

Statistics for Experimentalists
Prof. Kannan. A
Department of Chemical Engineering
Indian Institute of Technology - Madras


Lecture - 11
Random samples: Sampling distribution of the mean (Part B)

Welcome back after a brief break. Statistics and mathematics is indeed a heady combination for those who are mathematically inclined and to top it or an icing in the cake is there is lot of application in whatever we are doing. It has a lot of implications in decision making. In order to show to the outside world that our decisions are made in a scientific manner in an impartial manner, we resort to statistics.

The decisions are made in a manner, which is not arbitrary but that can be defended by sound scientific principles that is why numerous journal publications are insisting on statistical analysis of the experimental data that are being reported. They do not really care about the scatter or the lack of coincidence of the experimental data but they just want to assure that a proper statistical analysis has been carried out.

Now coming back to our lecture, we were talking about joint probability density functions without too much of a preamble let us straight away go to the expected value of a combination of random variables X and Y.

(Refer Slide Time: 02:04)


$$E(XY) = \iint_{-\infty}^{\infty} xyf_{X,Y}(x,y)dx dy$$
$$\text{Cov}(X,Y) = \iint_{-\infty}^{\infty} (x - E(X))(y - E(Y))f_{X,Y}(x,y)dx dy$$
$$\text{Cov}(X,Y) = \iint_{-\infty}^{\infty} xyf_{X,Y}(x,y)dx dy - E(X)E(Y)$$
$$\text{Cov}(X,Y) = E(XY) - E(X)E(Y)$$

So let us look at the expected value of X times Y , which is given by $\int x \cdot y \cdot f(x, y) \, dx \, dy$. The small x is the representative of capital X and small y is the representative of the random variable Y . So the value of X is not specified to be a point value but it is within a certain interval so that the probabilities can be calculated. Here we are putting the limits as $-\infty$ to $+\infty$.

If you did not have x and y here, the probability would be such that the multiple integral $\int f(x, y) \, dx \, dy$ would have been $=1$ but since you are multiplying it with x and y , the integral need not be 1 , it can take any other value. Even if you multiply it with only x into $\int x \cdot f(x, y) \, dx \, dy$ as in the case of expected value of X , the integral would not be $=1$ because you are multiplying it with the x .

Similarly with y , $\int y \cdot f(x, y) \, dx \, dy$ will not be $=1$ so what I am trying to say is expected value of X and expected value of Y need not be 1 all the time. Now we are talking about covariance represented by Cov between the variables X and Y . Now to understand covariance let us think in terms of something we already know. Instead of covariance if we have only variance right and we also have X and X instead of X and Y we have X, X .

So the covariance of X, X there is no sense in talking about the covariance between the same random variable. So X, X would be rather variance okay since we are talking about a single random variable X , it becomes variance and here it would be $\int (x - E[X])^2 \cdot f(x, y) \, dx \, dy$ so that would be $(x - E[X])^2$, which reminds us of the original definition of the variance for continuous probability distribution functions.

Similarly, covariance of X, Y may be defined as $\int (x - E[X]) \cdot (y - E[Y]) \cdot f(x, y) \, dx \, dy$. This is very interesting okay. We are drawing back on our knowledge of variance to understand covariance but since here we are having 2 independent different random variables X and Y , we do not call it as variance of X and Y , we call it as covariance of X and Y and so we have $\int (x - E[X]) \cdot (y - E[Y]) \cdot f(x, y) \, dx \, dy$.

This can be simplified. You will have $\int x \cdot y \cdot f(x, y) \, dx \, dy$ is what you have here and then you have $E[X] \cdot E[Y]$ and nothing else. This is interesting, what really happened expected value of X is a value okay since you are defining

expected of X between the lower limit to the upper limit after the integration has been carried out and the dust has settled, you will have a number.

So the expected value of X is a number, expected value of Y is also a number. So with that background when you multiply E of X*E of Y and then f of x, y dx dy E of X and E of Y are nothing but constants and then you have *1 because the area under the curve or the multiple integral -infinity to +infinity f of x, y dx dy=1. I would like or request you to expand this particular expression and carry out the necessary steps to arrive at the final answer.

I am deliberately missing out on these steps hoping that you would do them and understand it better. So E of X*E of Y times f of x, y dx dy after the integration is done since integration of f of x, y dx dy=1 you simply have E of X and E of Y. So you should have 4 terms, we have accounted for 2 terms. What happened to the remaining 2 terms? It is very interesting. If you look at it, x*E of Y f of x, y dx dy will become expected value of Y*x of f of x,y dx dy.

So that would have become E of X*E of Y and this combination multiplied by this function would lead to again E of X*E of Y. So you have x*-E of Y that is a negative, y*-E of X which is again a negative so you have -2 E of X E of Y and then you have 1+E of X*E of Y and so you have -E of X*E of Y. So the covariance between X and Y finally is E of XY-E of X*E of Y.

(Refer Slide Time: 09:13)

The image shows a greenboard with handwritten mathematical derivations for the covariance of two random variables X and Y. The derivation starts with the definition of covariance as a double integral of (x - E(X))(y - E(Y)) multiplied by the joint probability density function f(x, y). It then expands the product inside the integral to show four terms. Two terms are simplified using the fact that the integral of f(x, y) over all x and y is 1. The final result is E(XY) - E(X)E(Y).

$$\begin{aligned} \text{Cov}(X, Y) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [x - E(X)][y - E(Y)] f_{XY}(x, y) dx dy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [xy - xE(Y) - yE(X) + E(X)E(Y)] f_{XY}(x, y) dx dy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy f_{XY}(x, y) dx dy - E(Y) \int_{-\infty}^{\infty} x f_{XY}(x, y) dx dy \\ &\quad - E(X) \int_{-\infty}^{\infty} y f_{XY}(x, y) dx dy + E(X)E(Y) \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x, y) dx dy \\ &= E(XY) - 2E(X)E(Y) + E(X)E(Y) = E(XY) - E(X)E(Y) \end{aligned}$$

So rather than leaving the derivation to yourself, I thought I will use the board for a change and do the steps myself. Hopefully, I have not made any mistakes here. So what we do here is

x -E of $X \cdot Y$ - E of Y f of x, y dx dy so I am multiplying first these 2 terms or these 2 expressions in the 2 brackets, $xy - x$ E of $Y - y$ E of $X + E$ of $X \cdot E$ of Y f of XY x, y dx dy. Even though in the slide I have put $-\infty$ to $+\infty$.

X can vary from $-\infty$ to $+\infty$ - ∞ to $+\infty$ for Y also. As a general representation, it has been put as $-\infty$ to $+\infty$ in the slide okay. So now we can multiply each and every term in the bracket with f of x, y dx dy. We get this term and then we know that E of Y is a constant, it can be taken outside the integral. So we have $x \cdot f$ of x, y dx dy.

Similarly, E of X is a constant so it can be taken outside the integral $-\infty$ to $+\infty$ - ∞ to $+\infty$, here you will have $y \cdot f$ of x, y dx dy and this is interesting, these 2 are constants. So you get E of $X \cdot E$ of $Y \cdot f$ of x, y dx dy. So this is = 1. By definition this becomes E of XY , this becomes E of Y and this becomes E of X . This is E of XY and this is E of $Y \cdot E$ of X E of $X \cdot E$ of Y so this becomes $-2 E_X E_Y$.

And this is E of $X \cdot E$ of Y so once you subtract E of $X \cdot E$ of Y from $-2 E$ of $X \cdot E$ of Y , you get $-E$ of $X \cdot E$ of Y so you have E of $XY - E$ of $X \cdot E$ of Y . So this completes the derivation. Even though it looks very cluttered and highly mathematical, it is basically very simple. This is a very important result, which we will be using pretty frequently.

Those of you who are curious may wonder what will happen to the covariance between X and Y if X and Y are independent. If X and Y are independent, they may not have a combined action, a similar action, one variable determining the other variable or influencing the other variable. So intuitively you will have to question what will be the covariance if X and Y are independent.

What will be the value? Will it be $-\infty$, 0, 1 or $+\infty$ and can that be proved from E of $XY - E$ of $X \cdot E$ of Y . If X and Y are independent what will happen to E of XY ? Will E of XY be E of $X \cdot E$ of Y . So please look at these, look at the covariance between X and Y . If X and Y are independent what will happen to the covariance and what would happen to E of XY and what will be the relation between E of XY to E of $X \cdot E$ of Y ?

(Refer Slide Time: 14:07)

Background

❖ If X and Y are independent, $E(XY) = E(X)E(Y)$ and $\text{cov}(X, Y)$ becomes zero.

❖ Note that $\text{cov}(X, Y)$ is also denoted by σ_{XY}

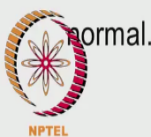


So to save you the long wait, I am giving you the answers right away if X and Y are independent, it can be shown that $E(XY) = E(X)E(Y)$ and the covariance becomes 0 between X and Y random variables. Since X and Y are independent they behave independent of each other, one does not depend on the other. The covariance within X and Y is also denoted by σ_{XY} .

(Refer Slide Time: 14:44)

Properties of the Sampling Distribution

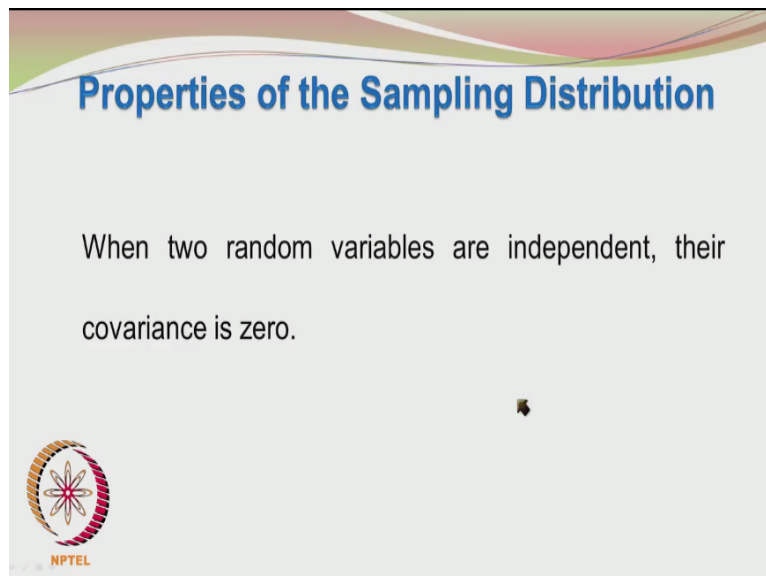
Actually if there are two independent normal distributions, the linear combination of the random variables based on these two populations will also be



Now let us assume that we have the distributions being normal or Gaussian. Let us say that we have 2 independent normal distributions. It can be shown that the linear combination of the random variables based on these 2 populations will also be normal. If there are 2 independent normal distributions, important thing to note here are independency and normalcy okay.


So if there are 2 independent normal distributions, the linear combinations of the random variables based on these 2 populations will also be normal.

(Refer Slide Time: 15:33)



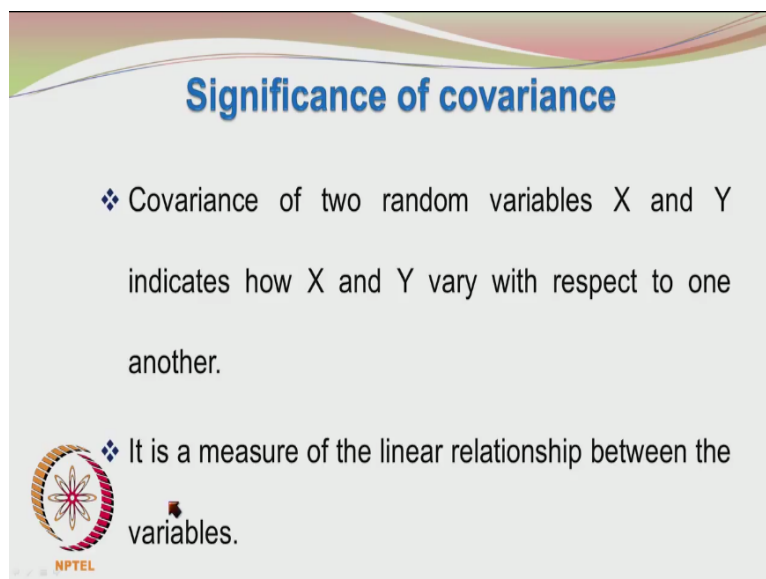
Properties of the Sampling Distribution

When two random variables are independent, their covariance is zero.




When 2 random variables are independent, their covariance is 0.

(Refer Slide Time: 15:38)



Significance of covariance

- ❖ Covariance of two random variables X and Y indicates how X and Y vary with respect to one another.
- ❖ It is a measure of the linear relationship between the variables.



So what is the significance of covariance? Covariance of 2 random variables X and Y indicates how X and Y vary with respect to each other. It is a measure of the linear relationship between the variables.

(Refer Slide Time: 15:57)

Significance of covariance

The correlation between variables X and Y (ρ_{XY}) is defined as follows

$$\rho_{XY} = \frac{\text{Cov}(X, Y)}{\sqrt{V(X)V(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$



So the correlation between variables X and Y denoted by a rho XY is defined as covariance between X and Y/square root of the variance of X*variance of Y, which is sigma XY/sigma X*sigma Y right.

(Refer Slide Time: 16:24)

Properties of the Sampling Distribution

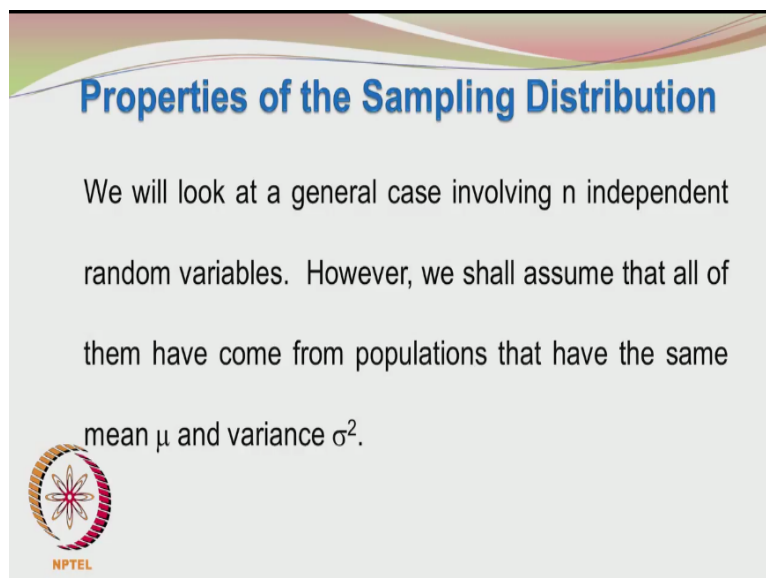
If the parameters of the two normal distributions are (μ_1, σ_1^2) and (μ_2, σ_2^2) what are the parameters of the resulting normal distribution arising out of the linear combination of the two?



So what we have understood is if the 2 random variables are independent, the covariance vanishes or becomes 0. Now we are going to talk about 2 independent normal distributions. So if you combine 2 normal distributions that the resulting distribution is also normal. If the parameters of the 2 normal distributions are μ_1, σ_1^2 and μ_2, σ_2^2 , what are the parameters of the resulting normal distribution arising out of the linear combination of the 2?


So when you are having 2 normal distributions and you are combining them, it also becomes a normal distribution, what are their properties? The properties of the original normal distributions where μ_1 , σ_1^2 , μ_2 , σ_2^2 . So the resulting normal distribution what mean and what variance would it have? That is the question we have to answer now.

(Refer Slide Time: 17:36)



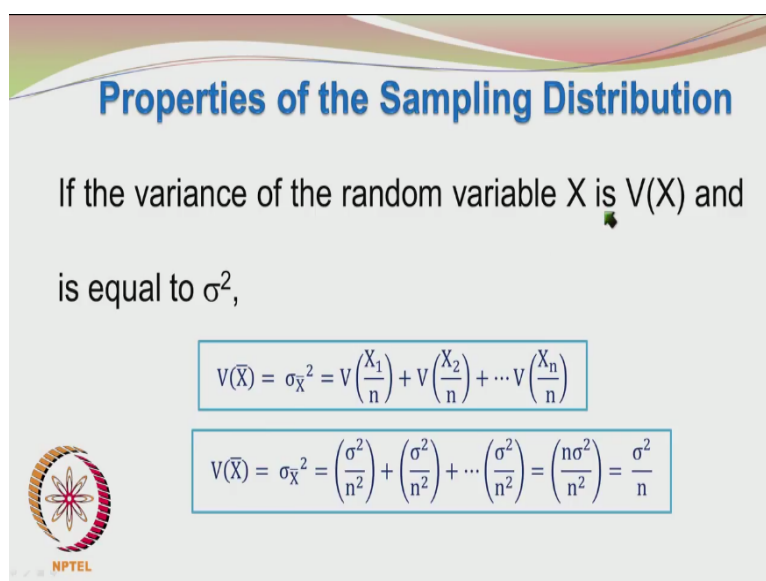
Properties of the Sampling Distribution

We will look at a general case involving n independent random variables. However, we shall assume that all of them have come from populations that have the same mean μ and variance σ^2 .



So we look at a general case involving n independent random variables. We will assume that all of them have come from populations that have the same mean μ and variance σ^2 . So we are talking about n independent random variables and all of them have come from identical populations of the same mean μ and same variance σ^2 .

(Refer Slide Time: 18:09)




Properties of the Sampling Distribution

If the variance of the random variable X is $V(X)$ and is equal to σ^2 ,

$$V(\bar{X}) = \sigma_{\bar{X}}^2 = V\left(\frac{X_1}{n}\right) + V\left(\frac{X_2}{n}\right) + \dots + V\left(\frac{X_n}{n}\right)$$

$$V(\bar{X}) = \sigma_{\bar{X}}^2 = \left(\frac{\sigma^2}{n^2}\right) + \left(\frac{\sigma^2}{n^2}\right) + \dots + \left(\frac{\sigma^2}{n^2}\right) = \left(\frac{n\sigma^2}{n^2}\right) = \frac{\sigma^2}{n}$$



So let us say that the variance of the random variable X is V of X and is σ^2 . So X is taken out of a probability distribution function and normal probability distribution function and the variance of this random variable X is σ^2 . So what would be the variance of \bar{X} ? \bar{X} we know is defined as $(X_1 + X_2 + \dots + X_n) / n$. σ^2 of \bar{X} will be for independent random variables X_1, X_2, \dots, X_n .

Simply, variance of $X_1/n + \text{variance of } X_2/n + \dots + \text{variance of } X_n/n$. So we have already seen one of the example sets if I remember the first example set if you take a variance of a quantity X_1/n , you cannot put n directly outside, it will be $1/n^2$ okay and then variance of X_1 will be σ^2 and variance of X_2/n will be again $1/n^2 * \sigma^2$ because X_1 and X_2 have come from identical distributions of the same mean μ and same variance σ^2 .

So we have so on to X_n will also be represented by σ^2/n^2 . Variance of X_n/n will be represented by σ^2/n^2 . So when you add up all these things, you have n entities $n \sigma^2/n^2$, which is nothing but σ^2/n . This is a very important result. What is the implication or meaning of this result? Do not look at the mathematics.

What is the inference you get out of this particular result? You are having a sample and that sample is having a mean \bar{X} . If I take many such samples not all of them would have the same sample mean okay. They will not have the same sample mean. So there is also a distribution of the sample means. Different samples will have different means and so there will be a distribution of the sample means.

It is hardly surprising because \bar{X} is also a random variable and it is also associated with the probability distribution. We are talking about a distribution of the sample means. What is the variance of that distribution? If the random variable X came from a population of variance σ^2 , what is variance of \bar{X} okay? From now on, we will be shifting to a slightly higher level.

Instead of talking about X , we will be talking more about \bar{X} . We know that X is a random variable which came from a population of mean μ and variance σ^2 . It might have been a normal distribution or a not normal distribution but properties are mean and

variance μ and σ^2 respectively. Now we are shifting gears or moving to a slightly higher level.

We are now talking instead of X , we are talking about \bar{X} . \bar{X} is also a random variable. It will also have its own mean. It will also have its own variance because it has a probability distribution. There is a distribution of the sample means. Hence, that distribution will have a variance. It will also have a mean. What is the variance of the distribution of sample means okay?

The variance of distribution of sample means is not σ^2 but σ^2/n okay. Variance means spread if I am taking a large number of samples then there will be a distribution of the sample means okay and that spread if I want to curtail that spread I do not want that much uncertainty, I want the values to be precise, what should I do? I will increase a sample size n .

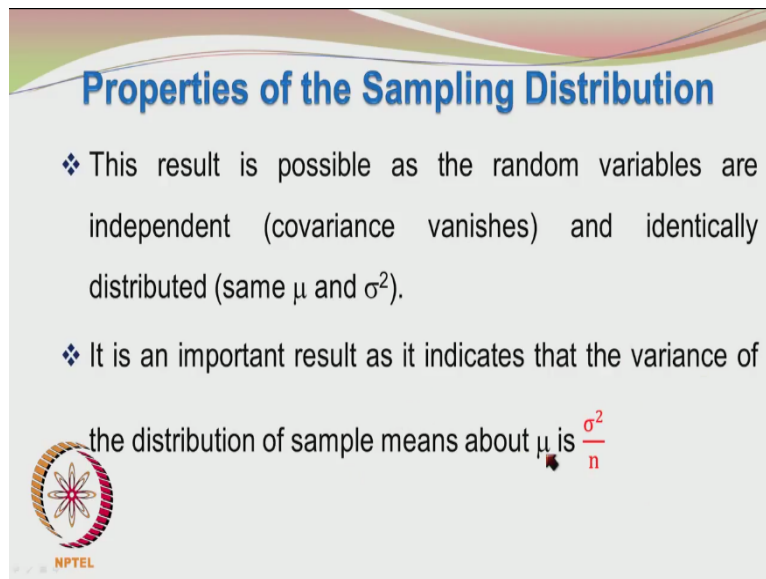
If I increase a sample size n , you can see that the variance of \bar{X} will reduce okay. So the spread of the different possible sample means will reduce if I increase the sample size. So lot of physical basis is there in this seemingly simple derivation. We have taken a linear combination of random variables, which is what we stated at the outset and we try to find its variance and we wrote \bar{X} as $X_1/n + X_2/n + \dots + X_n/n$ and then divided by n .

It looks very simple. It looks too easy to be true okay. Variance is an operator if you consider it as an operator and we are operating it on a combination or a function of random variables. It appears to be linear operator because it is $= X_1/n + X_2/n + \dots + X_n/n$ okay. It looks very simple but this is only applicable when the random variables were independent of each other. If they had not been independent of each other, what would have happened?

That would lead to again a cluttering the slide or the board with more of these multiple integrals but all of you may not have the time or patience to do these integrations. I am sure there will be many of you who would like to carry out the integrations on paper using pencil and paper but it is not necessary to do all those to understand the simple basic concepts.

If the random variables are independent, this variance of \bar{X} can be represented by V of $X_1/n + V$ of $X_2/n + \dots + V$ of X_n/n and that becomes σ^2/n eventually okay.

(Refer Slide Time: 25:51)



Properties of the Sampling Distribution

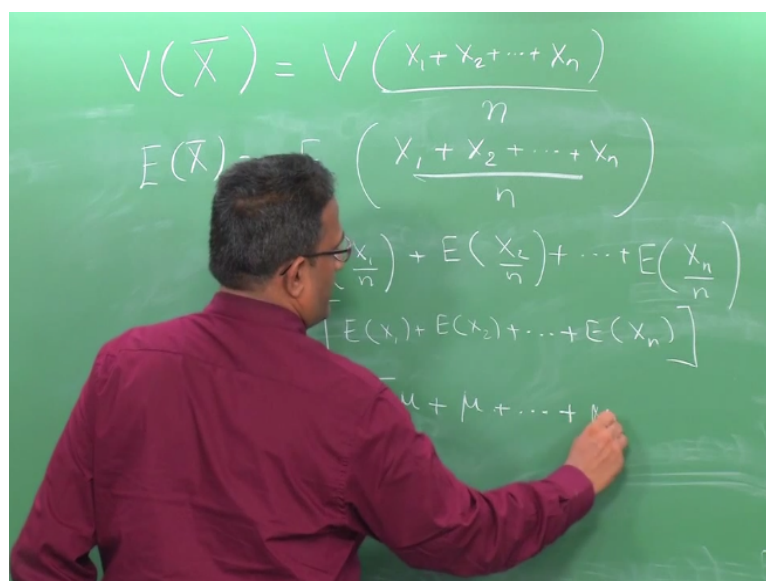
- ❖ This result is possible as the random variables are independent (covariance vanishes) and identically distributed (same μ and σ^2).
- ❖ It is an important result as it indicates that the variance of the distribution of sample means about μ is $\frac{\sigma^2}{n}$.

NPTEL

The variance of the distribution of sample means about μ is σ^2/n . So the question is where did this μ come from okay? We are talking about variance of \bar{X} . You know that X is coming from a probability distribution, expected value of $X = \mu$, what is expected value of \bar{X} ? We will be looking at that derivation also in one of the slides but I request you to write down expected value of \bar{X} .

And try to see what would be the resulting value okay. Expected value of \bar{X} would be expected value of $X_1 + X_2 + \dots + X_n/n$ and what is that value going to be okay. I will just use the board again.

(Refer Slide Time: 27:02)



$$V(\bar{X}) = V\left(\frac{X_1 + X_2 + \dots + X_n}{n}\right)$$
$$E(\bar{X}) = E\left(\frac{X_1 + X_2 + \dots + X_n}{n}\right)$$
$$= E\left(\frac{X_1}{n}\right) + E\left(\frac{X_2}{n}\right) + \dots + E\left(\frac{X_n}{n}\right)$$
$$= \left[E(X_1) + E(X_2) + \dots + E(X_n)\right]$$
$$= \mu + \mu + \dots + \mu$$

So we have been looking at variance of X bar, variance of X bar was $X_1+X_2+\dots+X_n/n$, expected value of X bar would be expected value of $X_1+X_2+\dots+X_n/n$. So this will be expected value of $X_1/n+\dots+X_n/n$. So unlike the variance where when you take it outside the bracket, it became $1/n$ squared it will become $1/n$ similarly for all other expected values.

And you will have expected value of $X_1+\dots+X_n$ okay. This is expected value of X bar. What this means is all these random variables are coming from identical distributions of mean μ and variance σ^2 . So we have $1/n*\mu+\dots+\mu$ to expected value of X bar is also $=\mu$, a very interesting result and much more simpler than the multiple integrals we did earlier and also simpler than the variance of X bar.

So it indicates that the variance of the distribution of the sample means about μ okay is σ^2/n . The mean of the random variable X probability distribution function is μ . The mean of the distribution of the sampling means is also μ okay. The variance of the random variable X is probability distribution function is σ^2 ; however, the variance of the sampling distribution of means is not σ^2 but σ^2/n .


So these are very important results and will be applying them in many problems from now on.

(Refer Slide Time: 29:51)

HOW?

$$V(\bar{X}) = \sigma_{\bar{X}}^2 = V\left(\frac{X_1}{n}\right) + V\left(\frac{X_2}{n}\right) + \dots + V\left(\frac{X_n}{n}\right)$$

$$V(\bar{X}) = E\{[\bar{X} - E(\bar{X})]^2\}$$

$$V(\bar{X}) = \iiint_{-\infty}^{\infty} \{[\bar{X} - E(\bar{X})]^2\} f_{X_1, X_2, \dots, X_n}(X_1, X_2, \dots, X_n) dx_1 dx_2 \dots dx_n$$


I told you that it is not a simple linear operator, there is something more involved here. So variance of \bar{X} = expected value of \bar{X} - expected value of \bar{X} squared okay. There are 2 expectations but one expectation is within the bracket and another expectation is outside the bracket. I hope I am not expecting too much out of your mathematical knowledge. These are pretty straight forward and terminology.

Variance of \bar{X} = expected value of any variable about the mean. So now putting it in terms of the original probability distribution function, multiple distribution or joint probability distribution, we have \bar{X} - expected value of \bar{X} whole squared*into this form.

(Refer Slide Time: 31:05)


HOW?

$$V(\bar{X}) = \iiint_{-\infty}^{\infty} \{[\bar{x} - E(\bar{X})]^2\} f_{X_1, X_2, \dots, X_n}(X_1, X_2, \dots, X_n) dx_1 dx_2 \dots dx_n$$

$$= \frac{1}{n^2} \iiint_{-\infty}^{\infty} \{[x_1 + x_2 + \dots + x_n] - [E(X_1 + X_2 + \dots + X_n)]\}^2 f_{X_1, X_2, \dots, X_n}(X_1, X_2, \dots, X_n) dx_1 dx_2 \dots dx_n$$

where

$$E(\bar{X}) = \frac{E(X_1) + E(X_2) + \dots + E(X_n)}{n}$$

 NPTEL


So when you look at this, it is a matter of expanding the terms inside the brackets or parenthesis. \bar{X} becomes $X_1 + X_2 + \dots + X_n / n$ and since we are having a squared term that would become $1/n$ squared. So we write $X_1 + X_2 + \dots + X_n - E$ of $X_1 + X_2 + \dots + X_n$. This E of \bar{X} also had a $1/n$ term, \bar{X} also had a $1/n$ term and when you are squaring it, it became $1/n$ squared and that was removed outside the integral.

So you have this form just more convenient to handle and then you have this probability distribution function for X_1 to X_n . I am as of now not assuming independence between the random variables.

(Refer Slide Time: 32:34)

HOW?

$$V(\bar{X}) = \sigma_{\bar{X}}^2 = V\left(\frac{X_1}{n}\right) + V\left(\frac{X_2}{n}\right) + \dots + V\left(\frac{X_n}{n}\right)$$

$$= \frac{1}{n^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \{[x_1 + \dots + x_n] - [E(X_1 + \dots + X_n)]\}^2 f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n) dx_1 dx_2 \dots dx_n$$



This can be written as $\frac{1}{n^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \{[x_1 + \dots + x_n] - [E(X_1 + \dots + X_n)]\}^2 f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n) dx_1 dx_2 \dots dx_n$

(Refer Slide Time: 32:57)

HOW?

$$V(\bar{X}) = \frac{1}{n^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \{[x_1 - E(X_1)] + \dots + [x_n - E(X_n)]\}^2 f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n) dx_1 dx_2 \dots dx_n$$

When expanding this, we will collect individual group terms² plus the binary product of the individual group terms



What we do here is we will collect individual group terms squared + the binary product of the individual group terms okay. So I am having this $X_1 - E(X_1)$ I am just simplifying this term, I am collecting terms of the deviations. So $X_1 - E(X_1)$, $X_2 - E(X_2)$, $X_n - E(X_n)$ within the bracketed term that is eventually squared.

So you have $X_1 - E(X_1)$ deviation of the first random variable about its mean, the second deviation, the nth deviation and so on into the joint probability distribution function.

(Refer Slide Time: 34:00)

HOW?

When expanding this, we will individual group terms² plus the binary product of the individual group terms

$$= \frac{1}{n^2} \int \int \dots \int_{-\infty}^{\infty} \{ \dots + [x_i - E(X_i)]^2 + [x_j - E(X_j)]^2 + \dots + 2[x_i - E(X_i)] * [x_j - E(X_j)] \} \dots$$



So when you do that, after taking the square you will have the squared of the deviations and also the cross product of the deviations this is very important. You are having the squared of the deviations and the cross product of the deviations. The squared of the deviation times the probability density function will represent the variance and the cross product term times the probability density function would represent the covariance.

(Refer Slide Time: 34:41)

HOW?

Hence it may be shown that

$$V(\bar{X}) = \frac{V(X_1) + V(X_2) + \dots + V(X_n) + 2 \sum_{i < j} \text{cov}(X_i X_j)}{n^2}$$



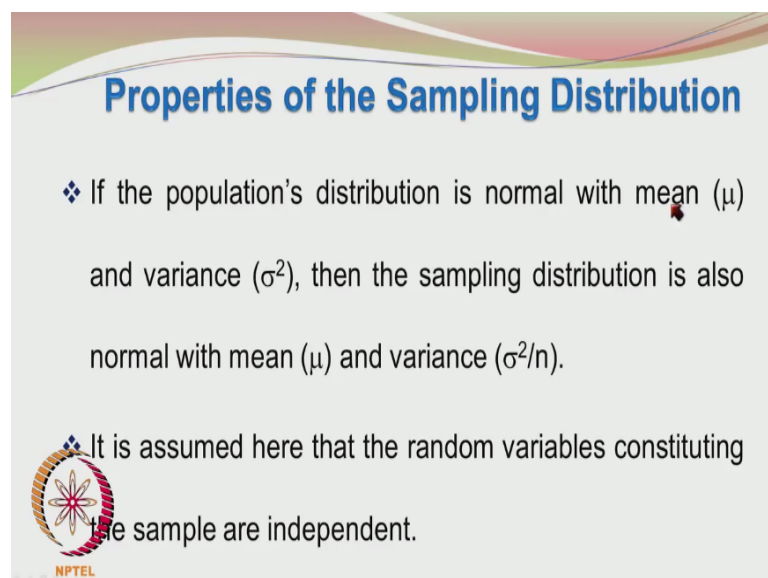
So we get V of $\bar{X} = V$ of $X_1 + V$ of $X_2 + \dots + V$ of $X_n +$ twice the sum of all the covariance terms between X_i and X_j okay. This may be a bit difficult for some people to follow. You may carry out the same derivation with 2 random variables X_1 and X_2 and you can see that it will reduce to variance of $X_1 +$ variance of $X_2 + 2$ times the covariance between X_1 and X_2 .

When you have more random variables, it is a simple extension and you will get the sum of the variances+the sum of the cross product terms or the sum of the covariances/n squared. If the covariance between X_i and X_j was 0 because X_i and X_j were independent random variables and if all the random variables were independent of each other, so that any combination between them will lead to a 0 covariance.

Then the entire sum of cross product terms or the covariance terms will vanish and you will have V of X_1+V of X_2+V of X_n/n squared. So, so much of mathematical background is behind the simple expression for variance of \bar{X} okay. Suppose you had difference of random variables V of X_1-X_2 , then the covariance term here would have a negative coefficient okay.

But the actual variance terms would all be having positive coefficients or positive unity in this case so if you had V of X_1-X_2 , it will be variance of X_1 +variance of X_2 not –variance of X_2 but +variance of X_2 okay. The negative sign corresponding to that X_1-X_2 would have come in the coefficient of the covariance okay. This is very important and follows from regress mathematical background.


(Refer Slide Time: 37:15)



Properties of the Sampling Distribution

- ❖ If the population's distribution is normal with mean (μ) and variance (σ^2), then the sampling distribution is also normal with mean (μ) and variance (σ^2/n).

It is assumed here that the random variables constituting the sample are independent.

 NPTEL

So if the population distribution is normal with mean μ and variance σ^2 then the sampling distribution is also normal with mean μ and variance σ^2/n . We assume that the random sample entities or the random variables constituting the random sample are independent of one another.

(Refer Slide Time: 37:44)

DOUBT ?!

What is the difference between sample variance S^2 and $V(\bar{X})$?

Note that $V(\bar{X})$ may also be referred to as $\sigma_{\bar{X}}^2$



Now we have been talking about a lot of variances okay. So we need to be sure that we have understood them properly. What is the difference between sample variance S squared and variance of X bar? The sample variance is the variance of the sample you have taken okay and that sample is averaged to give the sample mean X bar okay. S squared is the variance of the sample okay.

And variance of X bar is the variance of the sample mean okay. S squared refers to a specific sample. Variance of X bar refers to the many different samples that have been taken okay. So each sample would have its own sample mean. Each sample mean may be different from one another and hence V of X bar denotes the variability of the sample mean okay.

(Refer Slide Time: 39:12)

S^2 and $V(\bar{X})$

- ❖ Simply, the term S^2 is a random variable and is an estimator of the population variance σ^2 .
- ❖ It is based on the random sample of size n drawn from the population. It is given by



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


S squared is a random variable and is an estimator of the population variance sigma squared. We do not know about the population variance sigma squared, so we take the sample find its variance because a sample will have X1, X2, so on to Xn. So we take the values of the random variables and find the S squared and the sample variance is based on a single random sample of size n drawn from the population.

This is defined as for that particular sample i=1 to n Xi-X bar whole squared/n-1 okay.

(Refer Slide Time: 39:57)

What is then $V(\bar{X})$ or $\sigma_{\bar{X}}^2$?

From sampling, n random variables (X_1, X_2, \dots, X_n) and the sample mean \bar{X} is defined as follows

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


Now when you are talking about sample mean, we add up all the random variables and divided by the sample size to get X bar.

(Refer Slide Time: 40:08)


What is then $V(\bar{X})$ or $\sigma_{\bar{X}}^2$?

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$$

- ❖ A function of a random variable is also a random variable.

It is then associated with a probability distribution.

It is the variance of the distribution of the means of the random sample.



Now if you draw many samples from the population, you calculate the sample mean for each of those samples using the entities of those samples using the same formula X_1+X_2 so on to X_n/n okay. So each sample will be calculated for its sample mean using the same formula but each sample will involve its own entities okay. You are calculating sample mean for the first sample as that first sample's entities X_1+X_2+ so on to X_n/n .

The second sample will have again n entities and using those n entities, you will use the same formula to find the second sample mean. The first sample mean and the second sample mean need not be the same. Similarly, if you draw many such samples, those sample means will not be the same. So there will be a distribution, so the variance of that distribution of sample means is denoted by V of X bar or sigma X bar squared okay.


So we have now distinguished between the sample variance S squared and the variance of the distribution of sample means variance of X bar or sigma X bar squared.

(Refer Slide Time: 41:45)

What is then $V(\bar{X})$ or $\sigma_{\bar{X}}^2$?

If each random variable X_i ($i=1,2,\dots,n$) has a variance σ_i^2 ($i=1,2,\dots,n$) then we get $V(\bar{X})$ to be

$$V(\bar{X}) = \left[\frac{V(X_1) + V(X_2) + \dots + V(X_n) + 2 \sum_{i < j} \sum \text{cov}(X_i X_j)}{n^2} \right]$$



How did that come about? We have already seen, we assume that the sample means are comprising of entities that were independent of each other and they were identically distributed. So taking a particular sample mean, we saw that variance of X bar was defined as $1/n$ squared*variance of X_1 +variance of X_2 +so on to variance of X_n +the sum of the covariances between X_i and X_j taken in turn okay.


So since all the random variables were independent of each other, all the covariances disappeared and you had variance of X bar as sigma V of X_i/n squared.

(Refer Slide Time: 42:32)

What is then $V(\bar{X})$ or $\sigma_{\bar{X}}^2$?

$$V(\bar{X}) = \left[\frac{V(X_1) + V(X_2) + \dots + V(X_n) + 2 \sum_{i < j} \text{cov}(X_i X_j)}{n^2} \right]$$


If the individual X_i values were independent the covariance term vanishes and we get

$$V(\bar{X}) = \left[\frac{\sigma_1^2 + \sigma_2^2 + \dots + \sigma_n^2}{n^2} \right]$$


(Refer Slide Time: 42:51)

What is then $V(\bar{X})$ or $\sigma_{\bar{X}}^2$?

If the individual X_i values are not only independent but identically distributed we have


$$V(\bar{X}) = \left[\frac{n\sigma^2}{n^2} \right] = \frac{\sigma^2}{n}$$


And we also made the further assumption that all these entities have the same variance, so $\sigma_1^2 + \sigma_2^2 + \dots + \sigma_n^2$ will become $n\sigma^2/n^2$ or σ^2/n okay.

(Refer Slide Time: 42:53)

Interpretation of $V(\bar{X})$

- ❖ Many random samples may be drawn from a population and each of them may have a different mean.
- ❖ Hence there will be a distribution of the sample means.

 The variance of this distribution is $V(\bar{X})$.

So finally to interpret variance of \bar{X} , many random samples may have been drawn from a population, each of them may have a different sample mean. So there will be a distribution of sample means and the variance of this distribution is V of \bar{X} .

(Refer Slide Time: 43:08)

Interpretation of $V(\bar{X})$

Now, we saw that

$$V(\bar{X}) = \frac{\sigma^2}{n}$$

where n is the sample size. Hence, if the samples are larger in size, n is high.




V of \bar{X} is given by σ^2/n , larger the sample size smaller would be the variance of \bar{X} .

(Refer Slide Time: 43:18)

Interpretation of $V(\bar{X})$

- ❖ There is lesser variation among the means of the different random samples.
- ❖ Hence, $V(\bar{X})$ will decrease and the distribution in the sample means will become narrower.




So it means that if you take large enough samples, there will be less difference between the different samples that you have taken. So the distribution of sample means will become more narrow if you increase the sample size.

(Refer Slide Time: 43:36)

Statistics

- ❖ A function of random variables X_1, X_2, \dots, X_n is also a random variable
- ❖ A function of these random variables is called as a statistic

Hence \bar{X} and S^2 are statistics



We also are now defining another term called as statistics, any function of the random variables X_1 to X_n is termed as a statistic. Since \bar{X} and S^2 are taken from the sample random variables by using a mathematical definition for each case, they are referred to as statistics.

(Refer Slide Time: 44:05)

Sampling Distributions

- ❖ The statistics \bar{X} and S^2 also have a probability distribution associated with each of them
- ❖ They are referred to as the **sampling distributions** of the sample mean and sample variance respectively



Since they are also random variables, they have their probability distribution associated with them. For the time being, our focus is on the sampling distribution of the sample mean \bar{X} okay. Later on, we will be looking at the distribution of the sample variance S^2 . Each is described in terms of unique probability distribution functions, which constitutes the fascinating variety in the field of statistical analysis.

(Refer Slide Time: 44:46)

SUMMARY

If we draw a sample of size 'n' from a population, the resulting sample is classified as a random sample if

- ❖ Each of the X_i s has the same probability distribution
- ❖ The X_i values are independent of one another



So this sort of concludes our discussion on the distribution of sample means. We have seen what is meant by a sample, what is to be done in order to make the sample random and what are the properties of the random sample and if there are many random variables in the random sample, they have to be considered mathematically together in terms of a joint probability distribution function.

Fortunately, in our case the random variables were independent so some simplification was possible to the joint probability distribution, multiple integration and we were able to find that the expected value of the sample is μ and the variance of the sample is σ^2/n . Here the sample size n also plays a very important role in determining the shape of the distribution.

So we will conclude at this point and we will proceed to the next phase in a very short period of time. Thank you.