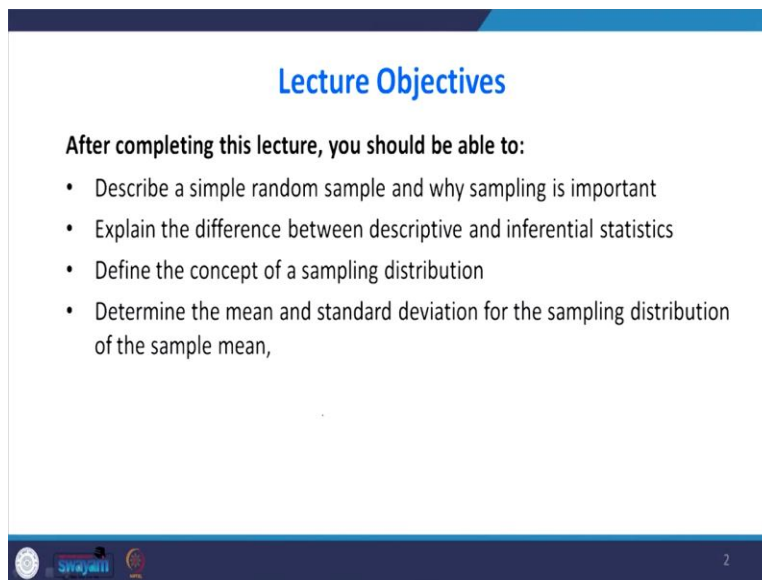


Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Management Studies
Indian Institute of Technology – Roorkee

Lecture – 11
Sampling and Sampling Distribution

Dear students we are going to the next lecture that is sampling and sampling distributions. The objective here is; objective of this lecture is describing a simple random sample and why sampling is important.

(Refer Slide Time: 00:40)



Lecture Objectives

After completing this lecture, you should be able to:

- Describe a simple random sample and why sampling is important
- Explain the difference between descriptive and inferential statistics
- Define the concept of a sampling distribution
- Determine the mean and standard deviation for the sampling distribution of the sample mean,


2

Explain the difference between descriptive and inferential statistics and defining the concept of sampling distribution. Determining the mean and standard deviation of the sampling distribution of the sample mean that very important theorem that we are going to see in this class, the central limit theorem and its importance and determining the mean and standard deviation of the sampling distribution of the sample proportions then at the end we will see the sampling distribution of sample variances.

(Refer Slide Time: 01:17)

Descriptive vs Inferential Statistics

- **Descriptive statistics**
 - Collecting, presenting, and describing data
- **Inferential statistics**
 - Drawing conclusions and/or making decisions concerning a population based only on sample data




4

The whole statistics can be classified into 2 categories one is the descriptive statistics another one is the inferential statistics. The descriptive statistics is only for collecting and presenting describing the data as it is it is very low-level statistics. Whereas the inferential statistics drawing conclusions are making decisions concerning a population based on sample data, in the inferential statistics with the help of sample data we are going to infer something about the population. So, when you say population you should know what is the population what is the sample?

(Refer Slide Time: 01:58)

Populations and Samples

- A **Population** is the set of all items or individuals of interest
 - **Examples:**
 - All likely voters in the next election
 - All parts produced today
 - All sales receipts for November
- A **Sample** is a subset of the population
 - **Examples:**
 - 1000 voters selected at random for interview
 - A few parts selected for destructive testing
 - Random receipts selected for audit

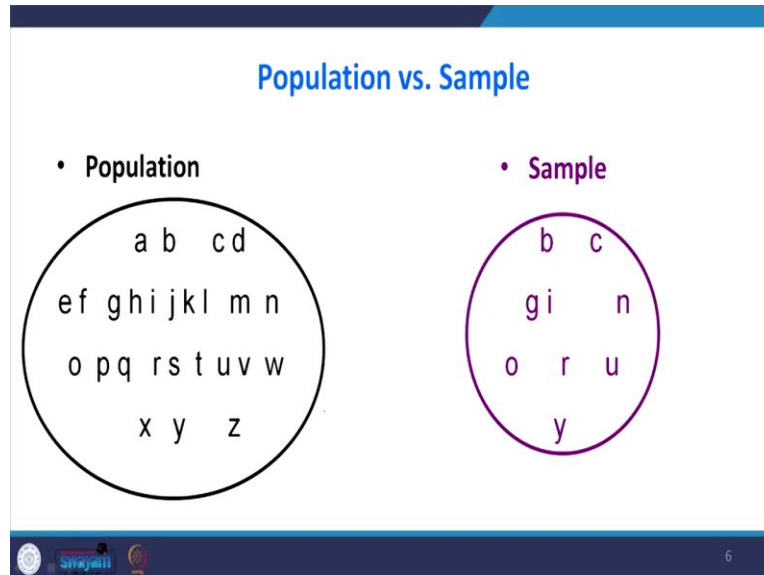


5

Population is the set of all items are individual of interest for example all likely voters in the next election, all parts produced today, all sales received for November. The sample is the subset of

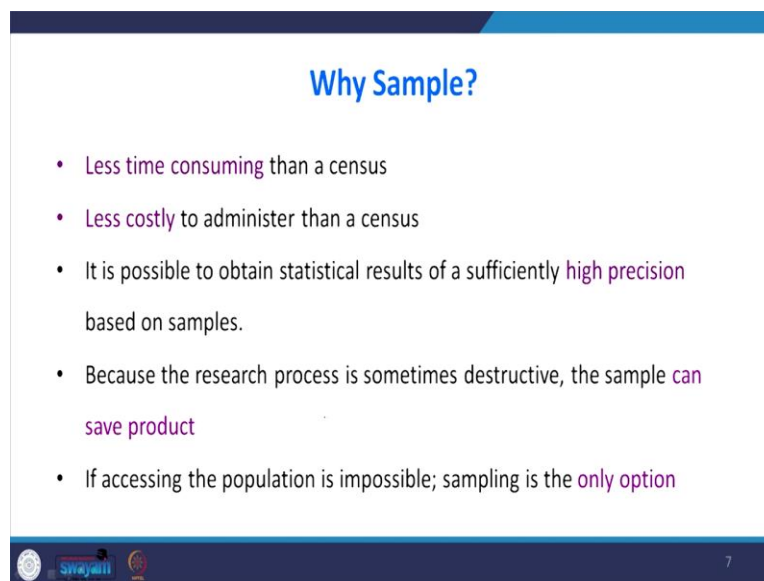
the population like 1000 voters selected at random for interview, a few parts selected for destructive testing, random received selected for audit. This is an example of sample.

(Refer Slide Time: 02:33)



When you look at the left hand side there is a bigger circle that is the population from there some numbers are bigger the collection of that picked of the values is called a sample.

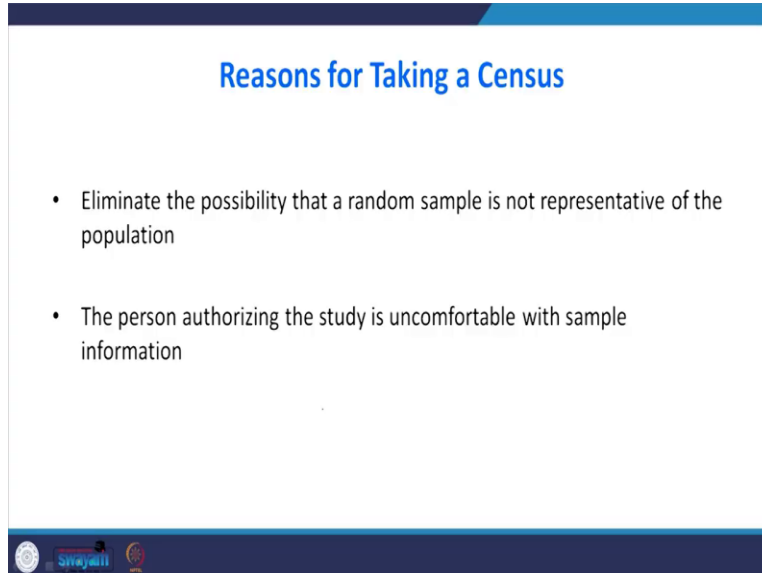
(Refer Slide Time: 02:35)



The question may come why we out to sample it is less time-consuming than a census less costly to administer then your census. It is possible to obtain statistical result of your sufficient the high precision based on the samples. Because of the research process sometimes destructive the sample can save the product. If accessing the population is impossible sampling is the only

option. Sometimes you have to go for census also we are in census we will examine each and every item in the population.

(Refer Slide Time: 03:17)



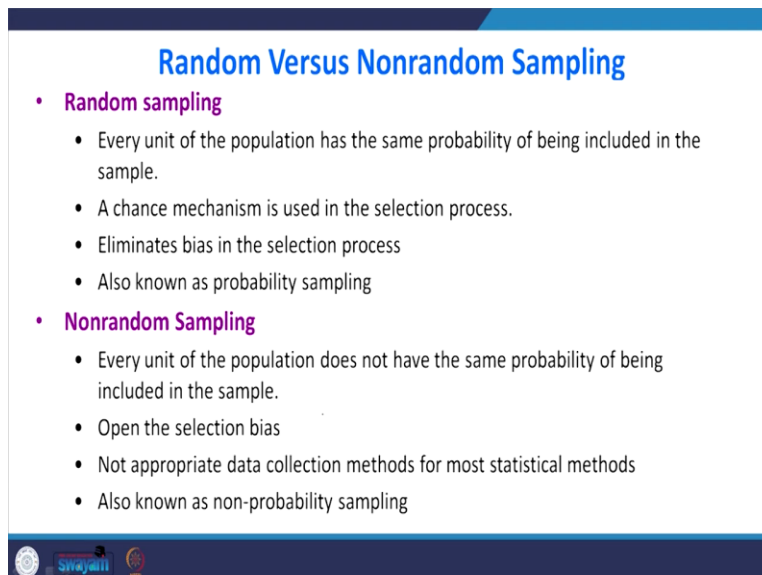
Reasons for Taking a Census

- Eliminate the possibility that a random sample is not representative of the population
- The person authorizing the study is uncomfortable with sample information

The slide features a blue header and footer with the Swayam logo. The main content is on a white background with a blue border.

Suppose if we need to have higher accuracy and you are not comfortable with the sample data then used to go for census. The reasons for taking a census because census eliminates the possibility that random sample is not representative of the population many time there is a chance that the sample which you have taken may not represent the population. Otherwise the person authorizing the study is uncomfortable with the sample information then you go for census.

(Refer Slide Time: 03:40)



Random Versus Nonrandom Sampling

- **Random sampling**
 - Every unit of the population has the same probability of being included in the sample.
 - A chance mechanism is used in the selection process.
 - Eliminates bias in the selection process
 - Also known as probability sampling
- **Nonrandom Sampling**
 - Every unit of the population does not have the same probability of being included in the sample.
 - Open the selection bias
 - Not appropriate data collection methods for most statistical methods
 - Also known as non-probability sampling

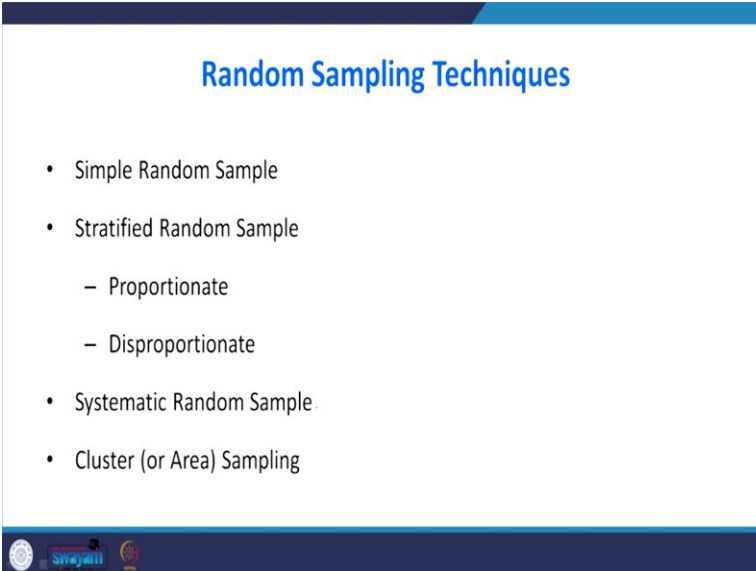
The slide features a blue header and footer with the Swayam logo. The main content is on a white background with a blue border.

We will see what is sampling? Sampling is generally selecting some items from the population that is a sampling. So, there are that can be classified into two way one is random sampling another one is a non random sampling in the random sampling. The concept of randomness is taken care non random sampling the randomness is not there. Sometimes we may go for non random sampling even though it is not so comfortable that is not good for doing many statistical analysis sometimes we have to go for non random sampling.

But they random sampling the outcome or the generalization which you provide with the help of random samplings are highly robust. So, we will go for what is the random sampling? Every unit of the population has the same probability of being included in the sample that is the concept of your randomness. A chance mechanism is used to selection of the process because the chance of mechanism is we can use a random table to choose someone, you can use your calculator you can choose someone, choose someone randomly that eliminates the bias in the selection process also known as the probability sampling.

They will go for non random sampling every unit of the population does not have the same probability of being included in the sample. It is open the you know selection bias there is a possibility selection bias not appropriate data collection methods for most statistical methods. So, it is not good method for doing some statistical analysis also known as non-probability sampling.

(Refer Slide Time: 05:18)



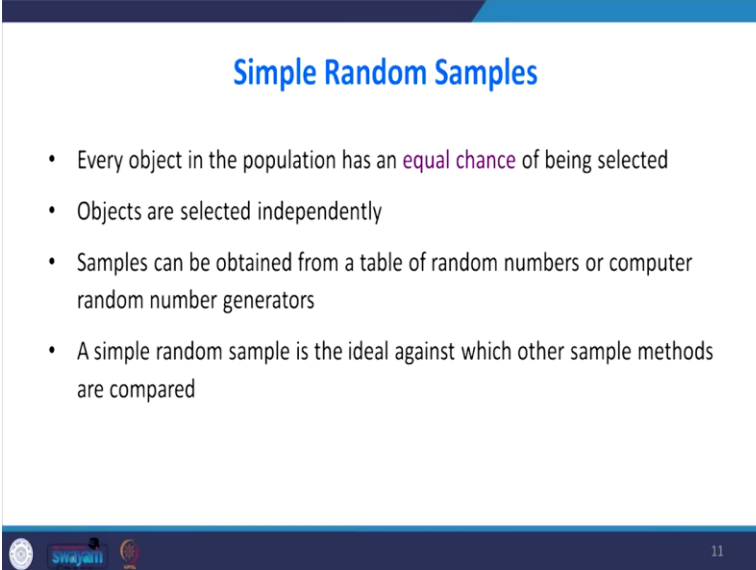
The slide is titled "Random Sampling Techniques" in blue text. It lists the following sampling methods:

- Simple Random Sample
- Stratified Random Sample
 - Proportionate
 - Disproportionate
- Systematic Random Sample
- Cluster (or Area) Sampling

At the bottom left of the slide, there are logos for "Sri Jayanti" and "SRM Institute of Science and Technology".

Random sampling techniques there are 4 way we can say of selecting random one is the simple random sample second one is a stratified random sample with the proportion disproportionate third one is a systematic random sample fourth one is cluster or area sampling. Simple random samples every object in the population has an equal chance of being selected objects are selected independently.

(Refer Slide Time: 05:48)



Simple Random Samples

- Every object in the population has an **equal chance** of being selected
- Objects are selected independently
- Samples can be obtained from a table of random numbers or computer random number generators
- A simple random sample is the ideal against which other sample methods are compared

11

Samples can be obtained from your table of random numbers or computer random number generators. A simple random sample is the ideal against which the sample methods are compared this is a best method.

(Refer Slide Time: 05:59)



**Simple Random Sample:
Numbered Population Frame**

01 Andhra Pradesh	11 Madhya Pradesh
02 Himachal Pradesh	12 Uttar Pradesh
03 Gujrath	13 Bihar
04 Maharashtra	14 Rajasthan
05 Nagaland	15 J & K
06 Goa	16 Tamil Nadu
07 West bengal	17 Karantaka
08 Haryana	18 Kerala
09 Punjab	19 Orissa
10 Delhi	20 Manipur

11

Suppose we will see there are 20 states have ever taken suppose I want to choose some states randomly for some studies. Suppose first task is I have given some number 2-digit number 01, 02 for example up to this one, it is only for illustrate the purpose it is not counting the number of states are more.

(Refer Slide Time: 06:22)

**Simple Random Sampling:
Random Number Table**

99	437	879	61	457	37	375	52	979	69	390	94	344	75	31	618
50	656	00	127	683	67	668	82	08	156	800	16	782	24	58	326
80	880	63	171	428	77	668	35	605	15	702	96	500	26	45	587
86	420	408	53	537	98	894	54	681	30	912	53	881	04	74	319
60	097	864	36	018	69	477	58	895	35	994	00	482	68	30	606
52	587	719	65	854	53	468	34	009	91	997	29	769	48	15	941
89	155	905	53	906	89	486	37	079	55	470	62	711	82	64	493

Next I am using the random table to choose the States randomly. For example you can start from you can see this is a random table you say see the table you can follow any 2 digit 99, 43, 78, 79, 61 because the random table can be read at any direction. so suppose if I am reading left to right 99, 43, 78, 76, 61, 45 and so on so 53 next is 16 so 16 is I have to choose the serial number 16 and corresponding states I am going back so the 16 is Tamilnadu. So, one state is chosen the next random number is 18 so the 18 is Kerala next state is chosen.

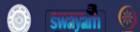
Next 50 there is no number 65 there is no number 60 there is no number but 01 number is a 01 is Andhra Pradesh then 27, 27 is not there 68 not there's 36 not there 76 not there 68 not there 82 is not but 08 is there 08 it is Haryana. So, like this I this is the way to use a random table to choose from the population. Here the population is the number of states suppose I want to choose some states randomly for my study so I can use the this random number table.

Suppose so the capital N is a 20 n is 4 so capital N represents n the population n represents the sample size.

(Refer Slide Time: 08:08)

Stratified Random Sample

- Population is divided into non-overlapping subpopulations called strata
- A random sample is selected from each stratum
- Potential for reducing sampling error
- Proportionate -- the percentage of these sample taken from each stratum is proportionate to the percentage that each stratum is within the population
- Disproportionate -- proportions of the strata within the sample are different than the proportions of the strata within the population



Then we will go for stratified sampling so the population is divided into non-overlapping subpopulations called strata. Random sample is selected from each stratum potential for reducing sampling error. We can go for proportionate the percentage of these samples taken from each stratum is proportion to the percentage that each stratum is within the population. We can go for disproportionated also the proportion of strata within the sample are different than the proportion of the strata within the population.

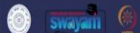
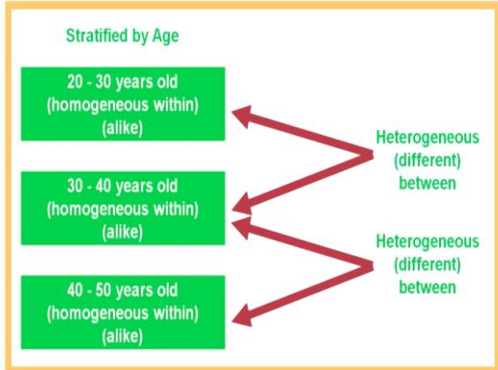
(Refer Slide Time: 08:41)

Stratified Random Sample: Population of FM Radio Listeners

Stratified by Age

20 - 30 years old (homogeneous within) (alike)	Heterogeneous (different) between
30 - 40 years old (homogeneous within) (alike)	
40 - 50 years old (homogeneous within) (alike)	

Heterogeneous
(different)
between

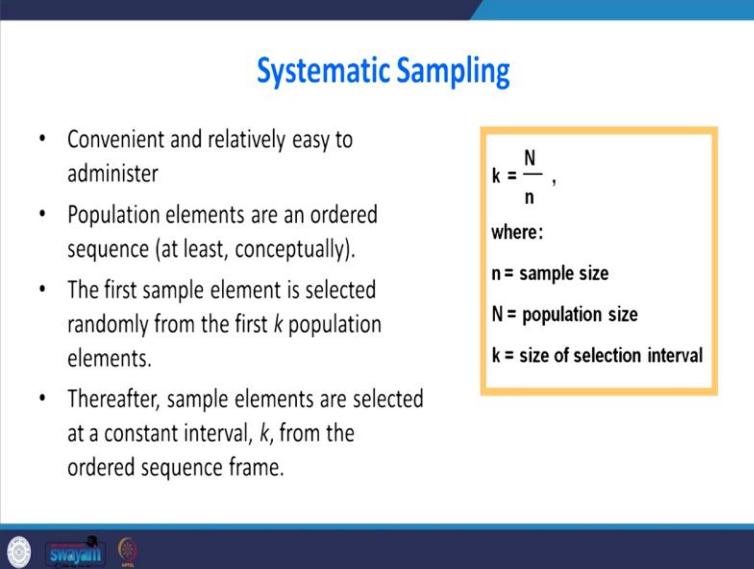


For example stratified random sample population of FM radio listeners so what I have done the whole population is divided into 3 stratum one is 20 to 30, 30 to 40, 40 to 50 you see that each

stratum are homogenous within between the stratum there may be a difference maybe there is a different variance but the same stratum will have is homogeneous the similar kind of behavior are dataset it will be there. Why it is reducing the sampling error that you if you choose 20 to 30 if you choose something from this strata so all we will we have the similar characteristics.

If you choose number some numbers 40 to 50 these sample will have similar characteristics. See that between the stratum it is a heterogeneous within the strata it is homogeneous.

(Refer Slide Time: 09:39)



Systematic Sampling

- Convenient and relatively easy to administer
- Population elements are an ordered sequence (at least, conceptually).
- The first sample element is selected randomly from the first k population elements.
- Thereafter, sample elements are selected at a constant interval, k , from the ordered sequence frame.

$$k = \frac{N}{n},$$

where:
 n = sample size
 N = population size
 k = size of selection interval

Swayamii

Then next method is the systematic sampling it is convenient and relatively easy to administer the population elements are ordered in sequence. The first sample element is selected randomly from the first K population element. Thereafter the sample elements are selected at a constant interval k from the ordered sequence of frame. What is the k is, k is the population size divided by sample size. The k represents the size of selection interval we will see an example.

(Refer Slide Time: 10:12)

Systematic Sampling: Example

- Purchase orders for the previous fiscal year are serialized 1 to 10,000 ($N = 10,000$).
- A sample of fifty ($n = 50$) purchases orders is needed for an audit.
- $k = 10,000/50 = 200$
- First sample element randomly selected from the first 200 purchase orders. Assume the 45th purchase order was selected.
- **Subsequent sample elements: 245, 445, 645, . . .**

Suppose the purchase order is from the previous fiscal year serialized one to 10,000 so capital N is 10,000 a sample of 50 n equal to 50 purchases orders need to be selected for an audit so here K is 10,000 will be 50 that is a 200, K is the interval so the first sample element randomly selected from the first 200 purchases assuming that we have chosen 45th the purchase order from the 45th you have to add 200, so 45th plus 200 245 245 + 200 445, 440 + 645 645 and so on.

(Refer Slide Time: 10:12)

Cluster Sampling

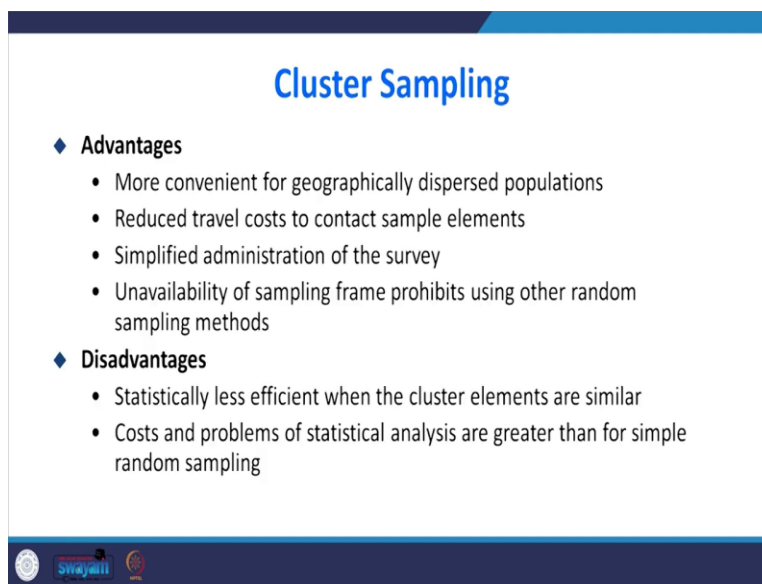
- Population is divided into non-overlapping clusters or areas
- Each cluster is a miniature of the population.
- A subset of the clusters is selected randomly for the sample.
- If the number of elements in the subset of clusters is larger than the desired value of n , these clusters may be subdivided to form a new set of clusters and subjected to a random selection process.

Then we will go for the cluster sampling here the population is divided into non-overlapping clusters or areas. Each cluster is miniature of the population the subset of cluster is selected randomly from the sample if the number of elements in the subset of cluster is larger than the desired value of n these clusters may be subdivided into form a new set of clusters and subjected

to a random selection process. Because each cluster will behave like your population now you may ask the difference between stratified sampling and cluster sampling.

In stratified sampling the things are homogeneous in each stratum the items within the in Stratham of homogenous but in cluster sampling it is highly heterogeneous and each cluster will act like your population. For example say a upwardly cluster Ludhiana a upwardly cluster Tirupur or these are the example of clusters because each cluster will have similar characteristics but will have different variants.

(Refer Slide Time: 12:04)



The slide is titled "Cluster Sampling" in blue text. It lists two categories: "Advantages" and "Disadvantages", each marked with a blue diamond icon. The advantages include being more convenient for geographically dispersed populations, reduced travel costs, simplified administration, and being the only method when a sampling frame is unavailable. The disadvantages include being statistically less efficient when cluster elements are similar and having higher costs and analysis problems compared to simple random sampling. The slide footer contains logos for Swayam and other educational institutions.

Cluster Sampling

- ◆ **Advantages**
 - More convenient for geographically dispersed populations
 - Reduced travel costs to contact sample elements
 - Simplified administration of the survey
 - Unavailability of sampling frame prohibits using other random sampling methods
- ◆ **Disadvantages**
 - Statistically less efficient when the cluster elements are similar
 - Costs and problems of statistical analysis are greater than for simple random sampling

So, we will go for advantages of cluster sampling it is more convenient for geographically dispersed a population, reduced travel cost to contact the sample elements, simplify the administration of the survey because the cluster itself will act as a population. Unavailability of sampling frame prohibits using other random sampling methods because there is no other method we can go for a cluster sampling. The disadvantage is statistically less efficient when the cluster elements are similar.

Because that cannot be generalized cost and problem of static analysis are greater than simple random sampling.

(Refer Slide Time: 12:40)

Nonrandom Sampling

- **Convenience Sampling:** Sample elements are selected for the convenience of the researcher
- **Judgment Sampling:** Sample elements are selected by the judgment of the researcher
- **Quota Sampling:** Sample elements are selected until the quota controls are satisfied
- **Snowball Sampling:** Survey subjects are selected based on referral from other survey respondents

The next kind of sampling technique is non-random sampling the first one is the convenience sampling because based on the convenience of the researcher the sample is selected. Next one is the judgement sampling sample elements are selected by the judgement of the researcher for example suppose you administering a questionnaire suppose that questionnaire can be understood only by a manager then you have to look for only the managers. So, the researcher is judging that who should fill this questionnaire so judgment sampling.

Then quota sampling sample elements are selected until quota controls are stratified. Suppose say some Uttarakhand there are some districts and each distinct I have to collect some sample so I may have some quota for example in Haridwar district how much sample has to be collected some other district how many sample has to be collected. So, there is a quota sampling. Snowball sampling is a very familiar that survey objects are selected based on the referral from other survey respondents.

Suppose you may approach one respondent out ever the survey is over you can ask him to refer his friends, so that is a snowball sampling. It is a very common method in the research.

(Refer Slide Time: 13:56)

Errors

- Data from nonrandom samples are not appropriate for analysis by inferential statistical methods.
- Sampling Error occurs when the sample is not representative of the population
- Non-sampling Errors
 - Missing Data, Recording, Data Entry, and Analysis Errors
 - Poorly conceived concepts , unclear definitions, and defective questionnaires
 - Response errors occur when people do not know, will not say, or overstate in their answers

Then there are some errors when we go when we go for sampling. Data from non-random samples are not appropriate for analysis of inferential statistical methods that was there a very important drawback because you cannot generalize because there is no randomness. Sampling error occurs when the sample is not the representative of the population, if the sample is not representing the population then whatever analysis you do that will become futile.

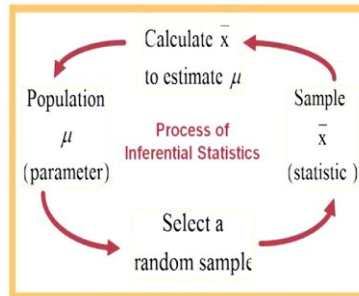
So, non sampling error suppose if you go for apart from this sampling procedure sometime there may be missing data that may be problem and recording there may be problem with the data entry there may be analysis error. Sometime the poorly consumed concepts unclear definition and defective questionnaires that also lead to error. Sometime response error occurs when the people may not understood what is the questionnaire.

Suppose there is option that not know will not say sometimes the respondent may water state their answers these are the possible error when you go for sampling. There is one more error type 1 and type 2 error that we will see in the coming classes.

(Refer Slide Time: 15:19)

Sampling Distribution of \bar{X}

Proper analysis and interpretation of a sample statistic requires knowledge of its distribution.



So, now is to go to the sampling distribution of mean here Expo represents the mean so the proper analysis and interpretation of your sample statistic require knowledge of its distribution that is a sampling distribution. For example we start from population say population is μ select a random sample from the sample you select the sample statistic into statistic it is not statistic yes there is no s, so whatever things would you say about the sample it is called a statistic T statistic Z statistic X-bar these are; since we you calculated from the sample we are calling it to statistic.

With the help of sample mean you can calculate or estimate the operation mean this is the process of we were inferential statistics. So, what is happening something we are going to assume about the population once we assume that population that is generally called hypothesis then we will take a sample randomly we will do some sample statistic with the help of sample statistic we can estimate the population mean or we can estimate the population variance. In this contest currently we are estimating the population mean.

(Refer Slide Time: 16:37)

Inferential Statistics

- Making statements about a population by examining sample results

Sample statistics (known) $\xrightarrow{\text{Inference}}$ Population parameters (unknown, but can be estimated from sample evidence)

Sample Population

24

This picture shows the inferential statistics there is a see there are bigger circle that is the population. So, the population parameter is unknown but can be estimated from the sample evidence see the red one shows that the sample statistic. So, what is the inferential statistics is making statements about a population by examining sample result that is the inferential statistic.

(Refer Slide Time: 17:04)

Inferential Statistics

Drawing conclusions and/or making decisions concerning a population based on **sample** results.

- **Estimation**
 - e.g., Estimate the population mean weight using the sample mean weight
- **Hypothesis Testing**
 - e.g., Use sample evidence to test the claim that the population mean weight is 120 pounds

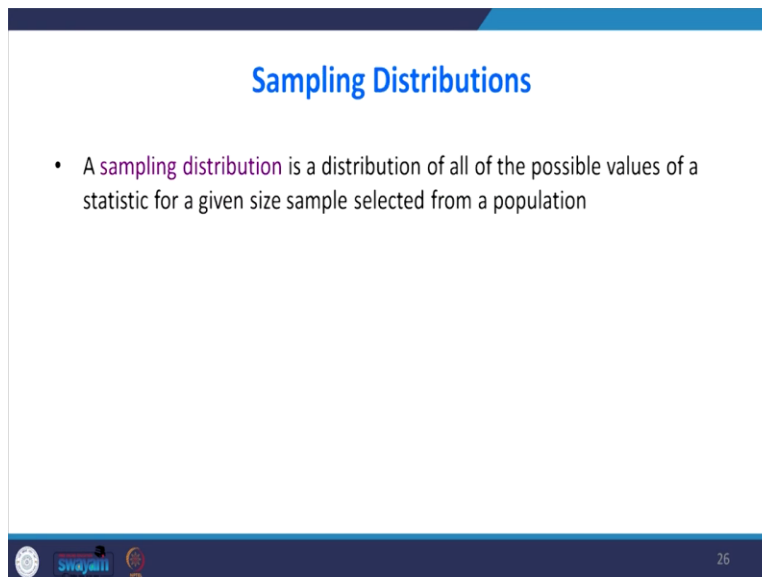
25

See another example of inferential statistics drawing conclusions or making decision concerning a population based on these sample results. You see there are different red color is there. So, these 1 2 3 4 5 6 7 this as the sample the whole things in the population, the inferential statistics is used for estimation estimating the population mean weight using the sample mean weight. For example if you want to know the weight of the population that can be estimated with the help of

weight of the sample mean then this inferential statistics are another application verse for hypothesis testing.

We can use sample evidence to test the claim that the population mean weight is for example 120 pounds are not. We will go in detail about the statistics in coming lectures.

(Refer Slide Time: 17:57)



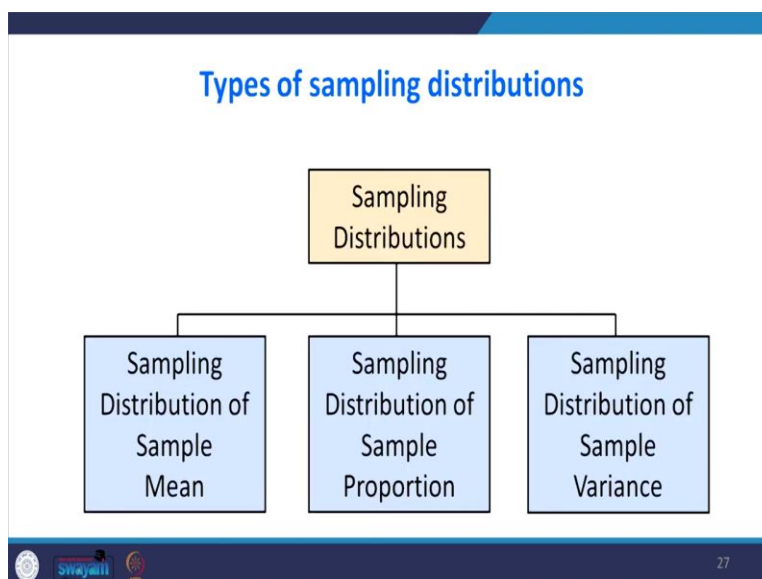
Sampling Distributions

- A **sampling distribution** is a distribution of all of the possible values of a statistic for a given size sample selected from a population

26

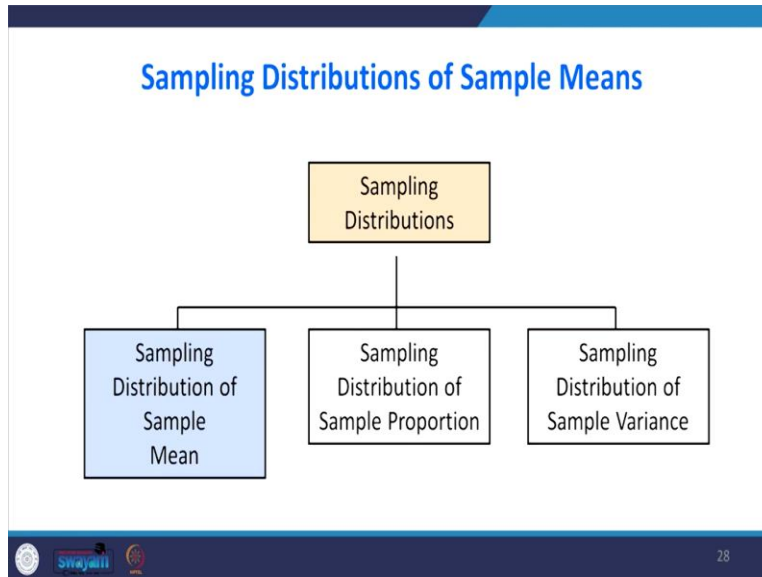
Now we are entering into the sampling distribution sampling distribution is a distribution of all of the possible values of your statistic for a given size sample selected from the population.

(Refer Slide Time: 18:14)



So, what will happen we can say type of sampling distributions we can do the sampling distribution for the sample mean. We can do the sampling distribution for the sample proportion. We can do the sampling distribution for sample variance.

(Refer Slide Time: 18:30)



First we will see the sampling distribution of sample mean.

(Refer Slide Time: 18:34)

- Assume there is a population ...
- Population size $N=4$
- Random variable, X , is age of individuals
- Values of X :
18, 20, 22, 24 (years)

Suppose assume that there is a population there are 4 people in a population that is age random variable is x is age of individuals. So, the value of x may be 18, 20, 22, 24 it is the population.

(Refer Slide Time: 18:54)

Developing a Sampling Distribution *(continued)*

Summary Measures for the Population Distribution:

$$\mu = \frac{\sum X_i}{N}$$

$$= \frac{18 + 20 + 22 + 24}{4} = 21$$

$$\sigma = \sqrt{\frac{\sum (X_i - \mu)^2}{N}} = 2.236$$



First you will find out the population mean population mean is Sigma of capital X i divided by N generally whenever you see a capital alphabet that is for the population. The smaller one is for the variance. So, 18, 20, 22, 24 divided by 4 is 21 similarly the population variance is 2.236. What is happening there are 4 element is there so the probability of getting each element that is choosing 18, 20 it is 1 by 4 so 0.25 + 0.25 and 0.25 it this follow uniform distribution.

Suppose if we choose only one sample when you plot it the chances for selecting each person from the population is 0.25.

(Refer Slide Time: 19:49)

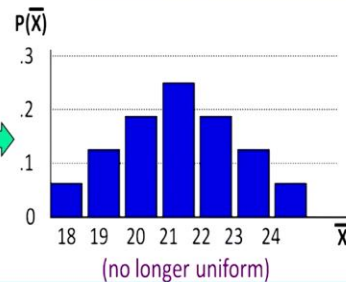
Developing a Sampling Distribution *(continued)*

- Sampling Distribution of All Sample Means

16 Sample Means

1st Obs	2nd Observation			
	18	20	22	24
18	18	19	20	21
20	19	20	21	22
22	20	21	22	23
24	21	22	23	24

Sample Means Distribution



Suppose if you consider all possible sample of size n, size n here means we are going to select 2 people with the replacement there is a possibility first observation may be 18 20 22 24 second observation may be 18 20 22 24 so possibility is 18 18, 18 20, 18 22, 18 24, 20 18, 20 20, 20 22 and so on. So, there are 16 possible samples here we are doing sampling with replacement that is why it is coming 20 20, 22 22, 24 24.

If we find the mean of this, so right side picture shows the mean of that 18 18 is 20, 18 20, 19 when you plot this me what is happening that mean of this sample is following normal distribution. |Previously when you take only one sample when you plot it we are getting uniform distribution. When you increase the sample size 1 to 2 what is happening you are getting here normal distribution it is no longer uniform.

(Refer Slide Time: 21:00)

Developing a Sampling Distribution
(continued)

- Summary Measures of this Sampling Distribution:

$$E(\bar{X}) = \frac{\sum \bar{X}_i}{N} = \frac{18+19+21+\dots+24}{16} = 21 = \mu$$

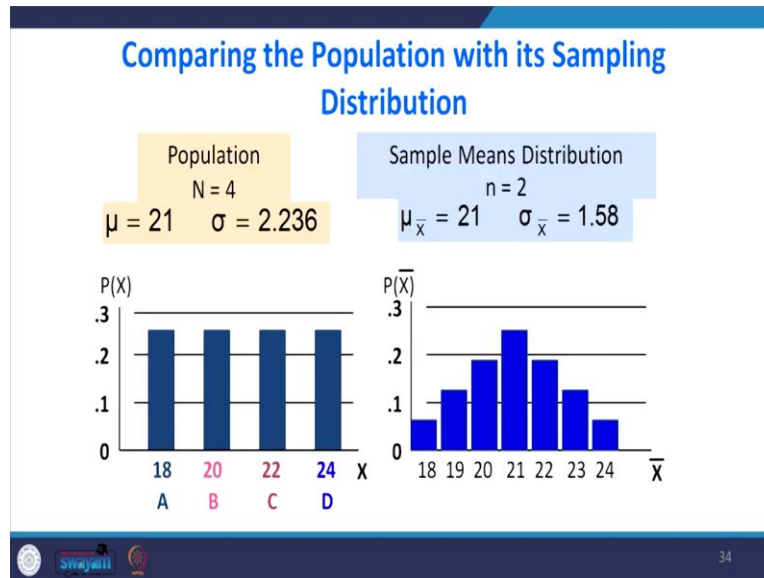
$$\sigma_{\bar{x}} = \sqrt{\frac{\sum (\bar{X}_i - \mu)^2}{N}} = \sqrt{\frac{(18-21)^2 + (19-21)^2 + \dots + (24-21)^2}{16}} = 1.58$$

33

Now summary measure of this sampling distribution where we selected to with replacement you see that and going back there are 16 elements 4 class 4, 4 4 times 4 16 element. So, the mean expected value of x bar is 18 19 21 up to 24 out of 16 mu equal to 21. Then the standard deviation of this sampling distribution is Sigma of x - mu whole square by n so the formula for standard deviation is first to find the variance mu is 21, so 18 - 21 whole square + 19 - 21 whole square up to 24 - 21 whole square it is 1.58.

Please look at and going back look at the population mean. The population mean is 21 and population standard deviation is to 0.236 when we select 2 with replacement mean of the sampling distribution is 21 but the standard deviation of the sampling distribution is 1.58 when you go for selecting 2 samples with replacement.

(Refer Slide Time: 22:16)

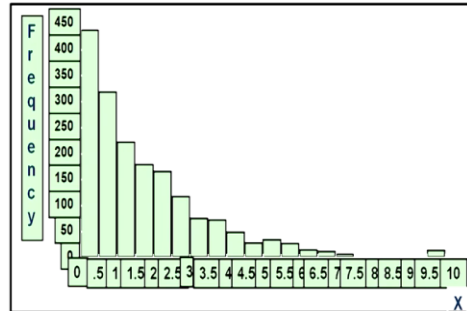


Next what we have to do we are going to select the 4 at a time we are going to construct the same table which have constructed previously. After constructing when you find the mean it will be 21 so we have found these summary measures for the sampling distribution where the mean of the sampling distribution is 21 and the standard deviation of sampling distribution is 1.58, so when we compare population data versus sample.

For population there are 4 element in the population in the sample there are 2 element. The mean of the population is 21 the mean of the sampling distribution is also 21 but the standard deviation of the population is 2.236 but the standard deviation of sample distribution is 1.58.

(Refer Slide Time: 23:08)

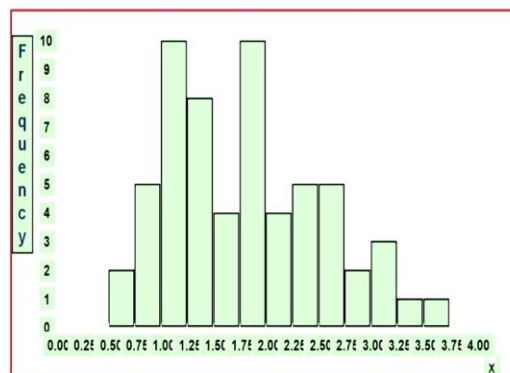
1,800 Randomly Selected Values from an Exponential Distribution



You will go for another example that there is a population which follows an exponential distribution. Now from this exponential distribution we are going to select 2 at a time with replacement. When you select two at a time then if I find the mean then if you construct frequency distribution then if I plot that frequency distribution when n equal to 2 we are getting this kind of distribution you see that the parent distribution is exponential when the sample size is 2 if I plot the mean of the sample mean that is following this kind of similar to uniform distribution.

(Refer Slide Time: 23:50)

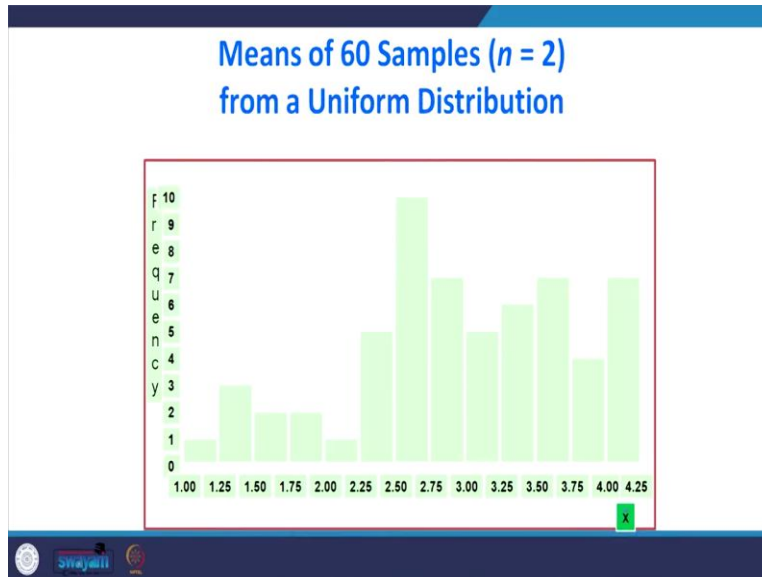
Means of 60 Samples ($n = 5$) from an Exponential Distribution



If I increase the sample size to 5 what is happening it is changed. so when n equal to 30 it is looking like here normal distribution. So, what is happening whatever may be the nature of the

population if you select any sample from the population then if you plot that the sample mean that will follow normal distribution. So, for example another example you take the population follow a uniform distribution.

(Refer Slide Time: 24:21)



You select 2 at a time and plot the sample mean that follow this kind of distribution increase sample size to 5 it is approaching normal distribution. When n equal to 30 it is looking like a normal distribution initially it was the uniform distribution when the sample size is increasing then it is following it is behaving like a normal distribution.

(Refer Slide Time: 24:43)

Expected Value of Sample Mean

- Let X_1, X_2, \dots, X_n represent a random sample from a population
- The **sample mean** value of these observations is defined as

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

So, expected value of sample mean let X_1, X_2, \dots, X_n represent a random sample from the population. The sample mean of these observation is defined as \bar{X} equal to $\frac{\sum X_i}{n}$ then standard error of the mean different sample of the sample size from the same population yield different sample means. A measure of variability in the mean from the sample to sample is given by standard error of the mean.

So standard error is $\frac{\sigma}{\sqrt{n}}$ note that the standard error of the mean decreases when the sample sizes increases.

(Refer Slide Time: 25:31)

If sample values are not independent
(continued)

- If the sample size n is not a small fraction of the population size N , then individual sample members are not distributed independently of one another
- Thus, observations are not selected independently
- A correction is made to account for this:

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{n} \frac{N-n}{N-1}$$

or

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

45

See if the sample values are not independent what will happen if the sample size is n and not a small fraction of the population size capital N then the individual sample members are not distributed independently of one another thus observations are not selected independently. So, a correction is made to account for this. So, σ^2 divided by n that was the variance of the sampling distribution that has to be multiply by $\frac{N-n}{N-1}$. You take square root of it σ by root or root of $\frac{N-n}{N-1}$.

(Refer Slide Time: 26:09)

If the Population is Normal

- If a population is normal with mean μ and standard deviation σ , the sampling distribution of \bar{X} is also normally distributed with

$\mu_{\bar{X}} = \mu$

and

$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$
- If the sample size n is not large relative to the population size N , then

$\mu_{\bar{X}} = \mu$

and

$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$

46

Ok if the population is normal with the mean μ and standard deviation σ the sampling distribution of \bar{X} is also normally distributed with the $\mu_{\bar{X}}$ equal to μ $\sigma_{\bar{X}}$ equal to σ by root n . When the sample size is not a large relative to the population then $\mu_{\bar{X}}$ equal to μ $\sigma_{\bar{X}}$ equal to σ by root n multiplied by correction factors.

(Refer Slide Time: 26:36)

Z-value for Sampling Distribution of the Mean

- Z-value for the sampling distribution of :

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

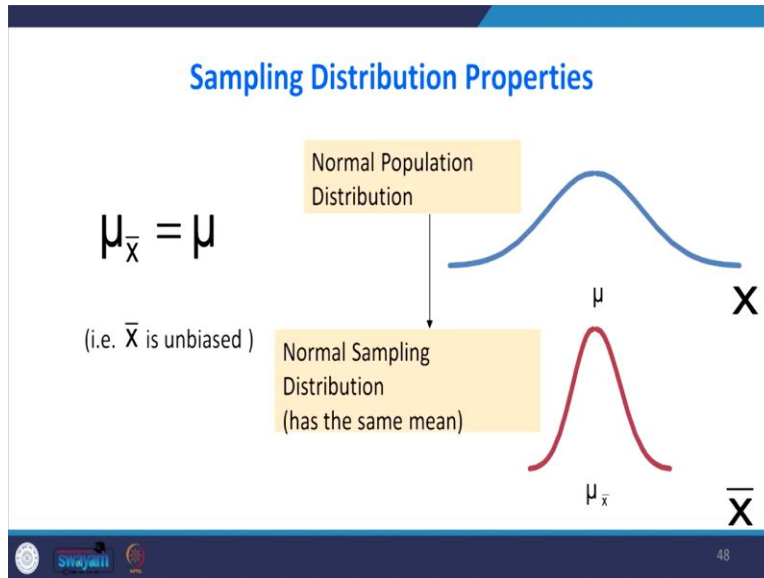
where:

- \bar{X} = sample mean
- μ = population mean
- $\sigma_{\bar{X}}$ = standard error of the mean

47

So, the Z value for the sampling distribution of the mean is ZT equal to $\bar{X} - \mu$ divided by $\sigma_{\bar{X}}$.

(Refer Slide Time: 21:00)



We look at the sampling distribution properties see the **the** top one it is a normal population distribution but the normal sampling distribution as the same mean then sampling distribution properties. For sampling with replacement when n increases sample size increases the standard deviation of sampling distribution decreases. So, what is happening look at the red color there is a large sample size that is the standard smaller standard deviation. Look at the blue one smaller sample size larger standard deviation.

(Refer Slide Time: 27:20)

If the Population is not Normal- Central Limit Theorem

We can apply the Central Limit Theorem:

- Even if the population is not normal,
- sample means from the population will be approximately normal as long as the sample size is large enough.

Properties of the sampling distribution:

$\mu_{\bar{X}} = \mu$

And

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

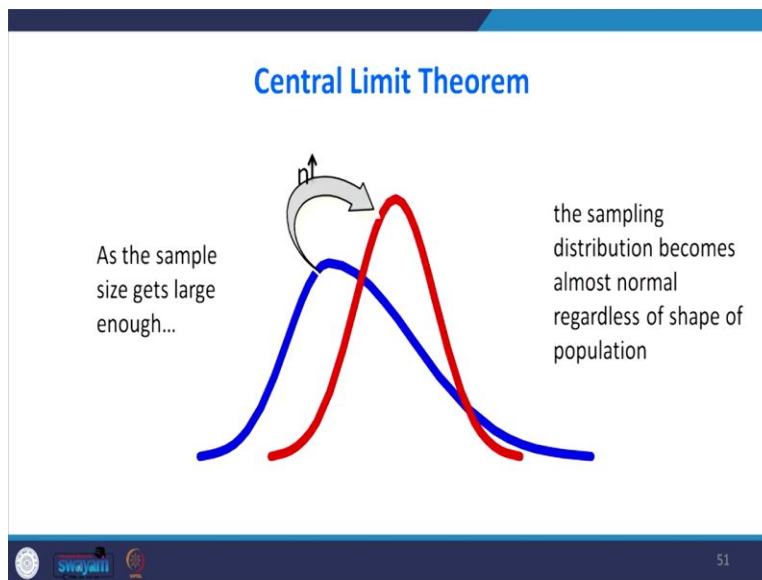
50

The population is not normal we can apply the central limit theorem even if the population is not normal. Sample means from the population will be approximately normal as long as the sample size is large enough. The properties of sampling distribution is $\mu_{\bar{X}} = \mu$ $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$

bar equal to σ by \sqrt{n} . This theorem is very important theorem that is the central limit theorem, why it is important through this theorem the concept of sample and population is connected. What is the result the mean of the sampling distribution is population mean.

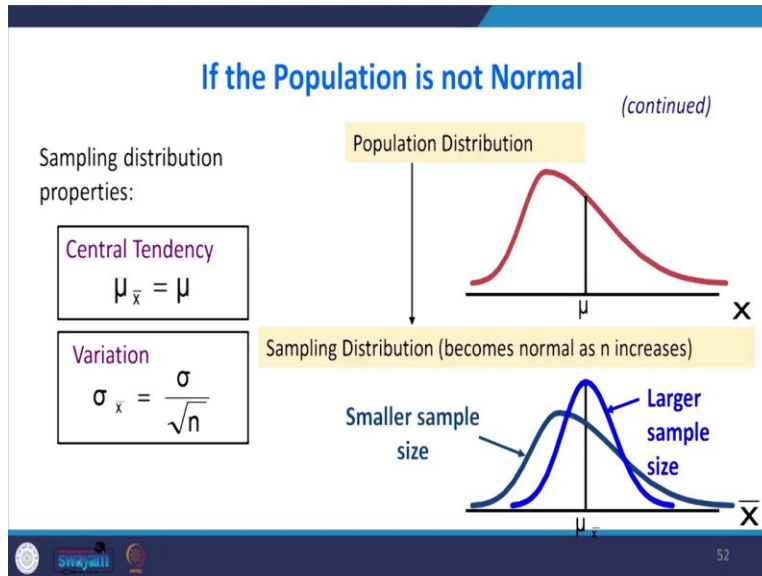
The standard deviation of sampling distribution is σ by \sqrt{n} where the σ represents the population standard deviation n represents the sample size. It is very powerful it is the very fundamental theorem for inferential statistics.

(Refer Slide Time: 28:19)



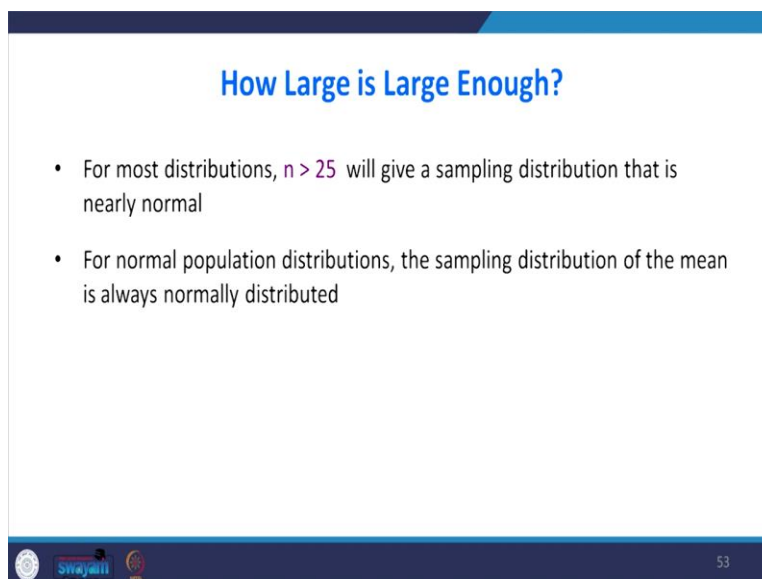
What is happening as the sample size get large enough the sampling distribution becomes almost normal regardless of the shape of the population. So, what is the meaning is suppose there is a population you take some sample if you plot the sample mean that will follow normal distribution provided n is large enough. So, when you keep on increase n then the sampling distribution will be exactly like your normal distribution.

(Refer Slide Time: 28:52)



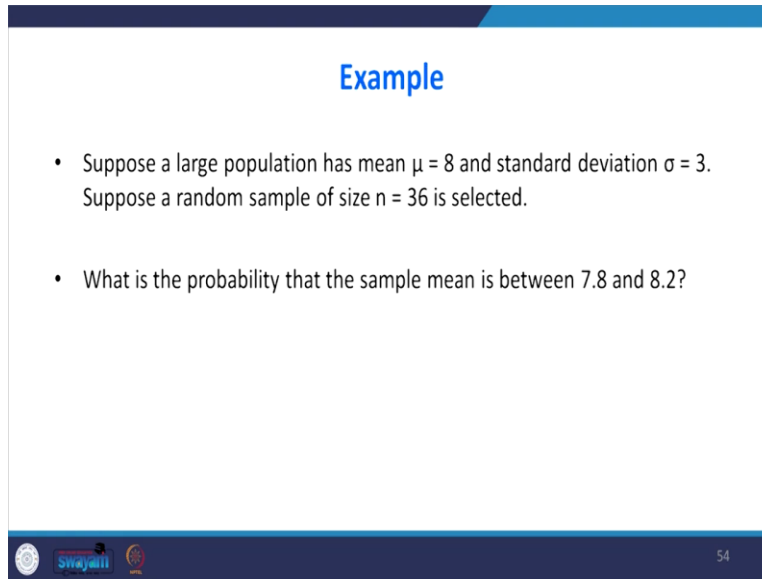
Even this is applicable even the population is not normal the parent population may be may follow any distribution but the sampling distribution will always will follow a normal distribution so the $\mu_{\bar{x}} = \mu$ the standard deviation is $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$. You look at this case also the population is not following a normal distribution but the sampling distribution will follow normal distribution.

(Refer Slide Time: 29:21)



So how large is large enough for most distribution when n is greater than 25 will assembling distribution that is nearly normal. For normal population distributions the sampling distribution of the mean is always normal normally distributed very important result. What is the meaning the sampling distribution of the mean is always normally distributed.

(Refer Slide Time: 29:46)



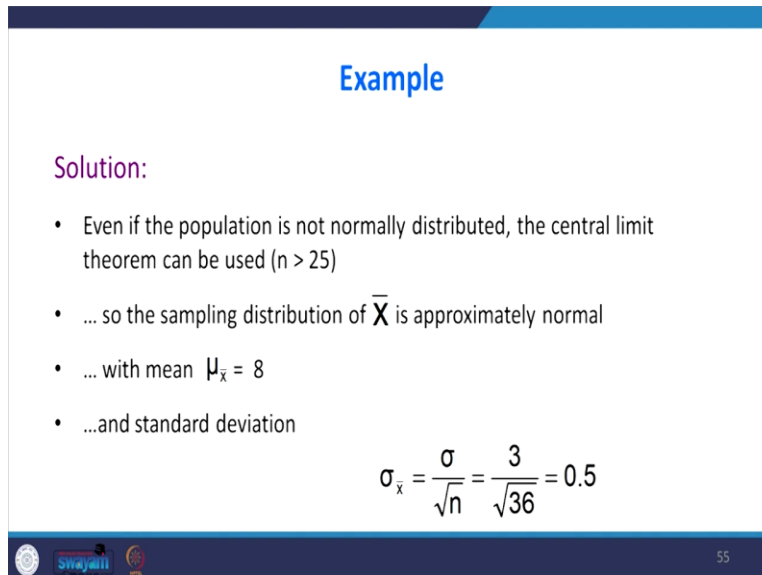
Example

- Suppose a large population has mean $\mu = 8$ and standard deviation $\sigma = 3$. Suppose a random sample of size $n = 36$ is selected.
- What is the probability that the sample mean is between 7.8 and 8.2?

54

Suppose we will see an example a large population has mean equal to 8 and standard deviation is 3 suppose a random sample of n 36 is selected what is the probability that the sample mean is between 7.8 and 8.2, we will see an example.

(Refer Slide Time: 30:06)



Example

Solution:

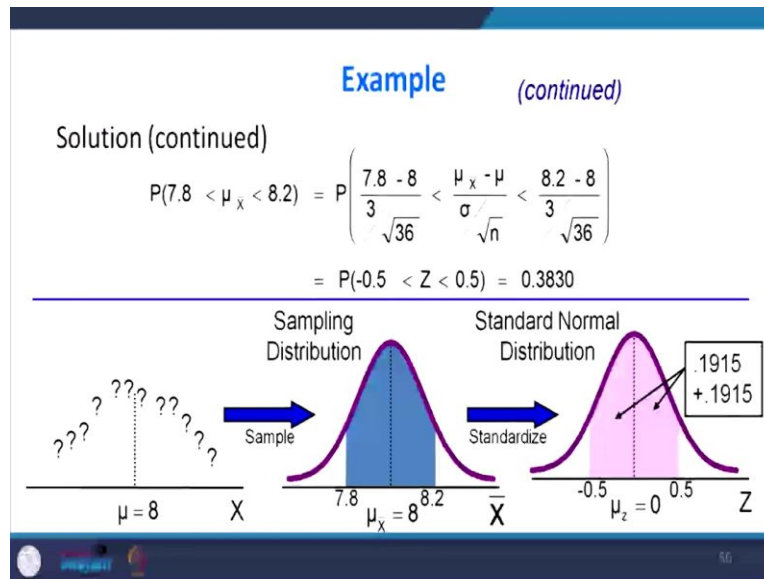
- Even if the population is not normally distributed, the central limit theorem can be used ($n > 25$)
- ... so the sampling distribution of \bar{X} is approximately normal
- ... with mean $\mu_{\bar{x}} = 8$
- ...and standard deviation

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{3}{\sqrt{36}} = 0.5$$

55

Even if the population is not normally distributed the central limit theorem can be used when n is greater than 25. So, the sampling distribution of x-bar is approximately normal that is the result which have seen Mu X bar equal to 8 and this standard because the MU X bar is mean of the sampling distribution is 8 the standard deviation of the sampling distribution is Sigma by root n Sigma is 3 n is 36 so 0.5.

(Refer Slide Time: 30:33)



So what will happen we were asked p of 7.8 less than MU X bar less than 8.2 so this 7.8 has to be converted to Z scale the conversion factor the conversion formula from converting to Z it is $X - \mu$ by σ by root n the X is given 7.8 mu is 8.2 Sigma is 3 sample sizes 36 that will be the middle one that is μX power - μ into d σ by root n that is nothing but your Z value less than equal to the upper limit.

So X is 8.2 - μ 2 by 3 by root of the 36 so, when you simplify this P of - 0.5 less then Z less than 0.5 that will give you the probability of 0.3830 so what is happening the extreme left shows the picture of your population there is a question mark that means the population may follow any distribution. if you select some sample when you find the sample mean then you draw the sampling distribution that will follow normal distribution.

So what is the area of the sampling distribution between 7.8 and 8.2 that was asked otherwise what is the probability of that the mean of the sampling distribution is between 7.82 to 8.2 so that 7.8 has to be converted into Z scale so that we can refer the table that conversion is done with the help of formulas Z equal to $X - \mu$ 2 by σ by root n after converting the 7.8 corresponding Z values - 0.5 8.2 corresponding Z values 0.5.

We can look at the statistical table or we can use Python to find the area between - 0.5 to 0.5 that will give the area of $.1915 + 0.1915$ with that we will close this one. So, I m concluding this lecture so what we have seen in this lecture is different sampling techniques. We have seen the importance of the sampling then we have seen the probability sampling non-probability sampling. Next we have seen an explanation for our central limit theorem.

What is the central limit theorem the nature of the population may be anything if you take some sample from the population if you plot that sample mean that will always follow a normal distribution that is the central limit theorem because the central limit theorem is very important theorem we have seen one problem also by using central limit theorem. We will continue in the next class.