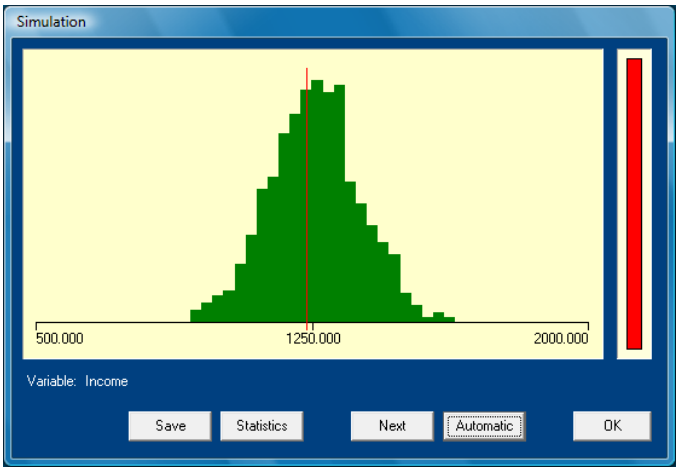
SimSam

Simulating samples from a finite population



Jelke Bethlehem

February, 2009

1. SimSam

1.1. About SimSam

SimSam is a program to simulate samples from a finite population. By repeating the selection of the sample a large number of times, and computing an estimate for the population mean for each sample, the distribution of the estimate can be portrayed on the screen. By taking a look at this distribution, it becomes clear how well an estimation procedure is able to estimate the population mean.

1.2. An introduction to SimSam

SimSam is a program to simulate samples from finite populations. The program assumes the objective of the sample selection to be estimation of the population mean of a variable.

By repeating the selection of the sample a large number of times, and computing an estimate for each sample, the distribution of the estimate can be portrayed on the screen. By taking a look at this distribution, it becomes clear how well an estimation procedure is able to estimate the population mean.

Two types of sample designs can be simulated:

- Simple random samples without replacement, and with equal probabilities.
- Samples with replacement, and with probabilities proportional to the values of some auxiliary variable.

Four different estimators can be computed:

- The simple sample mean (the direct estimator).
- The ratio estimator.
- The regression estimator
- The post-stratification estimator

To study the effects of non-response on an estimator, it is possible to generate non-response in the samples.

The result of a simulation is displayed on the screen in the form of a histogram of the distribution of the estimator. This histogram can be saved in a bitmap file for use in other applications. It is also possible to produce a numerical summary with theoretical and simulated means and standard errors.

Figure 1.2.1 shows an example of a histogram of the distribution of an estimator. The red vertical line indicates the population parameter to be estimated.

Figure 1.2.1. The distribution of an estimator

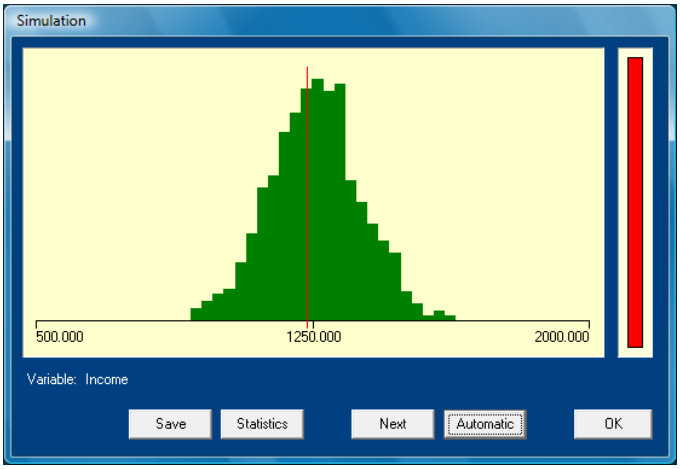
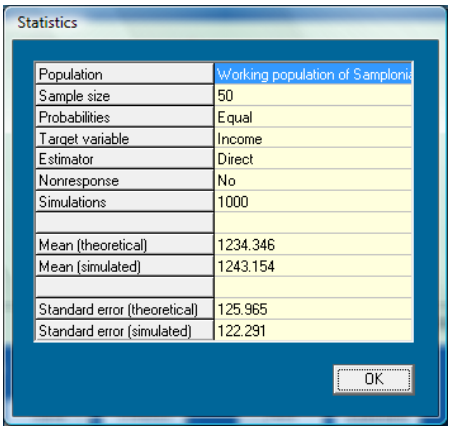


Figure 1.2.2 shows an example of a numerical summary of a simulation.

Figure 1.2.2. An example of a numerical summary



1.3. Carrying out a simulation, step by step

To carry out a simulation with SimSam, you have to go through a number of steps:

Step 1: Read a population file. You do that with the option *Read population* in the *Population menu*.

Step 2: Specify a sampling design. You do that with the options in the *Sample menu*. You can set the target variable, the sample size, the sampling design and the estimator.

Step 3: If required, you can specify the parameters for generating non-response. You do that with the options in the *Nonresponse menu*.

Step 4: Set the simulation parameters. You do this with the option *Parameters* in the *Simulation menu*.

Step 5: Start the simulations. You do this with the option *Begin* in the *Simulation menu*.

2. Overview of the SimSam menus

SimSam is a menu-driven program. This section describes the options in the various menus.

2.1. The Population menu

The *Population menu* contains three options: *Read population*, *Display variable* and *Stop*.

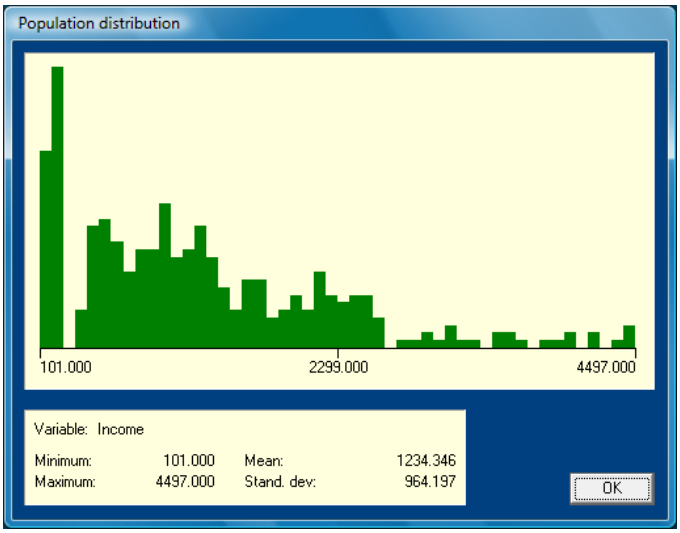
Read population

You use the option *Read population* to read an existing population file. If you activate this option you will see a list of available population files. Population files always have the extension *pop*. Select a population from this list. You can also activate this option with **<Ctrl-R>**.

Display variable

You use the option *Display variable* to take a look at the distribution of a variable in the population file. If you activate this option you will see a list of available variables. Select a variable from this list. You can also activate this option with **<Ctrl-D>**. Figure 2.1.1 shows an example of the type of output SimSam will produce.

Figure 2.1.1. The distribution of a variable



Display the population distribution of a variable in the population. Select a variable from the list of available variables.

Stop

You use the option Stop to stop execution of the program. You can achieve the same effect with **<Alt-X>**

2.2. The Sample menu

You use the *Sample menu* to select the target variable for which you want to estimate the population. Furthermore, you can select the sampling design and the type of estimator to be used. The menu has three options: *Target variable*, *Sampling design* and *Estimator*.

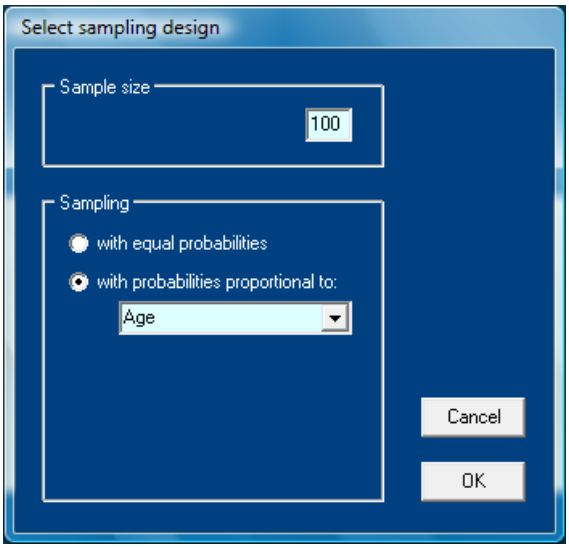
Target variable

You use the option *Target variable* to select a target variable. This is the variable for which you want to estimate the population mean. If you activate this option, a list with available variables will appear. Select the variable of your choice from this list. You can also activate this option with **<Ctrl-T>**.

Sampling design

You use the option *Sampling design* to specify the sampling design. If you activate this option, you will see a window like the one in figure 2.2.1. You can also activate this option with **<Ctrl-S>**.

Figure 2.2.1. The sampling design



You enter the sample size in the sample size field. Note that by default the sample size is set to 10.

Next, you indicate whether you want to select samples with equal or unequal probabilities. You do that with the radio buttons. If you select *with equal probabilities*, simple random samples without replacement will be selected. If you select with *probabilities proportional to*, you also have to select an auxiliary variable from the list. Elements will be selected with replacement and with probabilities proportional to the values of this auxiliary variable. Note that an auxiliary variable can only be used for this purpose of all its values are strictly positive.

Estimator

You use the option *Estimator* to select and estimator. If you activate this option, you will see a window in which you can choose from four estimators: the *direct estimator*, the *ratio estimator*, the *regression estimator* and the *post-stratification estimator*.

If you select your samples with equal probabilities, all four estimators can be used. However, if you select sampling with unequal probabilities, you can only do that in combination with the direct estimator.

The ratio and regression estimator use a continuous auxiliary variable. You have to select such a variable from the list of available variables.

The post-stratification estimator uses a categorical auxiliary variable. You have to select such a variable from the list of available variables. Note that only variables can be used that only assume at most 10 different values, and these values have to be numbered from 1 to 10. These numbers indicate the strata.

2.3. The Nonresponse menu

You can use the *Nonresponse menu* to generate nonresponse in the selected samples. The menu has two options: *Generate nonresponse* and *Full response*.

Generate nonresponse

Use the option *Generate nonresponse* to generate nonresponse. You can also activate this option with **<Ctrl-G>.** You must first select a nonresponse variable. The probability of nonresponse will have a linear relationship with this variable. You choose the nonresponse variable by selecting it in the list of available variables.

Furthermore, you must specify the percentages of nonresponse corresponding to the highest and the lowest values of this variable. For all values between these two bounds, the probability of nonresponse is obtained by means of interpolation. Note that the entered percentages must be at least 1% and at most 99%.

Figure 2.3.1. Generating nonresponse

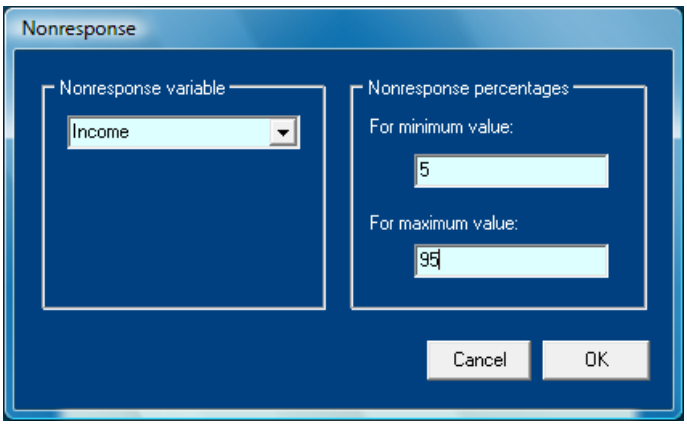


Figure 2.3.1 contains an example. The nonresponse probability depends on the variable income. The probability of nonresponse is 5% for the lowest income, and 95% for the highest income. So, persons with a high income will be under-represented in the sample.

Full response You use the option *Full response* to switch off the generation of nonresponse in the selected samples. You can also activate this option with **<Ctrl-F>**.

2.4. The Simulation menu

You can use the *Simulation menu* to set the options for the simulation process, and also to start the simulation process. The menu contains two options: *Parameters* and *Begin*.

Parameters You use the option *Parameters* to specify the settings for the simulation process. You can also activate this option with **<Ctrl-P>**.

Figure 2.4.1. Setting the simulation parameters

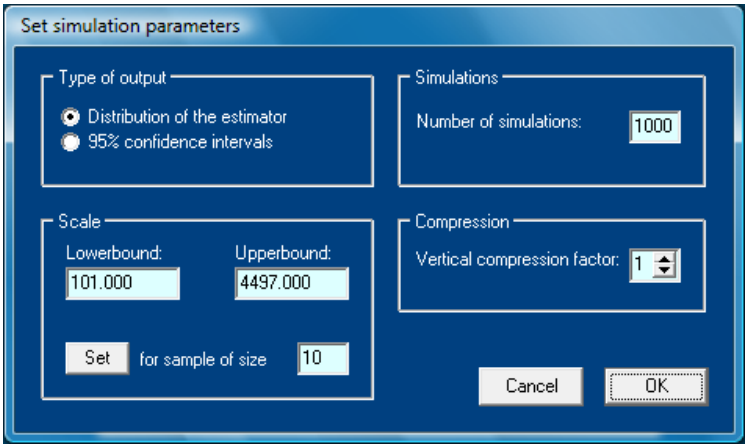


Figure 2.4.1 shows the dialog box in which you can set the parameters. The first parameter that can be set is the *type of output*. You can either produce a histogram of the sampling distribution of the estimator, or you can obtain a graph containing a visual representation of the 95%-confidence interval for each sample.

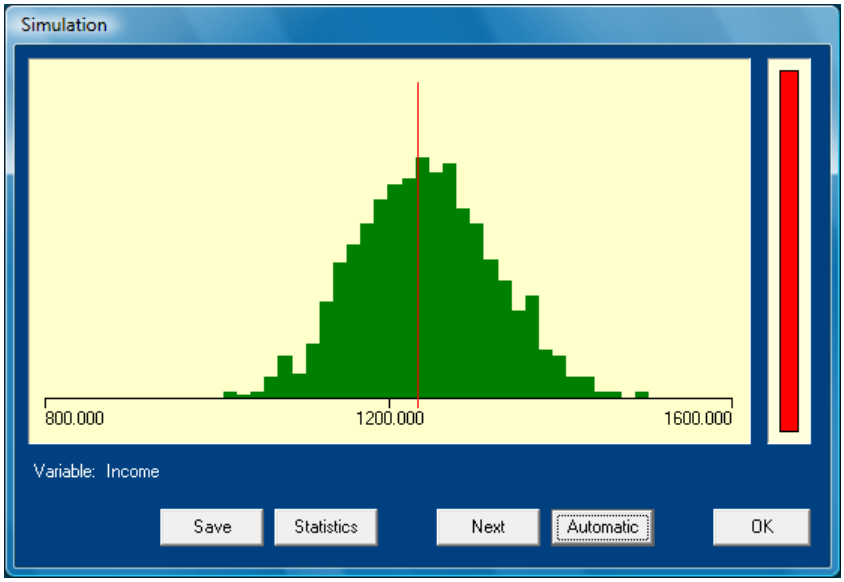
The second parameter that can be set is the *number of simulations*. The default value is 1000, but you can set any number between 2 and 1000.

The third parameter controls the horizontal *scaling* in the histogram of the distribution of the estimator. By setting a sample size, the lower bound of the x-axis is taken to be the smallest value the estimator can assume for that sample size. Likewise, the upper bound is taken to be equal to maximum value of the estimator for this sample size. By increasing the value of this parameter, the lower bound is increased and the lower bound decreased. It is also possible to directly enter a lower bound and upper bound for the x-axis.

The fourth parameter controls the amount of vertical *compression* of the histogram. You can select a value from 1 to 4. A value of 1 means no compression. A value of 4 will compress the histogram by a factor 4.

Begin After having set the simulation options, you start the simulation process by activating the option *Begin*. You also press **<Ctrl-B>**. A window like the one in figure 2.4.2 will appear on the screen.

Figure 2.4.2. Simulation



The vertical red line indicates the population mean of the target variable. This is the population parameter you attempt to estimate.

The simulation window contains a number of buttons:

- By clicking on *Next*, the program is instructed to generate the next sample of the simulation. This is the step-wise way of carrying out the simulation.
- By clicking on *Automatic*, you switch to continuous simulation. All samples are processed one after the other without the possibility of interruption.
- After the simulation has been completed, you press *Statistics* to get a numerical overview of the simulation results.
- You can save the graph in a bitmap file. You do that by clicking on the *Save* button.
- You close the simulation window by clicking on *OK*.

2.5. The Help menu

The Help menu is there to help you running the SimSam program. The menu has two options: *SimSam help* and *About this program*.

SimSam help

The option *SimSam help* gives you access to the help system of SimSam. You can also activate this option with **<Ctrl-H>.**

About this program

This option *About this program* produces some information about this program, such as its version number and the author. You can also activate this option with **<Ctrl-A.**

3. Population files

The program SimSam is distributed with a number of population files. These files contain data on the country of Samplonia:

- *Samplon.pop* contains the complete population of Samplonia. There are 1000 records and 6 variables. The variables are town of residence (there are 6 towns), province (there are 2 provinces), sex (1=male, 2=female), age (a value between 0 and 100), employment status (1=employed, 2=unemployed), and income (a value between 0 and 5000).

- *Working.pop* contains the working population of Samplonia. It is a subset of the file *Samplon.pop*. There are 341 records. The 6 variables are identical to those of the file Samplon.pop. Note that the variable employment status always has the value 1 (employed).

- *Milk.pop* contains the daily milk production (in litres) by cows on 200 farms in Samplonia. There are 3 variables: milk production (litres/day), size of the farm (hectares) and number of cows on the farm.

You can make your own population file. The maximum number of records is 1000. To explain the structure of the file, the first part of the file *Samplon.pop* is shown in figure 3.1.

Figure 3.1. Structure of a population file

```
Population of Samplonia
1000 6
Town
Province
Sex
Age
Employed
Income
5 2 1 65 2 0
6 2 1 36 2 0
7 2 2 73 2 0
6 2 1 6 2 0
3 1 2 33 1 158
. . .
```

The first line must contain a short description of the population. The second line contains the size of the population (number of records) followed by the number of variables. The two numbers must be separated by one or more blanks.

The next lines contain the names of the variables. Each variable name must start on a new line. The data records start after the last variable name. Each record must start on a new line. Each line must contains values for all variables in the order specified.. Numbers must be separated by one or more blanks.