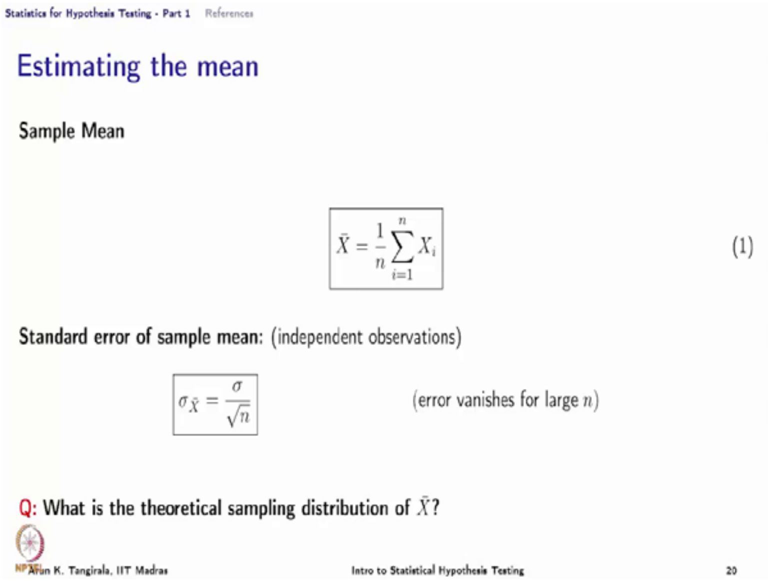
Introduction to Statistical Hypothesis Testing
Prof. Arun K. Tangirala
Department of Chemical Engineering
Indian Institute of Technology, Madras

Lecture – 07
Statistics for Hypothesis Testing

(Refer Slide Time: 00:01)



Statistics for Hypothesis Testing - Part 1 References

Estimating the mean

Sample Mean

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

Standard error of sample mean: (independent observations)

$$\boxed{\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}}$$

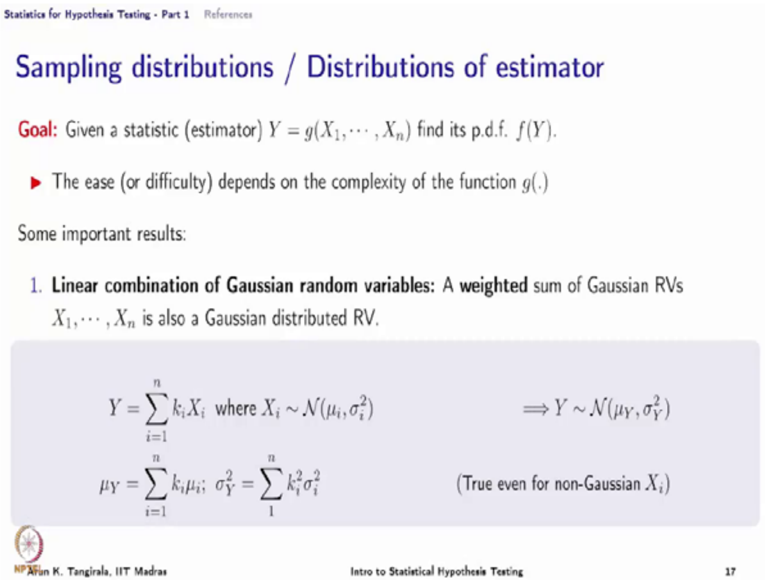
(error vanishes for large n)

Q: What is the theoretical sampling distribution of $\bar{X}$?

Arun K. Tangirala, IIT Madras Intro to Statistical Hypothesis Testing 20

So, let us now move on. Those are the standard results that we have.

(Refer Slide Time: 00:07)



Statistics for Hypothesis Testing - Part 1 References

Sampling distributions / Distributions of estimator

Goal: Given a statistic (estimator) $Y = g(X_1, \cdots, X_n)$ find its p.d.f. $f(Y)$.

- The ease (or difficulty) depends on the complexity of the function $g(.)$

Some important results:

1. **Linear combination of Gaussian random variables:** A **weighted** sum of Gaussian RVs $X_1, \cdots, X_n$ is also a Gaussian distributed RV.

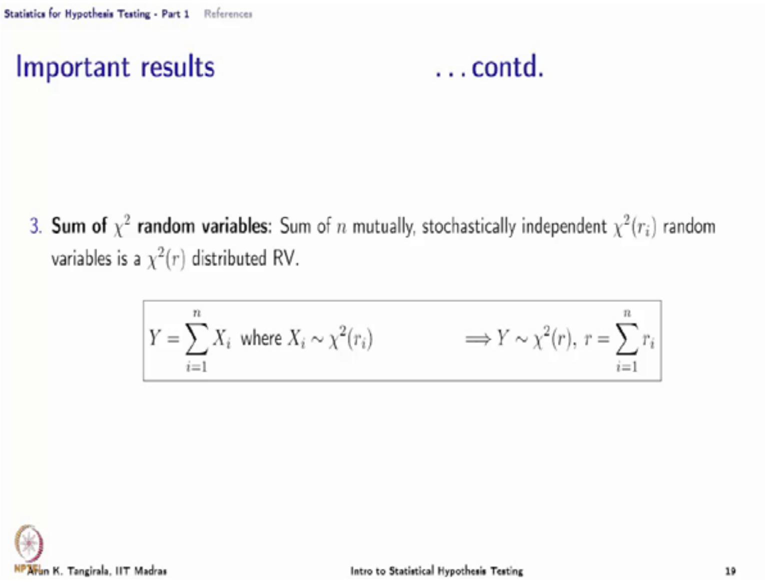
$$Y = \sum_{i=1}^{n} k_i X_i \text{ where } X_i \sim \mathcal{N}(\mu_i, \sigma_i^2) \qquad \Longrightarrow Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$$

$$\mu_Y = \sum_{i=1}^{n} k_i \mu_i; \; \sigma_Y^2 = \sum_{1}^{n} k_i^2 \sigma_i^2 \qquad \text{(True even for non-Gaussian } X_i\text{)}$$

Arun K. Tangirala, IIT Madras Intro to Statistical Hypothesis Testing 17

One was the linear combination of Gaussian, other was a linear combination of squared Gaussian and the other third one was the linear combination of chi square distributed random variables. So, all of them are assumed to be independent.

(Refer Slide Time: 00:16)



Statistics for Hypothesis Testing - Part 1 References

Important results ...contd.

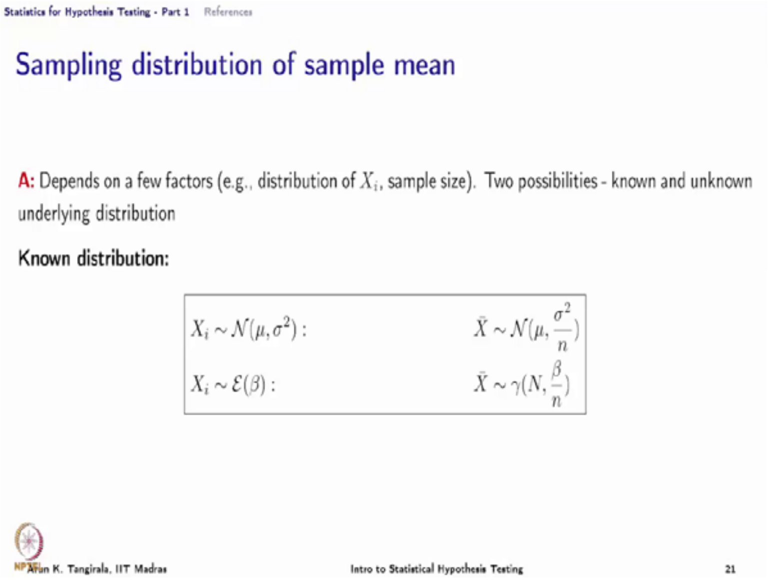
3. **Sum of χ^2 random variables:** Sum of n mutually, stochastically independent $\chi^2(r_i)$ random variables is a $\chi^2(r)$ distributed RV.

$$\boxed{Y = \sum_{i=1}^{n} X_i \text{ where } X_i \sim \chi^2(r_i) \qquad \Longrightarrow Y \sim \chi^2(r), \; r = \sum_{i=1}^{n} r_i}$$

Arun K. Tangirala, IIT Madras Intro to Statistical Hypothesis Testing 19

The thing that you should have seen straight away is in all these 3. We are only looking at linear operations, and therefore, we were able to derive the distribution results very easily. If I had performed some non-linear and of operation on these variables, then you could have been difficult to provide an answer, alright. So, let us look at the distribution of mean, a part of which sample mean; a part of which we have already discussed. So, some parts we will just breeze through and then we will conclude this lecture.

(Refer Slide Time: 01:01)



Statistics for Hypothesis Testing - Part 1 References

Sampling distribution of sample mean

A: Depends on a few factors (e.g., distribution of X_i, sample size). Two possibilities - known and unknown underlying distribution

Known distribution:

$X_i \sim \mathcal{N}(\mu, \sigma^2)$:	$\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$
$X_i \sim \mathcal{E}(\beta)$:	$\bar{X} \sim \gamma(N, \frac{\beta}{n})$

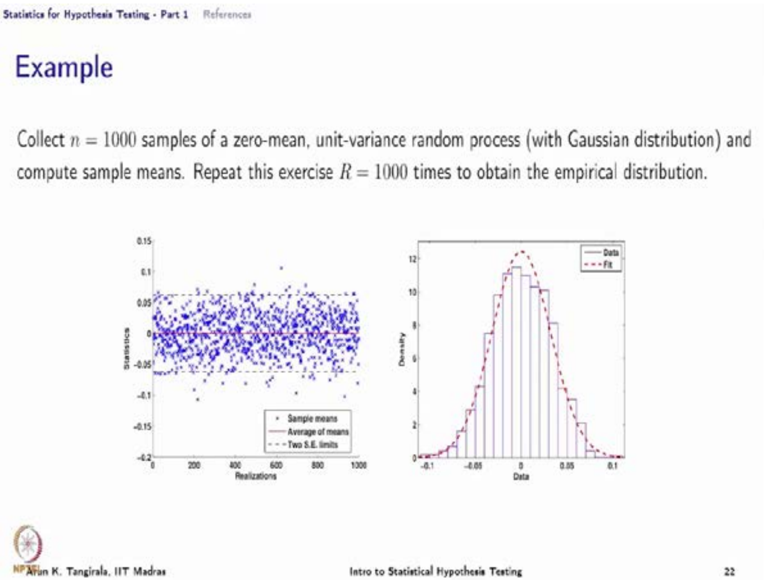
Arun K. Tangirala, IIT Madras Intro to Statistical Hypothesis Testing 21

Suppose, so now assume that we have decided to use sample mean as a test statistic for all our hypothesis test. In fact, that is going to be the case in this course. Then, we have already derived the results that I am showing you here that the standard error of sample mean under the independence assumption is sigma over root n. Again what, left hand side is sigma X bar and on the right hand side you have sigma X. One is that of the estimate, the left hand side of the statistic. And, on the right hand side we have sigma for the process. So, be careful with that.

Of course, now what is the theoretical sampling distribution of X bar? We have answered part of that question. For the case of X bar, sorry, X falling out of a Gaussian distribution. So the first equation here, the first result is, sorry, has been derived already on the board for us.

What if X falls out of an exponential distribution? Then X bar does not follow the Gaussian distribution, in fact it follows a gamma distribution. That is the interesting part of it. So you can see that for small, that is, now we are looking at finite n. When we go to large n, again all distributions tend to Gaussian. That is the beauty of Gaussian. Everything, it pulls all these distributions towards itself; whether it is a binomial or whether it is a chi square or whether it is a (Refer Time: 02:28) or a gamma, whatever it is, all distributions tend to take the shape of a Gaussian as n becomes large. So, these results are for finite n, alright. So now, remember this is the case when I know the distribution of X. Remember that. For unknown distributions, is something we will discuss it shortly.

(Refer Slide Time: 02:50)



Statistics for Hypothesis Testing - Part 1 References

Example

Collect $n = 1000$ samples of a zero-mean, unit-variance random process (with Gaussian distribution) and compute sample means. Repeat this exercise $R = 1000$ times to obtain the empirical distribution.

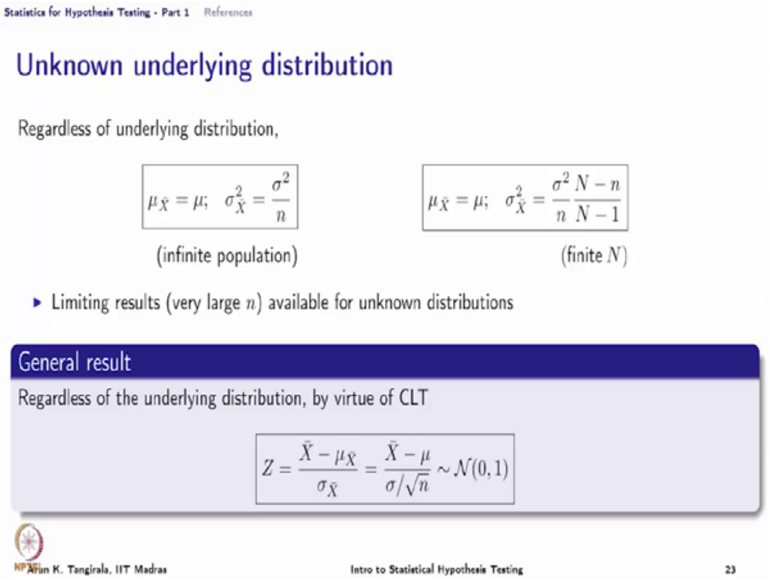
Arun K. Tangirala, IIT Madras Intro to Statistical Hypothesis Testing 22

This is just an illustration of the result for the case of Gaussian distributed observations. This is something that we have gone through already. On the left-hand side, in the plot I am showing you the estimates that I have obtained from 1000 replicates. And, they all seem to fall within this band. This band is the 99 percent probability band.

And, on the right-hand side you see a histogram. And, I would fit a Gaussian distribution there to show that yes, X bar follows a Gaussian distribution. You can now go and check for finite n and small n. Do not take 1000 samples. If you take 1000 samples, again whereas the observations fall out of a Gaussian or non-Gaussian distribution, you will always see a Gaussian distribution. In order to check or verify the second result here, generate observations from an exponential distribution; may be take 30 samples, 30 observations or 20 observations. And, go repeat that exercise with the use of replicate, plot a histogram and you will see that follows a gamma distribution one.

Observation that you may see here is that the mean here is a mu, alright; whereas, a gamma distribution is not given by the mean and variance. So, you have to be really careful. The numbers here are not the mean and variance of the gamma distribution. The gamma distribution is characterized by some other parameters. Eventually expectation of X bar, regardless of the distribution will still be the mean. That is something that you should remember. Do not get confused between the notations here for different distributions.

(Refer Slide Time: 04:34).



Statistics for Hypothesis Testing - Part 1 References

Unknown underlying distribution

Regardless of underlying distribution,

$$\boxed{\mu_{\bar{X}} = \mu; \quad \sigma^2_{\bar{X}} = \frac{\sigma^2}{n}}$$

(infinite population)

$$\boxed{\mu_{\bar{X}} = \mu; \quad \sigma^2_{\bar{X}} = \frac{\sigma^2}{n}\frac{N-n}{N-1}}$$

(finite N)

- Limiting results (very large n) available for unknown distributions

General result

Regardless of the underlying distribution, by virtue of CLT

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

NPTEL Arun K. Tangirala, IIT Madras Intro to Statistical Hypothesis Testing 23

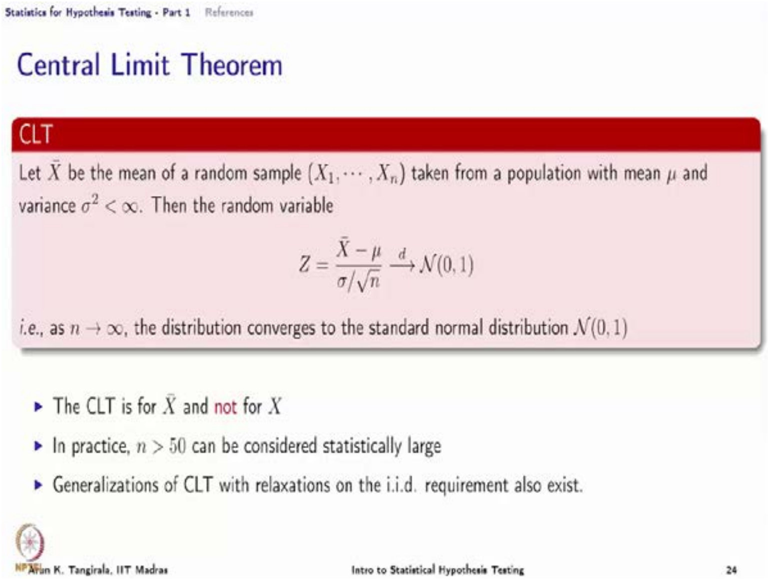
What about now unknown underlying distribution? Now, we have already said regardless of the underlying distribution, mean is mean. And variance, mean of X bar is a mean of X and variance of X bar is variance of the process by n. There we do not need to know the distribution. This is for the infinite populations and independence assumption as usual.

Suppose, the population size is infinite; we have been assuming that the population size is infinite. That means there are so many possibilities, infinite number of possibilities. If the population size is finite, there are many situations such as that and then the mean expression does not change. But, the variance expression takes a different route.

And, in fact you can perhaps see that as n goes to infinity, that is, the size of the population goes to infinity, you will recover the expression on the left. So, normally we will deal with infinite populations. But, you have to be watchful. If your application belongs to the right-hand side, then you should use this variance expression.

Now, what about the distribution of X bar when X falls out of an unknown distribution? Typically, that is a scenario. You may not know. But sometimes, yes, you may know up front. For example, if you have collected large data and you want to know what distribution the data is coming out of, you can actually take a histogram. Plot a histogram of the data; get a feel of the distribution. Let us say that I do not know the distribution. What happens to the distribution of X bar?

(Refer Slide Time: 06:14)



Statistics for Hypothesis Testing - Part 1 References

Central Limit Theorem

CLT

Let $\bar{X}$ be the mean of a random sample $(X_1, \cdots, X_n)$ taken from a population with mean μ and variance $\sigma^2 < \infty$. Then the random variable

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

i.e., as $n \to \infty$, the distribution converges to the standard normal distribution $\mathcal{N}(0, 1)$

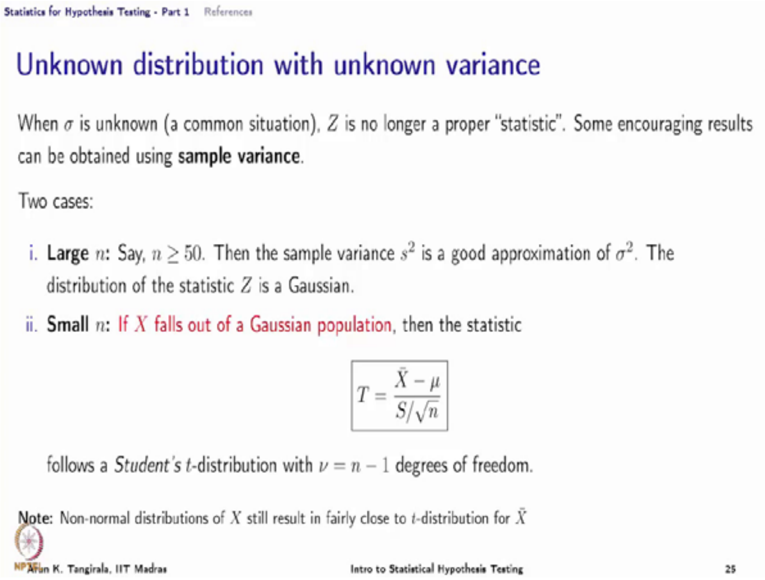
- The CLT is for $\bar{X}$ and not for X
- In practice, $n > 50$ can be considered statistically large
- Generalizations of CLT with relaxations on the i.i.d. requirement also exist.

NPTEL Arun K. Tangirala, IIT Madras Intro to Statistical Hypothesis Testing 24

Here is where the central limit theorem kicks in, which says that when you add up; we have already seen this before. When you add up n observations n coming out of a random sample then and each of them are identically distributed; that means, every observation falls out of the same distribution with the same mean and same variance. Then, the standardized X bar will follow a Gaussian distribution; standard Gaussian distribution. Now, there are some relaxations to this central limit. This is the original version of central limit theorem; where X 1 to X n is set to be falling out of an i.i.d. independent and identically distributed family. But, what if they are not coming out of an identical distribution?

What if X's are not independent? Still, the result mildly holds the Gaussian distribution. Still, it is valid. The only difference is the variance expressions would be different and perhaps the mean expression would be a bit different. But, in terms of distribution it still becomes X bar still tends to have a Gaussian distribution. And, a CLT is for X bar and not for X. Just as a reminder. And, what you have here is, therefore the same results straight away applied to the sample mean. So, CLT actually can be applied straight away to the sample mean, so that the standardized sample mean follows a Gaussian distribution.

(Refer Slide Time: 07:47)



Statistics for Hypothesis Testing - Part 1 References

Unknown distribution with unknown variance

When σ is unknown (a common situation), Z is no longer a proper "statistic". Some encouraging results can be obtained using **sample variance**.

Two cases:

i. **Large** n: Say, $n \geq 50$. Then the sample variance s^2 is a good approximation of σ^2. The distribution of the statistic Z is a Gaussian.

ii. **Small** n: If X falls out of a Gaussian population, then the statistic

$$\boxed{T = \frac{\bar{X} - \mu}{S/\sqrt{n}}}$$

follows a *Student's* t-distribution with $\nu = n - 1$ degrees of freedom.

Note: Non-normal distributions of X still result in fairly close to t-distribution for $\bar{X}$

Arun K. Tangirala, IIT Madras Intro to Statistical Hypothesis Testing 25

Now, finally, what about unknown distribution with unknown variance; until now, all along we have been assuming that the variance is known, alright. But, what if I do not know the variance? Which is also the more practical situation; we started from an ideal case, where we said I may know the distribution of X and I may know the variance of X. What is the problem if I do not know the variance? Then, this X bar minus mu by sigma root n is no longer a statistic because for Z to be called a ‘statistic’, we should know both mu and sigma, alright. So, here we do not know sigma. And therefore, we have a problem. Of course we do not know mean, but we know that ultimately X bar is an unbiased estimator of the mean. So, that is not so much of an issue because we can say instead of saying X bar minus mu, we can say X bar by sigma or root n as a Gaussian distribution and so on.

But, the fact is now that sigma is not known, how do we do it? The solution there is to replace the variance, the true variance with the sample variance. We have not yet looked at the expression for sample variance. So, this is just jumping (Refer Time: 09:07). Assume that we know the expressions for sample variance. It is a very common expression that we all know. We see on our; we see a button with sigma square n minus one or sigma n minus 1 on our calculators. The sigma n minus one is the square root of the sample variance, sample standard deviation.

I can replace the theoretical variance with this standard one. And, there are two different scenarios here that we can look at; large n. What is large typically depends? Statistically, n greater than 5; for example, 30, 40, 50 is considered large. In certain other cases, it may be 100 and so on. But, generally speaking for large n, when you replace the true variance with the sample variance, then all the results that we have discussed earlier hold good; that means X bar still follows the Gaussian distribution; for small n and additionally. you assume now not unknown distribution, but we assume that if X bar X falls out of a Gaussian population, then what is the story? In the large sample case, we do not really worry about the distribution of the observations. But in this small sample case, we again restrict ourselves to a known distribution. And, the result is if X falls out of a Gaussian population, then the statistic X bar minus mu by S over root n. S is our sample standard deviation; that follows what is known as a student’s t-distribution with n minus 1 degrees of freedom.

Again, here we have degrees of freedom. Here, what is a student’s t-distribution? It also looks like a Gaussian distribution, but it is different from it. In the sense that it has the certain, it has an additional parameter which is called degrees of freedom. Again, that is got to do with the sources of variability. t is a random variable; why is it a random variable? Because X bar is a random variable, in addition the sample variance also is a random variable because it is an estimate after all. How is a sample variance calculated? Well, S is a sample standard deviation, but it is coming out of sample variance. So, it is a random variable. How is a sample variance computed? We know it is computed as 1 over n minus 1 sigma X i minus X bar to the whole square. So, there you have n terms that we are using in computing the sample variance.

We shall learn in the later lecture that sample variance has a chi square distribution with n minus 1 degree of freedom. Why this n minus one degrees of freedom, when I have n observations, n sources of variability because one degree of freedom has already been taken up in the computation of X bar. So, already that has been taken up. So truly speaking, there are n minus one sources of variability. I am just giving a qualitative explanation. We will again revisit that point when we discuss sample variance.

So, the bottom line here is when the variance is unknown, I replace it with sample variance. And then, if the sample size is small and X falls out of a Gaussian population, the resulting random variable follows a t distribution. The name “student” is only a pen name. It is not the name of the person who invented it. The person chose to use the name “student”, and thereafter, it came to be known about as student’s t-distribution.

Now the student’s t-distribution, it turns out when the degrees of freedom becomes large, greater than thirty, it actually becomes independent of the degrees of freedom and tends to have a Gaussian distribution as well. That is why only for the small sample case we will have to worry about this difference between known variance and the difference between a Gaussian case and non-Gaussian and so on.

Now, what about in general that is for the small sample case? What about non-Gaussian? Well, approximately you can use a t distribution. In that case, it is only an approximate result. So, we have to be a bit practical there. Or of course, you can use a more powerful Monte Carlo simulation or bootstrapping techniques to determine a more accurate distribution. For our purposes, we will assume that X falls out of a Gaussian population, and whenever the sample size is small and then you say student’s t-distribution. So, that brings us to the closure of this lecture where we learnt some very critical aspects, core aspects of hypothesis testing. We learnt that statistics plays central role and hypothesis testing. The sampling distribution has a critical role to play in the test of hypothesis. We will see later on how the sampling distribution plays a vital role in construction of confidence regions.

Remember, the problem of estimation is not, does not end with arriving at the estimate. Rather, it actually begins there. One should not stop at during the estimate. This is what typically we see around when, you know, when many of the students present the estimate. We have to go one step further and provide a confidence region for where the truth is; that is the purpose of estimation, to say something about the truth. And, giving an estimate is not alone sufficient. And, that is where we will discuss the notion of confidence region and how confidence region construction is an alternative to hypothesis testing is what we will see. There once again f of theta hat plays a central role in this construction of confidence regions. And then, of course we went through some standard results.

In this lecture, we have specifically learnt the sampling distribution of sample mean under different conditions. In the next lecture, we will look at distributions of sample variance. In fact, difference in sample means and then sample variances, ratios of sample variances and then sample proportion and differences in sample proportion. Now, you can all relate that to the examples that we talked about in the motivation lecture, OK. So, this has been perhaps a long lecture for you, but it is worth it. Please, go through it once again because the concepts are very important. In case you are not following a certain thing, go through it. Wherever we have used r please work out by yourself, pause the video lecture, work out the r lecture examples by yourself. And, as usual if you have any questions you always have the forum at your (Refer Time: 16:09) to raise of questions and we will be happy to answer.

See you in the next lecture.