SIAP

Statistical Institute for Asia and the Pacific

STRATIFICATION AND CLUSTERING

BY

ALICK MJUMA NYASULU-SIAP

Regional Training Course on Sampling Methods for Producing Core Data Items for Agricultural and Rural Statistics

Jakarta, Indonesia ,29 Sep-10 October 2014.

UNITED NATIONS
siap
Statistical Institute for Asia and the Pacific

Outline

1. Basics of strata and clusters
2. Objectives of stratification
3. Sample allocation to strata

Strata and Clusters

- Both stratification and clustering involve subdividing the population into mutually exclusive groups.
- Sub-divisions of the population are called ‘clusters’ or ‘strata’ depending upon the sampling procedure adopted.
- The term ‘cluster’ is used in the context of cluster sampling and multi-stage (cluster) sampling.
- To understand the application of these in different situations, let us take a simple example.

Clustering and Stratification

NATURALLY OCCURRING CLUSTERS

Clusters are usually defined as groups of units that are found naturally 'clustered' together - by location or socially defined entities like households or by institutions like schools and enterprises.

Cluster	Population Unit
Census Enumeration Area	Dwelling
household	Person
Day	Hour
School	Student
Employer	Employee

CLUSTERING AND STRATIFICATION IN SAMPLE DESIGN

Typically, sample surveys conducted by NSOs involve subdividing the population into strata and clusters.

Usually, the technique of stratifying the clusters and then further stratifying the units within clusters are applied to obtain the final sample.

The sampler's objective is to get the right combination of stratification and clustering to get the required estimates at the desired level of accuracy with the given resources.

CLUSTERING AND STRATIFICATION IN SAMPLE DESIGN (CONTD.)

The reliability or precision of the estimates depends on the degree to which the sample is *clustered.*

Generally, *clustering* increases the *sampling variance* considerably.

Usually, stratification is applied to decrease the *sampling variance,* but its effect is often not significant.

Effects of *clustering* and *stratification* is measured by the design effect, or *deff.*

Primarily, *deff* indicates, how much *clustering* there is in the survey sample.

Objectives of Stratification

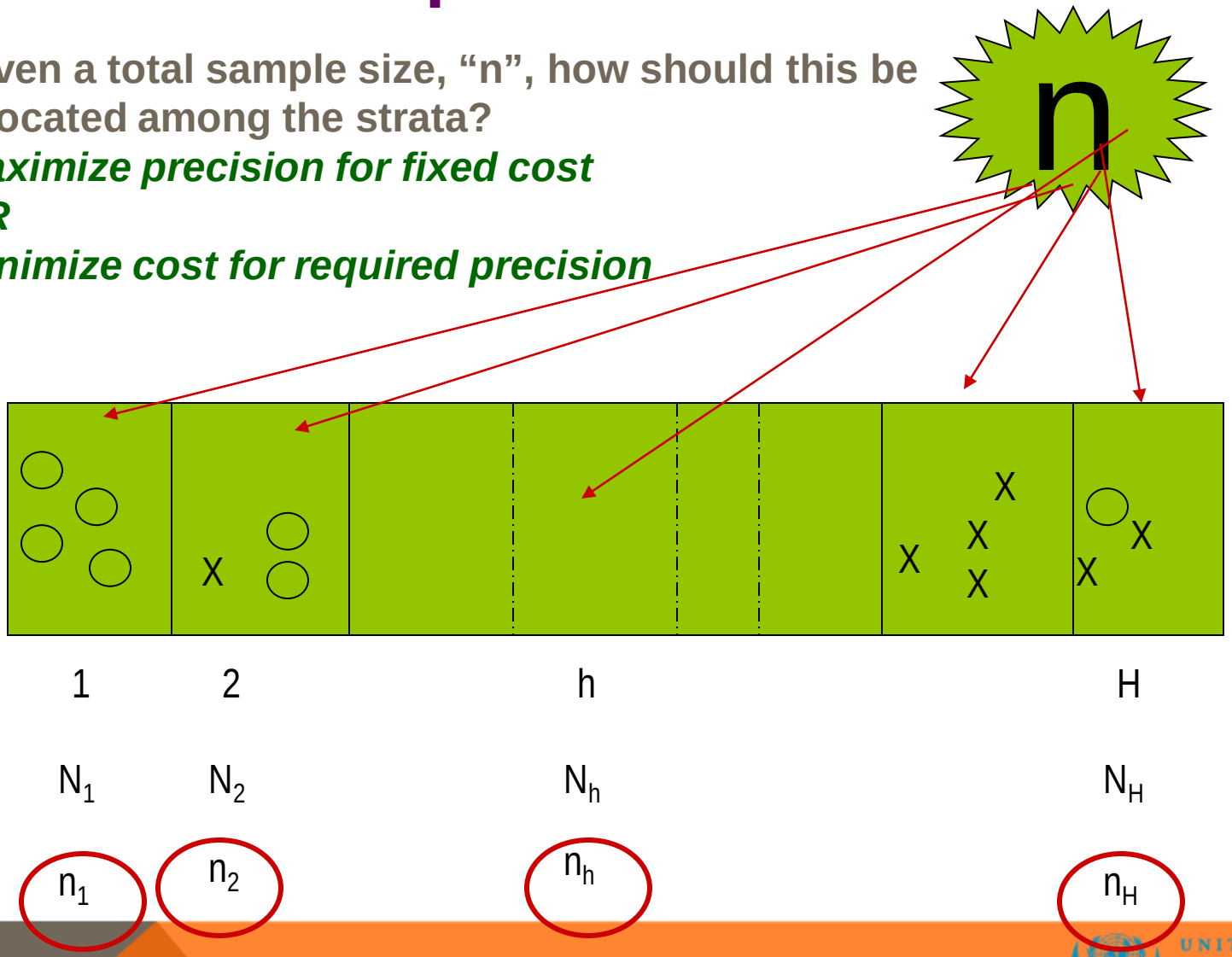
- To obtain estimates of higher efficiency for given per unit of cost
- Providing separate estimates required for each sub-division of the population – “domain” estimates
- Using different sampling procedures for different sub-population, to (i) increase efficiency of the estimates (ii) organize the field work

DEFINING STRATA

1. **Choice of stratification variables (location, output etc):**
   - Homogeneous within strata; Heterogeneous across strata
   - Highly correlated with study variables (output with profit or number of employees etc)
2. **Number of strata**
   - Depends on availability of stratifying information in sampling frame: less information, fewer strata
   - At least two sampling units per stratum to be able to compute sampling error

Allocation Sample over Strata

- Given a total sample size, “n”, how should this be allocated among the strata?
  Maximize precision for fixed cost
  OR
  Minimize cost for required precision



SAMPLE ALLOCATION TO STRATA

Alternatives Methods:

- Proportionate allocation
  - Uniform or equal allocation
- Disproportionate allocation
  - Optimum allocation (minimum variance), fixed sample size
  - Cost optimum allocation (not discussed!)

SAMPLE ALLOCATION TO STRATA

In *proportionate stratification*, an uniform sampling fraction is applied to each strata; that is, the sample size selected from each stratum is made proportionate to the population size of the stratum

In *disproportionate stratification*, different sampling rates are used deliberately in different strata

PROPORTIONATE ALLOCATION

In *proportionate stratification*,

$\frac{n_h}{N_h}$ is specified to be the same for each stratum.

This implies that the overall sampling fraction is $\frac{n}{N}$

$$\frac{n_h}{N_h} = \frac{n}{N}$$

The number of elements taken from the h^{th} stratum is

$$n_h = (N_h)\frac{n}{N}$$

PROPORTIONATE ALLOCATION

$$V_{SRS} \geq V_{prop}$$

Thus, for proportionate stratified *deff < 1*

For a given total variability in the population, the gain is greater if:

- the *strata mean are more heterogeneous* (more unequal strata mean)

OR

- *the element values within the strata are more homogeneous*

OPTIMUM ALLOCATION

Uses widely different sampling rates for the various strata.

Objective: to achieve the *least variance for the overall mean* for the given sample size (Neyman's allocation); as well as given per unit of *cost in different strata.*

Without cost consideration, the allocation is

$$n_h = n\frac{N_h\sigma_h}{\sum N_h\sigma_h}$$

This gives better efficiency as compared to proportionate allocation:

$$V_{SRS} \geq V_{prop} \geq V_{opt}$$

THANK YOU