actually Chi square function. So, F inverse is Chi square inverse function and therefore, I am applying Chi square inverse function with a probability, this is now a probability, remember it is coming from uniform random variable, uniform 0 to 1.

So, this is probability and this is the, what we have taken as degrees of freedom of Chi square distribution and we regenerate Chi square five random variables X_i . So, this if I had said that this is X_i Chi squared 5 then, I am generating $X_1, X_2, X_3 \dots X_{500}$ for X_5 . Now I calculate \bar{X} and I also calculate Z is equal to $\bar{X} - 5$; Remember that, expected value of X in this case will be k , which is 5 and variance of X is equal to $2k$, which is 10.

When X sorry, X is distributed as Chi square 5. So, when that happens, this is the, your expected value, which is μ , and this is sigma square. So, I am taking exactly this ratio, I am carrying it out here. So if this is a , this is what I am calculating here. Repeat the steps 1 to 500 times. So, this 500 data point is n , this we are doing hundred times so that we can have a nice histogram to generate $Z_1, Z_2, Z_3 \dots Z_{100}$ and then plot the histogram and let us see how does it look like?

(Refer Slide Time: 21:11)



So, you see here, I have shown one, two and three sample histograms out of hundred, we generated hundred histograms, actually, hundred histograms of Chi square 5. So I have taken a sample of one, two and three and you can see that how skewed they look, you see the distributions are all skewed. Please note that distribution is heavily skewed. I have not drawn it beautifully. Let me correct it myself so that it looks better, let us do it again, this is better.

So, this is a typically skewed Chi square distribution. Same is true here, but when you take the Z variable which is, summation or which is

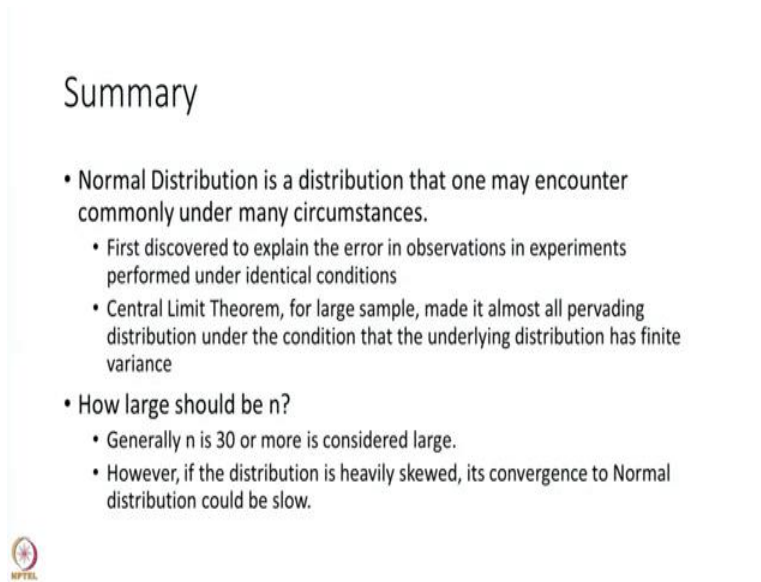
$$\frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

then this Z variable, this is a plot of a Z1 to Z100 and this distribution is looks, a very much of a Normal distribution.

You see how this skewness has slowly, here also I have not drawn it well, let me do that, so if you draw a bit carefully so, this skewed distribution, this skewed distribution as well as this skewed distribution are turning into a Normal distribution, which is not skewed.

You can carry out this exercise further when you learn different graphical ways of testing the two distributions are same. So you can plot this particular graph, which is what is known as the PP plot or a QQ plot. I believe you might have already been introduced to this in your R sessions during Descriptive Statistics, if you have then you can generate this data and plot it like one of those QQ or PP plots and you will see how close it is to normal.

(Refer Slide Time: 24:08)



Summary

- Normal Distribution is a distribution that one may encounter commonly under many circumstances.
 - First discovered to explain the error in observations in experiments performed under identical conditions
 - Central Limit Theorem, for large sample, made it almost all pervading distribution under the condition that the underlying distribution has finite variance
- How large should be n ?
 - Generally n is 30 or more is considered large.
 - However, if the distribution is heavily skewed, its convergence to Normal distribution could be slow.

So, let us summarize. We started this session by saying that why Normal distribution plays a very central and important role, because we encounter it very commonly under many circumstances. It was first discovered to explain the error in observations in the experiments performed under identical condition and we still observe that.

The Central Limit Theorem, for a large sample made it almost all pervading distribution because as far as the underlying distribution had a finite variance and you had a large number of observations coming from the underlying distribution which are identical in nature, independent in nature, in that case, if you take either the mean value or the summation of it, it tends to follow normal, standard normal variate with common mean, mean as a summation of mean and standard deviation as a summation of standard deviation.

The question is how large should be n ? Because when we think about it practically, mathematically it is wonderful to say n tends to infinity and you show the, or you prove

the theorem, but in reality, the question comes, what is really a large n ? Well the answer is that for generally, naturally less skewed distributions. If n is 30 or more, it is considered large enough, but if your distribution is heavily skewed in that case convergence to distribution could be slow and you may need very large number of samples.

So, that time 30 may not be a good large enough, it may go even further to something like 50, 75, 100 you have to find out how fast it converges. So, with this we complete this session on normal distribution and next we want to take all those distributions which we encounter most commonly in the engineering data and in particular for materials engineering and material science metallurgical data. So, with this we end this session,

thank you.