STRATIFIED RANDOM SAMPLING.

Stratified random sampling is a technique which attempts to restrict the possible samples to those which are ``less extreme'' by ensuring that all parts of the population are represented in the sample in order to increase the efficiency (that is to decrease the error in the estimation). In stratified sampling the population of N units is first divided into disjoint groups of N_1, N_2,...,N_h,..N_L, units, respectively. These subgroups, called strata, together they compromise the whole population, so that $N_1+N_2+...N_L=N$ From each stratum a sample, of pre-specified size, is drawn independently in different strata. Then the collection of these samples constitute a stratified sample. If a simple random sample selection scheme is used in each stratum then the corresponding sample is called a stratified random sample.

Reasons for stratification.

- To obtain estimates of known precision for certain subdivisions of the population by treating each subdivision as a stratum. Since sampling is done independently in each stratum, separate stratum estimates and their precision can be obtained by treating each stratum as a ``population'' in its own right. For example, in household surveys estimates may be required by province, income group, occupation, age group, etc. In business surveys, estimates are often required by Standard Industrial Classification(SIC).
- For administrative convenience; for example stratification can provide survey organization to control the distribution of fieldwork among its regional offices.
- Sometimes different parts of the population may call for different sampling procedures. With human populations, people living in institutions (e.g. hotels, hospitals) are often placed in a different stratum from people living in ordinary homes. In household surveys, since households are sparsely populated in rural areas compared to distribution of households in urban areas, separate sampling schemes have to be employed.
- Stratification may often produce a gain in precision of the estimates of characteristics of the whole population. The amount in the gain depends on the type of stratification. If the population is heterogeneous and if it can be divided, using prior information about the population, into subpopulations (strata), each of which is internally homogeneous. If each stratum is homogeneous, that is characteristic under consideration vary little from one unit to another, a precise estimate (an estimate with smaller variance) of any stratum parameter can be obtained from a small sample in that stratum. These estimates can then be combined to obtain a precise estimate for the whole population.

In this course, only simple random sampling selection plan within each stratum will be discussed. But, since stratification is a technique for structuring the population before taking the sample, it can be used with any of the sampling technique that will be discussed later in this course.

NOTATION.

The suffix h (h=1,2,...,L) denotes the stratum and i the unit within the stratum.

N_h:- Total number of population units in stratum h.

n_h:- Total number of sample units in stratum h.

$W_h=N_h/N$: The h-th stratum weight.

X_{hi} :- Value of the characteristic for the i-th unit in stratum h.

$X_{h+} = \sum_{i=1}^{N_h} X_{hi}$ =Population total of X-values for units belonging to stratum h.

$\overline{X}_h = \frac{\sum_{i=1}^{N_h} X_{hi}}{N_h} = \frac{X_{h+}}{N_h}$ =Population mean of X-values for units belonging to stratum h.

$\sigma_h^2 = \frac{1}{N_h}\sum_{i=1}^{N_h}(X_{hi} - \overline{X}_h)^2$ = Population variance of X-values for units belonging to stratum h.

$\overline{X} = \frac{\sum_{h=1}^{L}\sum_{i=1}^{N_h} X_{hi}}{\sum_{h=1}^{L} N_h} = \sum_{h=1}^{L} W_h \overline{X}_h$ =Population mean of X-values.

$x_{h+} = \sum_{i=1}^{n_h} x_{hi}$ =sample total of X-values for units belonging to stratum h.

$\bar{x}_h = \frac{\sum_{i=1}^{n_h} x_{hi}}{n_h} = \frac{x_{h+}}{n_h}$ =Sample mean of X-values for units belonging to stratum h.

$s_h^2 = \frac{1}{n_h - 1}\sum_{i=1}^{n_h}(x_{hi} - \bar{x}_h)^2$ = Sample variance of X-values for units belonging to stratum h.

$\bar{x} = \frac{\sum_{h=1}^{L}\sum_{i=1}^{n_h} X_{hi}}{\sum_{h=1}^{L} n_h}$ =Sample mean of X-values.

Estimation of Population Mean under Stratified Random Sampling

Note that the Population Mean is given by

$\overline{X} = \frac{\sum_{h=1}^{L}\sum_{i=1}^{N_h} X_{hi}}{\sum_{h=1}^{L} N_h} = \sum_{h=1}^{L} W_h \overline{X}_h$ and since within each stratum sample data are obtained using SRS,

an unbiased estimator of $\overline{X}$ is given by

$\hat{\overline{X}} = \sum_{h=1}^{L} W_h \overline{x}_h$.

Also, since sampling is done independently within each stratum

$\text{Var}(\hat{\overline{X}}) = \sum_{h=1}^{L} W_h^2 \frac{N_h - n_h}{N_h - 1} \sigma_h^2$.

Note that $\text{Var}(\hat{\overline{X}})$ can not be computed since it involves X-values for all the units in the population. However, based on X-values for the sampled units we can estimate $\text{Var}(\hat{\overline{X}})$ by using the following formula:

$\widehat{Var(\hat{\overline{X}})} \sum_{h=1}^{L} W_h^2 \frac{N_h - n_h}{N_h n_h} s_h^2$ which estimates $\text{Var}(\hat{\overline{X}})$ unbiasedly.

Sample Size Allocation (Chapter 6)

How to allocate over all sample size n among L strata? That is how to find n_h such that $n=n_1+n_2+...+n_L$?

1. Proportional allocation: Take $n_h=nW_h$
2. Optimum allocation (with equal cost): Take $n_h= n\frac{W_h\sigma_h}{\sum_{h=1}^{L} W_h\sigma_h}$
3. Optimum allocation (with unequal cost): Take $n_h= n\frac{W_h\sigma_h/\sqrt{C_h}}{\sum_{h=1}^{L} [W_h\sigma_h/\sqrt{C_h}]}$