Essentials of Data Science with R Software - 2
Sampling Theory and Linear Regression Analysis
Prof. Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology, Kanpur

Sampling Theory with R Software
Lecture - 28
Stratified Random Sampling
Drawing the Sample and Sampling Procedure

Hello, friends. Welcome to the course Essentials of Data Science with R Software – 2, where we are trying to learn the topics of Sampling Theory and Linear Regression Analysis. In this module on the Sampling Theory with R software we are going to begin with a new chapter on Stratified Random Sampling.

So, first let me give you the background what is this and where it is used. When the objective is to draw a representative sample from a population then, as I said whenever we use the word sample that goes without saying that it is going to be a representative sample. So, when you are trying to use the simple random sampling technique, you had assumed that variance of every observation is the same as $\sigma\iota\gamma\mu\alpha$; that means, the entire population is homogeneous.

So, now, the advantage is that if the amount of variability across the population is uniform, then either you try to take a sample from here or from here or from here or from here is anywhere. Or in simple words, if you try to draw any sample that will represent a population and the variability is going to be truly represented in the sample.

But, suppose the population is varying a lot with the variability with respect to the variability of the characteristic under study, then in case if you try to use the simple random sampling then it is possible that some part of the population is under represented or over represented and consequently your sample will not remain representative and once and once you try to use a statistical tool, then possibly they may be wrong.

So, now the situation is that population is varying a lot with respect to the characteristic under study or in simple word population is heterogeneous. Some part of the population has got some variability, different part of the population has got entirely different variability and the difference in the variability of two parts is also different.

For example, let me take a simple example. Suppose if you want to take a sample on say consumption pattern of the food from a country like India. India has got different states. Different states have got different type of consumption pattern on food. For example, if you go to a state like Punjab, there people are usually taking wheat more, but on the other hand if you go to West Bengal or if you go to southern part of India like Tamil Nadu, Kerala, etc. there people are using rice more. So, Punjab is using wheat more, these states in the south and in the east they are using rice more.

Whereas, there are some states like Uttar Pradesh, Bihar, Madhya Pradesh possibly they are using both wheat and rice. The type of food which is popular in different states is also different. If you go to southern part of India they are taking more like idli, sambar, dosa etc. But, if you come to a state like Jammu and Kashmir people are using more like rajma rice more and some of their local vegetables.

So, now, if you have to draw a sample from this country and if you choose only the simple random sample, there is a possibility that suppose most of the sample observations may come only from some part of the population, some part of the country. It is possible that suppose I draw a sample of 100 observations and suppose 90 observations are from rice dominated states.

So, now if you try to draw the inference from them it will appear as if that 90 percent of the people in the country like India are eating rice, that may not be correct and if you try to takes a sample that may misguide the Government of India.

And there can be a wrong policy formulation for the entire country. That you have to tell the government that 90 percent people are taking wheat because the government will try to arrange say this 90 percent wheat more or 90 percent rice more and once the things start coming then the situation will be very different.

People wanted suppose more wheat, but they are getting more rice or vice versa. So, in such a situation one option is that I try to divide the entire country into different segments. Those segments can be defined in such a way that every segment has got a similar type of food pattern. For example, I can create a groups of say this UP, Bihar, MP then I can create another group of say Kerala, Tamil Nadu, Karnataka etc. And, similarly

I can create another group of where people are eating more wheat like as Punjab, Haryana etc.

Once I create those groups then those groups become within homogeneous; that means, within group variability will be very small, but the variability among the groups will be larger and then I try to take the sample from each and every group. So, now, when once I have a sample in my hand the sample is coming from group 1, group 2 and group 3 also. So, at least in this sample there is a representation of all possible groups. So, then my sample will become more representative. This can be achieved by stratified random sampling.

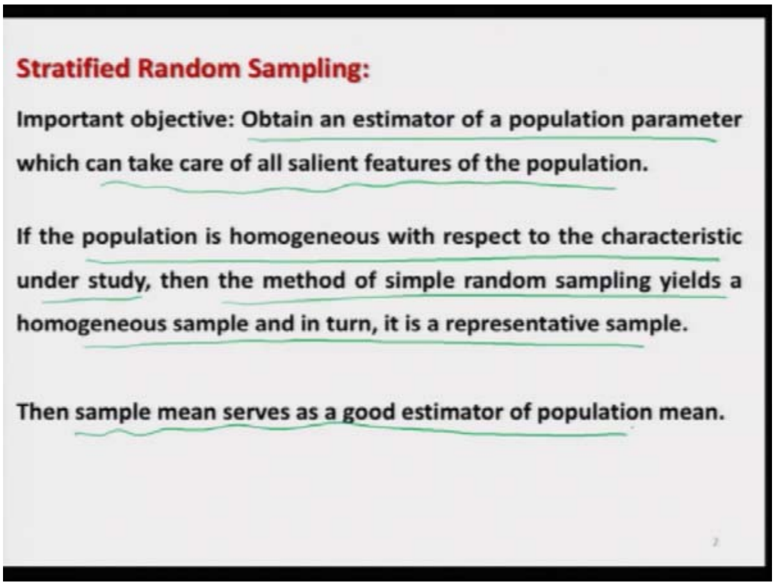
So, the first question which I have answered that if you ask me under what type of condition you have to use simple random sampling and under what type of condition one has to use a stratified random sampling I will say if you feel that the variability across the population is almost the same then use simple random sampling.

But, if you feel that the variability across the population is changing a lot different segment different groups may have different variability, then you use stratified random sampling. And, similarly there are many other types of sampling, they also have a reason that under what type of condition those sampling scheme can be used – the condition under which the cluster sampling can be used.

The condition under which systematic sampling can be used etc. they are different from each other. So, the moral of the story is that when you want to draw a representative sample please do not close your eyes and simply try to draw the sample by simple random sampling. You can think of all other sampling scheme also.

Try to match whatever are the real life condition or real experimental condition are matching with the conditions of the sampling scheme. And then you try to choose the sampling scheme and then you follow the procedure of drawing the sample under that scheme, right. So, now let me begin this lecture with our slides and let us see what is there in the slides, ok.

(Refer Slide Time: 09:08)



Stratified Random Sampling:

Important objective: Obtain an estimator of a population parameter which can take care of all salient features of the population.

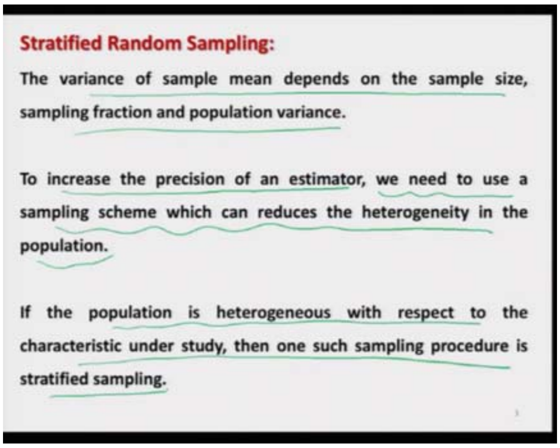
If the population is homogeneous with respect to the characteristic under study, then the method of simple random sampling yields a homogeneous sample and in turn, it is a representative sample.

Then sample mean serves as a good estimator of population mean.

So, whenever we are trying to consider a sample survey or we are; or we want to draw a representative sample, the basic objective is that we want to obtain an estimator of the population parameter like a population mean, population variance and the expectation is that this estimator should take care of all salient features of the population.

So, in case if the population is homogeneous with respect to the characteristic under study, then the method of simple random sampling will yield a homogeneous sample and in turn, it will be a representative sample. And, all the statistical inferences drawn from there they will give us a good outcome and in those situation we have seen that the sample mean serves as a good estimator of the population mean.

(Refer Slide Time: 10:02)



Stratified Random Sampling:

The variance of sample mean depends on the sample size, sampling fraction and population variance.

To increase the precision of an estimator, we need to use a sampling scheme which can reduces the heterogeneity in the population.

If the population is heterogeneous with respect to the characteristic under study, then one such sampling procedure is stratified sampling.

And, you have seen that the variance of the sample mean depends on the sample size, sampling fraction and the population variance. So, another way of looking is that whenever we draw the sample, we have to estimate a population parameter that population parameter is going to be estimated by some statistic that is your estimator that is a function of sample values.

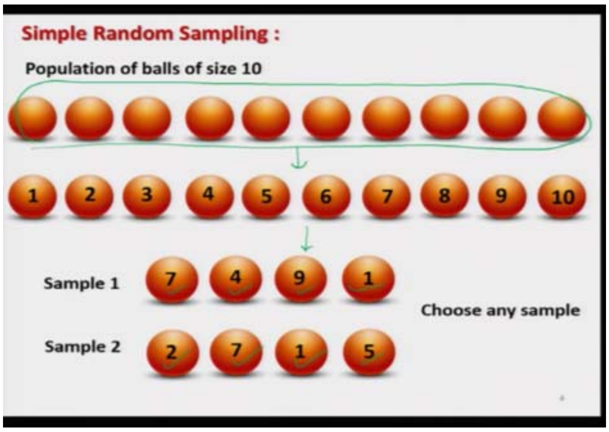
And, you have seen that an estimator also has its own variance. You have seen you have used $\bar{y}$ sample mean. So, you and then you have found the variance of $\bar{y}$ as well as you have found the estimate of the variance of $\bar{y}$ and you would always like from the statistics point of view that the variance of the statistics or the variance of the $\bar{y}$ should be as minimum as possible.

So, you also can recall that in the beginning of the lecture I had explained you that there are three stages in which the variability of an estimator can be controlled before the selection of sample, during the selection of sample and after the selection of sample. So, now once we have say decided that ok, this is an estimator for estimating a particular parameter so, then we try to find out its variance and we also try to see that how this variance can be made as small as possible.

For example, in the case of simple random sampling the variance was depending on three factors small n sample size, capital N population size and capital S^2 which is the population variability, right. So, here also in stratified sampling you will see that we will try to propose an estimator of the population mean and that estimator will be proposed in such a way such that its variance is as minimum as possible that cannot be 0 because if the variance becomes 0 that means, there is no use of statistics and everything become then constant. So, this I will try to show you in the further lectures.

So, in order to increase the efficiency or the precision of an estimator we would like to use a sampling scheme which can reduce the heterogeneity in the population and if the population is heterogeneous and if you want to reduce the variability in your estimator, then one such sampling procedure is stratified sampling.

(Refer Slide Time: 12:59)



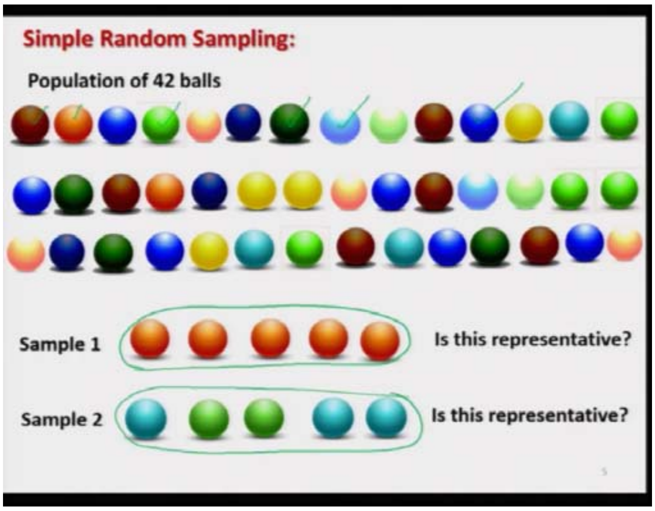
Simple Random Sampling :

Population of balls of size 10

So, let me try to take an example over here and try to show you that how the concept of stratified sampling comes into picture. Suppose, here I try to take a simple random sample that I this is the same example which I took during the simple random sampling that I have a population of 10 balls which are all the same.

So, there is no variability among the balls. Yes, the their size may vary, but the colours are the same, right. It may be possible that some are light red and some are dark red. So, I try to create here the sampling frame by giving them a number 1 to 10 and from there I try to draw a sample. Suppose the first sample comes out to be 7, 4, 9, 1 and the second sample comes out to be 2, 7, 1, 4, 2, 7, 1, 5 number of balls you can choose any of the sample and they will give you good outcome, right.

(Refer Slide Time: 13:59)



Simple Random Sampling:

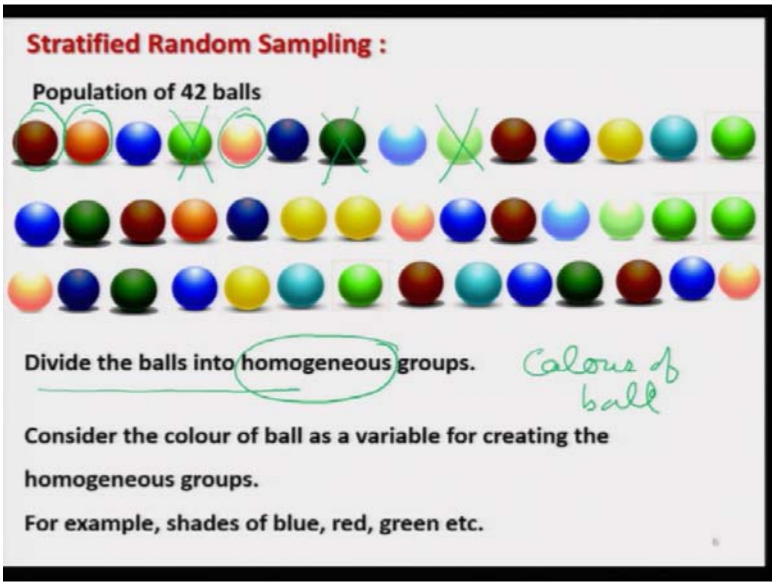
Population of 42 balls

And, you can choose actually any ball, but now suppose I try to take another example where I have got here 42 balls and you can see here that these balls are of different colors say some are light red, some are dark red, some are light green, some are dark green and some are light blue some are dark blue and so on. You can see there are different colors different shades of a color.

Now, if you try to consider your variable as the color of the ball, then you can see now in this case compared to the earlier case there is some variability. Colors are not the same, but some colors are close to some color and whereas, some colors are entirely different than other than the other colors.

Now, suppose you decide that you would like to use the simple random sampling and suppose you draw the sample and suppose you get here these 5 balls all are here red do you think that is it a representative sample or suppose you get here these 5 balls of blue and green colors only. Do you think that is it a representative sample? The answer is no. In both the cases the sample is not representative, right. But, then the question comes how the representative sample will look like.

(Refer Slide Time: 15:36)



Stratified Random Sampling :

Population of 42 balls

Divide the balls into homogeneous groups.

Colour of ball

Consider the colour of ball as a variable for creating the homogeneous groups.

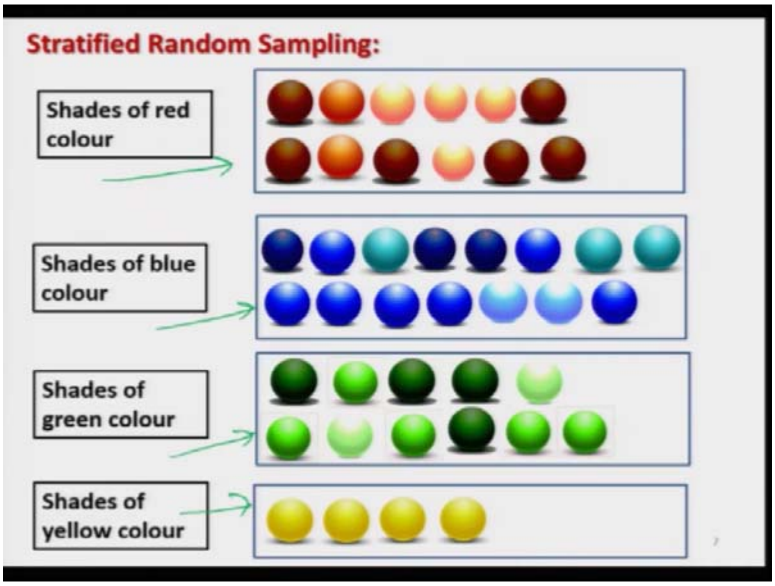
For example, shades of blue, red, green etc.

So, what I try to do in such a situation instead of using the simple random sampling, I try to use the stratified random sampling and I try to create groups of balls and those groups have to be; have to be created in such a way such that they are within homogeneous.

So, now my variable here is color of the ball. So, now, what I can do? I can look into the colors and then I will try to create different group. So, I try to divide the balls into different homogeneous groups, right. For example, I have chosen here say different shades of blue, red and green.

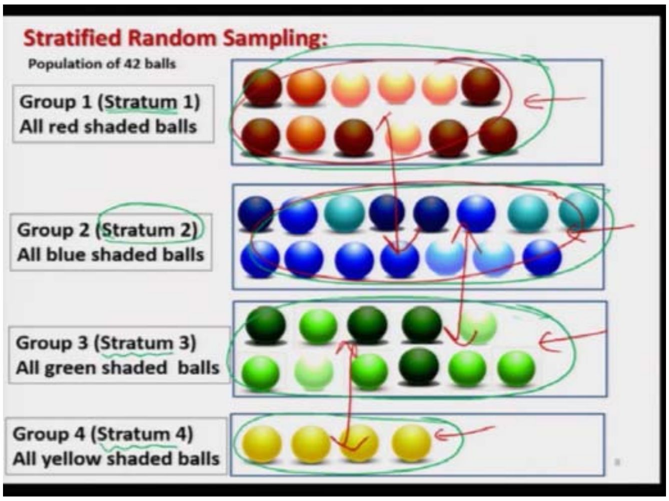
For example, if you see here this ball, this ball, this is red and this is another here red. Well, they are light red dark red or darker red so, but they are the different shades of color red. Similarly, if you try to see here green you can see here this is one green color, this is another green color, this is also another green color. So, there are three shades of green color and similarly, there are other colors. So, what I try to do, I try to club them together means the balls of similar color are grouped together.

(Refer Slide Time: 17:15)


Stratified Random Sampling:

So, you can see here this is how I am proceeding that here I have taken the balls of different shades of red color, in the second group I have taken different balls of shades of blue color, similarly in the third case green color and finally, in the fourth case means I have taken all the ball of yellow color. Well, in yellow color there are no shades there is only one color that may possible. This is the most perfect group which has got the least variability, right.

(Refer Slide Time: 17:51)



Stratified Random Sampling:

Population of 42 balls

Group 1 (Stratum 1)
All red shaded balls

Group 2 (Stratum 2)
All blue shaded balls

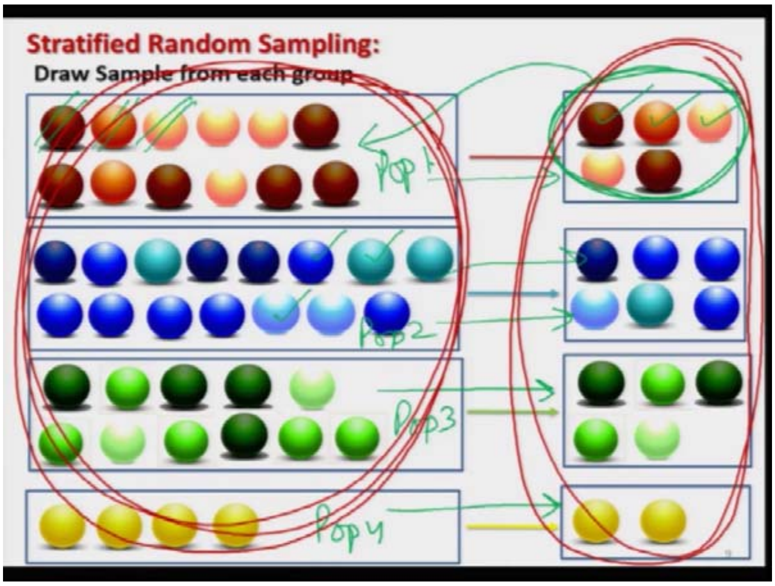
Group 3 (Stratum 3)
All green shaded balls

Group 4 (Stratum 4)
All yellow shaded balls

So, these groups they are actually called as stratum in the language of stratified random sampling. So, this is now this group is my stratum number 1 in which we have red color balls and those red color balls are varying with respect to their color within a group also. Similarly, the second group which is here this is called as stratum number 2 which consists of balls which have got approximately the bluish shades.

Similarly, the third group of green color ball, this is called as stratum number 3, consist consisting of balls of different shades of green color. And, similarly the group of yellow color ball this is my stratum number 4 which contains all the balls of yellow color.

(Refer Slide Time: 18:54)



Stratified Random Sampling:

~~Draw Sample from each group~~

Now, what I can do before that if you try to see here that if you try to compare the variability within the group there is some variability, but this variability among between the group is also varying.

But, if you try to compare the variability between the groups or among the groups then you can see that the variability is changing a lot the variability between the balls of red color and balls of blue color this is entirely different. These are two different colors, but within the groups the color is the same only the shade is changing.

So, you can see here the groups have been constructed in such a way such that within a group all the balls are of the same color, but they are varying only with respect to the shade. But, when they are coming from one group to another group, the color of the ball is also changing. So, red color green color blue color and yellow color they are entirely different. So, the variability among the group is more, right.

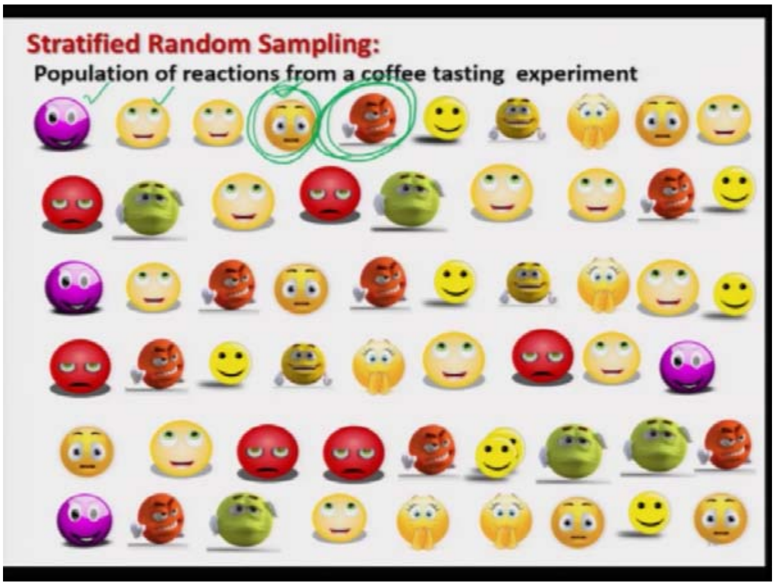
Now, what I can do? That if I want to draw here a representative sample one option is this from the stratum one of red color ball I try to draw a sample, from similarly from the blue color ball stratum I try to draw one sample. Then from the green color ball stratum I try to draw another sample and from yellow color stratum I try to draw another sample of yellow color ball.

So, you can see here if you try to see here just try to mark my pen. This sample is coming from this population. So, you can see here this is a representative sample, right. So, you can see here there are three shades light shade of red, darkest shade of red and another shade which is lying between the two and all these three shades are present in the sample.

And, similar is happening in the blue color also light blue, medium blue and dark blue. So, they are also present in the second sample and the same story is continued in the third and fourth sample. So, you can see here, here we are trying to draw the sample from every group. So, we are trying to consider this as a population 1, the second group has the population 2, third group has the population 3, fourth group has the population 4 and then from each of the population we are trying to draw the sample.

So, now my, the entire sample try to see my the movement of my pen in color red, this is my here complete sample, right and this here is my complete population. So, if you try to draw the directly from this population the sample may consist only of say blue balls or only of say green ball or something like that. So, that may not remain as a representative sample.

(Refer Slide Time: 22:17)



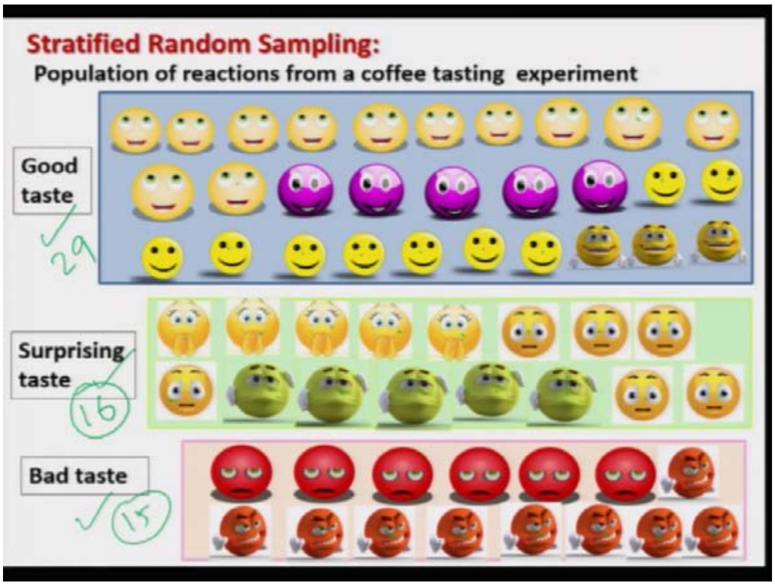
Stratified Random Sampling:

Population of reactions from a coffee tasting experiment

Similarly, I try to take here one more example that I also considered earlier. Suppose there was an experiment to know that taste of a coffee. Different people gave different types of expression or reactions after drinking the coffee. You can see here there are some faces like the first face, second face, which are saying yes, we are happy. There is a face like this one which is saying I am surprised and there is another say here group like this one which is saying that no I am not happy and so on.

So, different expressions if you try to see on the slides are indicating different type of reaction. So, now, the thing is this how you can draw the sample from here.

(Refer Slide Time: 23:14)



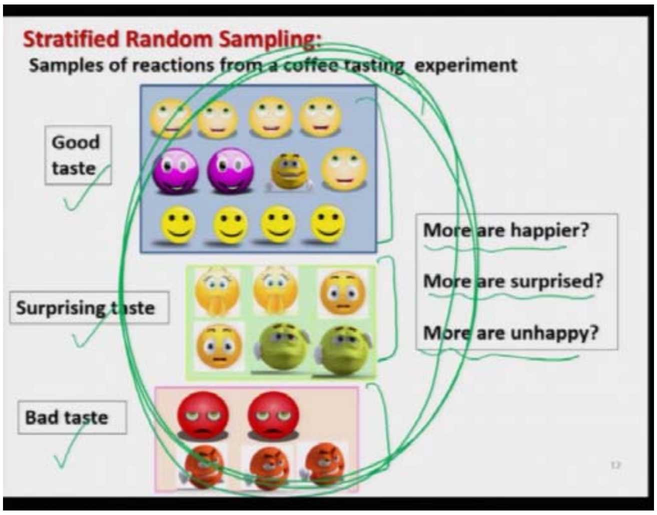
Stratified Random Sampling:

Population of reactions from a coffee tasting experiment

So, what I try to do? I try to divide the entire population with respect to different types of reactions. I have taken here three parts one is here good taste, another here is surprising taste and third group is here bad taste.

Well, that is my choice if you want you can choose any other criteria like the color of the smiley also and then you can group them together. So, that is essentially the choice of the experimenter who is considering who is conducting the experiment that what should be the random variable for consideration.

(Refer Slide Time: 23:52)



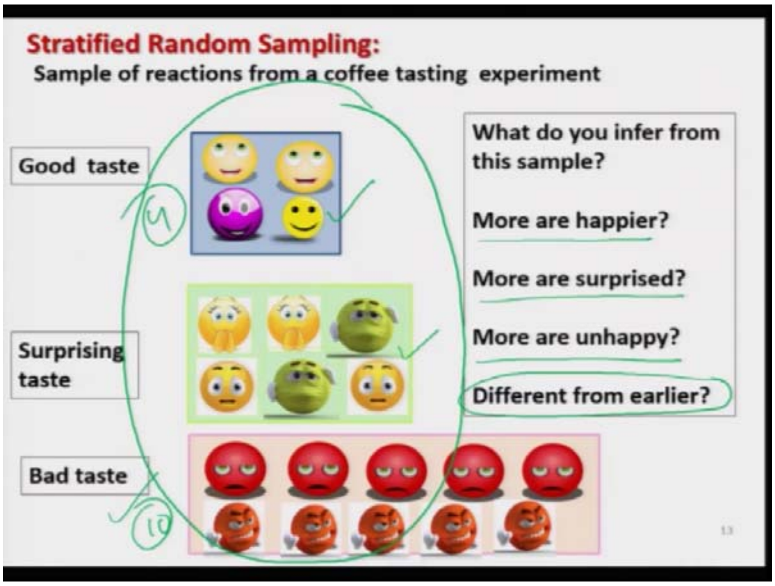
Stratified Random Sampling:

Samples of reactions from a coffee tasting experiment

So, now what I do? I try to take a sample from these three groups. So, I try to take it from a the population of good taste reaction surprising taste reaction and bad taste reaction. Now, if you try to see here well this is my sample. So, these are my three samples which will constitute one sample. Now, there is another thing means if you try to count the number of faces which are selected in the in these three groups they will try to give you different type of conclusions.

What type of conclusions you can have? For example, can you conclude that more people are happier or more people are surprised or more people are unhappy? This is not holding uniformly. You can see here in your sample which consists of this entire population sorry, entire group of this smiley's they are constituting your sample. And now, you have to devise something so that you can take a correct conclusion.

(Refer Slide Time: 25:08)



Stratified Random Sampling:

Sample of reactions from a coffee tasting experiment

Good taste

Surprising taste

Bad taste

What do you infer from this sample?

More are happier?

More are surprised?

More are unhappy?

Different from earlier?

13

For example, means I can take one more sample and after this sample I will ask you can you really take this sample to be a representative sample? I have means you can see here the number of people in the group which are saying good taste, number of people in the surprising taste and number of people in the bad taste right. You can count yourself.

Or now I try to take this thing. You can see here there are four persons in the sample indicating a good taste, six persons in the sample indicating a surprise surprising taste and there are 10 number of people who are taking the saying that the taste is bad. So, this is overly represented by the people who are saying the taste is bad. So, by looking at this sample can I conclude that the taste is really bad?

But, on the other hand if you try to see in this population, do you think that most of the peoples are not happy with the coffee taste? That is not correct actually. Means if you try to count here 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29, there are 29 people who are happy. There are 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16, there are only 16 people who are surprised over the taste and there are 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15, 15 people who are not happy.

So, the number of people who are not happy with the taste they are just 15 lowest and they are nearly the half of the people who are happy at the taste. But, if you try to look at this sample, this is not indicating the same thing. There are 10 people who are saying that the taste is bad and there are only 4 people who are saying that the taste is good.

So, this is going to be a bad sample. That is not really representing what is really happening in the population. So, you always have to take care of the quotients from the sample that whether more people are happier or more or more people are surprised or more people are unhappy or it is on you have to compare different sample and you have to see whether different samples are giving you the same information or different information.

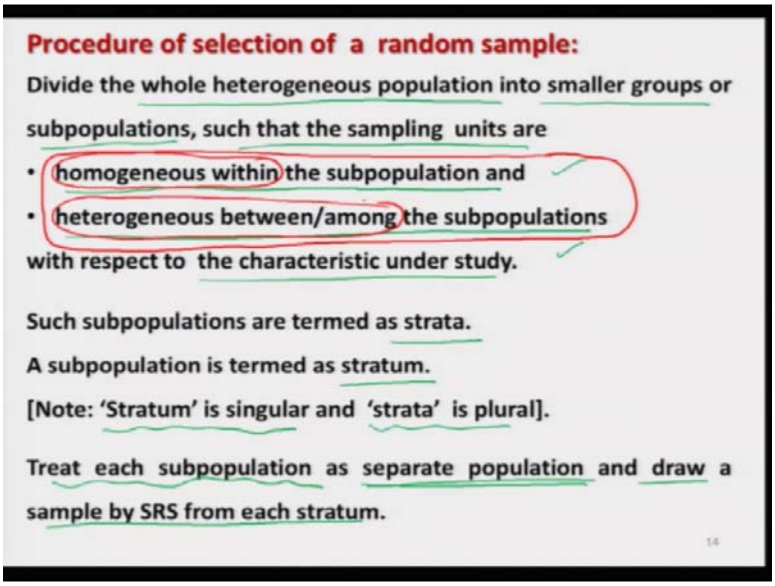
So, that will actually represent the concept of representative sample. The sample has to represent the same behaviour which is existing in the population. So, in this case at least, in my opinion this sample is not representative. So, now, I have here different types of questions. The first question is now you can see that the representative sample has got another dimension that it has to be a it has to take care of the variability of the characteristic under study first thing.

Second thing is this, what should be the sample size which you have to draw from every population such that the total sample looks representative? There can be different ways. For example, one can say I can draw equal number of observations from every group from every stratum. Somebody may say no, I can draw the number of observations proportionately.

Those group which are bigger from there I will try to draw bigger sample size and those group which are smaller I will try to draw a smaller number of sampling units and third thing can be also the variability in the units. But, when you are trying to draw the samples you cannot increase the sample size to be extremely large because every observation has some cost. And, in any survey the cost is not infinite, this is also fixed.

So, these are the different factors, these are the different questions, these are the different challenges which crop up when we are trying to consider the stratified random sampling for drawing a representative sample. So, that will be our objective when we try to study this stratified random sampling in the further lectures.

(Refer Slide Time: 30:20)



Procedure of selection of a random sample:

Divide the whole heterogeneous population into smaller groups or subpopulations, such that the sampling units are

- homogeneous within the subpopulation and
- heterogeneous between/among the subpopulations

with respect to the characteristic under study.

Such subpopulations are termed as strata.

A subpopulation is termed as stratum.

[Note: 'Stratum' is singular and 'strata' is plural].

Treat each subpopulation as separate population and draw a sample by SRS from each stratum.

14

So, now let us try to first understand what will be a procedure for the selection of the random sample from a stratified random sampling. The first step is this that you have to first divide the entire population into smaller groups and we are assuming that my population is heterogeneous with respect to the characteristic under study.

So, first step is to divide the whole heterogeneous population into smaller groups or say also called as sub populations which are called as a stratum I will introduce it later. And, these sub population or groups or stratum have to be constructed in such a way such that they are within homogeneous and they are between or among heterogeneous with respect to the characteristic under study.

And, this sub population or group whatever you have obtained here they are called as strata. But, you may recall that up to now I have used that word stratum. So, just to clarify these two words, stratum is a singular word and strata is a plural. So, when you have only one group that is called as a stratum, when you have more than one group that is called as strata.

Now, you try to treat each of this sub population or a strata as a separate population and draw a sample by simple random sampling from each stratum. Now, there are two things which I would like to address. If you follow my pen, here I am writing that the strata or the subgroups have to be created such that they are within homogeneous and they are between or remain heterogeneous with respect to the characteristic under study.

The question arises who told you and how do you came to this conclusion that these sub groups or the strata have to be constructed in such a way means if I ask you at this stage can we take the opposite? Can we make the sub groups in such a way so that within a subgroup they are as heterogeneous as possible and between subgroups they are as homogeneous as possible?

And, between these two options which will give you a good outcome or a good statistical inference at this today at this stage today it is not really possible for us to answer this question and I will try to answer this question in the next lecture.

And, you will see that in statistics every construction every rule has a reason. These rules are not coming from sky. They are not coming by do some by doing some magic. I want to teach you in this course that such conclusions are based on pure statistics and once again I will synthesize if you want to become a good data scientist you have to follow these rules.

The only thing is this the rules which I am explaining you here they are they were developed for a smaller sample, now you are working with a much bigger sample. So, that is the challenge what we have to face and what we have to solve that how to extend these results in such a big data set situations, right.

So, that is the first question which I will try to address that how we have taken the decision that if you want to construct a good strata, then it has to be within homogeneous and between heterogeneous. But, at this moment you can take my words that this is the correct approach, after that I will try to show you how it comes.

But, at the same time I would also like to address that if you do the opposite; that means, within group they are heterogeneous and between group they are homogeneous, then this is the setup of cluster sampling, right. But, the situation under which we are going to use the stratified sampling and the condition under which we are going to use the cluster sampling these are two different conditions.

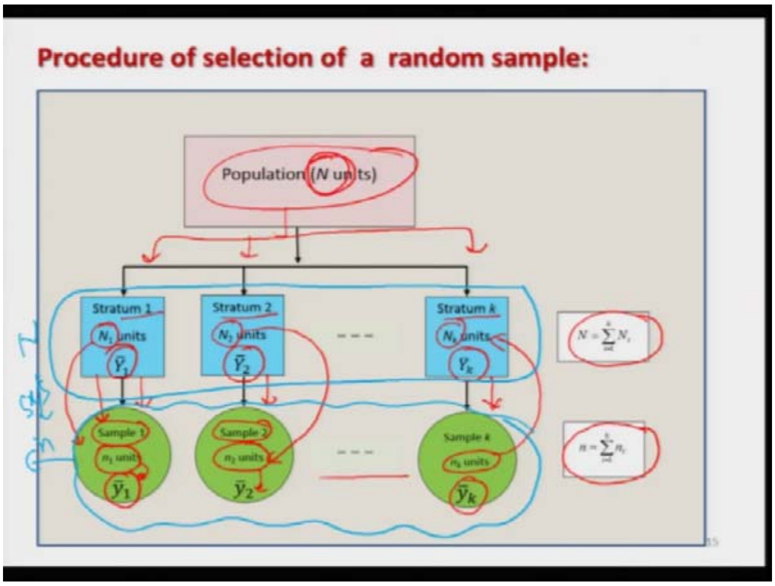
And, as an instructor of statistics, I am not trying to tell you here whether this sampling scheme is better than this or that. I am simply trying to tell you well these are sampling scheme this sampling scheme can be used under condition 1 number 1; this sampling scheme can be used in condition number 2.

Now, you understand it and whenever you go to the field you have to take a call that whether you want to use here the sampling scheme 1 or sampling scheme 2 based on the conditions which are present in the situation present during the experiment or the survey right. Second thing is, this I have told you that you try to create the strata and from strata you try to draw the sample by simple random sampling. Well, at this stage I am asking you to draw the sample by simple random sampling because you have done only one sampling scheme, but this thing can be generalized.

And, it is also possible that different types of concept from different sampling scheme can be clubbed together also that in the first step you try to use this sampling scheme try to draw a sample, then in the second step you can involve some other sampling scheme and then you can progress further.

The objective is the same that at the end of the day I want to have a representative sample. So, that is what I have I wanted to clarify on this screen that do not get confused between the two and these are the concepts.

(Refer Slide Time: 37:38)



Procedure of selection of a random sample:

Population (N units)

Stratum 1 — N_1 units — $\bar{Y}_1$

Stratum 2 — N_2 units — $\bar{Y}_2$

Stratum k — N_k units — $\bar{Y}_k$

Sample 1 — n_1 units — $\bar{y}_1$

Sample 2 — n_2 units — $\bar{y}_2$

Sample k — n_k units — $\bar{y}_k$

$N = \sum_{i=1}^{k} N_i$

$n = \sum_{i=1}^{k} n_i$

And, now I would try to show you here just graphically very quickly, so that you can understand it. So, I have taken here a population of size capital N units, I have divided it into different groups.

Now, I am calling them as a stratum; stratum number 1, stratum number 2 and say there are a small k number of stratum. And, once there are capital N number of units in the population, so, this stratum will have different number of units. Suppose, the stratum 1 has capital N_1 number of units; stratum 2 has capital N_2 number of units the stratum k has N_k number of units.

And, suppose, the sampling units which are present in the stratum number 1 they have got a population mean, but within the group within the stratum 1 which is denoted as $\bar{Y}_1$. Similarly, the arithmetic mean of the units in the stratum 2 is denoted by capital $\bar{Y}_2$ and similarly in the stratum k the population mean of the k-th strata is $\bar{Y}_k$.

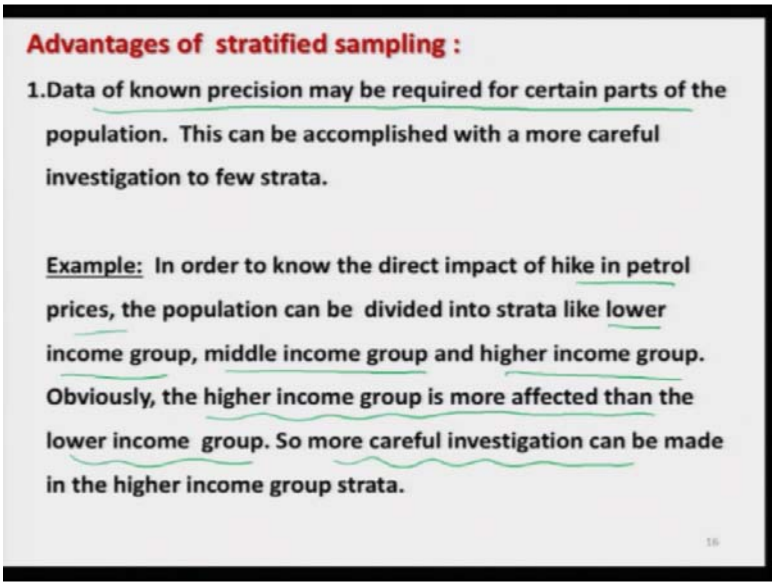
Now, from there we are trying to draw the samples. So, this is my sample number 1 which is drawn from the stratum 1. So, suppose I decide that from this N_1 number of units I try to draw here is small n_1 number of units. And whatever are the units are obtained I try to find out the sample mean which is denoted by a $\bar{y}_1$.

The same process I try to repeat for stratum 2, stratum k and so on. So, from the stratum 2, I try to draw from capital N_2 number of units, small n_2 number of units and this will create my sample number 2 and the mean of the units in the sample number 2 will give me $\overline{y}_2$ which is sample mean of second stratum.

And, the same thing can happen in the k-th stratum also that we have got the small n_k number of units collected from capital N_1 number of units and the sample mean of this unit is given by $\overline{y}_k$. So, you can now see here that this population size capital N can be expressed the sum of all the stratum size and the total sample size will be the sum of all the samples which are drawn from.

So, for a Y it appears that we have a population like this one here of size capital N and from there we have drawn here a sample of size small n like this one by S R S. For example, that can be S R S or this can be anything else. So, this is how we are going to collect the samples, right.

(Refer Slide Time: 40:30)



Advantages of stratified sampling :

1.Data of known precision may be required for certain parts of the population. This can be accomplished with a more careful investigation to few strata.

Example: In order to know the direct impact of hike in petrol prices, the population can be divided into strata like lower income group, middle income group and higher income group. Obviously, the higher income group is more affected than the lower income group. So more careful investigation can be made in the higher income group strata.

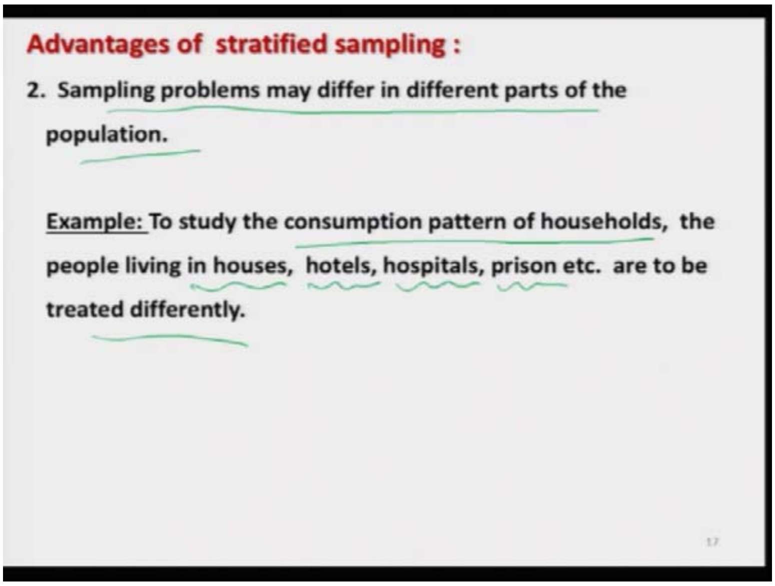
16

Now, after this I just give you a quick review of the advantages of a stratified sampling. So, whenever the data of some known precision is required from certain parts of the population, then this can be achieved by the by the stratified random sampling. For example, if there is a hike in the price of the petrol then definitely and if you want to study from the responses of different people, then the impact of the price of the petrol on different income groups people is different.

So, I would like to divide the entire population into say several groups. For example, lower income group, middle income group, higher income group and so on. So, now, there can be different conclusions what you can take. Somebody can say this obviously, the higher income group is more affected than the lower income group because they are using the petrol more.

But, on the other hand, somebody may also conclude that higher income group has got a higher income. So, they may not be affected more, but those people who are getting lower income they may be affected more. So, now I myself have given you two possible conclusion and now, how to verify that which one is correct. So, for that we have to go into more careful investigation and then we have to draw the sample of the data, and data will let us know what is really happening.

(Refer Slide Time: 42:04)



Advantages of stratified sampling :

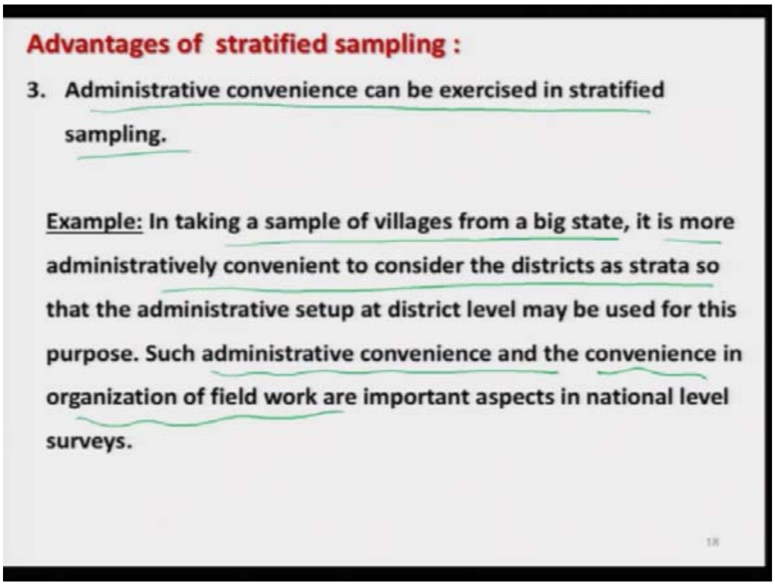
2. Sampling problems may differ in different parts of the population.

Example: To study the consumption pattern of households, the people living in houses, hotels, hospitals, prison etc. are to be treated differently.

17

The sampling problems in different parts of the population may differ. For example, if you want to study the consumption pattern of households in some city, then in some then in the city the people living in houses people, living in hotel, people living in hospitals and people living in prison etc. they have to be treated differently, right. Their consumption patterns are entirely different. So, if you try to take one sample possibly that may not be a representative sample.

(Refer Slide Time: 42:35)



Advantages of stratified sampling :

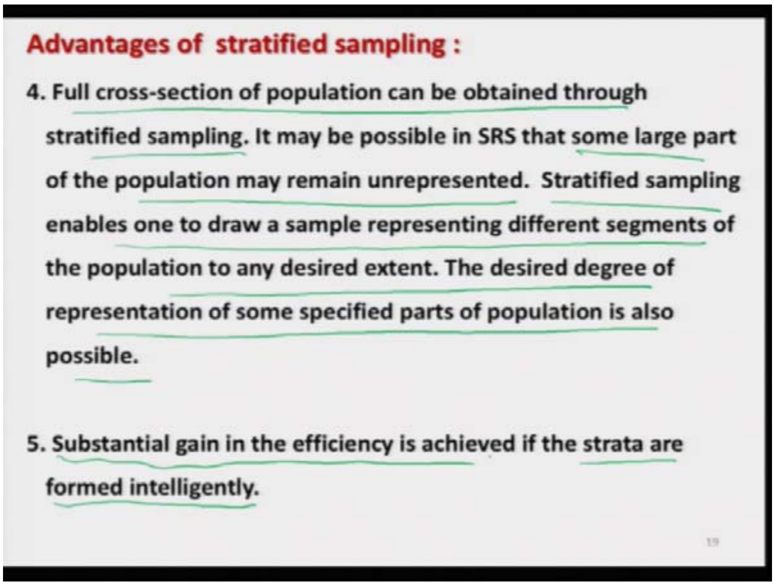
3. Administrative convenience can be exercised in stratified sampling.

Example: In taking a sample of villages from a big state, it is more administratively convenient to consider the districts as strata so that the administrative setup at district level may be used for this purpose. Such administrative convenience and the convenience in organization of field work are important aspects in national level surveys.

Thirdly, the administrative convenience is a great help in the stratified sampling. For example, taking a smaller sample from a smaller population is much easier than taking a bigger sample from a bigger population directly. So, for example, if you want to take a sample of villages from a bigger state, then it is administratively more convenient to consider the districts as a strata, and from the district one can draw the villages.

So, first type of administrative convenience and convenience in the organization of the field work, they play a very important role and help a lot when we are trying to do with this stratified sampling particularly, when we are trying to go for this national level survey.

(Refer Slide Time: 43:26)



Advantages of stratified sampling :

4. Full cross-section of population can be obtained through stratified sampling. It may be possible in SRS that some large part of the population may remain unrepresented. Stratified sampling enables one to draw a sample representing different segments of the population to any desired extent. The desired degree of representation of some specified parts of population is also possible.

5. Substantial gain in the efficiency is achieved if the strata are formed intelligently.

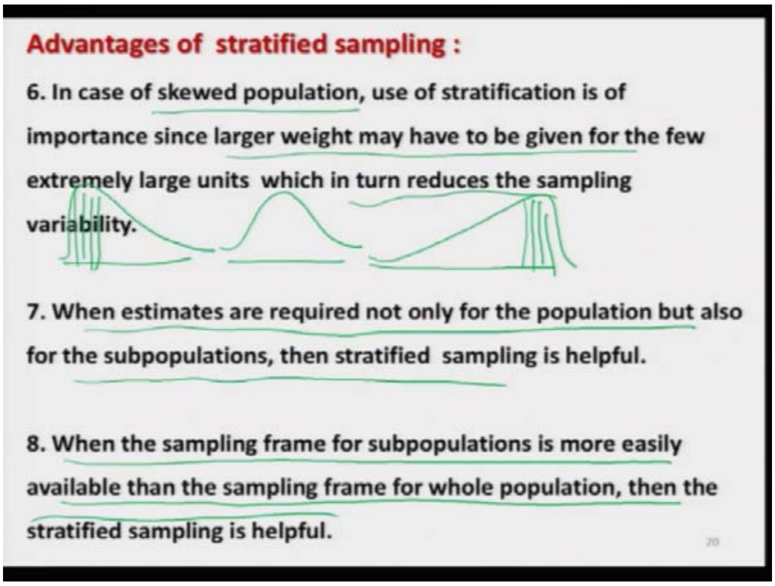
19

And, another advantage is that when we are going for stratified sampling, then the full cross section of the population can be obtained because if you are going for simple random sampling it is possible that the some part of the population or some large part of the population may remain unrepresented.

But, in the case of stratified sampling you are trying to divide the entire population into different segments, into different smaller groups and then you are trying to take the sample from each of the subgroup or the stratum. And, the desired degree of representation of some specified part of the population is also possible. You can decide that I want a smaller sample from this group or a bigger sample from this group.

And, when the strata are from intelligently; intelligently means they are within homogeneous and between heterogeneous and this degree is. Actually higher then there is a substantial gain in the efficiency of a simple of stratified random sampling over the simple random sampling.

(Refer Slide Time: 44:36)



Advantages of stratified sampling :

6. In case of skewed population, use of stratification is of importance since larger weight may have to be given for the few extremely large units which in turn reduces the sampling variability.

7. When estimates are required not only for the population but also for the subpopulations, then stratified sampling is helpful.

8. When the sampling frame for subpopulations is more easily available than the sampling frame for whole population, then the stratified sampling is helpful.

20

Particularly, when you have a skewed population; that means, the observation are more concentrated on one side of the population like for example, if I make here three conditions suppose this is here skew symmetric, this is here skewed; that means, more data is on the left hand side and this is also skewed more data is on the right hand side.

Under this type of condition the stratified sampling plays a very important role and the use of stratification gives you an advantage that larger weight may be given to few extremely large units which actually in turn will reduce the sampling variability in the sample. And, obviously, when the estimates are required not only for the population, but also for the sub population then stratified sampling is also helpful.

For example, if I also want the estimate for different states like Punjab, Haryana say Jammu Kashmir, Kerala, Tamil Nadu etc., then a certified sampling will be more helpful that we can have a single estimate for the entire country as well as for the for all the states. And, particularly when the sampling frame for the sub population is more easily available than the sampling frame for the whole population, then use of stratified sampling is more helpful.

And, this is obvious means getting a sampling frame for the entire country like India and getting the sampling frame for a state or for a city or for a district or for a villages much more easier. So, I stop now with this lecture and I have given you a good background that that how stratified sampling is useful, under what type of condition it can be used and what are the advantages of stratified random sampling over the simple random sampling.

Now, it is your turn you try to settle down this concept inside your brain, try to think about some situations where these things can be employed. For example, if you ask me in the case of data sciences suppose, I take the same example of online shopping websites, millions of people are hitting the websites and they are going to different parts of the website related to utensil, related to clothing, relating to laptop etc. etc.

Just by using their login ID you can always find that what type of age group is trying to hit a particular section of the website. Whether younger people are hitting more on the clothing section or the elder people are hitting more on the clothing section and based on that you can improve your marketing strategies.

Now, definitely if you go for all the hits millions and billions that will take a much longer time to react and nowadays the marketing people have to react very quickly. Somebody has gone to the site and some and he or she has look into the clothing

, but suppose he or she has not bought.

So, you have seen that immediately you will get a response like as e-mail or SMS, some coupon code, some e-mail you have not bought it, this item is going to finish soon, only few quantities are left and if you do not do it sometime you get a discount also when whenever the price is reduced you will get a message from the web website.

So, if you want to prepare these type of strategies how do you can you do it? You cannot take the entire data and you cannot make such strategies working with the data of millions and billions, it is more difficult than working with the data of size 100 or say 1000. So, if you can come up with a strategy based on a sample of observation of only 1000 or say 100, do not you think that this is a much better option than working with the millions and billions of observations? So, this is how the stratified random sampling can help you.

It can divide your population into different groups. Yes, you have to write a program. So, that these things are automated, but for that I already told you that if you want to become a data scientist you should have a good knowledge of statistics, as well as computer science and programming.

I am doing my job here as a statistician. Now, this is your job to get acquainted yourself with the good programming language, and then this stratified sampling is also going to help you a lot in the real data situation. So, you think, practice and I will see you in the next lecture.

Till then good bye.