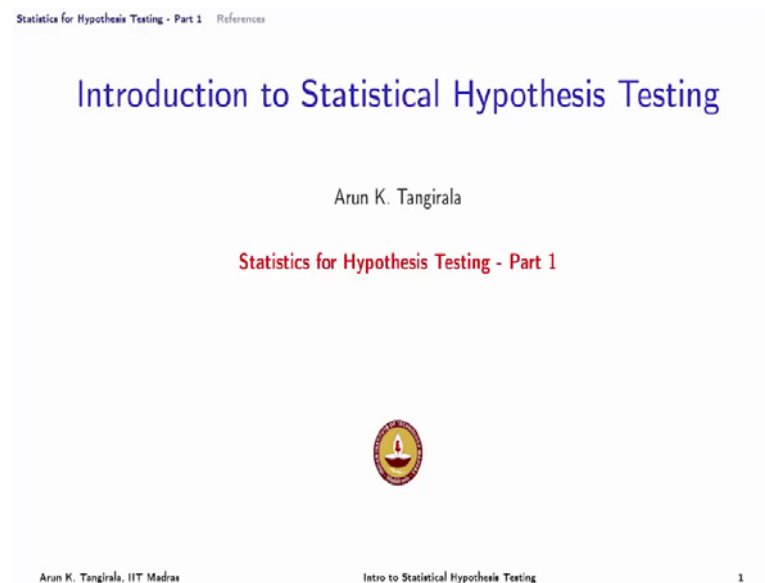


**Introduction to Statistical Hypothesis Testing**  
**Prof. Arun K. Tangirala**  
**Department of Chemical Engineering**  
**Indian Institute of Technology, Madras**

**Lecture – 06**  
**Statistics for Hypothesis Testing – Part 1**

(Refer Slide Time: 00:09)



So, welcome back after a short break. In the example that we just went through, we learnt several things. And, let us list a few of them. The first thing is that the sample mean has a Gaussian distribution. Of course, we have not yet proved it theoretically. But, we have seen it through simulations, and at least for the case of Gaussian distribution.

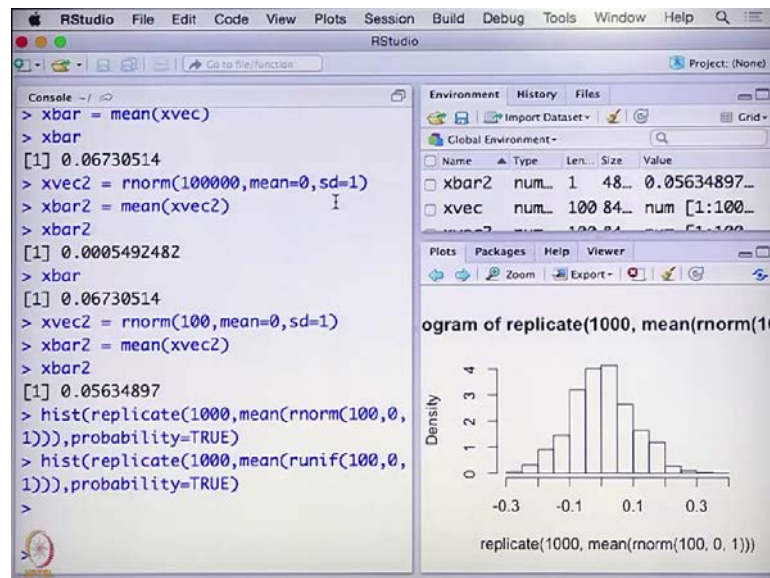
And, we have also seen that in the case of the samples falling out of the, the observations falling out of a uniform distribution. The sample mean still follows a Gaussian distribution. But, there is a small catch there. If we look at the case of fewer observations, that is, when  $n$  is small and the samples followed of a non-Gaussian distribution there is no need or necessity. In fact, the sample mean does not follow Gaussian distribution. It is only when the  $n$  becomes large. When the sample size becomes large, the sample mean tends to follow the Gaussian distribution, regardless of

the distribution of the observations. And this is; what is the essence of central limit theorem as we will shortly review; that is the first point.

The more important point that we have learnt is the role of the distribution of the statistic in hypothesis testing. Of course, you have shown this only in the case of, not shown; I would say we have discussed this in the case of mean testing. But, that is true of the general hypothesis test of the form  $\theta = \theta_0$  and that  $f(\hat{\theta})$  plays a critical role in hypothesis testing.

The way we used  $f(\hat{\theta})$  is to look at or to compute the probability of the test statistic taking on a certain value or within the interval of the observed value. And, if that probability turns out to be quite lower than what we can accept, which we call as a critical value; then we reject the null hypothesis. So, as you can see  $f(\hat{\theta})$  plays a critical role.

(Refer Slide Time: 02:32)



And, the other thing that we have seen is that this going to be certain error in any hypothesis test because it is based on probability and the critical value that I said. That is where we end up with type one and type two errors. A formal definition of which will be (Refer Time: 02:51) later on.

So, now let us continue with our discussion on sampling distributions where we will now try to derive theoretically the distributions of the statistics of interests namely the sample mean, sample variance and so on.

In this lecture, we will look at sample mean. In the next lecture, we will look at difference in sample means, sample variance, ratios of sample variance and sample proportion ratios of or differences in sample proportion. And eventually when we talk of correlation, we will also look at distributions of sample correlation.

(Refer Slide Time: 03:19)

Statistics for Hypothesis Testing - Part 1 References

### Sampling distributions / Distributions of estimator

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

- ▶ The ease (or difficulty) depends on the complexity of the function  $g(\cdot)$

Some important results:

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

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

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

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

In order to derive the theoretical results, some standard results that are available in the distribution theory can be invoked. And, the first result is right in front of us; which says that the linear combination of Gaussian random variables or a weighted sum of Gaussian random variables is also Gaussian distribution. This is the result that we have seen earlier. It is not something new.

But, let us use this opportunity to revisit that Y is a new random variable that has taken birth through a linear combination of n Gaussian random variables. And, each random variable in the sum follows a Gaussian distribution and has a same mean and a same variance, so that makes it easy to derive the mean or you can actually assume that each

random variable falls has a separate mean and separate variance. Even then, you can derive as I have shown here. The result that is shown here is for the general case. It is very easy to show this result. All you have to do is apply the expectation operator, that is, compute the expectation of Y and use the linearity property of the expectation operator and show that the mean of the resulting random variable Y also adds up in the same proportion as Y itself.

And the other assumption, of course that we are making here is that X i's are uncorrelated or independent. And, we can recall the definition of uncorrelatedness or independence. Basically that means, there is no covariance between any pair of random variables in the summation. In which case, the variance of Y simply turns out to be the weighting square in your sum, times the sigma square i. So, let me just quickly elaborate that a bit more for you.

(Refer Slide Time: 05:51).

The chalkboard contains the following derivations:

$$Y = \alpha_1 X_1 + \alpha_2 X_2$$

$$\mu_Y = E(Y) = \alpha_1 \mu_1 + \alpha_2 \mu_2$$

$$\sigma_Y^2 = E((Y - \mu_Y)^2) = \alpha_1^2 \sigma_1^2 + \alpha_2^2 \sigma_2^2 + 2\alpha_1 \alpha_2 \text{Cov}(X_1, X_2)$$

$$= \alpha_1^2 \sigma_1^2 + \alpha_2^2 \sigma_2^2$$

Other equations on the board include:

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

$$X_i \sim N(\mu, \sigma^2)$$

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

$$\sigma_{\bar{X}}^2 = \sum \frac{1}{n^2} \sigma^2 = \frac{\sigma^2}{n}$$

A boxed note at the bottom states:  $\lim_{n \rightarrow \infty} \sigma_{\bar{X}}^2 = 0$

So in general, when in general when I have a pair of, when I am adding up let us say 2 random variables X 1 plus X 2 to produce, let us say, alpha 1 X 1 plus alpha 2 X 2 to produce a new random variable. Of course, as I said it is straight forward to see that mu Y is alpha 1 mu 1 plus alpha 2 mu 2. This is regardless of the distribution of X 1 and X 2

and regardless of whether  $X_1$  and  $X_2$  are uncorrelated, correlated, independent, and dependent. It is got nothing to do with none of those requirements.

When it comes to the variance of  $Y$ , which is expectation of  $Y - \mu_Y$  to the whole square. And, once you plug in the values of, sorry, the expression for  $Y$  and expression for  $\mu_Y$ , you would end up with this expression; which is  $\alpha_1^2 \sigma^2 + \alpha_2^2 \sigma^2 + 2 \alpha_1 \alpha_2 \text{Cov}(X_1, X_2)$ . And, this is a generic expression. Once again, this is got nothing to do with  $X_1$  and  $X_2$  falling out of a Gaussian distribution and so on.

So, in deriving this results distribution has no role to play. Now, if you assume that  $X_1$  and  $X_2$  belong to a random sample that is they are independent; then independence means uncorrelatedness. And therefore, you can strike this off and you end up with this expression. And, now you can extend this result to the general case of  $n$  observations.

Now coming to the distribution of  $Y$  is what the other result states. That is where we are invoking the fact that each  $X_i$  falls out of a Gaussian distribution with  $\mu_i$  mean,  $\mu_i$  and variance  $\sigma^2$ . And, the result says when I add up 2 such Gaussian random variables then the resulting variable  $Y$  also has a Gaussian distribution with this mean and this variance. So, something that we can straight away use for the sample mean as an illustration.

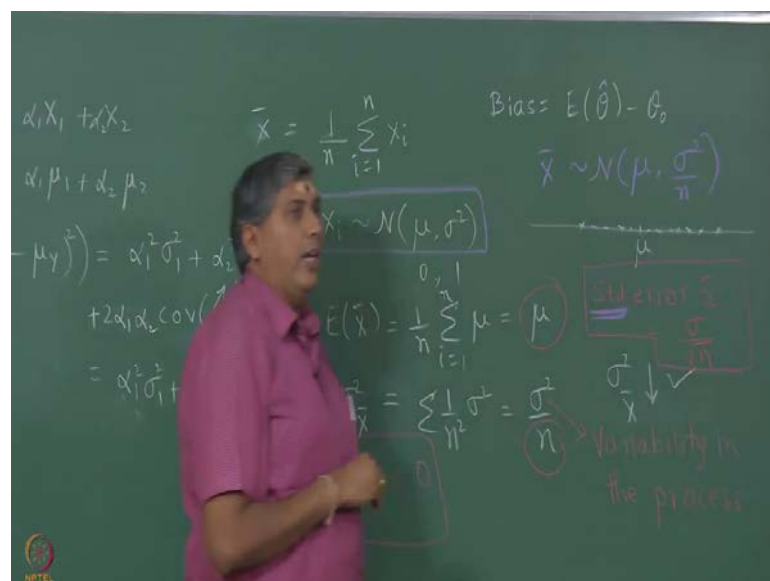
Suppose in place of  $Y$ , I have  $\bar{x}$  and in place  $\alpha_1 \alpha_2$  I have  $1/n$ , then we have  $\bar{x} = 1/n \sum_{i=1}^n X_i$ .  $X_i$ , of course it is a random in. At this stage, we are using the notation that  $X_i$  is random variable. In practice, what you do is you replace a random variable with the value that you have in the sample because we are now interested in the distribution of  $\bar{x}$ . We are using the random variable notation.

So, now we can straight away use this result; if  $X_i$  each observation is falling out of a Gaussian distribution. And, that is what we have essentially seen in the illustration in our  $\bar{x}$  follows a Gaussian distribution. Of course, in the example that we illustrated every observation fell out of the same Gaussian distribution mean and  $\sigma^2$ .

Specifically, we simulated for the case of mean two and variance one. But, this need not be true in general.

And, now you can figure out what is the mean of  $\bar{x}$  using this result here. Expectation of  $\bar{x}$ , I will also use this upper case notation, so that we are looking at random variables. Expectation of  $\bar{X}$  would be one over  $n$  times sigma expectation of  $X_i$ . Since each  $X_i$  falls out of the same distribution, I have  $1$  over  $n$  sigma  $\mu_i$  equals  $i$  running from  $1$  to  $n$ . And, therefore expectation of  $\bar{X}$  bar; that means, the average of  $\bar{X}$  bar is the same as the average of  $X$  or  $X_i$  you can say. Now, this is very important because this is a test for bias of the estimator. We say that any estimator is unbiased, if its average is the same as a true mean. Here, the estimator is sample mean.

(Refer Slide Time: 10:47).



And bias is; in general, for any parameter  $\theta$  bias is defined as an expectation of  $\hat{\theta}$  hat minus  $\theta_0$ .  $\theta$  is mean for us in this exercise and  $\theta_0$  is  $\mu$ . And therefore, we get this.

So, to give you an idea; to give you a pictorial idea of this, so let us say the  $\mu$  is here. The truth is fixed. And, our estimates are all over the place, not all over the place, but to the left and right of  $\mu$ . Each cross, here marked is an estimate from a data record. We

talked about this in the example and we also simulate it when we use the command replicate. In each of iteration of the replicate, it is actually generating a data record. From each data record, you have an estimate which is denoted by the cross arrow mark here. And, what this result says is that the average of all such estimates coincides with the truth. If it does not, then you have a biased estimator.

As a simple example, suppose in place of  $\frac{1}{n}$  I had used this estimator  $\bar{X}$ , let us call this  $\tilde{X}$  as a different estimator. With one over, let say,  $n - 1$ . Then, here clearly expectation of  $\bar{X}$  is  $\frac{n}{n - 1} \mu$ ; which means average of all such readings would fall to the right of  $\mu$ ; if you think of the origin being here. Essentially, right or left we are not so much worried about it. It does not coincide with  $\mu$ . And, in this case, sorry, it is not  $\bar{X}$ , but  $\tilde{X}$ . In this case,  $\tilde{X}$  is a biased estimator; which means there is a systematic difference between  $\tilde{X}$  and  $\mu$ .

Again, you should keep telling this to yourself that every estimate, that is, estimate computed from every data record will always been error that there is no denial. It is a fact to acknowledge. What is important is to see is the average of all such errors will be 0, so that expectation of  $\bar{X}$  coincides with  $\mu$ . So, we have now computed the mean of  $\bar{X}$ . Correct. And that turns out to be  $\mu$ . Now, we can also use this result to derive the variance of  $\bar{X}$ , alright.

So, let us now derive the expression for the variance. Now, I just notice that there was a small mistake. I am sure you would have noticed by now and would be eager to correct me. So, here you go. I stand corrected. We have here  $2 \times \alpha_1 \alpha_2$ ; 2 times the covariance. Anyway, this result would not change eventually, whenever the covariance is 0, alright.

So, now let us actually use this result to derive the variance of  $\bar{X}$ . We have not yet come to the distribution of  $\bar{X}$ . As I said earlier, the expression for the mean and variance do not really have anything to do with the distribution of  $X_i$ 's. So, this result that is on the unbiasedness holds good for any situation. That is, regardless of whether this holds or not.

Now, we are interested in variance of  $\bar{X}$ . Remember, now that we are looking at 2 random variables; 1 is  $X$  corresponding to your population, the variable of interest and other is  $\bar{X}$ . So, do not get confused between these 2. It is a common thing to be confused between  $X$  and  $\bar{X}$ . But, with some practice you can get over that confusion. So,  $\sigma^2_{\bar{X}}$  is what we want. And, we can straight away use this result. We know that  $\alpha_1$  or any  $\alpha_i$  is  $1/n$  and because each  $X_i$  falls out of the same distribution, I can therefore say that it is  $1/n^2$ . That is, if I extend this result to the  $n$  random variable case, I would have  $1/n^2$  times  $\sigma^2$ . And, that gets me  $\sigma^2/n$ . This is a very useful result in many different ways. Of course, we have derived this result assuming. Now, at this point we have assumed that each  $X_i$  is uncorrelated with the other. We do not even have to assume independence, but if the observations are fallen out of a random sample, then anyway they are independent and then they are uncorrelated. So, that is anyway true.

So, the only assumption that we have made here in deriving this is that  $X_i$ 's are independent. We have still not made use of the Gaussian distribution at all, whereas, here we have not made use of any kind of assumption; apart from the fact that the  $X_i$  has a mean  $\mu$ . That is all.

Now, this is a very important result in many different ways. Remember, we talked about precision of an estimate which tells me the variability of an estimate; which tells me how  $\bar{X}$  varies from data record to another data record. And larger the variability, worst is the situation because the estimate that I compute from one data record is not truly reliable. So, what we want to achieve is a high precision; which means I want the  $\sigma^2_{\bar{X}}$  or  $\sigma^2_{\hat{\theta}}$  I may say to be as low as possible. This is good.

So, when can I achieve a very low variability in the parameter estimate or high precision in the parameter estimate? Let us look at what affects  $\sigma^2_{\bar{X}}$ . One is  $\sigma^2$ ; which is the variability in the population. So, this is the variability in the process itself. Let us use a term process; Variability in the random process, which very rarely we have a control over. Of course if it is a manufacturing process, if it is a man made process, then by design I can actually try to minimize this. But, I can never



take this to 0. If it is natural process such as an atmospheric phenomenon and so on, then I do not have any control.

Therefore, this is something we will assume that it is fixed for an experiment because once the process is designed or already naturally in place, sigma square is fixed. So, the only factor that I have in control to control a precession is  $n$ . Of course, this is not necessarily part of hypothesis testing. But, this is something very important to know. Eventually, this number of observations, thus play a role in the goodness of the hypothesis tests. So, you can therefore think that this is a very relevant discussion.

As I increase the number of observations to infinity; infinity is a mathematical term; in engineering terms, very large. Let us say million observations. Then, sigma square anyway is a finite. So, you can expect sigma square  $\bar{X}$  to go to 0. So, the result that we have is in a limit as  $n$  goes to infinity, sigma square  $\bar{X}$  goes to 0. This is a very good result; which means that as I collect more and more samples, the precession of the estimate improves. And finally, at some point sigma square  $\bar{X}$  is 0 or an extremely low value.

And, if you recall when we discussed the notion of variance, we said when a random variable is such that its variance goes to 0 or its variance is 0, then it loses its randomness and becomes a deterministic variable; which means now it will reach a fixed value. That is,  $\bar{X}$  will reach a fixed value. And, what is that fixed value?  $\mu$ . So, what this tells us is that as  $n$  goes to infinity, the estimate converges to the truth. And, when this happens we say it is a consist estimator.

And, of course now we can ask another question which is typically asked in estimation. Whether this is the lowest variability that I can achieve? The positive square root of this is called standard error, which will be useful to us in confidence interval construction. So, the standard error for  $\bar{X}$  under these assumptions is  $\sigma / \sqrt{n}$ . Again, you have to ask yourself what assumptions we have made. We have made an important assumption which is that  $X_i$  are uncorrelated or may be stronger assumptions  $X_i$  are independent. Therefore, if  $X_i$ 's are not independent, which means if you have taken a

sample where the observations are biased, then this discussion and all the discussions we just had need not apply.

Now, the question that we asked earlier just now, is whether this standard error that we have achieved next, we call it as standard error? You should actually make note of this standard prefix there because this is not the error in  $\bar{X}$ . If this was the error in  $\bar{X}$ , I would simply go and add it to  $\bar{X}$  and get the truth, alright. So, this is standard; which means its average error in  $\bar{X}$ . Whether, this is the lowest possible error that I can achieve in the estimate of  $\mu$ .  $\mu$  is the truth that I am trying to estimate. Is there any other way of estimating mean which can give me error lower than this? Although, we would not prove anything, it is a widely known result that when the observations are uncorrelated. This is actually the so-called minimum variance estimator.  $\bar{X}$  is a minimum variance estimator of mean.

In other words, the most efficient estimator; there is no other estimator that can be more efficient than this; which means if you were to use sample median, it would have a larger error, than sample mean. That is a very interesting thing to know because when you are computing hypothesis test, sorry, when you are conducting hypothesis test we have to choose a test statistic. That is 1 of the key steps, if you recall in the general procedure.

So, when it comes to choosing test statistic, there are several options of which we choose an estimator that is efficient; that means, it has a lowest error and it is unbiased. These are very important things because if I choose a bias statistic to conduct a hypothesis test, the outcome of the hypothesis test is also going to be biased. So, it is important that when you choose a test statistic for hypothesis testing, you are sure that under the circumstances and the assumptions that you have made for the process, you are choosing an unbiased estimator. And if possibly, efficient estimator.

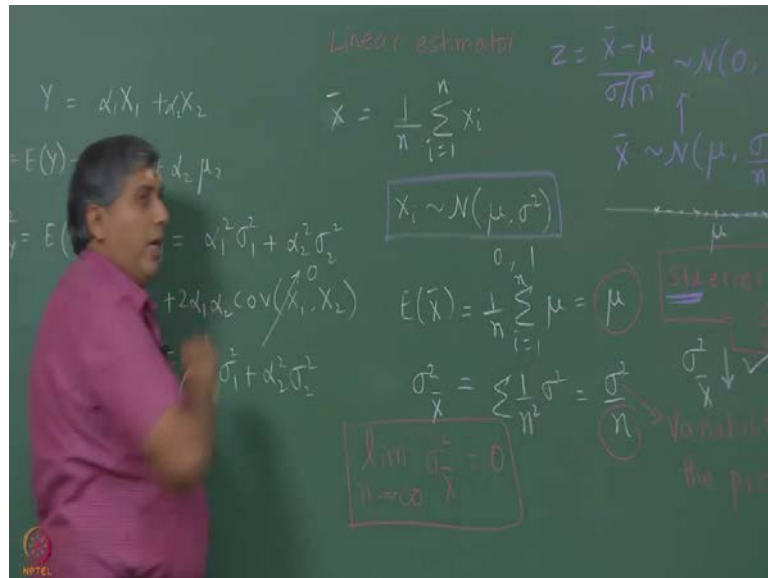
Now, here in this case we were able to verify very quickly that it is an unbiased estimator. And of course, we have not proved that it is the most efficient. But, even showing that it is fairly easy. In a general scenario, it may not be so easy. For example, you know if you are working with the fully efficient or the most efficient estimator and likewise, it may be also not so easy to determine whether the test statistic that you are

using is actually unbiased or biased. But, you can look up the literature and determine whether somebody has studied the biased nature of it. Fortunately, for us the kind of problems that we are going to look at involves test statistics such as sample mean, sample variance and so on. They are all biased. Of course, under some conditions nothing is universally unbiased. So, this is a story here.

Now coming back to the distribution, now we go back to asking where does  $X_i$  come from. What is the maternal place for the random variable? The maternal place happens to be Gaussian. And under these assumptions, this result that we have just seen on the screen tells me that  $\bar{X}$  follows a Gaussian distribution. So, that says all. So, now I can put together everything and write that  $\bar{X}$  follows a Gaussian distribution with mean being the same as they process but, the variance being  $\sigma^2/n$ . This is a fantastic result that we have. Now, we can use this to conduct hypothesis test. And, we will also show that this is very useful in constructing confidence regions.

Now, often this result is written in a standard form, so that we say  $\bar{X}$ . This is, it is not so appropriate to write this. We will not. The reason being when  $n$  goes to infinity, when  $n$  goes to infinity, the variance goes to 0 and  $\bar{X}$  no longer has any randomness in it. And therefore, we cannot talk of any distribution. So, there is a degenerate case there; situation there.

(Refer Slide Time: 25:21)



On the other hand, if I consider a standardized random variable constructed from  $\bar{X}$ , which we normally call or denote by  $Z$ . This remains a random variable. Even, of course when  $n$  goes to infinity. And, this we say follows a standard Gaussian distribution with mean 0 and variance 1.

That is why in many text books, you see this distribution being given more often than this. Of course, for finite  $n$ , this is all fine. It is only for infinite  $n$  that does not make much sense. This is the birth place of the statement, alright. Which means, now what we have understood is when the observations are obtained randomly and they fall out of a Gaussian distribution, the sample mean, first of all is an unbiased estimator. It is a most efficient estimator of the mean and follows a Gaussian distribution. Excellent, so, this is what we want.

Now, typically this is a procedure that we follow for every parameter estimate. But, the question is it going to be that easy to determine. Even in the sample mean case, suppose this was not true, then what about the distribution of  $\bar{X}$ ? This has got nothing to do with this here, the distribution of  $x$ . The central limit theorem comes to the rescue and says that for large  $n$ ;  $\bar{X}$  follows a Gaussian distribution. But, we may not have such privileges for other kinds of estimators because what we have here is a linear estimate.

This is said to be a linear estimator because as you can see the estimate which is  $\bar{X}$  linear (Refer Time: 27:10) or a linear function of the observations. Linearity has got nothing to do with the equal weight-age. I could have given different weights also. Linearity has got to do whether the right-hand side is a linear function. And, you all know the definition of linear functions.

So for linear estimators, typically it is easy to derive the distributions especially, using central limit theorem. But for non-linear estimator, it becomes difficult because there is no theorem (Refer Time: 27:40) to straight away help you, and even deriving some expression scientifically. Only for the variance part, it is a bit easy because we have random variable which follows a chi square distribution and so on. That is something that we will make use of. But, in general when you are looking at a complicated estimator, which is a non-linear function of the observations, becomes difficult to determine the distribution analytically. That is, by hand. So, then what is the natural recourse? A natural recourse is to turn to Monte Carlo simulations; something that we did in the example. Remember, we went through an illustrative example where we observed that sample mean follows a Gaussian distribution. Here, we have derived it theoretically.

So, in a general scenario we may have to turn to Monte Carlo simulations, where we repeatedly generate the data and then determine different values of the estimator. Estimate  $\hat{\theta}$  and then plot a histogram of it from where we try to (Refer Time: 28:40) distribution. What if I cannot perform simulations? What if it is an experiment? How do I determine the distribution of a statistic for an estimator that is a complicated function of the experimental data? That is something which I cannot simulate. Then, unfortunately we cannot perform repeated experiments; infinite number of experiments that could be a near impossible task. What is therefore done instead is a technique that is used called bootstrapping; which again falls within the (Refer Time: 28:56) of Monte Carlo methods.

And, we will not perhaps talk about it now. If I get an opportunity towards the end of the course, I will talk about bootstrapping which will allow us to determine the distribution of an estimate that I cannot determine analytically. And, but I can determine using a single record of data. That is very interesting. Is not it. So, let us wait and see for that.

Now, let us get back to the more comfortable world of deriving distributions analytically for the statistics of interest. So, we now learnt how this single result is so useful in deriving the distribution of sample mean. What are other results that we have? The second result that we have is the distribution on sum of squares of standardized Gaussian variables.

(Refer Slide Time: 30:06)


Statistics for Hypothesis Testing - Part 1 References

**Important results** . . . contd.

2. **Sum of squares of standard Normal variables:**  
 Sum of **standardized** squared variables of a random sample  $X_1, \dots, X_n$ , all possessing the distribution  $\mathcal{N}(\mu, \sigma^2)$  is a  $\chi^2(n)$  distributed RV.

$$Y = \sum_{i=1}^n \left( \frac{X_i - \mu}{\sigma} \right)^2 \text{ where } X \sim \mathcal{N}(\mu, \sigma^2) \implies Y \sim \chi^2(n)$$

where  $n$  refers to the *degrees of freedom*.

 Anon K. Tingirala, IIT Madras

Intro to Statistical Hypothesis Testing 18

So, as the expression shows here what I am doing is I am standardizing  $X_i$  and then squaring it. In this case the standardization, always the standardization of a random variable would mean that random variable minus mean of that random variable divided by the standard deviation. So here, earlier we standardized  $\bar{X}$ , where whose mean was  $\mu$  and standard deviation  $\sigma$  and  $\sqrt{n}$ . Here, we are standardizing  $X_i$ .

Now, when I take such standardized variables and square them and then sum them up, the resulting random variable that is born out of this operation follows what is known as a chi square distribution. This is something that we have learnt earlier when we went through some standard distributions. We define chi square variable as a sum of standardized squared Gaussian variables. And, there we said that the random variable follows a chi square distribution with  $n$  degrees of freedom. These degrees of freedom have got to do; you can say loosely with the sources of variability, alright. How many

independent sources of variability do you have? So, here we have  $n$  terms in the summation. And, each term is contributing to the variance of  $Y$ . And therefore, we say assuming that all of them are independent of each other.  $X_i$ 's are independent of each other. And, which is true when  $X_1, X_2, \dots, X_n$  constitutes a random sample. So in that case, each term has a unique contribution to the variance of  $Y$ . And, since we have  $n$  such terms, we say that it has  $n$  degrees of freedom.

Quite a few people have difficulty in understanding the degrees of freedom in statistics. Degree of freedom is a term that you will see in linear algebra, in engineering. It is a very common term that is used. Now, where do we use this result? Of course, we use this result in the derivation of sampling distribution of variance. And, I will talk about it later on. When we look at sampling distribution of variance, we will see how this result comes handy. And, the other result that we may use is when I add up  $n$  chi square distributed random variables, then the resulting variable also has a chi square distribution.

(Refer Slide Time: 32:26)

Statistics for Hypothesis Testing - Part 1 - References

**Important results** **... contd.**

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

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

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

So, if  $X_i$  falls out of a chi square distribution with  $r_i$  degrees of freedom, so here it is a different. So, what I am doing is I am actually computing the sum of  $n$  chi square distributed variables. Until now, we have been assuming normally distributed random variables.

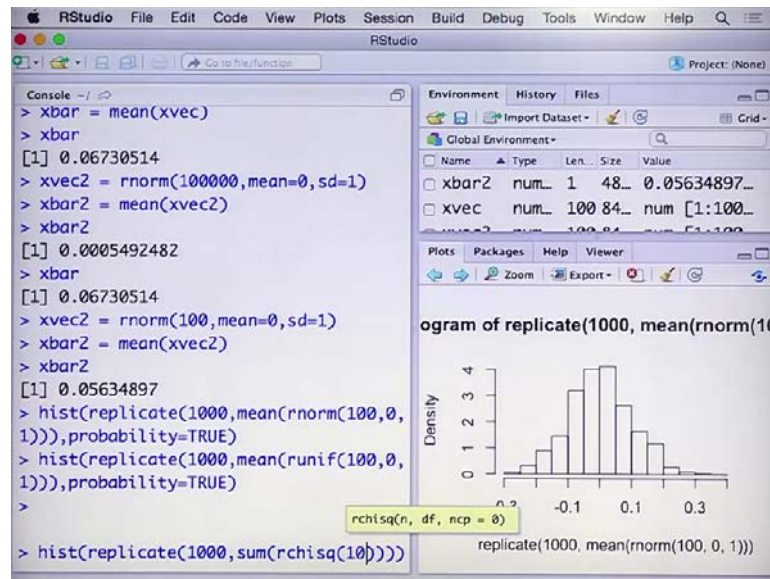
When I do that, the random variable that is born out of this operation is also chi square distribution with  $r$  degrees of freedom; where the  $r$  degrees of freedom, is simply the sum of the  $r_i$  or the sum of the degrees of freedom of the respective variables. Again here, the important requirement is that each  $X_i$  is independent. Only then, the degrees of freedom add up. A lot of times independence affects the degrees of freedom more and less on the distribution. So, for example, if you go back to the discussion on  $\bar{X}$  we have assumed that  $X_i$ 's independent and so on.

Suppose  $X_i$  were not independent, would that affect the expectation of  $\bar{X}$ ? Average of  $\bar{X}$ ? No, we have seen that that assumption is not required. Does it affect the variance of  $\bar{X}$ ? Well, it does because the expressions for computing the variance changes because the covariance, cross covariance terms (Refer Time: 34:09) and then the variance takes a different expression. What about the distribution? Is the distribution affected? Typically, not so much; the distribution may still remain Gaussian, but the variance would change. At least, this we are talking for the large  $n$  case. Here, as well when you add up chi square distributed random variables, we may ask; how does independence affect the distribution of  $i$  and so on? We cannot answer definitively because we need some theory. But, intuitively speaking, the first factor that the independence or lack of independence would affect is on the degrees of freedom.

Now of course, what I leave to you as a simple exercise is to verify even this result for the independent case through simulations  $r$ . Again using the same thing, you can actually instead of  $r$  norm, you use  $r$  chi square  $r$  chi sq and add up  $n$  such variables and see what distribution it follows. I will give a sample and then you can take it from them.



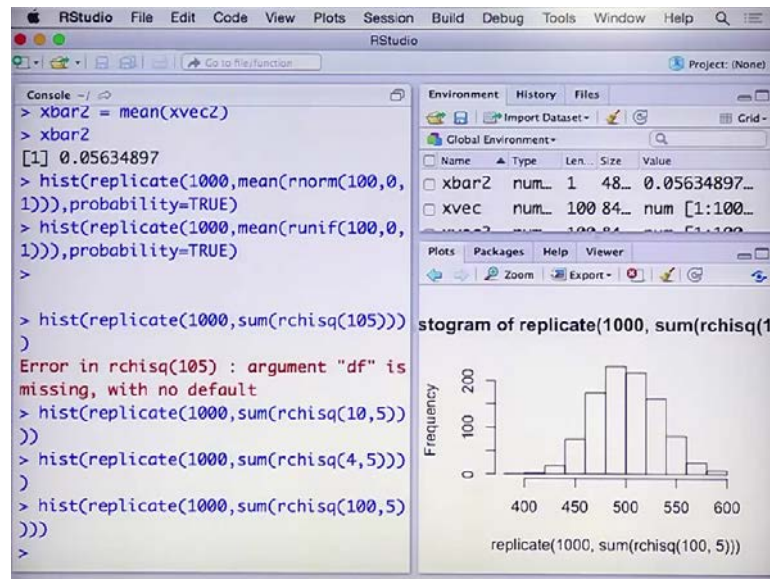
(Refer Slide Time: 35:25)



So, what we shall do is we say `hist`. Again, use the same idea and repeat the experiment thousand times. What do we do here? Sum up. Let say ten random variables; that are all coming out of the same distribution, which is a chi square distribution. In the result that I have given, we assumed each random variable comes out of a chi square with different degrees of freedom. Here, we will assume that they are all coming from chi square with the same degrees of freedom. Just to simplify things. So, now I have to specify the degrees of freedom as also nicely the help shows in our studio.

So, let us say that each random variable comes from a chi square distribution with five degrees of freedom. I am summing up those ten observations and I am going to repeat this and then I am going to plot a histogram, so that I get an idea of the density.

(Refer Slide Time: 36:23)



Sorry. OK, right. So you can actually see that in this case, there is a certain asymmetry. It looks like a bit of Gaussian, but it is not. It is actually a bit chi square. In fact, what we do is we can even reduce the number of variables that I am adding up. Let us say 4 such variables. Now, you can see it is tending towards a chi square.

So, on the other hand if I can, if I actually add up many such random variables which follow a chi square distribution, let us say hundred, you can see now the distribution is tending towards a Gaussian. So, is this a violation of what we have seen? No because any chi square distribution with large degrees of freedom will tend to a Gaussian distribution. It takes the shape of a Gaussian distribution. So, the result is not violated. The result is still that it is chi square distribution, but with a large number of things that we are adding up. So, what is degrees of freedom for this case? That is what we are doing is we are adding up hundred variables, each with five degrees of freedom and they are independent because r chi square samples randomly. So if you look at the result, it says the degrees of freedom for Y is the sum of the degrees of freedom of the respective random variables.

(Refer Slide Time: 37:49)

Statistica for Hypothesis Testing - Part 1 References

Important results . . . contd.

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

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

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

We have added up 100 of them; 100 times, say 500 degrees of freedom is a way too high for the chi square to remain a chi square. It becomes Gaussian. It takes the shape of a Gaussian distribution, just may be by about 30 at around 30 degrees of freedom itself. So, even when we used 10 random variables to add up, each with 5 degrees of freedom, the resulting degrees of freedom for Y was 50.

(Refer Slide Time: 38:28)

The screenshot shows the RStudio interface. The console on the left contains the following R code:

```
[1] 0.05634897
> hist(replicate(1000,mean(rnorm(100,0,1))),probability=TRUE)
> hist(replicate(1000,mean(runif(100,0,1))),probability=TRUE)
>
> hist(replicate(1000,sum(rchisq(105))))
)
Error in rchisq(105) : argument "df" is missing, with no default
> hist(replicate(1000,sum(rchisq(10,5))))
)
> hist(replicate(1000,sum(rchisq(4,5))))
)
> hist(replicate(1000,sum(rchisq(100,5))))
)
> hist(replicate(1000,sum(rchisq(10,1))))
)
>
```

The environment pane on the right shows the following table:

Name	Type	Len.	Size	Value
xbar2	num.	1	48..	0.05634897...
xvec	num.	100	84..	num [1:100...

The plot pane on the right shows an histogram titled "histogram of replicate(1000, sum(rchisq(10, 1)))". The x-axis is labeled "replicate(1000, sum(rchisq(10, 1)))" and ranges from 0 to 35. The y-axis is labeled "Frequency" and ranges from 0 to 400. The histogram shows a distribution that is roughly bell-shaped and centered around 10.

So, truly to see a chi square distribution for  $Y$ , we should either reduce the number of variables that we are adding up or reduce the degrees of freedom for each of them. So, we can do one of these. Here, of course we reduce the number of observations to 4. What we can do is we can say it adds up 10, but each with may be one degree of freedom. So, here you can clearly see that it is a chi square distribution. And, in fact the exercise for you is to check is this is something that you get as a chi square distribution with 10 degrees of freedom because that is what the theory says.

We are adding up 10 chi square distributed random variables; independent random variables each with 1 degree of freedom. You can either generate the density curve for a chi square distributed random variable with 10 degrees of freedom or you can use density fitting tools in  $r$ . There are wonderful density fitting tools to determine, what is the best chi square distribution fit for this? And, you may find that the answer is 10 degrees of freedom. So, it is a very nice simple excise that you can go through, very good. So, you can see simulations really help us a lot in understanding and sometimes corroborating the theory. Of course, theory always remains powerful.