1. Introduction

Welcome to the series of e-learning modules on Practical Difficulties in Adopting Optimum Allocations.
In this module we are going cover the basic problems of Stratified Sampling, Optimum Allocation, Need for Optimum Stratification, and Practical Difficulties in Adopting Optimum Allocation.

By the end of this session, you will be able to understand:
- The practical difficulties of stratified sampling
- Optimum Allocation
- The need for optimum allocation, and
- The practical difficulties in adopting optimum allocation

To begin with, let us recall a few points.
Any result produced from a sample can be used to estimate the corresponding result for a population.
It is essential that the sample taken is as representative as possible of that population.
Common sense rightly suggests that the larger the sample the more representative it is likely to be, but also the more expensive it is to take and analyze.
A random sample selected using Stratified sampling is ideal for statistical analysis. But, for various reasons, this method is not feasible.

In stratified sampling, the population of interest can be divided into 'k' non-overlapping sub-populations or strata of size Nh, where h equals to 1, 2 etc k, according to a stratification variable.
The total sample size 'n' is then allocated to the strata, so that 'n equals to summation nh'.
Samples of size 'nh' are drawn within each of the 'k' strata.

<u>Condition 1:</u>
A minimum of two elements must be chosen from each stratum so that sampling errors can be estimated for all strata independently.
- The survey commands will not work if there are less than 2 elements in any strata. This is often a problem for sub-group analyses.
- A solution to this is to combine adjacent strata.

<u>Condition 2:</u>
The population (elements) should be homogenous within stratum, and the population (elements) should be heterogeneous between the strata.
This condition suggests that: "the gains in variance precision is greatest when the strata are maximally heterogeneous between, but homogenous within".

Stratified Sampling has advantages like:
- It assures representation of all groups in sample population needed
- The characteristics of each stratum can be estimated and comparisons can be made, and
- It reduces the variability

2. Disadvantages of Stratified Sampling

At the same time, it also has its own disadvantages.
1. It is difficult for the researcher to decide the relevant criteria for stratification
2. Only one criterion can be used for stratification but it generally needs more than one criterion relevant for stratification, making it costly and time consuming
3. Selected sample may be representative with respect to the used criterion but not with respect to others and there is a risk in generalization
4. Stratified sampling is not useful when the population cannot be exhaustively partitioned into disjoint subgroups.
5. It would be a misapplication of the technique to make subgroups' sample sizes proportional to the amount of data available from the subgroups, rather than scaling sample sizes to subgroup .
6. On the other hand, if the variances vary so much among subgroups, that the data need to be stratified by variance, there is no way to make the subgroup sample sizes proportional (at the same time) to the its sizes within the total population.
7. Information on stratification variables is required for each element in the population. If such information is not readily available, they may be costly to compile. This technique obtains a more representative sample because it ensures that elements from each stratum are represented in the sample, but at the same time it is more expensive, time-consuming, and complicated than simple random sampling.

The precision and cost of a stratified design is influenced by the way the sample elements are allocated to strata.
One approach of stratified sampling is proportionate stratification.
With proportionate stratification, the sample size of each stratum is proportionate to the population size of the stratum.

The second approach is Optimum Allocation.
This can be a better choice in terms of cost and precision, if sample elements are assigned correctly to strata.
With Optimum stratification, the sample size of each stratum is proportionate to the population size of the stratum multiplied by the root mean square deviation for population.
Strata sample sizes are determined by the equation:
'nh' is equal to 'n into Nh into Sh' divided by 'summation Nh into Sh'.

To take complete advantage of Optimum stratification, researchers need to know the following, based on the variability of sampling:
- More variable strata should be sampled more intensely. (e.g. as variability increases sample size of stratum should increase)
- The number of sampling units drawn from each stratum is allocated according to analytical considerations

Hence Optimum Allocation needs "weighted analysis". That is:

- The number of sampling units drawn from each stratum is determined on the basis of both size and variation, and,
- Optimum Allocation has to be calculated statistically
- Comparison between Proportional & Optimum Allocations
- When there is high variation between strata, we get high precision under Optimum Allocation than Proportional Allocation.
- In practice the gain by switching from Simple Random Sampling to Proportional Allocation is much bigger than by switching from Proportional Allocation to Optimum Allocation.
- In Optimum Allocation, we need to know the stratum variances in order to obtain the samples sizes for each stratum, which could be more problematic.
- Although PA may yield smaller margins of error than Simple Random Sampling in estimating population parameters, it may still be possible to do better
- Optimum Allocation is designed to achieve even greater overall accuracy than that achieved using Proportional Allocation. It sets the sample size of the different strata, taking into account two important aspects of doing research: Cost and Precision.
- The sampling fraction varies according to the costs and variability within the various strata. Optimum allocation, may be more appropriate for a study than proportionate stratified sampling when the strata differ in terms of data collection costs and the variability of the variables of interest.

3. Optimum Allocation

This topic deals with the practical difficulties of optimum allocation in spite of its precision over the other techniques.
Suppose that we intend to use Optimum allocation for given ‘n’. The sample size n_h in stratum h should be:
‘n’ into ‘Nh into Sh’ divided by ‘summation Nh into Sh’.

The variance of the unbiased estimate of the population variance is given by:
‘Summation wh into Sh whole square by n’ minus ‘Summation wh into Sh square by N’.
In practice Since ‘Sh’ is not known, and we can only approximate this allocation.

Optimum allocation leads to the following rules of conduct:
In a given stratum, take a larger sample if:

- The stratum is larger
- The stratum is more variable internally
- Sampling is cheaper in the stratum

The most serious limitation of Optimum Allocation is the absence of the knowledge of Sh’s in advance.
In order to overcome this difficulty, a pilot survey of size ‘n’ may first be carried out in order to provide the estimates of Sh, where h is equal to 1, 2 etc k.
These estimated values of ‘Sh’s’ may be used in obtaining optimum values of ‘nh’ to be allocated in different strata .

However these estimates are subject to sampling errors and in case Sh’s happen to be estimated with low precision then the advantage of optimum allocation may be lost and we might even be in a worse position as compared to proportional allocation.

Sukhatme obtained the expression n dash, the size of the pilot survey, in order that ‘y-bar st’ under Neyman’s Allocation, based on the estimated Sh’s may not, on an average lead to loss of precision as compared to proportional allocation and is given by:

‘n dash’ is equal to ‘summation, h runs from 1 to k’ ‘Wh into Sh whole square’ minus ‘summation ,h runs from 1 to k’ ‘Wh into Sh square’ divided by ‘two into summation, h runs from 1 to k’ Wh into ‘Sh minus S bar whole square’.

From the above, it may be observed that, the larger the difference in the values of Sh’s, smaller the value of ‘n’. Hence unless Sh’s are homogeneous, even moderately small values of n-dash will give, on an average more precise estimates, than proportional allocation

The second difficulty with this method is that, if our study relates to the estimation of more than one population characteristic from the sample survey, then the Neyman's allocation of the sample to different strata on the basis of one characteristic may result in loss of precision on other characters as compared to the method of Proportional allocation

The third difficulty involved is that,

Sometimes it may happen that the optimum values of nh in any stratum may be greater than

Nh, the total number of units in that stratum.

In such situation we take nh as equal to Nh for the stratum requiring 100 percent sampling,

while for the other strata, the optimum sample is recalculated.

The additional difficulties associated include:
- High cost and low frequency of use
- It requires a sampling frame
- Does not make use of researchers' expertise

OA has a larger risk of random error.

In OA, or Disproportionate allocation, each stratum is proportionate to the standard deviation of the distribution of the variable.

Larger samples are taken in the strata with the greatest variability to generate the least possible sampling variance.

Optimum allocation may be applied focusing on:

- Cost only
- Precision only, or
- Both cost and precision jointly

Homogeneous strata with a smaller sample size can have the same level of precision as heterogeneous strata with a larger sample size.

Applying this principle, it may be useful to make the number of elements selected from each stratum directly related to the standard deviation of the variable of interest in the stratum.

- The greater the variability of the variable in a stratum, the higher the sample size of the stratum should be.
- Moreover, taking into account data collection costs, the higher the data collection costs

of a stratum, the lower the Marketing Region Population.

- If data collection costs for the various districts are unavailable or essentially the same, one may yet optimize the sample sizes of the various strata by allocating the sample size of each stratum by considering the variability of the strata.

○ The use of this optimization procedure is dependent on data on the variability of the variable of interests for the different strata.

○ Often such data are not available. Moreover, if the study has multiple purposes and more than one variable of interest, their optimization might conflict with each other.

If data are available for both the data collection costs and the variability of the variable of interest, one may optimize for both costs and precision.

A weighting factor taking into account both data collection costs and standard deviation may be computed as:

's by squareroot ch',

Where 's' represents the standard deviation within the stratum, and 'ch' represents the per-unit data collection costs within the strata.

- OA takes greater advantage of the knowledge the researcher has about the population.
- Selection of stratification variables may be difficult if the study involves a large number of variables.
- Data collection costs may be lower if the stratification variable breaks up the population into homogeneous geographical areas, or so as to facilitate data collection.

4. Drawbacks of Stratified Optimum Sampling

One of the primary drawbacks of stratified Optimum sampling arises when and if, survey designers change strata boundaries. Some reasons that designers might want to change strata boundaries include:

1. Rare habitats are not actually present in the ‘rare habitat’ strata
2. The rare habitat in the ‘rare habitat’ strata dies or moves
3. Access costs are not as anticipated, and
4. Access costs change

Once drawn, the strata definitions and boundaries must remain fixed forever.
Changing the strata definitions results in an entirely new survey, and it is complicated to compare parameters before and after the strata changes.
Such comparisons are likely to involve unequal probability analyses, and should be conducted by a qualified statistician.

For this reason, we recommended defining strata based on unchanging features and not a vegetation map, which is likely to change.
Even though we recommend stratification based on access or rare habitat, designers must think carefully about this option because access costs may change in future, and rare species might not be present or might emigrate from a certain stratum.

2. Another difficulty with Optimum stratification is perceptual, but nonetheless results in pressure to either change strata membership or boundaries.
In reality, strata are artificial constructs used to control and distribute sampling and may include a mix of habitat types despite efforts to the contrary.
However, because the same name is often given to both the habitat type and the stratum, it is easy to perceive habitat and strata as synonymous.

This perception can result in a misguided desire to change stratum membership of sample units if habitat is not in the strata where it was expected.
If habitat changes, or if sample units in certain strata do not share a characteristic with other units in the same stratum, the stratum membership of the affected units cannot be changed.
Doing so will bias estimate because similar ‘corrections’ cannot be made for the points that were not in the sample.

This means that if researchers define a stratum to include rare habitat, but upon arrival at a sample point in this stratum, the point is determined to be in common habitat, the point must not be changed or relabeled into the common habitat stratum. Estimation domains should be used to make estimates for habitat types, whenever habitats do not completely match stratum boundaries.

3. Another disadvantage of stratified optimum sampling is that stratum usually cannot be defined to improve estimation of all parameters.

For example, stratification that is appropriate for the vegetation component of monitoring is probably not appropriate for the large undulate or glacier monitoring components.
If strata are based on access, different components of the overall monitoring project are likely to have different access issues.

For example: Field work for a small mammal monitoring component may require more equipment than the bird monitoring component, and this fact may restrict the areas accessible to the small mammal component relative to the bird monitoring program.
In these cases, basing a single stratification scheme on access will be at a minimum difficult, and at a maximum, counterproductive for certain components.

4. The potential for stratification ‘errors’ (i.e., mis-classification) is also higher when many strata are defined. When many ‘errors’ occur, pressure to change strata boundaries increases and the analysis is continually complicated.

- In most situations researchers do not know enough about responses to effectively stratify in a way that increases precision.
- If pilot data are available, it may be possible to investigate different stratification schemes and their effects on both the complexity and precision of results.

5. Reducing bias and under-representation is the aim of stratification, however unless the strata are carefully chosen, the sampling error and bias can sometimes be increased via this method.
Researchers must take precautions to choose the factors correctly and assign the strata properly.
For example: Researchers might set strata based on race for an election poll, not foreseeing that race would not have a major effect on voting.
Optimum allocation is also particularly tricky as several calculations and estimations need to be made to choose the sample allocations.

5. Optimum Allocation with Reference to Proportional Allocation

We shall now look at Optimum Allocation with Reference to Proportional Allocation:
Stratified Optimum Sampling is used to eliminate possible errors that can occur with simple random sampling.
However, as with any sampling method, the researchers' knowledge and experience play an important role in how accurate the final results may be.
Proportional allocation is advisable when all we know of the strata is their sizes.
In situations where the standard deviations of the strata are known it may be advantageous to make a Optimum allocation.

Suppose that, once again, we have Stratum A and Stratum B, but we know that the individuals assigned to Stratum A were more varied with respect to their opinions than those assigned to Stratum B.
Optimum allocation minimizes the standard error of the estimated mean by ensuring that more respondents are assigned to the stratum within which there is greatest variation.

Optimum Allocation takes into account stratum sizes, different variances, and different costs of sampling in different strata.
Optimal Allocation is not used as much as proportional allocation but it can result in a gain in precision if the costs and variances are known or well-estimated from a prior study or a pilot survey.

When Sh is unknown we go for estimation of Sh.
Evans examined the effects of errors in the estimated Sh and developed an approximate rule showing whether an estimated optimum is likely to be more precise than proportional allocation.
He supposes that the coefficient of variations of the estimated Sh is the same in all the strata.

This assumption is appropriate when Sh has been estimated from preliminary samples of the same size in each stratum.
Evans shows how to compute the size of the preliminary samples needed to make an optimum allocation better, on the average than proportional allocation.
Previously Sukhatme showed that a small initial sample usually gives a high probability that optimum allocation will be superior to proportional allocation.

Here’s a summary of our learning in this session:

- Illustrated practical problems of Stratified sampling
- Discussed need for Optimum Allocation and its merits
- Addressed practical difficulties and drawbacks of optimum allocation