Taking into account sampling design in DAD

SAMPLING DESIGN AND DAD

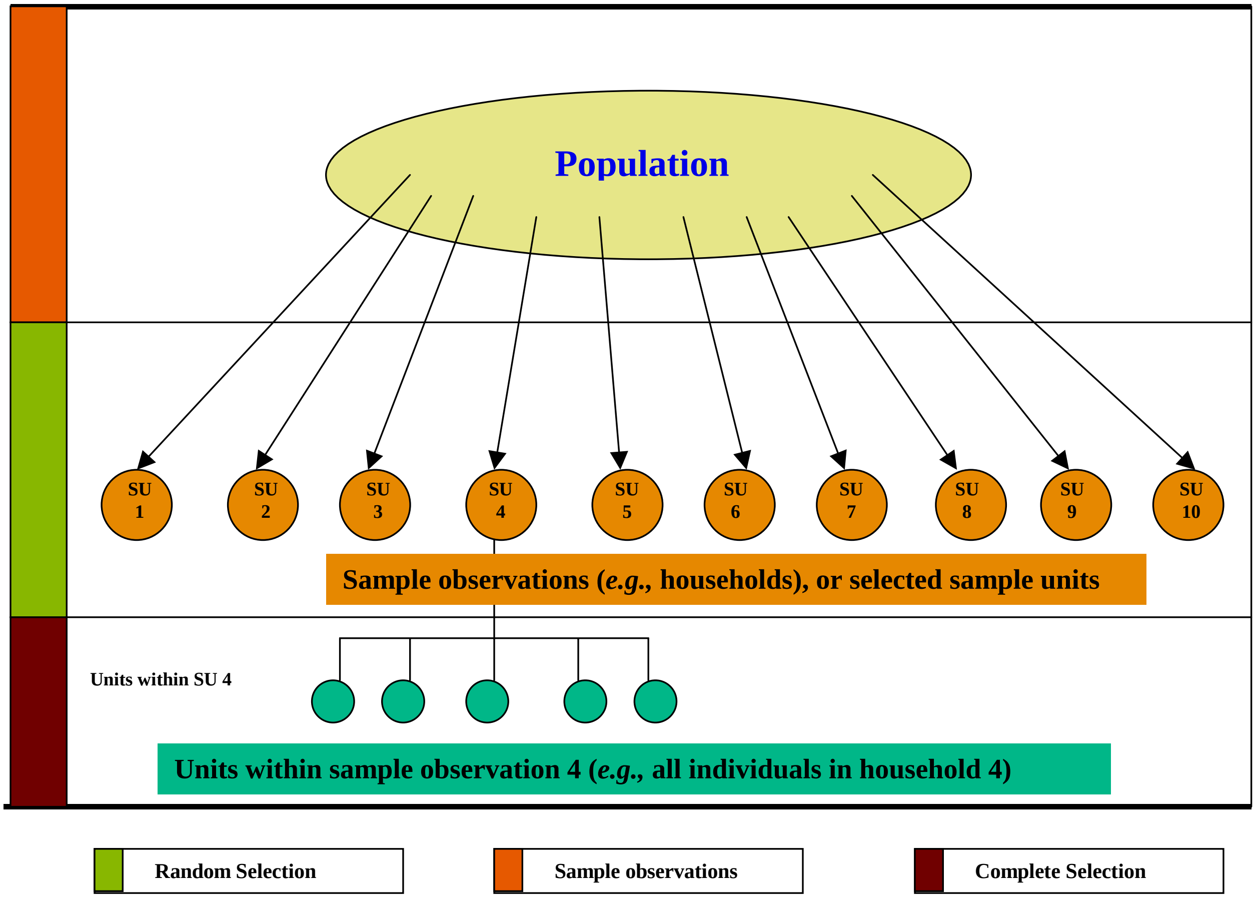
With version 4.2 and higher of DAD, the Sampling Design (SD) of the database can be specified in order to calculate the correct asymptotic sampling distribution of the various indices and statistics provided by DAD.

Data from sample surveys usually display four important characteristics:

1- they come with sampling weights (SW), also called inverse probability weights;
2- they are stratified;
3- they are clustered;
4- sample observations provide aggregate information (such as household expenditures) on a number of “statistical units” (such as individuals)

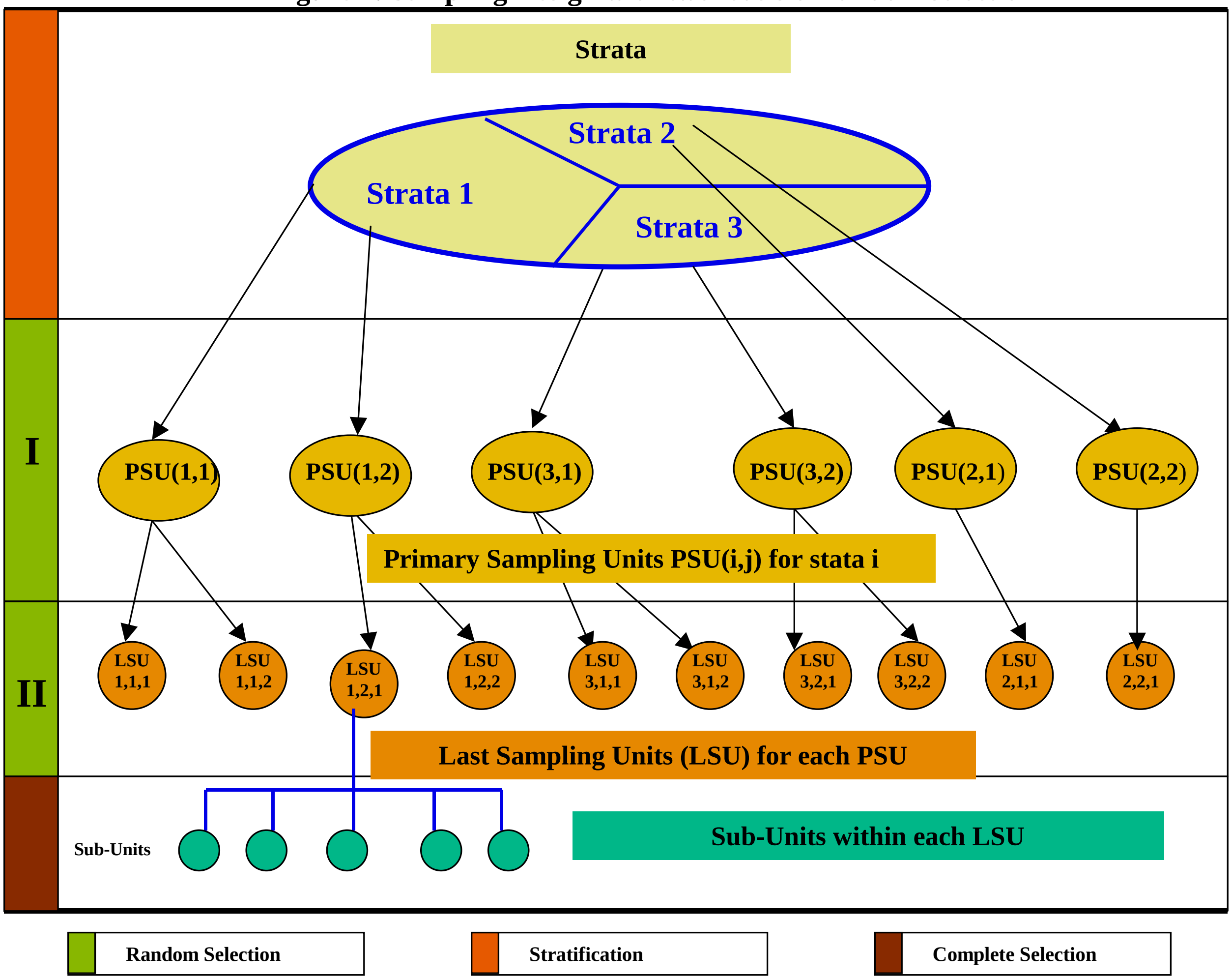
Figure 1 shows a graphical SD representation for the case of Simple Random Sampling (SRS), in which it is supposed that sample observations are directly and randomly selected from a base of sampling units (SUs) (*e.g.*, the list of all households within in a country).

Figure 1: Simple Random Sampling


Population
SU 1
SU 2
SU 3
SU 4
SU 5
SU 6
SU 7
SU 8
SU 9
SU 10
Sample observations (*e.g.*, households), or selected sample units
Units within SU 4
Units within sample observation 4 (*e.g.*, all individuals in household 4)
Random Selection
Sample observations
Complete Selection

SRS is rarely used to generate household surveys. Hence, most SD encountered in practice will not look like that in Figure 1. Most SD will look instead like that of Figure 2. A country is first divided into geographical or administrative zones and areas, called *strata*. Each zone or area thus represents a strata in Figure 2. The first random selection takes place within the Primary Sampling Units (denoted as PSU's) of each stratum. Within each stratum, a number of PSU's are randomly selected. This random selection of PSU's provides "clusters" of information. PSU's are often provinces, departments, villages, *etc.* Within each PSU, there may then be other levels of random selection. For instance, within each province, a number of villages may be randomly selected, and within every selected village, a number of households may be randomly selected. The final sample observations constitute the Last Sampling Units (LSU's). Each sample observation may then provide aggregate information (such as household expenditures) on all individuals or agents found within that LSU. These individuals or agents are *not* selected – information on all on them appears in the sample. They therefore do not represent the LSUs in statistical terminology.

Figure 2: Sampling Design with two levels of random selection



IMPACT OF SD ON THE SAMPLING ERROR OF DAD'S ESTIMATORS

a) Impact of stratification

Generally speaking, a variable of interest, such as household income, tends to be less variable within strata than across the entire population. This is because households within the same stratum typically share to a greater extent than in the entire population some socio-economic characteristics, such as geographical locations, climatic conditions, and demographic characteristics, and that these characteristics are determinants of the living standards of these households. Stratification ensures that a certain number of observations are selected from each of a certain number of strata. Hence, it helps generate sample information from a diversity of "socio-economic areas". Because information from a "broader" spectrum of the population leads on average to more precise estimates, stratification generally decreases the sampling variance of estimators. For instance, suppose at the extreme that household income is the same for all households in a stratum, and this, for all strata. In this case, supposing also that the population size of each stratum is known, it is sufficient to draw one household from each stratum to know exactly the distribution of income in the population.

b) Impact of clustering (or multi-stage sampling)

Multi-stage sampling implies observations end up in a sample only subsequently to a process of multiple selection. "Groups" of observations are first randomly selected within a population (which may be stratified); this is followed by further sampling within the selected groups, which may be followed by yet another process of random selection within the subgroups selected in the previous stage. The first selection stage takes place at the level of PSU's, and generates what are often called "clusters". Generally, variables of interest (such as living standards) vary less within a cluster than between clusters. Hence, multi-stage selection reduces the "diversity" of information generated by sampling. The impact of clustering sample observations is therefore to tend to decrease the precision of populations estimators, and thus to increase their sampling variance. *Ceteris paribus*, the lower the variability of a variable of interest within clusters, the larger the loss of information that there is in sampling further within the same clusters. To see this, suppose for instance an extreme case in which household income happens to be the same for all households in a cluster, and this, for all clusters. In such cases, it is clearly wasteful to adopt multi-stage sampling: it would be sufficient to draw one household from each cluster in order to know the distribution of income within that cluster. It would be more informative to draw randomly other clusters.

SAMPLING DESIGN IN DAD

By default, when a data file is loaded in DAD, the type of SD assigned to the data is the SRS presented in Figure 1. Once the data are loaded, the exact SD structure can nevertheless be easily specified. Up to 5 vectors can help specify that structure:

Table 1: Description of vectors used in DAD to specify the SD

Vectors	Description
Strata	Specifies the name of the variable (integer type) that contains stratum identifiers
PSU	Specifies the name of the variable (integer type) that contains identifiers for the Primary Sampling Units
LSU	Specifies the name of the variable (integer type) that contains identifiers for the Last Sampling Units
SW	Specifies the name of the variable for the Sampling Weights. Sampling weights are the inverse of the sampling rate. Roughly speaking, they equal the number of observations in the underlying population that are represented by each sample observation.
FPC	Specifies the name of the variable for the Finite Population Correction factor.
	With FPC, DAD derives an indicator f_h for each observation h, which is then used to compute SD-corrected sampling errors. • If the variable FCP is not specified, f_h=0 for all observations; • When the variable specified has values <= 1, it is directly interpreted as a stratum sampling rate f_h =n_h/N_h, where n_h = number of PSUs sampled from the strata to which h belongs and N_h = total number of PSUs in the population belonging to stratum h. • When the variable specified has values greater than or equal to n_h, it is interpreted as representing N_h; f_h is then set to n_h/N_h.

The following table contains an example of vectors used to specify the type of SD shown in Figure 2.

Table 2: Example of SD.

OBS	Strata	PSU	LSU	SW
1	1	1	1	6
2	1	1	2	6
3	1	2	1	6
4	1	2	2	6
5	3	1	1	5
6	3	1	2	5
7	3	2	1	5
8	3	2	2	5
9	2	1	1	3
10	2	2	1	3
SUM	**3**	**6**	**10**	**50**

Omitting SW will systematically bias both the estimators of the values of indices and points on curves as well as the estimation of the sampling variance of those estimators. Consider for instance the estimation of total population income from the data shown in table 2. 4 households appear in strata 1, but the population number of households in that strata is six times as large (that is, 24), and this is captured by the SW variable. Total population income for strata 1 would therefore be estimated to be six times that of total sample income for strata 1.

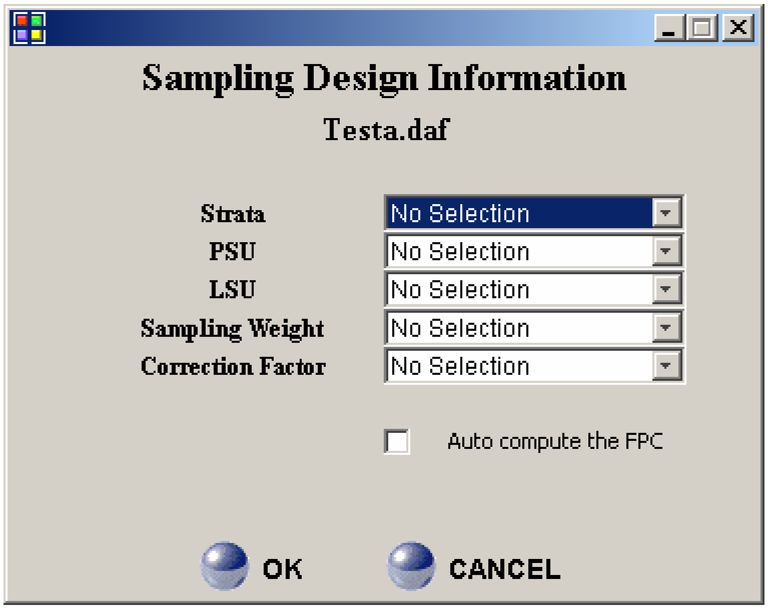
Table 3: Example of SD.

OBS	Strata	LSU	SW	N_h
1	1	1	6	24
2	1	2	6	24
3	1	3	6	24
4	1	4	6	24
5	3	1	5	20
6	3	2	5	20
7	3	3	5	20
8	3	4	5	20
9	2	1	3	6
10	2	2	3	6
SUM	**3**	**10**	**50**	**---**

The FPC factor accounts for the reduction in sampling variance that occurs when a sample is drawn *without* replacement from a finite population (as compared to sampling *with* replacement). According to table 3, the four LSU's of strata 1 were selected without replacement from a population of 24 LSU's. These fuor LSU's are then necessarily distinct by design. If sampling had been done with replacement, then multiple observations of the same population LSU's could have been generated. Because sampling without replacement guarantees that sample observations represent different sampling units, it therefore generates greater sampling information and leads to smaller sampling variances than with sampling with replacement. For strata 1 of Table 3, data from four distinct LSU's (or PSU's) out of 24 are necessarily generated after sampling. The f_h factor for that strata is then 4/24=0.1666.

REMARK: We can initialise and use the FPC correction just when the SD is based on one stage of random selection of LSU's. In this case PSU's and LSU's are equivalent.

To initialize the SD after loading the database, select from the main menu the item "Edit->Set Sample Design". The following window then appears.



Sampling Design Information

Testa.daf

Strata	No Selection
PSU	No Selection
LSU	No Selection
Sampling Weight	No Selection
Correction Factor	No Selection

☐ Auto compute the FPC

This allows DAD to take into account a wide variety of possible SD. This is made by selecting (or not selecting) vectors for any of the five choices offered above. In the case of SRS within a number of strata, there would be an indicator of a strata vector without any indication of a vector of PSU's. The following table presents some of these combinations.

Strata	PSU	LSU	SW	FPC	Indication
					SD is SRS without sampling weights
X			X		SD is stratified with SW
	X	X	X		No stratification, but multi-stage sampling and SW
		X	X		Random (one-stage) sampling of LSU's with LSU-specific selection probabilities. This can occur for instance if, once an individual is selected, all individuals in his household are also automatically selected. Implicitly, then, it is the household that is selected as a LSU
X	X		X		Stratification with only the first sampling stage specified by the user
X		X			Stratification with one-stage sampling and sampling weights (wrongly?) omitted
X		X	X		Stratification with one-stage sampling and sampling weights (wrongly?) omitted
X	X	X			Stratification with multi-stage sampling and sampling weights (wrongly?) omitted
X	X	X	X		Stratification with multi-stage sampling and sampling weights provided
X	X	X	X	X	Stratification with multi-stage sampling and sampling weights provided. The finite population correction factor is also provided; this supposes that sampling for the statistical inferences

X: Indicate that the variable is selected

Note that when DAD finds the values of the strata-psu-lsu variables to be the same across observations, it supposes that these observations comefrom just one LSU.

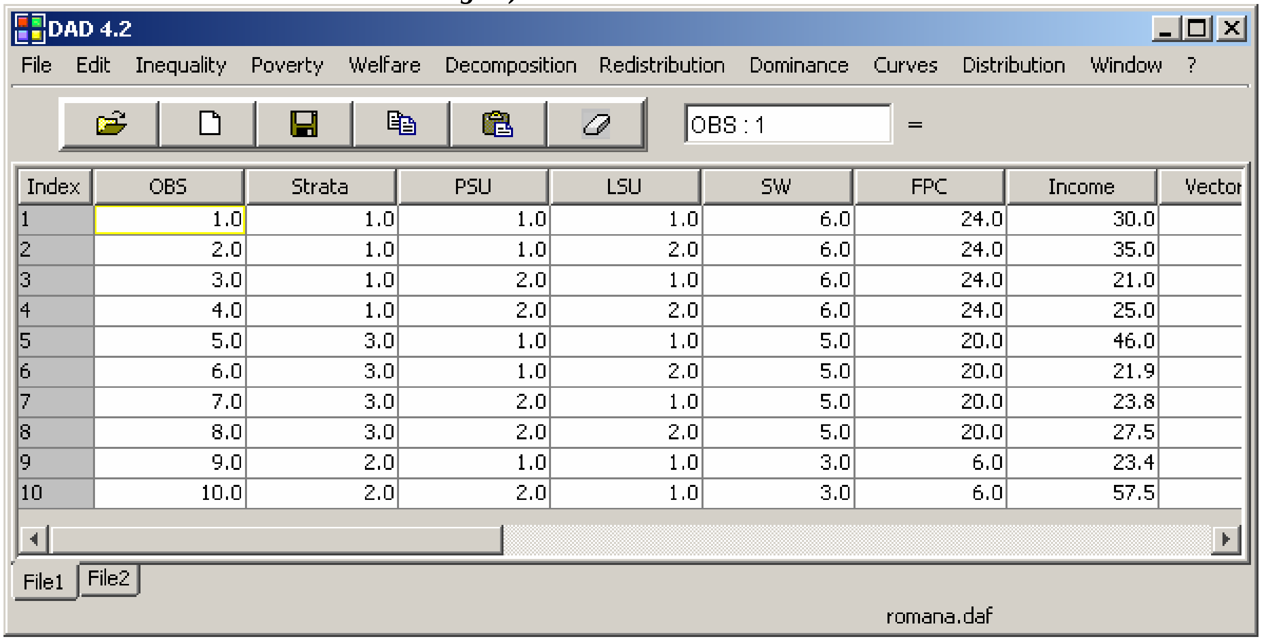
If the option “Auto-compute FPC” is activated, DAD generates implicitly the FPC vector.

REMARKS:

- After initialization of the SD information, the dataset is automatically ordered by (when specified) strata, PSU’s and LSU’s.
- There should be more than one PSU within each stratum.

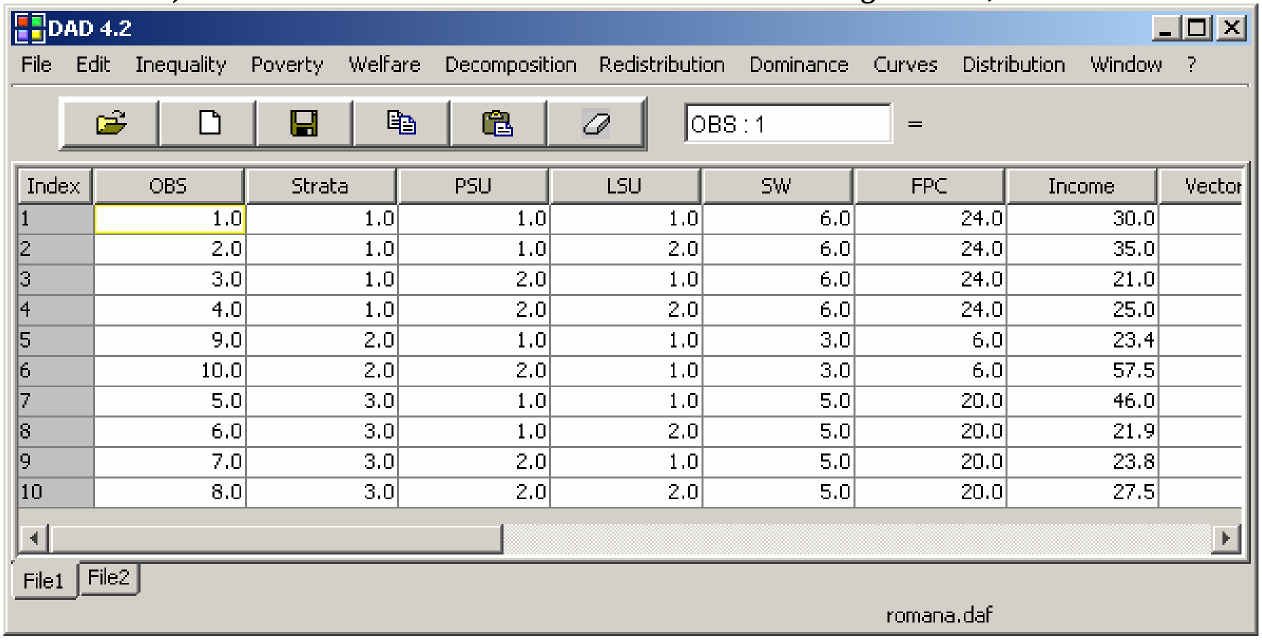
.

e.g.:**1) before initialization of the SD**



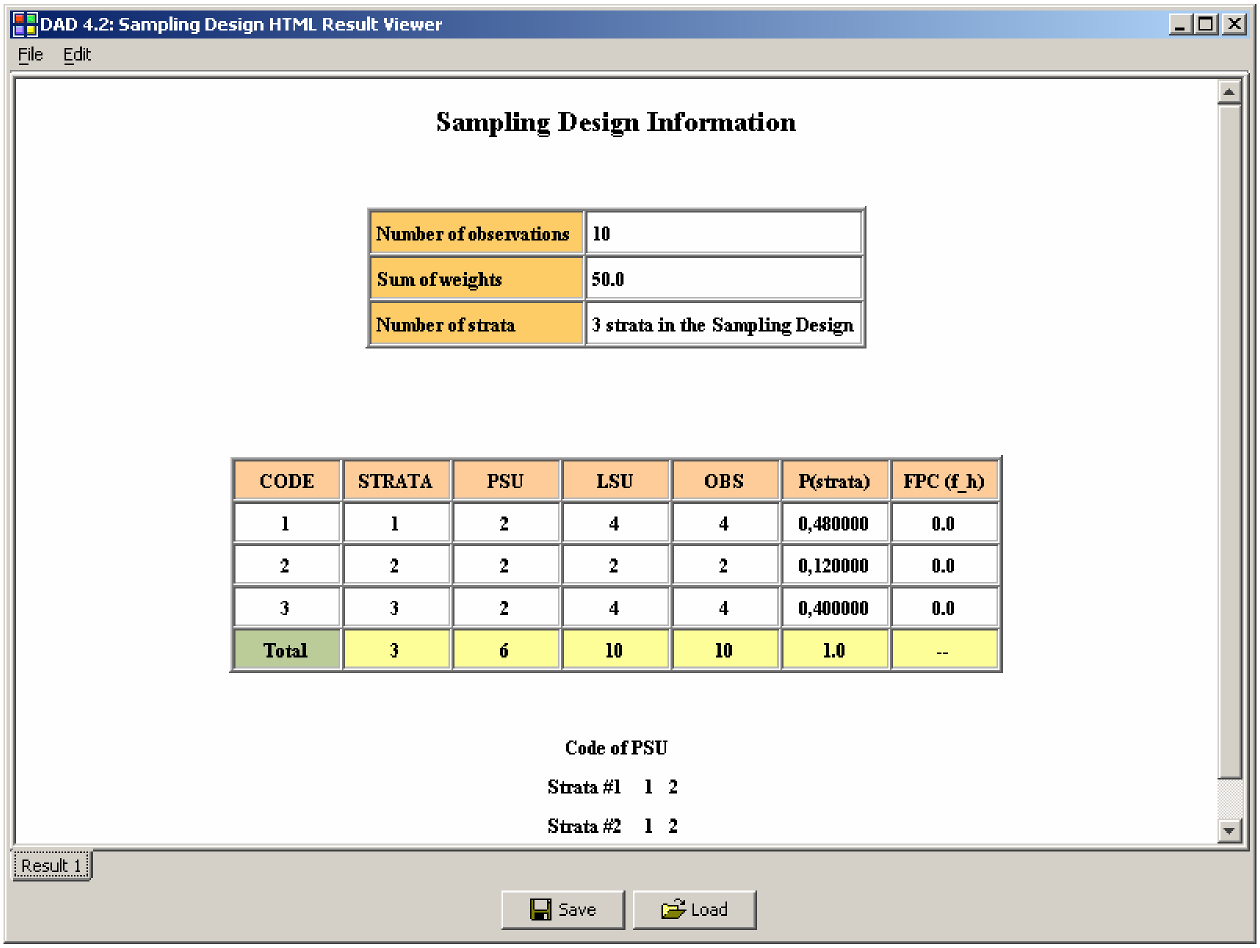
Index	OBS	Strata	PSU	LSU	SW	FPC	Income	Vector
1	1.0	1.0	1.0	1.0	6.0	24.0	30.0	
2	2.0	1.0	1.0	2.0	6.0	24.0	35.0	
3	3.0	1.0	2.0	1.0	6.0	24.0	21.0	
4	4.0	1.0	2.0	2.0	6.0	24.0	25.0	
5	5.0	3.0	1.0	1.0	5.0	20.0	46.0	
6	6.0	3.0	1.0	2.0	5.0	20.0	21.9	
7	7.0	3.0	2.0	1.0	5.0	20.0	23.8	
8	8.0	3.0	2.0	2.0	5.0	20.0	27.5	
9	9.0	2.0	1.0	1.0	3.0	6.0	23.4	
10	10.0	2.0	2.0	1.0	3.0	6.0	57.5	

2) after initialization of the SD: data is ordered according to strata, PSU and LSU



Index	OBS	Strata	PSU	LSU	SW	FPC	Income	Vector
1	1.0	1.0	1.0	1.0	6.0	24.0	30.0	
2	2.0	1.0	1.0	2.0	6.0	24.0	35.0	
3	3.0	1.0	2.0	1.0	6.0	24.0	21.0	
4	4.0	1.0	2.0	2.0	6.0	24.0	25.0	
5	9.0	2.0	1.0	1.0	3.0	6.0	23.4	
6	10.0	2.0	2.0	1.0	3.0	6.0	57.5	
7	5.0	3.0	1.0	1.0	5.0	20.0	46.0	
8	6.0	3.0	1.0	2.0	5.0	20.0	21.9	
9	7.0	3.0	2.0	1.0	5.0	20.0	23.8	
10	8.0	3.0	2.0	2.0	5.0	20.0	27.5	

To show the SD information, select from main menu the item “Edit->Summarize Sample Design”. The following window appears.



Sampling Design Information

Number of observations	10
Sum of weights	50.0
Number of strata	3 strata in the Sampling Design

CODE	STRATA	PSU	LSU	OBS	P(strata)	FPC (f_h)
1	1	2	4	4	0,480000	0.0
2	2	2	2	2	0,120000	0.0
3	3	2	4	4	0,400000	0.0
Total	3	6	10	10	1.0	--

Code of PSU

Strata #1	1	2
Strata #2	1	2