stratified sampling

Canonical name	StratifiedSampling
Date of creation	2013-03-22 15:17:17
Last modified on	2013-03-22 15:17:17
Owner	CWoo (3771)
Last modified by	CWoo (3771)
Numerical id	6
Author	CWoo (3771)
Entry type	Definition
Classification	msc 62D05
Defines	stratified random sampling
Defines	stratum
Defines	stratum weight

In sampling surveys, it is sometimes a good idea to break up the population into subdivisions before any sampling were to take place. For example, from a population U of automobile insurance policies, the claim frequencies (loosely speaking, this is the ratio of the number of claims to the number of policies in the U) are found in the following table:

male drivers	female drivers	all drivers
10%	7%	9%

Suppose that we would like to take a sample from U so that when the sample units are divided up into male drivers and female drivers, the respective sample claim frequencies are more or less 10% and 7%. How would we do this? If a simple random sample is taken directly from the population, we may get the total claim frequency (for all drivers) to be more or less 9%, but when the sample is broken down into two groups by gender, we may no longer guarantee that the claim frequencies by gender match (more or less) those calculated from the population. To insure that the sample taken preserves claim frequencies by gender, we would take a *stratified sampling*.

Formally, in *stratified sampling*, the following steps are taken, in order, from a population U of N units:

1. Decide what subdivisions are to be analyzed from within U and what information (or statistics) within the subdivisions should be "preserved". For example, if we want to analyze our data by gender, then we would have two subdivisions to study. If there is more than one categorical variable, then we would look at all the *possible* combinations of the these variables.

2. Make sure all the possible combinations are mutually exclusive events;

3. Divide U into k subdivisions, or *strata*, U_i, where k is the total number of possible combinations described above. From the first two steps, we have
$$U = U_1 \cup U_2 \cup \dots U_k \text{ such that } U_i \cap U_j = \varnothing,$$
for all $i \neq j$ and $1 \leq i, j \leq k$. In addition, if we let $N_i = |U_i|$, then

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

4. Draw a sample S_i from each stratum U_i.

Remarks.

- When each S_i is a simple random sample within each U_i, then we call this procedure a *stratified random sampling*.

- Each stratum corresponds to a number

$$W_i := \frac{N}{N_i},$$

  called a *stratum weight*.

- Suppose each sample S_i contains n_i units ($|S_i| = n_i$) and that $n = \sum_{i=1}^{k} n_i$. We call the stratified sampling *proportional* if, for each i,

$$\frac{n}{n_i} = W_i.$$