

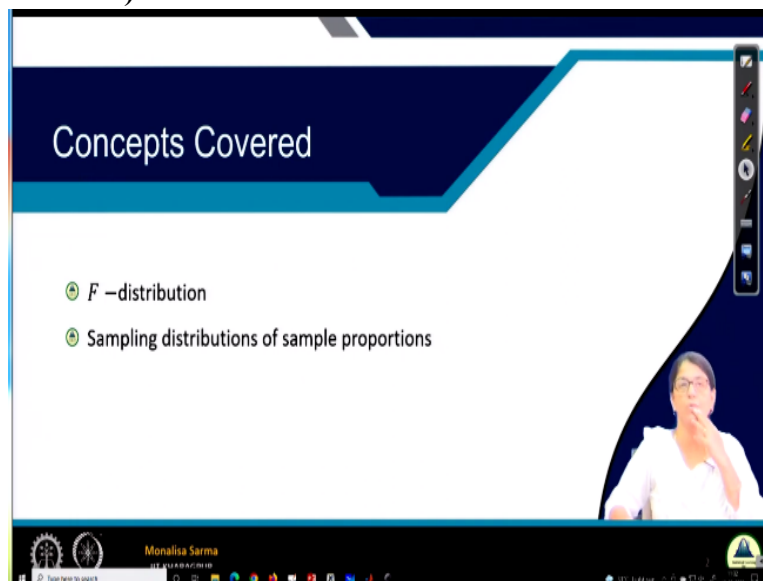
Statistical Learning for Reliability Analysis
Dr. Monalisa Sarma
Subir Chowdhury of Quality and Reliability
Indian Institute of Technology – Kharagpur

Lecture – 20
Sampling Distributions (Part 5)

Welcome back again. So, this is again in continuation of our earlier lecture on sampling distribution, where we have learned till now we have learned sampling distribution of mean that using both Z distribution and T distribution, then we have learned Z distribution and T distribution and we have learned chisquare distribution, this distribution we have learned is not it? Remember what is for sampling distribution of mean when the gradient population variance is known.

And we Z distribution when population variance is not known, we use T distribution and for sampling distribution of variance when we infer something about the population variance, then we use chisquare distribution.

(Refer Slide Time: 01:05)



Today, we will be learning 2 more distributions basically which we will be used our objective here is not for learn the distribution objective here is to learn that this distribution which is necessary for our sampling distribution. So, next to which we will be learning here is F distribution and sampling distribution of sample proportion.

(Refer Slide Time: 01:27)

The F -Distribution

The F -distribution finds enormous applications in comparing sample variances.

Definition: The F -distribution

The statistics F is defined to be the ratio of two independent Chi-Squared random variables, each divided by its number of degrees of freedom.

Hence,

$$F(v_1, v_2) = \frac{\chi^2(v_1)/v_1}{\chi^2(v_2)/v_2}$$

Monalisa Sarma
IIT KHARAGPUR

Now, F distribution first: F distribution, I think you remember when I have just mentioned about it in my first class or second class of sampling distribution, what is F distribution basically, we use when we want to compare 2 different populations, when you want to compare the variances of 2 different populations, comparing the mean of 2 populations that we have used Z distribution even we can use T distribution as well.

When we want to compare 2 different means, see, when we want to infer something about the population mean, we can use either Z distribution or we can use T distribution, again, the same 2 distribution we use when we want to compare 2 different populations. So, but in case of variance, when we want to infer about the population variance, we use chisquare distribution. And when we want to infer about when we want to basically compare 1 different population on variance, then we do not use chisquare distribution, we use F distribution.

And one more thing which I forgot to mention in my last class, chisquare distribution, I think I have mentioned that it is very much sensitivity to normality assumption, that means my parent population has to be normal. Same is the case with T distribution that also the parent population has to be normal. So, F distribution also same, let us parent population has to be normal population. So, now F distribution has one more application that we will be seeing while we will be discussing ANOVA.

Now, what is first let us come to see what is F distribution; what is the PDF of F distribution? So, PDF of F distribution as similar to chisquare and T distribution, there can be different representations. One of the representations which we; will be using, which we need us,

basically for our sampling distribution. So now, first we need to know the PDF of the F distribution. So, when we talk about the PDF of the F distribution, there can be again different representation.

So, we will be using the representation which will be easier for us to use in order to find out the sampling distribution fine. So, this is the PDF for the F distribution. So, what this is a PDF for the F Distribution? See, we have 2 chisquare distribution first in the numerator gives 1 chisquare distribution whose degrees of freedom is ν_1 divided by the degrees of freedom and in the denominator, we have another chisquare distribution with degrees of freedom ν_2 divided by the degrees of freedom. So basically, we will be using this.

(Refer Slide Time: 03:57)

The F-Distribution

Corollary: Recall that $\chi^2 = \frac{(n-1)S^2}{\sigma^2}$ is the Chi-squared distribution with $(n - 1)$ degrees of freedom.

Therefore, if we assume that we have an independent sample of size n_1 from a normal population with variance σ_1^2 and an independent sample of size n_2 from another normal population with variance σ_2^2 , then the statistics

$$F = \frac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2}$$

Handwritten red annotations on the slide show the derivation: $\frac{(n_1-1)S_1^2 / \sigma_1^2 / (n_1-1)}{(n_2-1)S_2^2 / \sigma_2^2 / (n_2-1)}$

Monalisa Sarma
IIT KHARAGPUR

Now, we know what is chisquare distribution we have seen, what is the statistics associated with chisquare this is $n - 1 S^2$ by σ^2 what is the degrees of freedom for this? Degrees of freedom for it is $n - 1$, so we will be using this, this use substitute this and F value. So, if we substitute this chisquare $n - 1 S^2 / \sigma^2$ n_1 would you say n_1 I am writing here say, if I use $n_1 - 1 S^2 / \sigma^2$, because these are 2 different chisquare distribution, is not it?

So, $n_2 - 1$ what is that this is my chisquare and I have here again what to say divided by the degrees of freedom, so degrees of freedom is $n_2 - 1$. So, here also similarly S_1^2 divided by σ_1^2 again divided by the degrees of freedom that is $n_1 - 1$. So, my degrees of freedom, degrees of freedom gets cut and if I get simplified this is my F distribution. So, what is the PDF of F distribution is S_1^2 / σ_1^2 that means the concerning the first population.

Because we are talking about comparing 2 different population so, this is the, S_1 is the standard deviation of the sample of the first population, σ_1 is the standard deviation of the first population, S_2 is the standard deviation of the sample of the second population, σ_2 is the standard deviation of the second population.

(Refer Slide Time: 05:36)

The F-Distribution

Characteristics of the F distribution

- The F -distribution is defined only for nonnegative values.
- The F -distribution is not symmetric.
- A different table is needed for each combination of degrees of freedom.
- The choice of which variance estimate to place in the numerator is somewhat arbitrary; hence the table of probabilities of the F -distribution always assumes that the larger variance estimate is in the numerator.

$d.f. = (10, 30)$

$d.f. = (6, 10)$

Monalisa Sarma
IIT KHARAGPUR

So, again F distribution since F distribution is like a division of 2 chisquare values, is not it numerator we have a chisquare denominator we have a chisquare so, definitely it will also have only nonnegative values, chisquare we already we have seen it always has nonnegative, it cannot have negative values because it is just 2 term. So, here f also will have only nonnegative values.

And here in F distribution there are 2 degrees of freedom and numerator we have 1 degrees of freedom that is the v_1 degrees of freedom denominator we have another degrees of freedom that is v_2 degrees of freedom. Basically, the numerator is the sample size that we have taken from the first population and denominator is the sample size that we have taken for the second population. So, here see there are 2 degrees of freedom 10, 30 6, 10 these are the degrees of freedom, the 2 parameters are the F distribution.

So, F distribution is defined only for non terminal and F distribution is also not symmetric. Chisquare distribution is not symmetric we have seen similarly F distribution is also not symmetric. And we need a different table for each combination of degrees of freedom. For each combination of the degrees of freedom, we need a different table. So that is why it is

very difficult to have a table for different probability values. So, in most of the standard textbook, you will find this of course, this lookup table for F calculating the F value.

Also, we have the lookup table like we have for binomial distribution for chisquare for T distribution at all. Similarly, for here, also, we have the lookup table, but we have to for very limited probability values, most of the standard textbooks have F distribution values, probability values for the 0.05 and 0.01 maybe this range. Maybe 05, I can say this is the probability, this whole is a 05 probability and 01 maybe this one this portion, only 2 probability value we have.

But there is 1 theorem, if we know this value, we can find out this value as well. How do we find that out? First, we will have to one more important point thus we have 2 variance estimate 1 in numerator, 1 in denominator; now which one will put in a numerator, which one will put in the denominator, this choice of which variance estimate to be placed in a numerator is somewhat arbitrary.

Hence the table of probabilities of the F distribution always assumed that a larger variance estimate in the numerator, the table in the standard textbook the table that we have always it puts with a larger variance in the numerator. But this is arbitrary, like you can put anything. If you can calculate the value, you can calculate yourself as well.

(Refer Slide Time: 08:23)

The F-Distribution

Writing $f_{\alpha}(v_1, v_2)$ for f_{α} with v_1 and v_2 degrees of freedom, we obtain

$$f_{1-\alpha}(v_1, v_2) = \frac{1}{f_{\alpha}(v_2, v_1)}$$

So, as I told you, you just know the value of f of alpha that means this portion. Similarly, if I want, there is a way if I can find out this portion as well, how do I find out this is there is a

theorem for that this is theorem, f of 1 - alpha if I know this value corresponding to area of 0.05. This is an area of probability means what this is nothing but this area, is not it? If I know the value of f a corresponding with a probability of 05, I can also know corresponding to the value of 1 - 0.05 that is 0.95, I can also know the value of f of 0.95.

But maybe that means this whole area, how do I find out this is the formula? But see, there is a difference f of 1 - alpha v 1 v 2 1 / f of alpha v 2 v 1, whatever I am using numerator that gets change when I am doing 1 - alpha numerator becomes denominator, denominator becomes numerator, we will be solving some problems then it will be things will be more clear.

(Refer Slide Time: 09:27)

The F-Distribution

- Writing $f_{\alpha}(v_1, v_2)$ for f_{α} with v_1 and v_2 degrees of freedom, we obtain

$$f_{1-\alpha}(v_1, v_2) = \frac{1}{f_{\alpha}(v_2, v_1)}$$

- Thus, the f-value with 6 and 10 degrees of freedom, leaving an area of 0.95 to the right, is

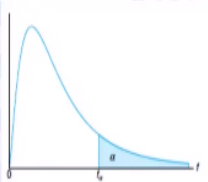
$$f_{0.95}(6, 10) = \frac{1}{f_{0.05}(10, 6)} = \frac{1}{4.06} = 0.246$$

So, if thus the f value which 6 and 10 degrees of freedom having an area of 0.95. So, I suppose we are interested in the area of 0.95 total, so to find an area of 0.95 what we will be doing, so 0.95 where we do not have in the F table, it is not given in the F table. So, what we will do I am interested in this I will calculate this f 0.05 10, 6 from the table I can get this value is not it? This is 0.05 values I will get in a table on the table I can get this value and 1 by of that I will give me this value.

(Refer Slide Time: 10:07)

The F-Distribution

		$f_{0.05}(v_1, v_2)$											
		v_1											
v_2		1	2	3	4	5	6	7	8	9	10	120	∞
1	161.45	199.5	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	...	253.25	254.31
2	18.51	19	19.16	19.25	19.3	19.33	19.35	19.37	19.38	19.4	...	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	...	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6	5.96	...	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	...	4.4	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.1	4.06	...	3.7	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	...	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.5	3.44	3.39	3.35	...	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	...	2.75	2.71
10	4.96	4.1	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	...	2.58	2.54
11	4.84	3.98	3.59	3.36	3.2	3.09	3.01	2.95	2.9	2.85	...	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3	2.91	2.85	2.8	2.75	...	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	...	2.25	2.21



So, here we have seen f of 0.05 this is 10 this is 6 numerator is v_1 denominator is v_2 I got this is 4.06. So, then my f of 0.95 6 10 will be $1 / 4.06$ that is 0.246 this value 0.246.

(Refer Slide Time: 10:34)

The F-Distribution: Example – 1

Problem

Consider the following measurements of the heat-producing capacity of the coal produced by two mines (in millions of calories per ton):

Mine 1: 8260 8130 8350 8070 8340 ✓ 5

Mine 2: 7950 7890 7900 8140 7920 7840 ✓ 6

Can it be concluded that the two population variances are equal?

So, how we will do from will to take a simple example, as I told you F distribution we use to compare 2 different variance of 2 different population this is the example, consider the following mechanism of the heat producing capacity of the coal produced by 2 mines that is in millions of calories per ton. So, this is one first mine, this is a second mine how many sample size this is 1 2 3 4 5 from the first population, we have taken a sample size of 5.

And from the second population 1 2 3 4 5 6 from the second population, we have taken a sample size of 6. So, degrees of freedom here the degrees of freedom is 4, here the degrees of freedom is 5 can it be concluded that the 2 population variances are equal? So, see what it is asking can it be concluded that the 2 population variances are equal that means, we are

estimating we are considering that the 2 population variances are equal, if the 2 population variances are equal than this data what we got from this data we will be getting a variance.

So, whatever this variance, the difference of these 2 variances whether is it, it has a higher probability basically that we need to find out, is not it? If we compare the variance of these 2 population and if we find out considering both the population variance is the same, if we find out the probability is very less that means, we can conclude that the 2 population variances are not equal, if we find a population is quite high, we can say yes, the population variance is equal like whatever we have done for all the other same type of things.

(Refer Slide Time: 12:14)

The F-Distribution: Example - 1

Solution

From previous discussion, we know, $f = \frac{s_1^2/m_1^2}{s_2^2/m_2^2} = \frac{\sigma_1^2 s_1^2}{\sigma_2^2 s_2^2}$

Considering the two population variances are equal, therefore, $\frac{\sigma_1^2}{\sigma_2^2} = 1$

Here, $s_1^2 = 15750$ and $s_2^2 = 10920$ which gives $f = 1.44$

Now, from Table $f_{0.05}(4,5) = ?$

Consider the following measurements of the heat-producing capacity of the coal produced by two mines (in millions of calories per ton):
 Mine 1: 8260 8130 8350 8070 8340
 Mine 2: 7950 7890 7900 8140 7920 7840
 Can it be concluded that the two population variances are equal?

1.44 5 19

Monalisa Sarma
IIT KHARAGPUR

So, from the previous discussion, we know this is the f value and we are considering that the both the population variances are equal, population variance is equal means, σ_1^2 / σ_2^2 squared = 1. So, from the sample given data we can calculate the variance so, you can calculate and see by yourself just S_1^2 / S_2^2 I can founded this S_2^2 founded this, then putting this in the numerator putting this in the denominator, then I got $f = 1.44$.

If when I put this upper one because higher value in a table, it is assumed that it will keep the higher value in that numerator. If I keep in higher this in the numerator, what are the degrees of freedom for here, it is 4. So, that means, I need to find out f of 1.44 for degrees of freedom 4 and 5, is not it? So, I am interested in finding out for a value of f 1.44 for degrees of freedom 4 and 5, but the f value also.

It is as I told you it is only given for some very limited value most of the books you will find that this value is given only for 0.05 and 0.01. So, just seeing the value 1.45 only we can very well make out that it is very near to 0. So, probability of that will be definitely a bigger probability, but still let us see what we have because we have only 0.5 that is 5% probability and 1% probability.

First let us check for 5% probability, what is the f value corresponding to the 5% probability first you see the figure this is the figure. So, from the figure you will be able to make it say this is if I consider this is the figure. So, this is for 5% what is the value?

(Refer Slide Time: 14:08)

The F-Distribution: Example - 1

The F table for Example 1

		$f_{0.05}(v_1, v_2)$									
		v_1									
v_2	1	2	3	4	5	6	7	8	9	10	
1	161.45	199.5	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	
2	18.51	19	19.16	19.25	19.3	19.33	19.35	19.37	19.38	19.4	
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6	5.96	
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.1	4.06	
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	
8	5.32	4.46	4.07	3.84	3.69	3.58	3.5	3.44	3.39	3.35	

$f_{0.05}(4, 5) = 5.19$

Let me find out what is the value for this say for 5% for degrees of freedom 4 and 5 it is 5.19 for that means this is 5.19 and my value calculated value f is 1.44. So, value is 1.44 is definitely to the left of this may be somewhere in this 1.44 and definitely this means this area is greater than 0.05 that greater than 5% when this probability much greater than 5% then it is obvious that whatever this is predicting that the both of the population parents are equal that is a reasonable estimate.

When we get a very less probability much lesser than 5%, much lesser than 1%, then we can say, there is no probability. But we are getting definitely more than 5% is not it? But we do not have the F distribution value to find out what is the exact probability for 1.44 but we can understand when this value 5.19 will be somewhere here and 1 will be somewhere here definitely this area will be much more.

(Refer Slide Time: 15:12)

The F-Distribution: Example – 1

Solution


From previous discussion, we know, $f = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} = \frac{\sigma_2^2 s_1^2}{\sigma_1^2 s_2^2}$


Considering the two population variances are equal, therefore, $\frac{\sigma_1^2}{\sigma_2^2} = 1$

Here, $s_1^2 = 15750$ and $s_2^2 = 10920$ which gives $f = 1.44$


Since, from Table $f_{0.05}(4,5) = 5.19$, the probability of $f > 1.44$ is much bigger than 0.05, which means the two variances can be considered as equal.

Consider the following measurements of the heat-producing capacity of the coal produced by two mines (in millions of calories per ton):
 Mine 1: 8260 8130 8350 8070 8340
 Mine 2: 7950 7890 7900 8140 7920 7840
 Can it be concluded that the two population variances are equal?





Monalisa Sarma
IIT KHARAGPUR




So, since from table the probability of f 1.44 is much bigger than 0.05 which means the 2 variances can be considered as equal. So, now, we will be discussing the next topic that is the sampling distribution of the sample proportion this is the last topic on the sampling distribution.


(Refer Slide Time: 15:38)

Sampling Distribution of the Sample Proportion


- In a random sample of 200 parts from a manufacturing process, 18 had major manufacturing defect.
- Here, \hat{p} the sample proportion

$$\hat{p} = \frac{18}{200}$$
- \hat{p} represents the proportion of the individuals or objects in the sample that has a certain characteristics.
- In statistical inference scenario, we use the sample proportion \hat{p} to estimate the population proportion p





Monalisa Sarma
IIT KHARAGPUR



So, first to start before starting sampling distribution and sample proportion let us take an analogy are very common feature which we get to see during election time exit poll you know what is an exit poll and exit poll what happens? We from the exit poll they try to predict which party will come to power or maybe which candidate will win the election. So, what is then so, in a locality suppose a particular political party affiliation or suppose there are 3 political affiliation party x, y and z 3 different parties.

So, in the election if we want to know which party will win let us make it simple a 2 party x and y I am just making it simple. So, which party will win the election? So, what is the exit poll what we do is that do we interview from we know exit poll means do we interview the people after when they both cast their vote and come out then we sort of take the information from them and accordingly based on that we give the information.

But is it that we try to find out from all the people, no we do not try to it is not possible many term lakhs and lakhs of people come and cast your vote we do not try to find out the information from all those lakhs and lakhs of people, but just we take a small sample of those lakhs and lakhs of people from the whole population we take a small people from a subset of this people, but it has to be unbiased, it should not be biased, unbiased.

In the sense maybe we have taken the feedback maybe some from different age category, different profession, different locality, that that may be an unbiased estimator, we try to take out the thing what to say there which political party they have cast a vote and based on whatever result we get, based on that we try to infer about the election result that means, based on that, we try to infer what the whole population might have voted, is not it? That is the exit poll.

So, now what is that is means is a big population from the population we want to infer something about the population we cannot do that. So, when we take a sample of the population that means we take a proportion of that, from that we try to infer about the population that is for that we need to do again this proportion is a random variable there can be different values if we take different samples there can be different values.

So, random variables means there will be different frequency of observation meaning it will also have a probability distribution that probability distribution is sampling distribution of sample proportion. With that diagram, let us now see. So, in a random sample of 500 parts from a manufacturing process 18 has major manufacturing defect. 2 from 200 parts 18 had major manufacturing defects it is that is a sample again, that is not a population. So, let me tell sample proportion, which how much proportion is defective?

How much proportions of the people have caused their 4 for political party x? That proportion I am telling it as a p cap so my p cap is $18 / 200$. So, what this p cap represents? p

\hat{p} represents the portion of the individuals or objects in the sample that has a certain characteristic here, what is the characteristic that \hat{p} has, that is the portion the defective portion of the whole sample as it poll the portion maybe the portion of the population who have voted for political party x or political party y that is \hat{p} .

In statistical inference scenario, we use the sample proportion \hat{p} to estimate the population proportion p . We now from this sample, we have to estimate the population proportion from the exit poll we have to make an estimate of the whole population. The whole population may have voted for which political party which person is basically may have voted for which political party, so that is from this \hat{p} we have to find out estimate of the population proportion let me take population proportion.

Let me take that I have written it is p . So, from \hat{p} I have to estimate for p . Now for doing that, I need a sampling distribution of \hat{p} , sampling distribution \hat{p} means I need to know \hat{p} represent what sort of distribution and accordingly whatever distribution I need to know the characteristics of the distribution, if it is a what is a normal distribution then I will be needing to know the mean and the variance of the distribution.

Similarly, whatever distribution is this I need to know the characteristics of the distribution then only once I have the characteristics of the distribution then only I can use the distribution to evaluate the probability of a particular occurrence. So, now, I have to first find out what is the \hat{p} we will have what distribution then once we know that people what distribution then accordingly what are the different parameters of this distribution, we need to find out the values of those parameters.

(Refer Slide Time: 20:47)

Characteristics of the sampling distribution of \hat{P}

- Let us assume we are sampling from a infinite population, or rather we are sampling a small fraction from a large population.
- We can view the sample proportion as \hat{p} , where $\hat{p} = \frac{X}{n}$

where

X is a random variable that represents the no of individuals in the sample with the characteristic of interest, and n represents the no of individuals in the sample.

Monalisa Sarma
IIT KHARAGPUR

So, in that angle first we will see, you know, what we will assume? Let us assume that we are sampling from an infinite population or rather we are sampling a very small fraction of the last population. So, we can view the sample proportion \hat{p} as X / n what is X ? X is a random variable that represents a number of individual in the sample with the characteristics of interest they are what is X and my example X is the number of defective components.

The characteristics of interest means here are characteristics of interest is defective component, characteristic of interest is people who have voted for X political party X and n represents the number of individual in the sample total sample what we have taken that is so, \hat{p} is X / n .

(Refer Slide Time: 21:43)

Sampling Distribution of the Sample Proportion

Characteristics of the sampling distribution of \hat{p}

- Let us assume we are sampling from a infinite population, or rather we are sampling a small fraction from large population.
- We can view the sample proportion as \hat{p} , where $\hat{p} = \frac{X}{n}$
- Here, X is a discrete random variable, and the value it can take is $0, 1, 2, \dots, n$; i.e. $(n+1)$ possible values.
- X can be thought as a binomial random variable with parameter n and p
- Recall that the binomial random variable X has a
 - mean = np
 - variance = $np(1 - p)$
 - It is approximately normally distributed for large sample size

Monalisa Sarma
IIT KHARAGPUR

So, now X what is X here? That is very important now, X is very much a discrete random variable it X cannot take any values in an interval, it will only take some discrete values is

not it? So, X is definitely not a continuous random variable it is a discrete random variable and find it is a description or we will not watch sort of district random variable and what the value it can take 0, 1, 2 up to n what value 0 X can take?

X can take value 0 no defective component that means p cap is $0/n$, 1 defective component in a sample $1/n$, 2 defective components in a simple $2/n$ and n defective component in a simple n/n . So, it can take total $n + 1$ values, is not it? Now, this X cannot we see that this X is very much a binomial random variable is not it? Because what does X indicates either the presence of the characteristics or the absence of the characteristic, in binomial random variable X what is that X ? X is either what is X ?

X is the total number of failure or the total number of heat, total number of miss, whatever it is success or failure heat miss, whatever it is like here similarly, X is the thing whether it contains the desired characteristics what is X here? Characteristics of interest. So, what is my characteristic of interest in our case in our example, if needed a defective and what value it can take? It can take 0, 1, 2 like the probability of 0 success, probability 1 success, probability 2 success.

Similarly, probability of 0 defective, probability of 1 defective, probability of 2 defective. So, X can take this total number of defective in n trials and n is the number of trials in total trails how many defectives can be there, is not it? And it is probability of each trial getting a defective in each trail is independent there is no any dependency in it yes or no? So, we can very well say that X is a discrete of course it is this thus the X is a binomial random variable.

When it is a binomial random variable, then we need to know binomial there are 2 parameters n and p what is n sample size, p is the probability of success or the probability failure whatever it is, so, we know n we know p , X can be thought as a binomial random variable parameter n and p . Now, this is X now what is p cap? p cap is X/n let us go there later. Now binomial random variable X as parameter n and p what is the mean of this?

Mean is np we know mean of binomial random variable is np variances npq let us $np \times 1 - p$ and moreover, binomial interval it is approximately normally distributed for large sample size. If we take a larger sample size it is approximately normally distributed we have seen this well we have discussed binomial distribution know when we have discussed normal

distribution then we have mentioned it even binomial distribution also when the sample size is larger number of trials is very large.

Instead of considering binomial distribution we consider normal because it is almost same it can be approximated. So, if sample size is large it can be approximated as normally distribution, we know what is mean we know what is variance?

(Refer Slide Time: 25:09)

Sampling Distribution of the Sample Proportion

Characteristics of the sampling distribution of \hat{p}

- Now to derive the characteristics of the sampling distribution of the sample proportion \hat{p} :
- The mean of the sampling distribution $= E(\hat{p}) = E\left(\frac{X}{n}\right) = \frac{1}{n} E(X) = \frac{1}{n} \times np = p$
 - And so, we can say on an average the sample proportion equals the population proportion.
- The variance of the sampling distribution of \hat{p}

$$= \sigma_{\hat{p}}^2 = \text{Var}(\hat{p}) = \text{Var}\left(\frac{X}{n}\right) = \frac{1}{n^2} \text{Var}(X) = \frac{1}{n^2} np(1-p) = \frac{p(1-p)}{n}$$
- The standard deviation $= \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$

Monalisa Sarma
IIT KHARAGPUR

Now, we need to derive the characteristics of p cap, we understood what is X ? X we can consider it a binomial random variable, we know the different mean and variance parameter as well. And we know that X can be remember if n is large X can be approximated by normal distribution. Now, what is p cap see p cap is what? p cap is X / n , p cap is nothing but X / n . Now, what values p cap can take? p cap can take $0 / n$, $1 / n$, $2 / n$, $3 / n$ there is also again discrete value.

Again p cap also cannot take any continuous as well a p cap also can take this discrete the values are $0 / n$, $1 / n$, $2 / n$, $3 / n$ and n / n that is a valid p cap can take similar to the values that extremity and the probability that p cap will take $0 / n$ is same as the probability that X will take value 0 probability that p cap will take $1 / n$ value is the same as the probability that X will take value 1.

So, say p cap is also when X is a binary random variable n is a fixed with a constant value, then definitely p cap also we can consider that it is a binomial random variable. So, if p cap is a binomial random variable now that means, we need to know the characteristics of n and p

that means the mean and the variance of the \hat{p} . So, mean of the sampling distribution E of \hat{p} is what for this \hat{p} ? \hat{p} is X / n we have initially when we have learned a mean of variable random variable.

So, if there is a constant number comes out mean of X / n is nothing but $1 / n E$ to the power X and E to the power X we have already seen it as np so, mean of \hat{p} is p see, mean of \hat{p} is p that means, we can say on an average the sample proportion equals the population proportion so, on an average the sample proportion equals the population proportion. Now, the variance of the sampling distribution of \hat{p} is where of \hat{p} nothing but we have X / n .

X remember where of X / n when there is a constant this is $1 / n^2$ of X . So, where of X is what npq then we got is the variance. Here also since \hat{p} is again and we can consider is a normal random variable if the size of N is large, we can \hat{p} as a binomial random variable, and it is the size of n is large we can consider it as a normal random variable.

(Refer Slide Time: 28:09)

Sampling Distribution of the Sample Proportion

Characteristics of the sampling distribution of \hat{p}

- The sampling distribution of \hat{p} is approximately normal if the sample size is large
- So, finally we can say, for large sample size, the sampling distribution of \hat{p} is approximately normal, represented as

$$N\left(p, \frac{p(1-p)}{n}\right)$$

Monalisa Sarma
IIT KHARAGPUR

So, therefore, and we can write so, sampling distribution of \hat{p} is approximately normal if the sample size is large and this is how we can represent, the sampling distribution of \hat{p} is approximately normal distribution we need mean and variance. So, this is mean, this is variance so, we have the sampling, we have the distribution, we have the mean we are the variance now, we can calculate a probability.

(Refer Slide Time: 28:40)

Sampling Distribution: Example – 2

Problem

It is believed that 20% of voters in a certain city favor a tax increase for improved schools. If these percentage is correct, what is the probability that in a sample of 250 voters 60 or more will favor the tax increase?

When we will do 1 example before complete this topic. So, it is believed that 20% of the voters in a certain city favour a tax increase for improved school it is believed that is estimated it is predicted 20% of the voters in a certain city favour a tax increase. So, that is the population proportion is given what is the proportion is 20%, if this percentage is correct, what is the probability that in a sample of 250 voters 60 or more will favour the tax increase. If this is correct, what is the probability of this happening? So, my p cap is what 60 / 250 I have to find out probability that of p cap greater than 60 / 250.

(Refer Slide Time: 29:21)

Sampling Distribution: Example – 2

Solution

The sampling distribution of proportion is approximately normal with mean $\mu_{\hat{p}} = 0.2$

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.2 \times 0.8}{250}} = 0.025$$

$$\therefore P\left(\hat{p} > \frac{60}{250}\right) = P\left(Z > \frac{\frac{60}{250} - 0.2}{0.025}\right)$$

$$= 1 - P(Z < 1.6)$$

$$= 1 - 0.9452$$

$$= 0.05$$

For that first I found out first I know this is a normal distribution first I need to know the mean I need to know the standard deviation what will be my mean? Mean is on an average we have seen here on and every sample proportion equals the population proportion, is not it? So, my mean is 20%. So, my mean of p cap is 20% that is the population proportion. My

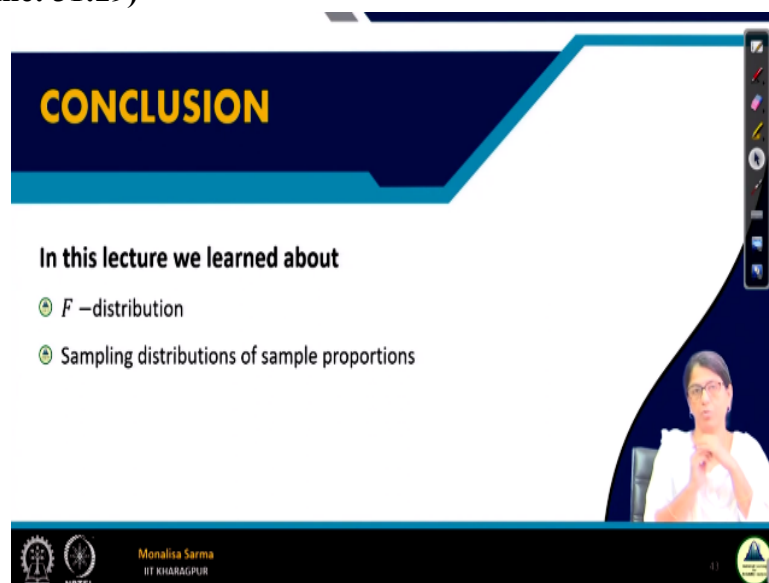
variance this is the formula for variance and standard deviation is the formula for standard deviation.

Then what I need to find out p cap is greater than this is the probability I need to find out what is the probability given, what is the probability that in a sample of 250 voter 60 or more will favour a tax increase? So, p cap greater than $60 / 250$, normal distribution that means we will have to convert it to Z , because we do not have the normal table lookup table here, the Z lookup table, so we have converted it to Z probability of Z . So, what is this, Z is what? X bar $- \mu / \sigma$.

So, what is X bar here X bar is nothing but p cap my X bar value is nothing but a p cap here. So, this is p cap -0.2 , this is the σ value and what I got is 1.6 . So, probability of Z greater than 1.6 is $1 -$ probability of Z less than 1.6 . 1.6 and this from the table, I am not showing it I have to show it many places from the Z table I can find out Z probability corresponding to Z less than 1.6 this probability. So, this is the probability. If that is correct probability then a sample of 250 voters 60 or more will favour the tax increase that probability is 5% probability.

Now, it is up to you whether you will accept that what it is believed that 20% of voters is correct or not if you find this probability is very small, then we will say that no whatever is believed it is not correct, if you find that there is this 5% this may be some variation, then we can consider that.

(Refer Slide Time: 31:29)



CONCLUSION

In this lecture we learned about

- F -distribution
- Sampling distributions of sample proportions

Monalisa Sarma
IIT KHARAGPUR

NPTEL

The slide features a dark blue header with the word 'CONCLUSION' in yellow. Below the header, the text 'In this lecture we learned about' is followed by two bullet points: '• F -distribution' and '• Sampling distributions of sample proportions'. At the bottom, there is a footer with the NPTEL logo, the name 'Monalisa Sarma', and 'IIT KHARAGPUR'. A small video inset in the bottom right corner shows a woman with glasses speaking.

So, with this I conclude this topic on sampling distribution. In this lecture, we have learned distribution and sampling distribution of sample proportion. So, here I conclude the portion on sampling distribution. So, sampling distribution, we have learned Z distribution, sampling distribution of mean using both Z distribution and T distribution, sampling distribution of variance using chisquare distribution, sampling distribution for what is a combination of 2 population variance, the product we have used F distribution, then again we have seen sampling distribution of proportion.

(Refer Slide Time: 32:10)



These are my references and thank you guys.