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

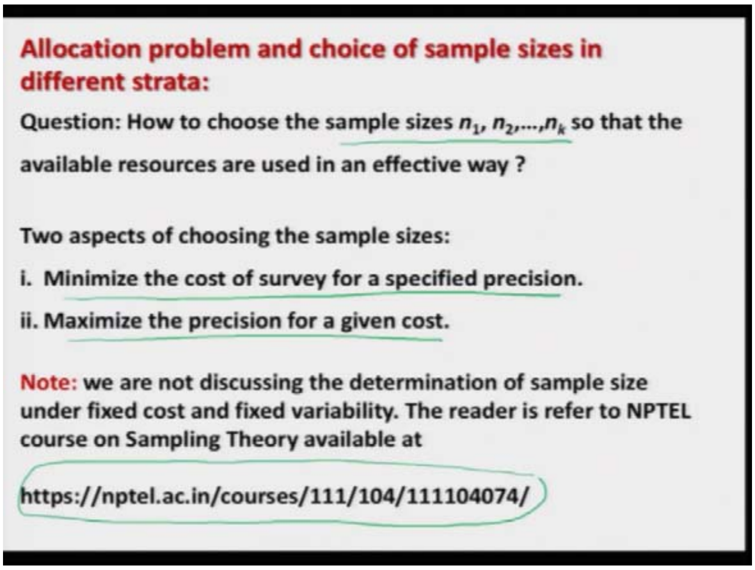
Sampling Theory with R Software
Lecture - 30
Stratified Random Sampling
Sample Allocation and Variances Under Allocation

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. And in this module, we are going to continue with the Sampling Theory with R Software and we will try to learn the topics of Stratified Random Sampling.

And you may recall that in the earlier lecture, I had discussed an issue that how do you decide that, whether the strata to be constructed should be within homogeneous and between heterogeneous. In that process, I had shown you analytically that the variance depends on sample size as well as the variability of the strata.

So, we had control the variability of the strata and I had discussed that we have one more chance to control the variability that is through the collection of sample size, by the choice of sample size. So, people have given different types of criteria to choose the sample size to be drawn from each and every strata. So, today we are going to discuss about those topic. So, let us begin our discussion with this slide.

(Refer Slide Time: 01:40)



Allocation problem and choice of sample sizes in different strata:

Question: How to choose the sample sizes $n_1, n_2, \ldots, n_k$ so that the available resources are used in an effective way ?

Two aspects of choosing the sample sizes:

i. Minimize the cost of survey for a specified precision.

ii. Maximize the precision for a given cost.

Note: we are not discussing the determination of sample size under fixed cost and fixed variability. The reader is refer to NPTEL course on Sampling Theory available at

https://nptel.ac.in/courses/111/104/111104074/

So, now, the problem is this. How we are going to allocate the sampling units in the sample and what should be the my sample size, which has to be chosen from different strata, right? So, essentially we want to answer that, how to choose the sample sizes n_1, n_2,.., n_k. And our objective is that, ok if we want to make the variability of the strata mean as minimum as possible, we want to utilize our resources also effectively; because the experimental, material, time, labour, cost everything is involved when you are trying to conduct a survey.

So, when you are trying to choose the sample size, then everybody would like to choose the sample size in such a way such that, the cost of the survey should be minimum and the sample size should be maximum. So, ideally everybody would like to have minimum cost and maximum sample size.

But you have to understand one thing, both the things cannot be met at the same time. Why? Because when you try to minimize the cost, then the sample size will also decrease; on the other hand, in case if you try to increase the sample size, then the cost will also increase. So, somewhere we have to strike a balance between the cost and sample size.

So, in order to achieve this, various strategies have been presented in stratified sampling and in other types of sampling scheme. So, at this stage I am not really going to discuss all the things, because otherwise the flavour of the decision sciences will be lost and we will be simply precipitating to the say mathematical statistics.

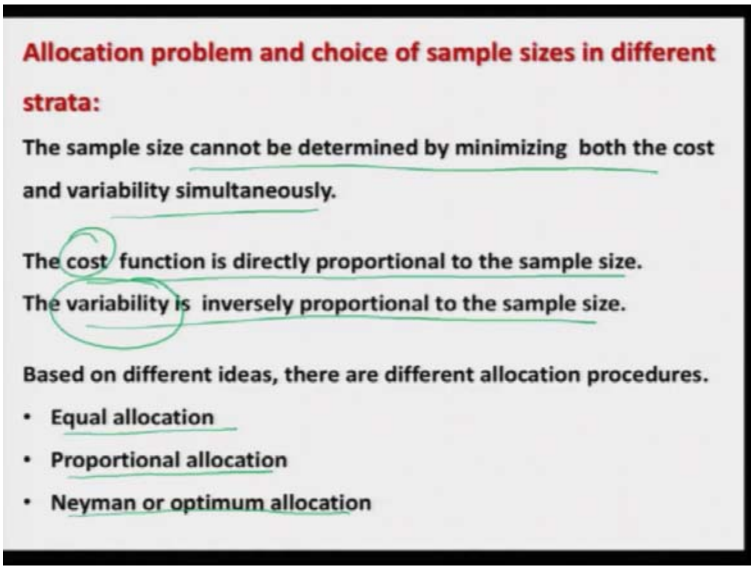
But as I said without mathematical statistics, you cannot implement the decision sciences or the data sciences. So, I would request you that you please try to go through with the chapter on stratified sampling theory in any good book and I also have given the address of the web link, where an NPTEL course on sampling theory is there. So, you can also take help from that course also, right.

So, in this lecture one, two possible approaches as I said are that, we would like to minimize the cost of survey for a specified precision and we would like to maximize the precision for a given cost. These are the two popular approaches which have been considered in stratified sampling that, you try to fix the variance and then you try to minimize the cost or you fix the cost and you try to minimize the precision.

Because you have to understand one thing that, the sampling variability is directly connected to the sample size; as the sample size increases, the variability of the estimator also decreases. So, there is an inverse relationship; so smaller sample size, higher variability; higher sample size, lower variability, right.

So, we have to strike a balance among all those things. So, you can see here that, I am not going to really discuss all the things over here; but this is the web address from where you can have a look on the course on sampling theory, right.

(Refer Slide Time: 05:52)



Allocation problem and choice of sample sizes in different strata:

The sample size cannot be determined by minimizing both the cost and variability simultaneously.

The cost function is directly proportional to the sample size.
The variability is inversely proportional to the sample size.

Based on different ideas, there are different allocation procedures.

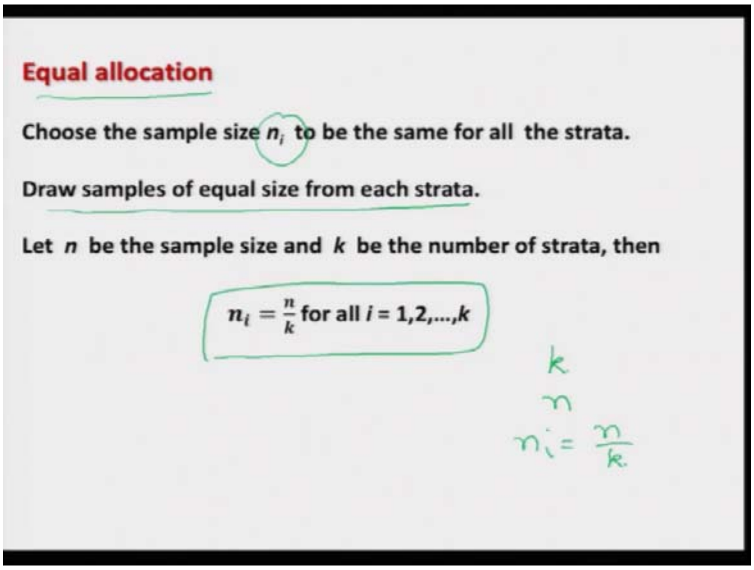
- Equal allocation
- Proportional allocation
- Neyman or optimum allocation

But I will discuss here few simple things. So, now as I said that the sample size cannot be determined by minimizing both the cost and variability simultaneously and the cost function is actually directly proportional to the sample size and the variability is inversely proportional to the sample size.

So, you can see here that, cost and variability, they have got a opposite relationship. So, that is what people have try, people have attempted; they try to fix cost and then they try to minimize variability or they try to fix variability and they try to minimize the cost. Based on this idea, there are three popular allocation schemes; one is equal allocation, second is proportional allocation, and third is Neyman or optimum allocation.

So, just for the sake of understanding to understand the basic concept, I will try to consider these three schemes and I will try to find out the variance of $\bar{y}_{st}$ and the these three type of sample allocation.

(Refer Slide Time: 06:56)



Equal allocation

Choose the sample size n_i to be the same for all the strata.

Draw samples of equal size from each strata.

Let n be the sample size and k be the number of strata, then

$$\boxed{n_i = \frac{n}{k} \text{ for all } i = 1,2,\ldots,k}$$

k

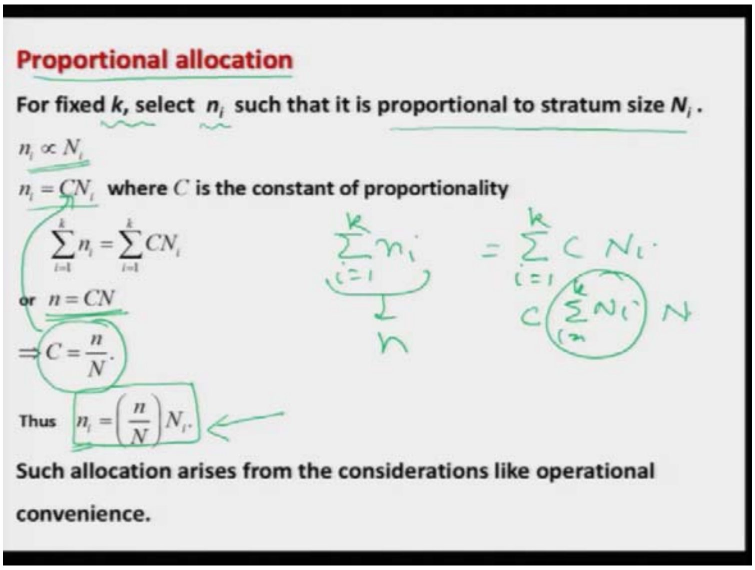
n

$n_i = \frac{n}{k}$

So, first we try to understand, what is equal allocation. So, in the case of equal allocation, we simply tried to choose the sample size n_i to be the same for all the strata; that means from every strata, draw the samples of equal size, as simple as that. So, we are not really bothered that, if the size of any particular strata is say smaller or larger; but we are simply trying to take equal number of observations from every strata.

So, in this case obviously, if there are k strata and your total sample size which you want here is suppose n; then obviously the number of units to be drawn from every strata that is small n_i will simply become here n upon k, right. So, this is what is the rule for the equal allocation.

(Refer Slide Time: 08:08)



Proportional allocation

For fixed *k*, select n_i such that it is proportional to stratum size N_i.

$n_i \propto N_i$

$n_i = CN_i$ where C is the constant of proportionality

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

or $n = CN$

$$\Rightarrow C = \frac{n}{N}.$$

Thus $\boxed{n_i = \left(\frac{n}{N}\right)N_i.}$

Such allocation arises from the considerations like operational convenience.

$$\underbrace{\sum_{i=1}^{k} n_i}_{n} = \sum_{i=1}^{k} C N_i$$

$$C\left(\sum_{i=1}^{k} N_i\right) N$$

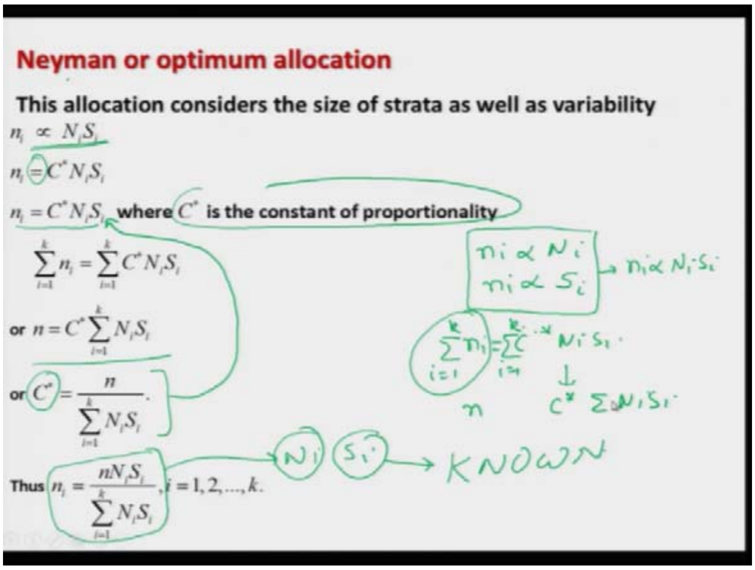
Then we consider the proportional allocation. In proportional allocation the basic concept is that, the sample size to be drawn from the strata depends on the size of the strata. So, the rule is simple, in case if the size of the strata is more; then you try to obtain a sample of bigger size and if the size of the strata is small, then you try to draw a sample of a smaller size, right.

And obviously, the number of strata are already actually fixed. So, for a fixed number of strata, we have to simply choose small n_i such that it is proportional to the stratum size capital N_i; so obviously this I can write it here say small n_i is proportional to capital N_i. And if you try to write down it mathematically, this will come out to be n_i is equal to C times capital N_i, where C is the constant of proportionality.

And if you try to take here sum on both the sides, i goes from 1 to k sum over all the strata; then this is your small n and this will become here C and then $\sum_{i=1}^{k} N_i$ and this is your here population size C, right. So, this is what is here and based on that the constant of proportionality can be obtained as small n upon capital N.

Now, I can substitute this constant over here and I can obtain that n_i, small n_i is n upon capital N into N i; this is actually the rule for proportional allocation. So, it is simply trying to tell you that, if your stratum size is big, try to obtain a sample of big size and if the stratum size is small, try to have a sample of small size, right.

(Refer Slide Time: 10:30)



Neyman or optimum allocation

This allocation considers the size of strata as well as variability

$n_i \propto N_i S_i$

$n_i \boxed{=} C^* N_i S_i$

$n_i = C^* N_i S_i$ **where** C^* **is the constant of proportionality**

$$\sum_{i=1}^{k} n_i = \sum_{i=1}^{k} C^* N_i S_i$$

or $n = C^* \sum_{i=1}^{k} N_i S_i$

or $C^* = \dfrac{n}{\sum_{i=1}^{k} N_i S_i}.$

Thus $n_i = \dfrac{n N_i S_i}{\sum_{i=1}^{k} N_i S_i}, i = 1, 2, \ldots, k.$

$\boxed{\begin{array}{l} n_i \propto N_i \\ n_i \propto S_i \end{array}} \rightarrow n_i \propto N_i \cdot S_i$

$\underbrace{\sum_{i=1}^{k} n_i}_{n} = \sum_{i=1}^{k} C^* N_i S_i \rightarrow C^* \sum N_i S_i$

$N_i \quad S_i \rightarrow$ KNOWN

And after this the there is one more allocation scheme, this is called as optimum allocation or Neyman allocation, right. So, you can see that in the case of proportional allocation, we had considered that a small n_i is proportional to capital N_i; but we also know that if you want to control the variability of the stratum, which is here given by capital S_i that is the standard deviation of the ith stratum.

Then if the variability is more, we need to take a larger sample size. So, I can also write down here that n_i is proportional to capital S_i. So, both these thing should be actually incorporated. So, if I try to combine these two concepts together; I can write down that n_i is proportional to say N_i into S_i, right.

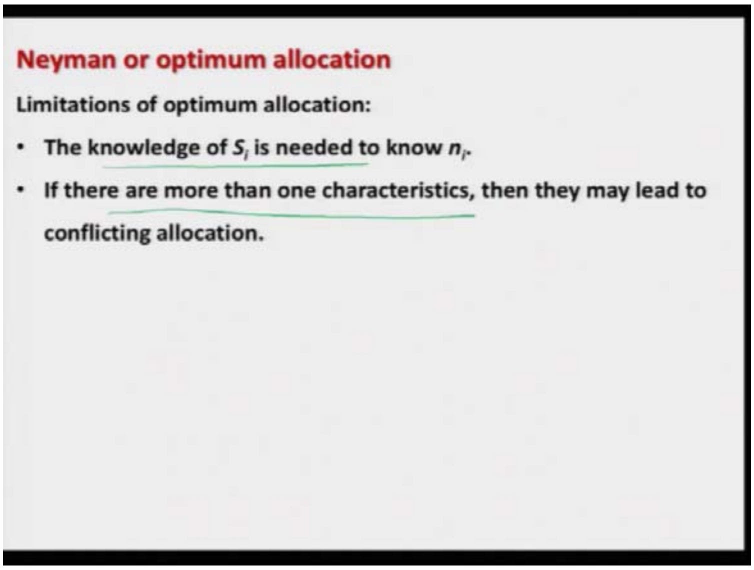
So, this is the basic concept behind the name and allocation that, it is trying to control the variability of the ith stratum as well as the size of the ith stratum. And if I try to write down here with an equality sign; then I have to use here the constant of proportionality and I can write down here $n_i = C^* N_iS_i$, ok.

And if I and then after this, what I try to do? I try to write down here say n i is equal to C_i* N_i S_i and then I try to take the summation over all the stratum on both the sides. So, this quantity becomes here small n and on the right hand side, this becomes here say here $C^*\sum_{i=1}^{k} N_i S_i$, right. So, from here I can obtain, this is here like this and from here I can obtain the value of the constant or proportionality C* here like this.

And if I try to substitute this value of C* over here; then I get $n_i = \frac{nN_i S_i}{\sum_{i=1}^{k} N_i S_i}, i = 1, 2, ..., k$, right. So, you can see here that, this allocation depends on the capital N_i as well as capital S_i. So, now here is a here one issue. Capital N_i will be known to us; but now we also have to assume that, capital S_i is known to us, right. So, this creates hindrance in the application in real data, but this problem can be solved.

For example, one option to know S_i is that, well the value of S_i is known from some past experience from some similar kind of studies which have been conducted in the past or means ultimately it has to be known. Or second option is this which I can propose here is that, if it is unknown; one can estimate it and use it here as if this is known to us. And in order to estimate it, there are various technique and one of the popular technique is bootstrapping.

(Refer Slide Time: 14:13)



Neyman or optimum allocation

Limitations of optimum allocation:

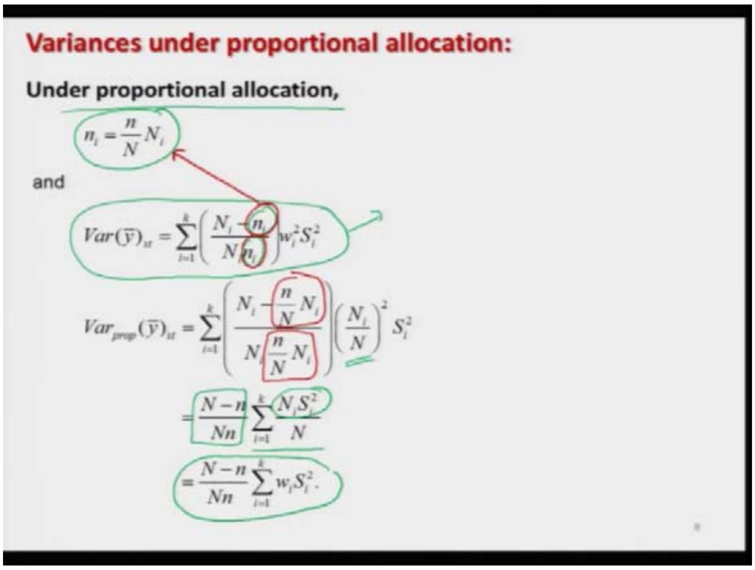
- **The knowledge of S_i is needed to know n_i.**
- **If there are more than one characteristics, then they may lead to conflicting allocation.**

So, what is bootstrapping that will be our next chapter, there I will try to show you, right. So, once you estimate the value of S_i through bootstrapping, that can be replace here and that will work as a reasonable solution; I am not saying that is going to be the best solution, because you are replacing a known value by some estimated value, so obviously it will introduce some variability in the estimator.

So, you can see here that, one of the limitation of Neyman or optimum allocation is that, the knowledge of S_i is needed to know the value of a small n_i. More trouble come if there are more than one characteristic under study, right. Then in that case that becomes a challenge that, which of the variable has to be chosen; because both the variables cannot be chosen or the strata sizes cannot be determined in such a way such that S_i for both the characteristics can really be controlled.

So, this may result into some conflicting allocation. For example, if one choice of variable gives some other value of S_i and another choice of variable gives another value of S_i, so, finally, that may lead to some contradictory results also. So, we have to be careful, ok.

(Refer Slide Time: 15:37)



Variances under proportional allocation:

Under proportional allocation,

$$n_i = \frac{n}{N} N_i$$

and

$$Var(\bar{y})_{st} = \sum_{i=1}^{k} \left(\frac{N_i - n_i}{N_i n_i} \right) w_i^2 S_i^2$$

$$Var_{prop}(\bar{y})_{st} = \sum_{i=1}^{k} \left(\frac{N_i - \frac{n}{N} N_i}{N_i \frac{n}{N} N_i} \right) \left(\frac{N_i}{N} \right)^2 S_i^2$$

$$= \frac{N-n}{Nn} \sum_{i=1}^{k} \frac{N_i S_i^2}{N}$$

$$= \frac{N-n}{Nn} \sum_{i=1}^{k} w_i S_i^2 .$$

So, after this next I try to do one thing that, one simple question arises that, yesterday we had obtained the variance of $\bar{y}_{st}$ and if you remember this was obtained here like this, this expression. So, a natural question arises; what will happen to this variance, when we are trying to choose n_i, a small n_i by different types of allocation schemes.

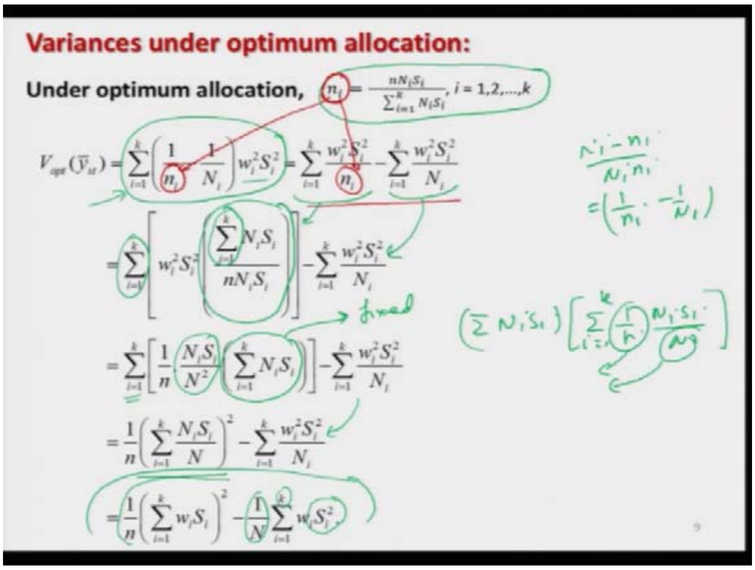
So, it can be equal, it can be proportional or it can be Neyman. So, that is what I am trying to do here that, I would try to find out the variance of $\bar{y}_{st}$ when n_i has been chosen by different types of sampling schemes. First I try to handle the situation, when then the sample size has been chosen by the proportional allocation.

So, what I have to do here that, under proportional allocation, this small n_i is given by this quantity. And what you have to do? You simply have to replace this small n_i in the variance of $\bar{y}_{st}$. with the n_i under proportional allocation. And this is exactly what I have done in the next step; you can see here, this is the n_i under proportional allocation, right.

So, if you simply try to write down here and w_i was say N_i upon N. So, and if you simply try to simplify this expression, this will come out to be $\frac{N-n}{Nn}\sum_{i=1}^{k}\frac{N_i S_i^2}{N}$ or this $N_i/N = w_i$. So, this expression of variance of $\bar{y}_{st}$ under proportional allocation can be written compactly as $\frac{N-n}{Nn}\sum_{i=1}^{k} w_i S_i^2$, right.

You have to be careful with my voice when I say Nn. So, you have to be careful and look into the slides that to which n I am referring to.

(Refer Slide Time: 17:49)



Variances under optimum allocation:

Under optimum allocation, $n_i = \frac{nN_iS_i}{\sum_{i=1}^{k} N_iS_i}, i = 1,2,\dots,k$

$$V_{opt}(\bar{y}_{st}) = \sum_{i=1}^{k}\left(\frac{1}{n_i}-\frac{1}{N_i}\right)w_i^2S_i^2 = \sum_{i=1}^{k}\frac{w_i^2S_i^2}{n_i} - \sum_{i=1}^{k}\frac{w_i^2S_i^2}{N_i}$$

$$= \sum_{i=1}^{k}\left[w_i^2S_i^2\left(\frac{\sum_{i=1}^{k} N_iS_i}{nN_iS_i}\right)\right] - \sum_{i=1}^{k}\frac{w_i^2S_i^2}{N_i}$$

$$= \sum_{i=1}^{k}\left[\frac{1}{n}\frac{N_iS_i}{N^2}\left(\sum_{i=1}^{k} N_iS_i\right)\right] - \sum_{i=1}^{k}\frac{w_i^2S_i^2}{N_i}$$

$$= \frac{1}{n}\left(\sum_{i=1}^{k}\frac{N_iS_i}{N}\right)^2 - \sum_{i=1}^{k}\frac{w_i^2S_i^2}{N_i}$$

$$= \frac{1}{n}\left(\sum_{i=1}^{k} w_iS_i\right)^2 - \frac{1}{N}\sum_{i=1}^{k} w_iS_i^2$$

fixed

$$\frac{N_i - n_i}{N_i n_i} = \left(\frac{1}{n_i} - \frac{1}{N_i}\right)$$

$$\left(\sum N_iS_i\right)\left[\sum_{i=1}^{k}\left(\frac{1}{n}\right)\frac{N_iS_i}{N^2}\right]$$

I will try my best every year. So, now, after this, we consider the optimum allocation and we try to find out the variance of $\bar{y}_{st}$ under this allocation. So, you may recall that the value of the sample to be drawn from the ith stratum is given by this n i. So, what I have to do here? The same thing what I did earlier; the variance of $\bar{y}_{st}$ that we had obtained earlier is given by this quantity, $\sum_{i=1}^{k}\left(\frac{N_i - n_i}{N_i n_i}\right)$.

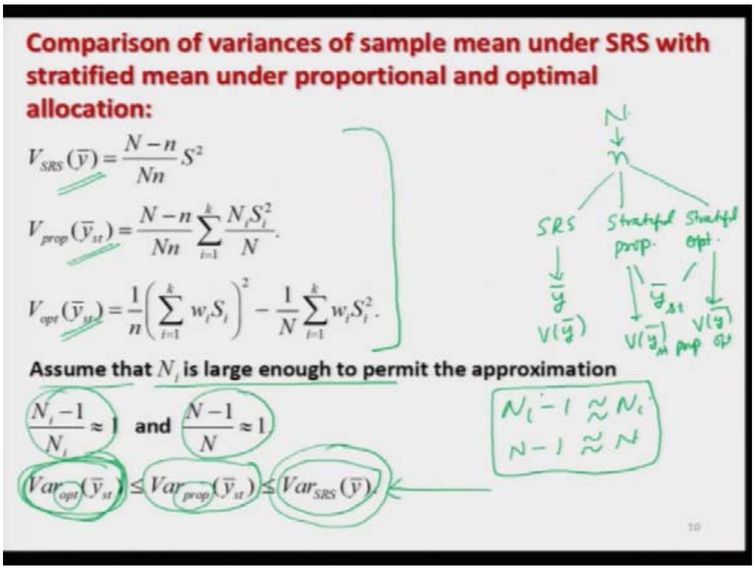
So, this I can write down here 1 upon n_i minus 1 over N_i. So, this is what I have written here. So, this variance is given by this $\sum_{i=1}^{k}\left(\frac{N_i - n_i}{N_i n_i}\right)w_i^2 S_i^2$. So, now what I have to do? I simply have to follow the same rule which I followed in the earlier case that, take this n_i small n_i and substitute it here and just simplify the entire expression, right.

So, in order to just make the algebra more simpler, I have written this expression in this format and I try to replace this n_i over here, right ok. So, if you try to do it here. So, you can see here that, this expression is here and I have simply replace this 1 upon n_i here and after this the second expression, this continues here as such. And if you try to simplify here, you have to be little bit careful; because you can see here there is here i goes from 1 to k and this is also here i goes from 1 to k.

So, do not try to cancel it, sometime people make this mistake and this expression can be as written here as say $\frac{1}{n}\left(\sum_{i=1}^{k}\frac{N_iS_i}{N}\right)^2$. And this is also here summation i goes from 1 to k; but remember one thing, once you are trying to sum all the N_iS_i from i goes from 1 to k, this quantity becomes just like a fixed quantity, right. So, this quantity comes out of the bracket say N_i S_i and then I can write down here $\left(\sum_{i=1}^{k}\frac{N_iS_i}{N}\right)^2$, right.

So, now this 1 upon n and N square they will also come out of the bracket and this quantity can be written as $\left(\sum_{i=1}^{k}\frac{N_iS_i}{N}\right)^2$ like this and this quantity will continue as such. So, now, the variance of $\bar{y}_{st}$ under the optimum allocation becomes 1 upon n inside the bracket $\left(\sum_{i=1}^{k}w_iS_i\right)^2$; mean the $\frac{1}{n}\left(\sum_{i=1}^{k}w_iS_i\right)^2 - \frac{1}{N}\sum_{i=1}^{k}w_iS_i^2$. So, this is the expression for the variance of $\bar{y}_{st}$ under the optimum, right.

(Refer Slide Time: 21:31)



Comparison of variances of sample mean under SRS with stratified mean under proportional and optimal allocation:

$$V_{SRS}(\bar{y}) = \frac{N-n}{Nn}S^2$$

$$V_{prop}(\bar{y}_{st}) = \frac{N-n}{Nn}\sum_{i=1}^{k}\frac{N_iS_i^2}{N}.$$

$$V_{opt}(\bar{y}_{st}) = \frac{1}{n}\left(\sum_{i=1}^{k} w_iS_i\right)^2 - \frac{1}{N}\sum_{i=1}^{k} w_iS_i^2.$$

Assume that N_i is large enough to permit the approximation

$$\frac{N_i-1}{N_i} \approx 1 \text{ and } \frac{N-1}{N} \approx 1$$

$$Var_{opt}(\bar{y}_{st}) \leq Var_{prop}(\bar{y}_{st}) \leq Var_{SRS}(\bar{y})$$

$$N_i - 1 \approx N_i$$
$$N - 1 \approx N$$

Now, the one simple question which arises to our mind is that, suppose somebody has got a population and the person has to draw a sample. Now, the person has got three options; he can draw the sample by SRS, he can draw the sample by stratified and he and the person uses the proportional allocation. And third option is that, the person uses the stratified random sampling, but with optimum allocation.

Now, from here the person estimate the population mean by sample mean $\bar{y}$ and from these two cases, the person will estimate the population mean by $\bar{y}_{st}$. So, a question comes, what will happen to the variance of $\bar{y}$ under SRS, the variance of $\bar{y}_{st}$ under proportional allocation and variance of $\bar{y}_{st}$ under optimum allocation?

So, what will be the inter relationship among these three variances? A popular result in the stratified sampling which is available to us; this says that, in case if I say capital N$_i$ is large and then the question comes, how large it should be? So, we simply assume that it is so much large that, N$_i$ minus 1; that means 1 unit less is approximately same as N$_i$.

And similarly for the capital N also N minus 1 is approximately equal to N, which is not a very strict assumption; means you are simply saying that one unit less and the complete population size. So, this condition is not very difficult to meet in practice. Under this type of condition, it is possible to compare the three variances.

So, you can see here I have written the three variances; variance of $\bar{y}$ under simple random sampling, variance of $\bar{y}_{st}$ under proportional allocation and variance of $\bar{y}_{st}$ under the optimum allocation. And under this assumption that a capital N i is large enough to permit the approximation that N$_i$ minus 1 upon N$_i$ is approximately equal to 1 and N minus 1 upon N is approximately equal to 1.

Under such a condition, the result turns out to be, that the variance of $\bar{y}_{st}$ under the optimum allocation is the least. And this is followed by the variance of $\bar{y}_{st}$ under the proportional allocation and this is further followed by the variance of $\bar{y}$ under simple random sampling. Well, I am not going to give you the proof of this result, which is available in almost all the books on the sampling theory in the chapter of stratified random sampling.

So, but my idea here is this; if you try to look into this expression, do not you think that this expression is intuitively clear? Because once you are trying to say that you have to take a decision that, whether you have to find out, that whether you want to use simple random sampling or the stratified sampling. So, obviously if you want to use the stratified sampling; then the variability of the population is going to be quite large.

And once the variability is going to be large, under that situation if you try to use the simple random sampling; that will obviously not result in a say a good amount of variability, the variability will be higher for this small $\bar{y}$. And obviously, when you are trying to use the stratified random sampling; then when you are trying to use the proportional allocation, you are using only one information that is small n$_i$ is proportional to capital N$_i$.

But when you are trying to use the optimum allocation, then you are using more information; you are using both the information that n$_i$ is proportional to capital N$_i$ as well as it is proportional to capital S$_i$. So, that is why the variance of $\bar{y}_{st}$ under the optimum allocation is using more information. So, that is intuitively expected to give you a better result, right ok.

So, now I will stop in this lecture. And so, what we have done? We have considered different types of allocation schemes; we have obtained the different types of choice of sample sizes which can be drawn from each of the stratum and then we also have computed their variances. There are some other types of allocation, which are available in the literature; there are several allocation which are based on some constraints, like as fixed cost, very fixed variability and so on.

This I am not considering here; but I will request you that you please study it study them yourself. And one thing is there that the question comes how are you going to estimate this variances on the basis of random sample, which you have obtained. So, you already have seen that, we already proved that expected value of small s square is equal to capital S square.

And when you are trying to consider the simple random sampling without replacement; so from every strata, so, you can assume that expected value of s_i^2 is equal to S_i^2 and you can use this result to obtain the estimates of different types of variances. In some cases it might be possible that you have to find out an approximate estimate, but that is the option what we have.

Second thing is this, what we have used here; we have used the simple random sampling without replacement to draw the samples from each and every stratum. Second option is this; well if somebody wants to use simple random sampling with replacement, then what will happen? So, I will say, the way I have presented all the results; the same results can also we obtain for the simple random sampling with replacement, but there will be some difference. So, be careful when you are trying to use it.

In practice when you are not given any choice; then obviously we have proved that, simple random sampling without replacement provides more efficient results. So, usually we will go only for SRSWOR, unless and until it is the need of the time, right. So, now I will stop here with the theory part of the stratified random sampling. And from the next lecture, I will try to come up to the software part; that means how to generate the samples using the R software.

So, you practice it, learn it and I will see you in the next lecture once again; till then good bye.