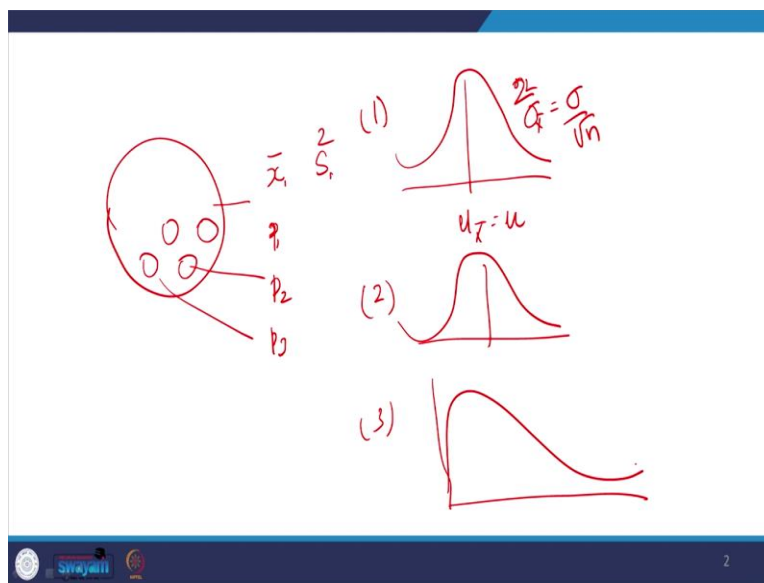


Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Management Studies
Indian Institute of Technology – Roorkee

Lecture – 13
Distribution of Sample Mean, Proportion, and Variance

Welcome Students, last class we have started various sampling distributions. We have seen the sampling distribution of mean. In this class, we will continue from the sampling distribution of mean then we will talk about the sampling distribution of proportions and sampling distribution of the variance. Then I will introduce the concept of chi-squared distribution. We will do some problems then with that we can close this lecture. Before going to this lecture just to recollect what we have done in the previous class.

(Refer Slide Time: 00:31)



This is a population from the population I am taking a sample 1, say, sample 2, sample 3, sample 4. For each sample I can find out the sample mean and sample variance for example if I write \bar{x}_1 bar this is for sample 1 its corresponding mean, if I say, s_1^2 it is the sample with sample variance. So what will you do this is for continuous variable. Continuous variable in the sense if I am measuring some length or height or something, suppose if I take the same sample assume that I am taking a discrete variable or the categorical variable, categorical, categorical variable in the sense, it can have only two values positive negative or good or bad.

Suppose I am taking so this is sample one out of the sample 1, e how many good product is there. So, what is the proportion so then I can call it this is P 1, P 1, another sample that will be P 2 another sample that is P 3. So, if I plot this P1, P2, P3 directly I cannot plot it. First I have to construct a frequency distribution then, I have to plot it I will get another distribution that is the sampling distribution of proportion.

So, there are three point here one is first you take the sample, you take the sample mean, if I plot that sample mean that will follow a normal distribution. So, mean of the sampling distribution is $\mu_{\bar{x}} = \mu$.

the variance of sampling distribution is I am writing $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ This is my first result. What is the first result? From the population I have taken the sample, if I plot that sample, we will mean that will follow normal distribution.

Similarly, in this lecture, what you are going to do, we are going to take some sample from the population, each population is you know, each sample is going, we are going to find out the proportion so proportion means the probability. So I make it P1, P2, P3 that also will follow normal distribution okay. The third one, which we are going to see in the class, so, we have taken the mean, if you take the variance of each sample, if I plot that variance, if I plot that variance which has come from normal distribution that will follow a special shape, this is called chi-square distribution okay. This is going to be summary of our class. We will continue yeah.

(Refer Slide Time: 04:01)

Acceptance Intervals

Goal: determine a range within which sample means are likely to occur, given a population mean and variance

- By the Central Limit Theorem, we know that the distribution of \bar{X} is approximately normal if n is large enough, with mean μ and standard deviation
- Let $z_{\alpha/2}$ be the z -value that leaves area $\alpha/2$ in the upper tail of the normal distribution (i.e., the interval $-z_{\alpha/2}$ to $z_{\alpha/2}$ encloses probability $1-\alpha$)

• Then

$$\bar{x} \Rightarrow \mu \pm z_{\alpha/2} \sigma_{\bar{x}}$$

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

is the interval that includes \bar{X} with probability $1-\alpha$

Before that from the previous class we have seen sampling distribution of the mean, with the help of sampling distribution of the mean, we can find out the lower, upper limit of a sample

mean that is done with the help of $\mu \pm \frac{z_{\alpha} \sigma_{\bar{x}}}{2}$.

We will see how it is Goal: Determine a range within which the sample means are likely to occur given a population mean and variance.

So, what they are asking? Population mean is given, population variance is given, we have to find out the range of sample means that is \bar{X} , lower limit, upper limit. By the central limit theorem, we know that the distribution of \bar{X} is approximately normal, if n is large enough with a mean μ and standard deviation. Let $Z_{\alpha/2}$ be the Z value that leaves area $\alpha/2$, in the upper tail of the normal distribution, that is in the interval $\pm Z_{\alpha/2}$, encloses probability $(1-\alpha)$.

$$\mu + \frac{z_{\alpha} \sigma_{\bar{x}}}{2}$$

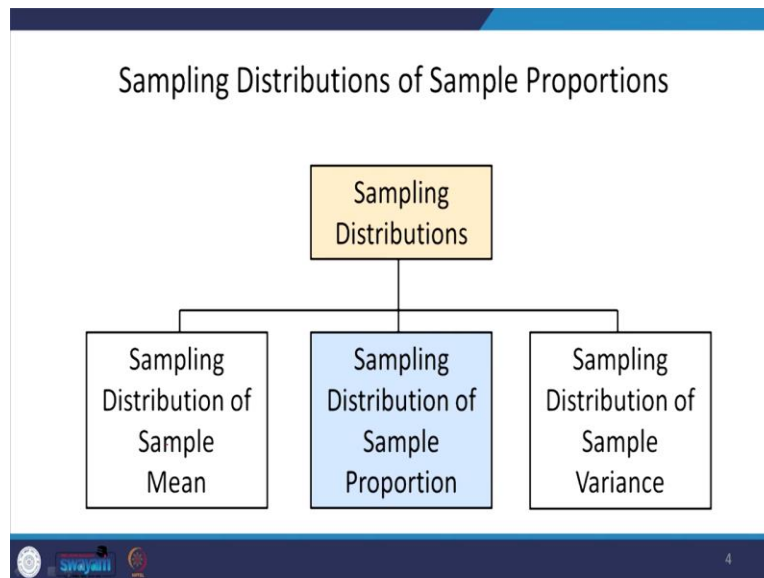
So we can find out the upper limit, the upper limit is

$$\mu - \frac{z_{\alpha} \sigma_{\bar{x}}}{2}$$

The lower limit is , actually this has come from this formula very, very famous $\bar{x} \pm \mu$ by σ by \sqrt{n} . From this relationship we can say μ if you re-adjust that you will get this equation okay. So, if you get you know you from this you can find out the \bar{X} . So,

this value is \bar{X} value we can get the upper limit and lower limit of \bar{X} is a sample mean okay.

(Refer Slide Time: 05:43)



This was what we have started in the last class. So, sampling distribution there are three things which you are going to see. One is sampling distribution of sample mean which I have seen. This class, we are going to see the sampling distribution of sample proportion and sampling distribution of sample variance. First you will see sampling distribution of sample proportion.

(Refer Slide Time: 06:06)

Sampling Distributions of Sample Proportions

P = the proportion of the population having some characteristic

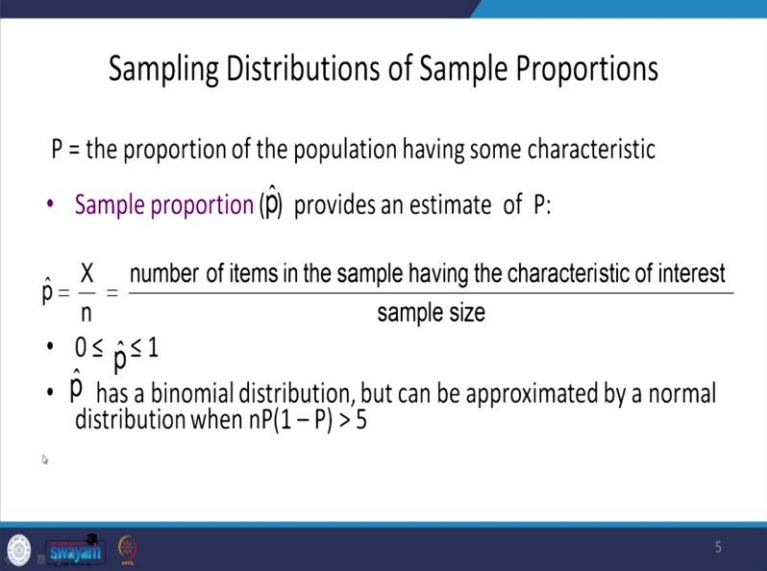
- Sample proportion (\hat{p}) provides an estimate of P :

$$\hat{p} = \frac{X}{n} = \frac{\text{number of items in the sample having the characteristic of interest}}{\text{sample size}}$$

- $0 \leq \hat{p} \leq 1$
- \hat{p} has a binomial distribution, but can be approximated by a normal distribution when $nP(1 - P) > 5$

P equal to the proportion of populations having some characteristics, we can call it as P is the population proportion. This sample proportion we are going to call it as a small \hat{p} . It provides an estimate of P.

(Refer Slide Time: 06:24)



Sampling Distributions of Sample Proportions

P = the proportion of the population having some characteristic

- Sample proportion (\hat{p}) provides an estimate of P:

$$\hat{p} = \frac{X}{n} = \frac{\text{number of items in the sample having the characteristic of interest}}{\text{sample size}}$$

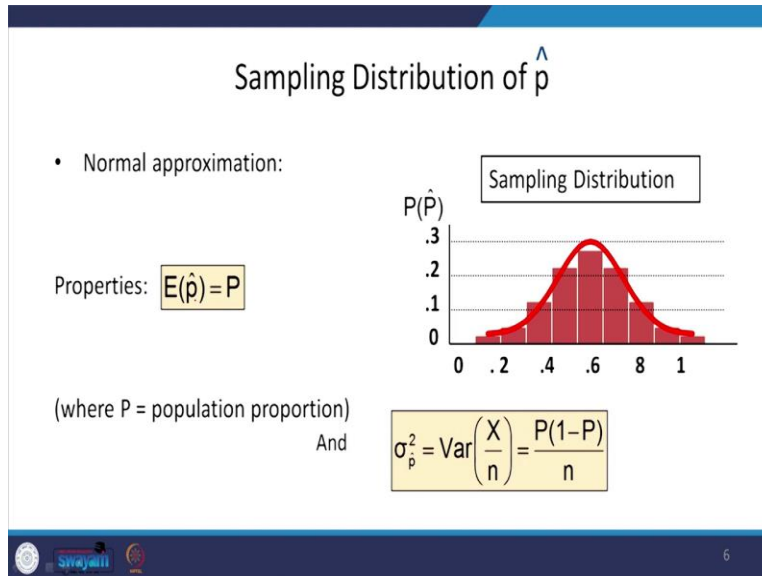
- $0 \leq \hat{p} \leq 1$
- \hat{p} has a binomial distribution, but can be approximated by a normal distribution when $nP(1 - P) > 5$

What is the meaning of this estimate of P is sampling distribution of sample proportion. We are going to use capital P, the proportion of population having some characteristics. Then, sample proportion we are going to call it as \hat{p} provides an estimate of capital P. so, what is the meaning of this one is, with the help of sample proportion we can find out the estimate of population proportion.

So, here how the sample proposed is found equal to X divided by n, X is number of items in the samples sample having the characteristics of interest divided by n is sample size the range of sample proportion is as usual $0 \leq \hat{p} \leq 1$. P has the binomial distribution, but can be approximated by a normal distribution when $n P Q$ is greater than 5. Here, Q is nothing but 1 minus P so here it is following binomial distribution.

As we know the binomial distribution having properties of having only two alternatives that is good are defective, pass or fail, yes or no. So, only two alternatives is there okay. So what will you do?

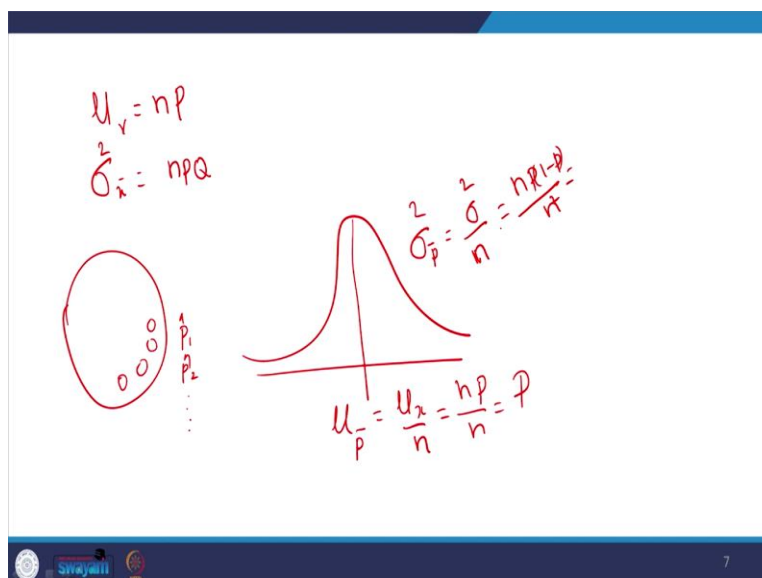
(Refer Slide Time: 07:41)



From the population, we will take sampling proportion so, when you plot the sampling proportion that will follow a normal distribution. So, what will happen? This picture shows the different sample as taken from the population for each sample we find out the sampling proportion, if you plot that sampling proportion that will follow a normal distribution. When we know that it is following normal distribution will has two parameters.

So, mean of that sampling proportion is that is the expected value of your P hat is nothing but P the population proportion. And the variance of this sampling distribution is PQ / n that is a $P.(1-P) / n$.

(Refer Slide Time: 08:27)



Actually this need not remember this formula we can derive it because we know, we have seen in the previous class, the mean of binomial distribution is nP , the variance of binomial distribution is nPQ . Actually we have to use capital P for the population so we will, I use capital P. Otherwise we can write $1 - P$. Suppose what happen there is a population I am taking proportion 1, proportion 2, proportion 3, proportion 4, like that I may get say \hat{P}_1 , \hat{P}_2 , I will get too many such proportion.

If I plot this, if I plot this sampling proportion that will follow, normal distribution so, we have to find out what is the mean of this sampling proportion distribution. Similarly, what is the variance of this sampling proportion distribution? We know that since we have taken n sample from the population so the $\mu_{\bar{p}} = \frac{\mu_x}{n}$ okay. We know that not μ not \bar{x} μ_x μ \bar{x} all know μ_x we can write it as nP/n so that is nothing but your population P .

So, what this result says that the mean of the sampling proportion is equal to population proportion. Similarly, according to central limit theorem, this is $\sigma_{\bar{p}}$ by now when you square into σ^2 by n σ^2 by n , so, if you substitute σ^2 is $nP(1-P)$ I am writing $1 - P$ divided by n this is n square. So, variance is because σ^2 by n .

(Refer Slide Time: 11:01)

Z-Value for Proportions

Standardize \hat{p} to a Z value with the formula:

$$Z = \frac{\hat{p} - P}{\sigma_{\hat{p}}} = \frac{\hat{p} - P}{\sqrt{\frac{P(1-P)}{n}}}$$

8

So, the Z value for the proportion is so small \hat{p} minus capital P divided by Sigma P we know that Sigma P is root of P into 1 minus P divided by n so P minus so it P minus capital P here capital P represents the population proportion, \hat{p} represents the sample proportion.

(Refer Slide Time: 11:23)

Example

- If the true proportion of voters who support Proposition A is $P = .4$, what is the probability that a sample of size 200 yields a sample proportion between .40 and .45?
- i.e.:

if $P = .4$ and $n = 200$, what is
 $P(.40 \leq \hat{p} \leq .45)$?

We do one small problem. If the true proportion of voters who support proposition A is P equal to 0.4, what is the probability that a sample of size 200 yields, a sample proportion between 0.40 to 0.45? What is asked here is the population proportion is given that is a 0.4 that is a 40%. What is the probability that the sample proportion will lie between 0.4 and 0.45.

(Refer Slide Time: 12:03)

Example (continued)

- if $P = .4$ and $n = 200$, what is
 $P(.40 \leq \hat{p} \leq .45)$?

Find: $\sigma_{\hat{p}}$

Convert to standard normal:

$$\sigma_{\hat{p}} = \sqrt{\frac{P(1-P)}{n}} = \sqrt{\frac{.4(1-.4)}{200}} = .03464$$

$$P(.40 \leq \hat{p} \leq .45) = P\left(\frac{.40 - .40}{.03464} \leq Z \leq \frac{.45 - .40}{.03464}\right)$$

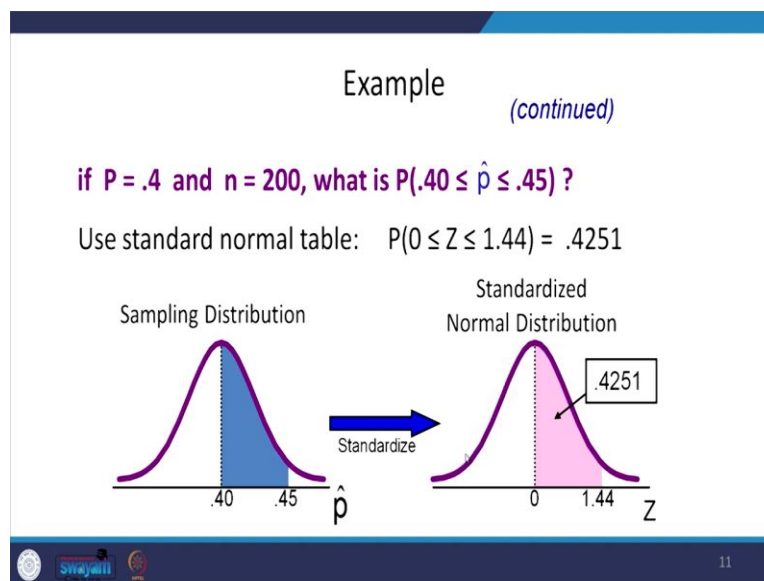
$$= P(0 \leq Z \leq 1.44)$$

So, n is taken equal to 0.4 and n equal to 200 what is the probability of $P(0.4 \leq \hat{p} \leq 0.45)$ here \hat{p} is sampling proportion less than or equal to 0.45. First you will find out the Sigma proportion that is the standard deviation of sampling proportion. Sigma \hat{p} equal to root of PQ by n so P is given 0.4, $1 - 0.4$ by 200 we got point 0.034. We have to convert this 1 Sigma \hat{p} equal to standard normal distribution by using $X - \mu / \text{Sigma } \hat{p}$ so Z is given.

Z is 0.4 minus so we have find out the standard deviation of sampling proportion that is the 0.03464 so that we will convert into standard normal so that we can refer the table. So, $P(0.4 \leq \hat{p} \leq 0.45)$. $P(0.4)$ this 0.4 is Z equal to the small $p - 0.4$ is that capital P divided by 0.034 because that is this your Sigma \hat{p} there is nothing but did this form is $\hat{p} - \mu / \text{Sigma } \hat{p}$ so Z is given.

So, P cap that is a lower limit is 0.4, capital P which is given 0.4 divided by Sigma \hat{p} so this portion will get 0 and right hand side see the another upper limit of P cap 1 is 0.45 minus 0.42 the divided by 0.034 we got the $P(0 \leq Z \leq 1.44)$.

(Refer Slide Time: 13:44)



When you look at the table $P(0 \leq Z \leq 1.44)$ we can get 0.425. So, will summarize what we have done the, it, was asked what is the, what is the probability sampling proportion to lie between 0.4 and 0.45. So, what we are done this 0.4 we are converting to corresponding Z scale it becomes zero. This 0.45, we converted to corresponding Z scale it is 1.44 then we found this area between

Z value is 0 to 1.44 which we got 0.4251. So, now we have seen this one we will go to the sampling distribution of sample variance.

(Refer Slide Time: 14:27)

Sample Variance

- Let x_1, x_2, \dots, x_n be a random sample from a population. The **sample variance** is

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

- the square root of the sample variance is called the **sample standard deviation**
- the sample variance is different for different random samples from the same population

Let X_1, X_2 and X_n be the random sample from a population, the sample variance is sample

variance $\frac{\sum (X_i - \bar{X})^2}{n-1}$. The square root of the sample variance is called the sample standard deviation. The sample variance is different for different random samples from the same population, because every time you may get different sample variance, okay, very important result which we are going to see.

(Refer Slide Time: 14:55)

Sampling Distribution of Sample Variances

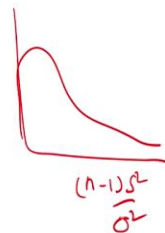
- The sampling distribution of s^2 has mean σ^2

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

- If the population distribution is normal then

$$\frac{(n-1)s^2}{\sigma^2}$$

has a χ^2 distribution with $n-1$ degrees of freedom

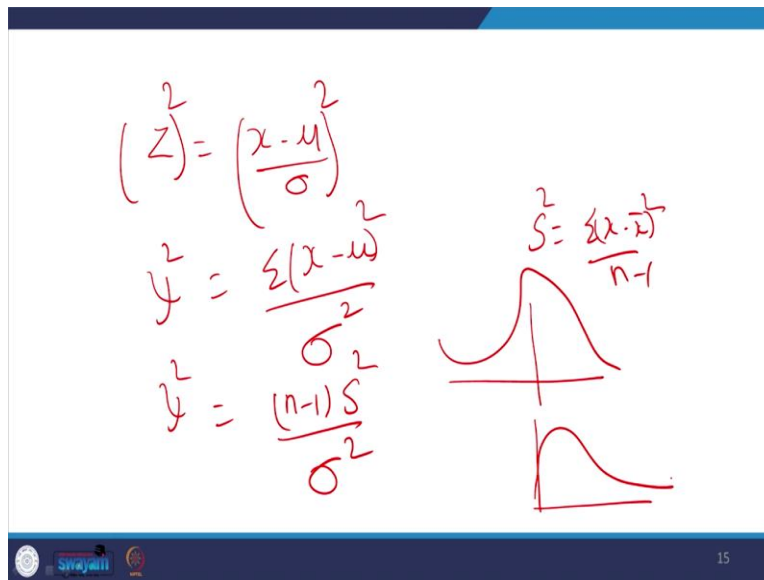


The sampling distribution of sample variance has the mean population variance. So, what is the meaning in that one is, from the population, you take different sample for that sample you find the sample variance we know of that sample variance is equal to population variance but when you take the from the normal population, if you take some sample, then, you find the sample variance.

If you plot that that will follow a particular distribution that shape of this will be like this, right skewed distribution. That distribution is called chi-square distribution that you will see in the next slide. So, another important result is if the population distribution is normal then there is a relationship between sample variance and population variance. That is that relation is $(n-1)s^2 / \sigma^2$ as a chi-squared distribution with the n minus 1 degrees of freedom.

So this x axis is nothing but $(n-1)s^2 / \sigma^2$. This is nothing but our chi-square distribution. You may see there is a similarity between, there may be intuitively you can connect with the normal distribution. For example, we say that we will see in the next slide.

(Refer Slide Time: 16:16)



The image shows handwritten mathematical derivations and a graph. On the left, the following equations are written in red ink:

$$Z = \frac{(X - \mu)}{\sigma}$$

$$Z^2 = \frac{\sum (X - \mu)^2}{\sigma^2}$$

$$Z^2 = \frac{(n-1)S^2}{\sigma^2}$$

On the right, there is a graph of a chi-square distribution, which is a right-skewed curve. Above the graph, the formula for sample variance is written:

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

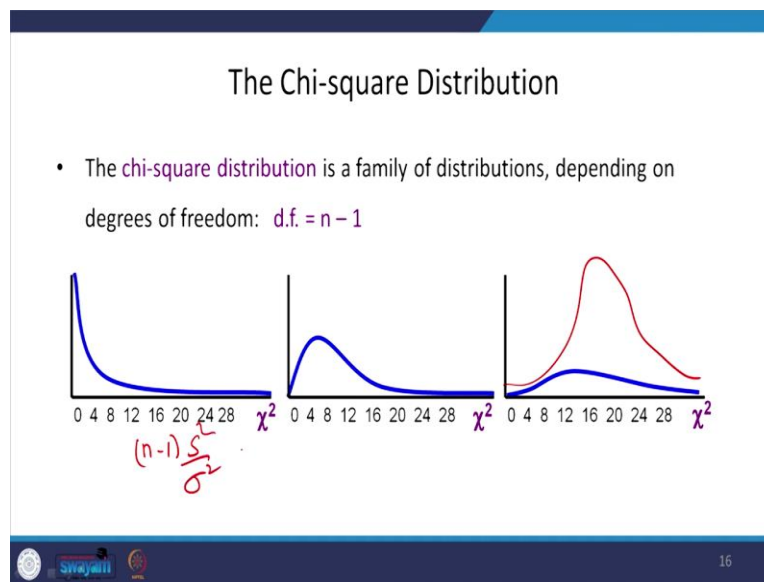
The graph shows a bell-shaped curve that is skewed to the right, with a vertical axis and a horizontal axis. The number 15 is visible in the bottom right corner of the slide.

For example $Z = (X - \mu)/\sigma$, you take different X_1, X_2, X_3 variable so you will get $\sum (X - \mu)$. So, what will happen when you square both side when you square both side for different degrees of freedom, so like that you take different sample different means different X_1, X_2, X_3 so this will become $\sum (X - \mu)^2 / \sigma^2$.

So, the square of Z this is will become a chi square. So, what is nothing but Chi square is nothing but $\Sigma(X - \mu)^2$. I think you know this formula the variance is $\Sigma(X - \bar{X})^2 / (n-1)$, sample variance. So, this numerator can be replaced by that is $\Sigma(X - \bar{X})^2$ can be replaced by $(n-1) s^2 / \sigma^2$. That is nothing but your chi-square distribution.

So there is a connection between your Z distribution and chi-square distribution the other thing since it is a squared you see that Z it is normal distribution this way, so chi-square distribution is like this, because you see that we have squared that Z value, so that there will not be negative Z; so Chi square will be always positive. That is the connection between your Z distribution and Chi-square distribution.

(Refer Slide Time: 18:00)



What will happen? From the sample, you would have taken the variance when you plot that sample variance that will follow this shape. So, this x axis is nothing but your chi-square value. So, the chi-squared distribution is your family of distribution depending on the degrees of freedom n-1. So, when the degrees of freedom it is increasing that means if you are started to take more samples from the population, then, you plot that the variance at the end that will follow a normal distribution.

What will happen? Your chi-square distribution if the degrees of freedom has increased, that will follow a normal distribution. What is the chi-square distribution? From the population, you take some sample, for that sample, you find the variance, like that you take many sample you will find different variance when you plot that variance that will follow this shape. This shape is nothing but the chi-square distribution. What is this chi-square distribution? This x-axis is $(n-1)s^2/\sigma^2$, okay.

(Refer Slide Time: 19:03)

Degrees of Freedom (df)

Idea: Number of observations that are free to vary after sample mean has been calculated

Example: Suppose the mean of 3 numbers is 8.0




Let $X_1 = 7$
 Let $X_2 = 8$
 What is X_3 ?

→

If the mean of these three values is 8.0,
 then X_3 must be 9
 (i.e., X_3 is not free to vary)

Here, $n = 3$, so degrees of freedom $= n - 1 = 3 - 1 = 2$

(2 values can be any numbers, but the third is not free to vary for a given mean)

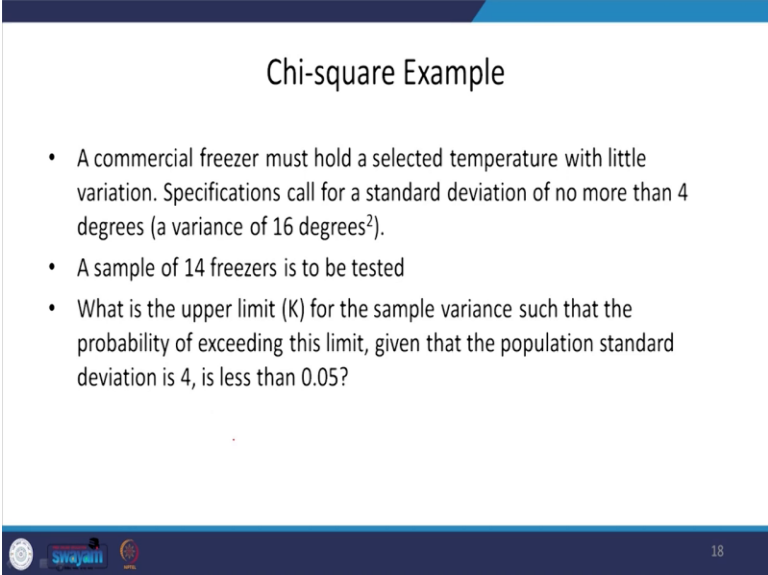



17

Then, another important concept is degrees of freedom because many of the time we will use this concept degrees of freedom, we will see, what is the degrees of freedom? Number of observations that are free to vary after a sample mean has been calculated. That is the degrees of freedom. Suppose that the mean of 3 numbers is 8, say 8 so x_1 equal to 7, x_2 equal to 8 what is the value of x_3 what will happen? Since already the mean is known to us we can supply any value to x_1 any value to x_2 .

But you cannot give any value to x_3 because we have lost one degrees of freedom because already we know, what is the mean of that? So what is the logic here is when n equal to 3 so the degrees of freedom is $n - 1 = 2$ values can be any numbers but the third is not free to vary from given mean. It is like, example like, assume that there are three chair is there we are asking three student to sit there. The first person who is entering will have three possibilities.

That is the three degrees of freedom because three chairs are available. The second person will have two possibilities there is a two degrees of freedom. The third one but there is only one chair there is no option for that so you are lost one degrees of freedom. There are if there are n values you will have only n minus 1 degrees of freedom just we have introduced what is the chi-square distribution and how it has connection with the normal distribution. We will do a small problem to understand the application of chi-square distribution.

(Refer Slide Time: 20:34)



Chi-square Example

- A commercial freezer must hold a selected temperature with little variation. Specifications call for a standard deviation of no more than 4 degrees (a variance of 16 degrees²).
- A sample of 14 freezers is to be tested
- What is the upper limit (K) for the sample variance such that the probability of exceeding this limit, given that the population standard deviation is 4, is less than 0.05?

A commercial freezer must hold their selected temperature with a little variation specification called for a standard deviation of no more than 4 degrees that is the variance 16 degree square you should not exceed 16, and the standard deviation 4. For a sample of 14 freezers is to be tested what is the upper limit of the sample variance such that the probability of exceeding this limit given that the population standard deviation is 4 is less than 0.05.

What is it asking, what is the probability of sample variance that the, the probability of exceeding this limit is less than 0.05? You will see the next slide what it says exactly.

(Refer Slide Time: 21:25)

Finding the Chi-square Value

$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$

Is chi-square distributed with $(n-1) = 13$ degrees of freedom

- Use the the chi-square distribution with area 0.05 in the upper tail:

$\chi^2_{13} = 22.36$ ($\alpha = .05$ and $14 - 1 = 13$ d.f.)

χ^2

$\chi^2_{13} = 22.36$

probability $\alpha = .05$

So, first thing is we have to find out the Chi square value for n minus 1 degrees of freedom. This is a chi-square distribution there are 14 sample is n the degrees of freedom is for 13. 14 minus 1 13, so, the corresponding alpha is equal to 0.05, is 22.36.

(Refer Slide Time: 21:46)

Chi-square Example (continued)

$\chi^2_{13} = 22.36$ ($\alpha = .05$ and $14 - 1 = 13$ d.f.)

So:

$$P(s^2 > K) = P\left(\frac{(n-1)s^2}{16} > \chi^2_{13}\right) = 0.05$$

or $\frac{(n-1)K}{16} = 22.36$

so $K = \frac{(22.36)(16)}{(14-1)} = 27.52$

(where $n = 14$)

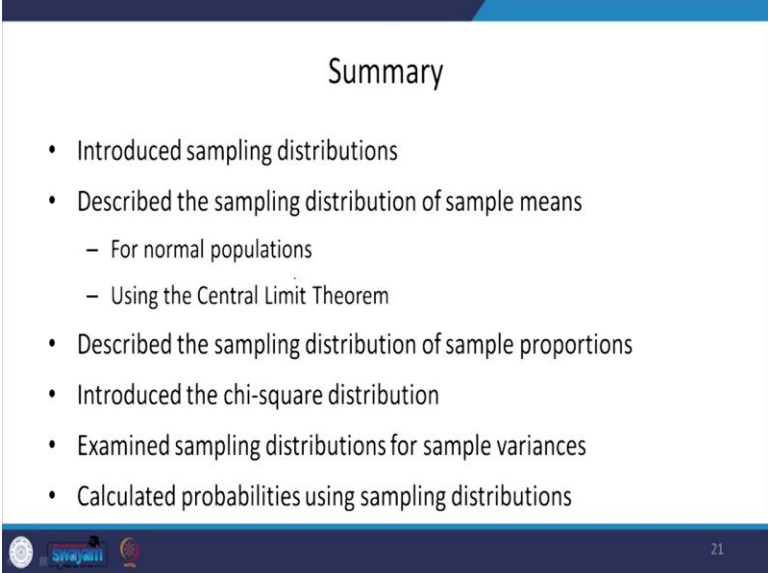
If s^2 from the sample of size $n = 14$ is greater than 27.52, there is strong evidence to suggest the population variance exceeds 16.

So, what is asked is, if, if the chi is 22.36 right we know that the P of $(n-1) s^2$ by, Sigma is 4 so σ^2 is 16 if the chi-square value is 22.36, what is the value of your sample variance that was the question. What is asked the chi-square value is known to us that is 22.36 when alpha equal to 0.05 when chi-square value is 22.36 what is the maximum value of your sample variance.

So, probability of $(s^2 > k)$ equal to $\{P(n - 1) s^2 / 16\}$ greater than chi-square 13, equal to 0.05. So, this value this value, this value, $(n - 1) s^2 / 16$ so with this s^2 between that is your K. So, $(n - 1) K / 16$ equal to 22.36 and you simplify we are getting 27.52. The result is give the sample variance from the sample size of 14 is greater than 27.52 there is a strong evidence to suggest that the population variance exceeds 16.

That is the application of this chi-square distribution. We will see in detail there are many applications for a chi-square distribution one is test of Independence, another one is good goodness of it that we will see in coming classes.

(Refer Slide Time: 23:11)



Summary

- Introduced sampling distributions
- Described the sampling distribution of sample means
 - For normal populations
 - Using the Central Limit Theorem
- Described the sampling distribution of sample proportions
- Introduced the chi-square distribution
- Examined sampling distributions for sample variances
- Calculated probabilities using sampling distributions

21

Now we will summarize in this class what we have seen we have introduced what is the sampling distributions described the sampling distribution of sample means for a normal population. Then we have explained what is the central limit theorem, then, we have seen the sampling distribution of mean, then we have seen the sampling distribution of variance, then, we have seen the sampling distribution of proportions.

Then I have introduced the concept of chi-square distribution how it has connection with normal distribution. Then, we have seen application of chi-square distribution. The next class will go to the next topic, Confidence Interval. We will continue in the next class. Thank you.

