Distribution

DESCRIPTIVE STATISTICS

This application provides basic descriptive statistics on variables in the database: the mean, the standard deviation, and the minimum and the maximum values of each of the vectors.

To reach this application:

1- From the main menu, choose: "Distribution ⇒ Statistics".
2- Choose the data bases if you have activated two databases.
3- Choose the weight variable if the observations must be weighted.
4- Choose the group variable and the group number if you would like to compute the statistics for a specific group.

The results are as follows:

Name of variable 1	Mean	Standard deviation	Minimum	Maximum
Name of variable 2	Mean	Standard deviation	Minimum	Maximum
:	:	:	:	:

STATISTICS

This application computes basic descriptive statistics for a given variable of interest, as well as the ratio of two such variables. The application also computes the effect of the sampling design on the sampling error of these basic statistics.

1- $\text{Total}=\sum_i sw_i x_i$

2- $\text{Mean} = \dfrac{\sum_i sw_i x_i}{\sum_i sw_i}$

3- $\text{Ratio} = \dfrac{\sum_i sw_i x_i}{\sum_i sw_i y_i}$

To activate this application for one distribution, follow these steps:

- From the main menu, choose: "Distribution ⇒ Statistics".
- Choose the different vectors and parameter values.

Parameters

k:	Group Number

Vectors

x	Size Variable 1
s(x)	Variable of interest 2
y	Size Variable 2
s(y)	Group Variable
c	Group Number

CONFIDENCE INTERVAL

When results of a given application only provide the standard deviation of an estimator of interest, one can use this *Confidence Interval* application to calculate a confidence interval or to perform statistical tests under asymptotic approximations. For more information, see the sub-section Standard deviation, confidence intervals and hypothesis testing ➔ Asymptotic approach.

DENSITY FUNCTION

The gaussian kernel estimator of a density function $f(x)$ is defined as:

$$\hat{f}(x) = \frac{\sum_i w_i K_i(x)}{\sum_{i=1}^{n} w_i} \quad \text{and} \quad K_i(x) = \frac{1}{h\sqrt{2\pi}} \exp\left(-0.5\,\lambda_i(x)^2\right) \quad \text{and} \quad \lambda_i(x) = \frac{x - x_i}{h}$$

where h is a bandwidth which acts as a "smoothing" parameter.

To reach this application:

- From the main menu, choose the item: "Distribution ⇒ Density function".
- Choose the different vectors and parameter values.

Parameters

y	Parameter
h	Smoothing parameter

Among the buttons, you find the following commands:

COMPUTE: to compute $f(x)$.

GRAPH: to draw the value of the function as a function of x.

CORRECTED BOUNDARY KERNEL ESTIMATORS

A problem occurs with kernel estimation when a variable of interest is bounded. It may be for instance that consumption is bounded between two bounds, a minimum and a maximum, and that we wish to estimate its density "close" to these two bounds. If the true value of the density at these two bounds is positive, usual kernel estimation of the density close to these two bounds will be biased. A similar problem occurs with non-parametric regressions. One way to alleviate these problems is to use a smooth "corrected" Kernel estimator, following a paper by Peter Bearse, Jose Canals and Paul Rilstone. A boundary-corrected Kernel density estimator can then be written as

$$\hat{f}(x) = \frac{\sum_i sw_i K_i^*(x) K_i(x)}{\sum_{i=1}^{n} sw_i}$$

where

$$K_i(x) = \frac{1}{h\sqrt{2\pi}} \exp\left(-0.5\,\lambda_i(x)^2\right) \quad \text{and} \quad \lambda_i(x) = \frac{x - x_i}{h}$$

and where the scalar $K_i^*(x)$ is defined as

$$K_i^*(x) = \psi(x)'\, P(\lambda_i(x))$$

$$P(\lambda) = \left(1 \quad \lambda \quad \frac{\lambda^2}{2!} \quad \cdots \quad \frac{\lambda^{s-1}}{(s-1)!}\right)$$

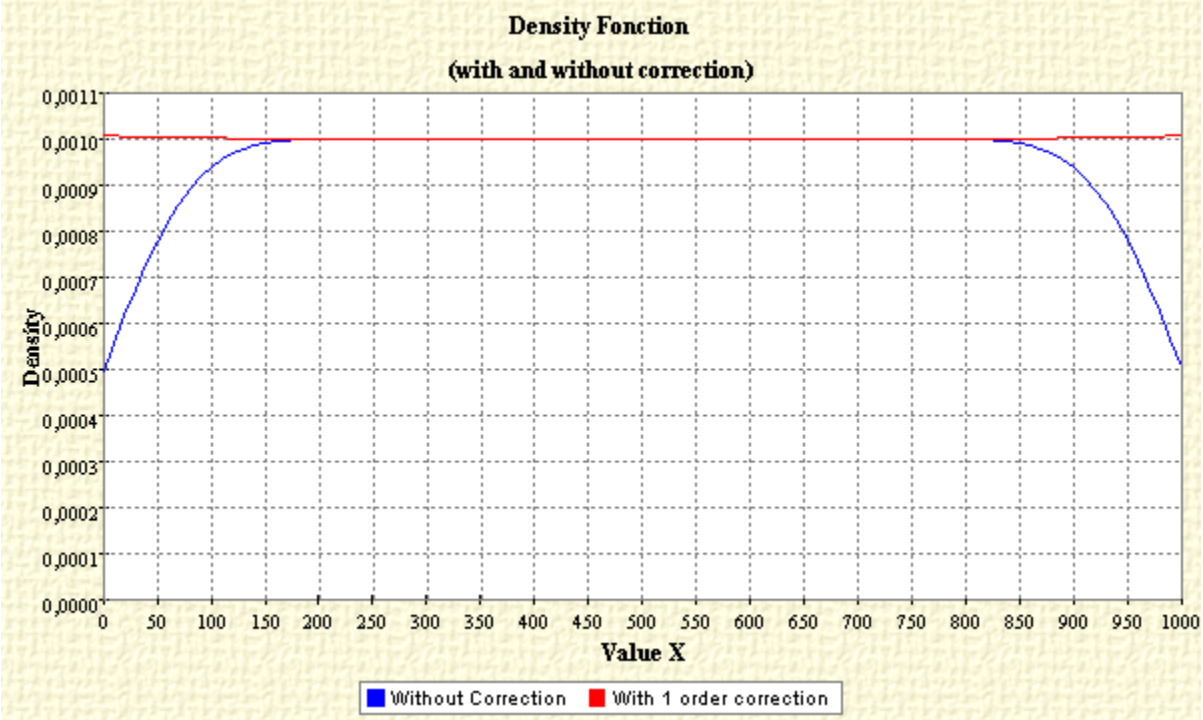
$$\psi(x) = M^{-1} l_s' = \left(\int_A^B K(\lambda)P(\lambda)P(\lambda)'d\lambda\right)^{-1} l_s \quad : A = \frac{x - max}{h}, \quad B = \frac{x - min}{h}, \quad l_s' = (1 \quad 0 \quad 0 \cdots 0)$$

min is the minimum bound, and *max* is the maximum one. *h* is the usual bandwidth. This correction removes bias to order h^s.

DAD offers four options, without correction, and with correction of order 1, 2 and 3.

Example 1:

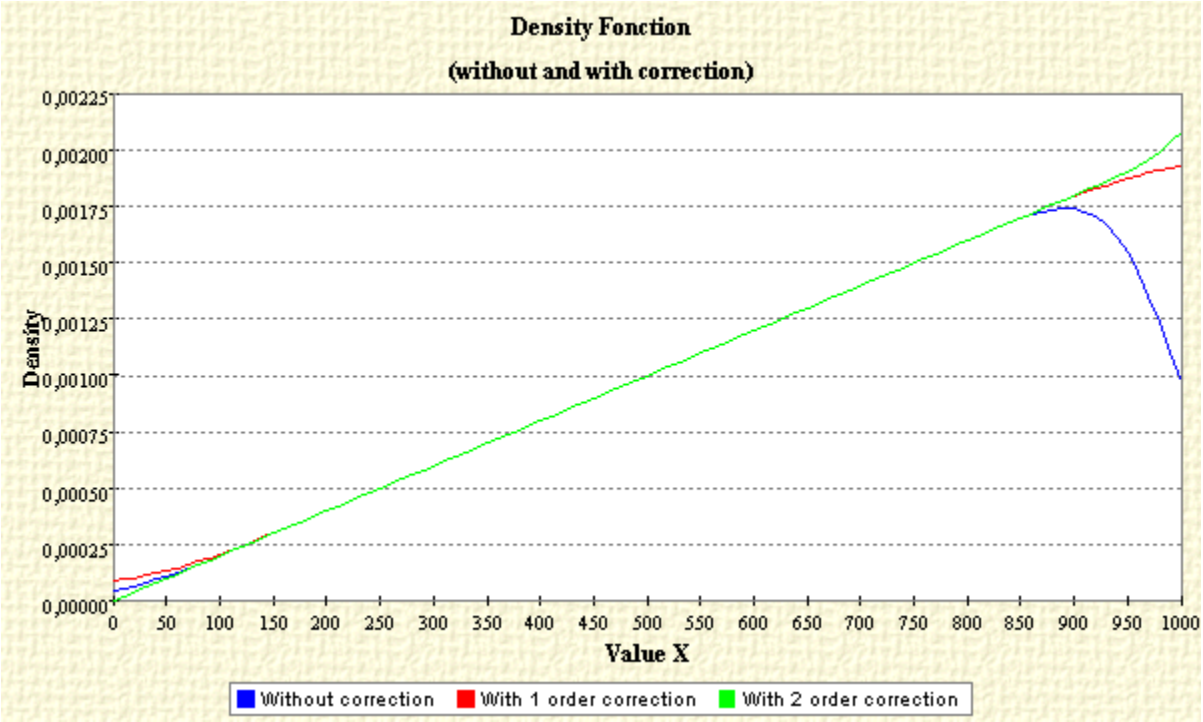
Suppose that an observed vector of interest y takes the form : y={1,2,3,…i+1….999,1000} because it is drawn from a uniform distribution. The density at any income between 0 and 1000 is the same and equals 1/1000. The following figure shows the impact of the above correction on the density estimation:



This shows that a correction of order 1 corrects well the boundary problem of estimating the density close to 0 and 1000.

Example 2:

Suppose that an observed vector of interest y takes the form : y={1,2,2,3,3,3,…,….1000,1000}. The total number of observations sums to N=1000*(1+1000)/2=50500. The population density equals f(x)=x/500.The following figure shows the impact of a correction of order 1 and 2 on the density estimation:



THE JOINT DENSITY FUNCTION

The gaussian kernel estimator of the joint density function $f(x,y)$ is defined as:

$$\hat{f}(x,y) = \frac{1}{\sum_{i=1}^{n} sw_i h^2} \sum_{i=1}^{n} sw_i \frac{1}{2.\pi} \exp\left(-\left(\frac{1}{2}\right)\left(\left(\frac{x-x_i}{h}\right)^2 + \left(\frac{y-y_i}{h}\right)^2\right)\right)$$

To reach this application:

- From the main menu, choose the item: "Distribution ⇒ Joint density function".
- Choose the different vectors and parameter values.

Parameters	
y	Parameter

x	Parameter
h	Smoothing parameter

Among the buttons, you find the following commands:

COMPUTE: to compute the estimate of the joint density function $f(x, y)$.

GRAPH: to generate an ASCII file that can be used with **GnuPlot 4.0** to plot the graph.

THE DISTRIBUTION FUNCTION

To reach this application:

- From the main menu, choose the item: "Distribution ⇒ Distribution function".
- Choose the different vectors and parameter values.

Parameters

y Parameter

Among the buttons, you find the following commands:

COMPUTE:	to compute the estimate of the distribution function $F(y)$.
GRAPH:	to draw the value of the function as a function of y.

PLOT_SCATT_XY

- This application plots a scatter graph of two variables. To activate this application, choose from the main menu the item: "Distribution ⇒ Plot_Scatt_XY". When the window of this application appears, choose the two X and Y variables and click on the button "**Graph**". You can also use the command **"Range"** to specify the range of the horizontal axis (X).

NON-PARAMETRIC REGRESSION AND NON-PARAMETRIC DERIVATIVE REGRESSION

The Gaussian kernel regression of y on x is as follows:

$$\hat{\Phi}(y \mid x) = \frac{\hat{\alpha}(x)}{\hat{\beta}(x)} = \frac{\sum_i sw_i K_i(x) y_i}{\sum_i sw_i K_i(x)}$$

From this, the derivate of $\Phi(y \mid x)$ with respect to x is given by

$$\frac{\partial \Phi(y \mid x)}{\partial x} = \frac{\alpha(x)'}{\beta(x)} - \frac{\beta(x)'\alpha(x)}{\beta(x)^2}$$

REMARK: The instructions for non-parametric derivative regression are similar to those for non-parametric regression.

To reach this application:

- From the main menu, choose the item: "Distribution ⇒ Non-parametric regression".
- Choose the different vectors and parameter values.

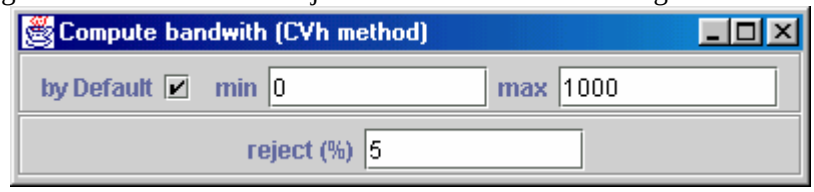
Parameters	
x	Level of (X) or (p)
h	Smoothing parameter
Vectors	
x_i	Exogenous Variable (X)
y_i	Endogenous Variable (Y)

REMARK:
- The option "Level" vs "Percentile" allows the estimation of the expected value of y either at a level of x or at a p-quantile for x.
- The option "Normalised" vs "Not normalized" by the mean or by x allows the estimation of the expected value of y normalized or not by x or by the overall mean of y.

You will find:

- The command **"Compute":** to compute $\Phi(y \mid x)$. To compute its standard deviation, choose the option for computing with standard deviation.

- The command **"Compute h":** to compute an optimal bandwidth according to the cross-validation method of Härdle (1990), p. 159-160. When you click on this command, the following window appears, giving you the option of choosing the min/max bands and the percentage of observations to be rejected on each side of the range of x.



- The command **"Graph":** to draw $\Phi(y \mid x)$ as a function of x. To specify a range for the horizontal axis, choose the item " Graph management ⇒ Change range of x " from the main menu.
- The command **"Range":** to specify the range of the horizontal axis.

BOUNDARY-CORRECTED NON-PARAMETRIC REGRESSION AND NON-PARAMETRIC DERIVATIVE REGRESSION

For the boundary-corrected non-parametric regression, the estimation is as follows:

$$\hat{\Phi}(y \mid x) = \frac{\sum_i sw_i K_i^*(x) K_i(x) y_i}{\sum_i sw_i K_i^*(x) K_i(x)}$$

The boundary-corrected non-parametric derivate regression is obtained by differentiating the above with respect to x:

$$\Phi'(y \mid x) = \frac{\sum_i sw_i \left(K_i^*(x)' K_i(x) y_i + K_i^*(x) K_i(x)' y_i \right)}{\sum_i sw_i K_i^*(x) K_i(x)} - \frac{\sum_i sw_i \left(K_i^*(x)' K_i(x) + K_i^*(x) K_i(x)' \right)}{\left(\sum_i sw_i K_i^*(x) K_i(x) \right)^2}$$

Note that:

$$K_i^*(x) = \psi(x)' \, P(\lambda_i(x)) \text{ and } P(\lambda) = \left(1 \quad \lambda \quad \frac{\lambda^2}{2!} \quad \cdots \quad \frac{\lambda^{s-1}}{(s-1)!} \right)$$

$$\psi(x) = M^{-1} l_s{}' = \left(\int_A^B K(\lambda) P(\lambda) P(\lambda)' d\lambda \right)^{-1} l_s \quad : A = \frac{x - \max}{h}, \quad B = \frac{x - \min}{h}, \quad l_s{}' = (1 \quad 0 \quad 0 \cdots 0)$$

$$K_i^*(x)' = \frac{\partial M^{-1}(x)}{\partial x} l_s{}' \, P(w) + \frac{\partial P(\lambda(x))}{\partial x} M^{-1}(x) l_s{}' \text{ where}$$

$$\frac{\partial M^{-1}(x)}{\partial x} = -M^{-1}(x) \left[\frac{\partial M(x)}{\partial x} \right] M^{-1}(x)$$

CONDITIONAL STANDARD DEVIATION

A kernel estimator for the Conditional Standard Deviation of y at x can be defined as:

$$\widehat{ST}(x) = \left[\frac{\sum_i sw_i K(x_i, x) \left(y_i - y(x) \right)^2}{\sum_i sw_i K(x_i, x)} \right]^{\frac{1}{2}}$$

where K is a kernel function and y(x) is the expected value of y conditional on x.
To reach this application:

- From the main menu, choose: "Distribution ⇒ Conditional Standard Deviation".
- Choose the different vectors and parameter values.

Parameters	
x	Level of (X) or (p)
h	Smoothing parameter
Vectors	
x_i	Exogenous Variable (X)
y_i	Endogenous Variable (Y)

REMARK: The option "Level" vs "Percentile" allows the estimation of the conditional standard deviation of y either at a level of x or at a p-quantile for x.

Among the buttons, you find the following commands:

COMPUTE:	to compute ST(x).
GRAPH:	to draw ST(x) as a function of x.

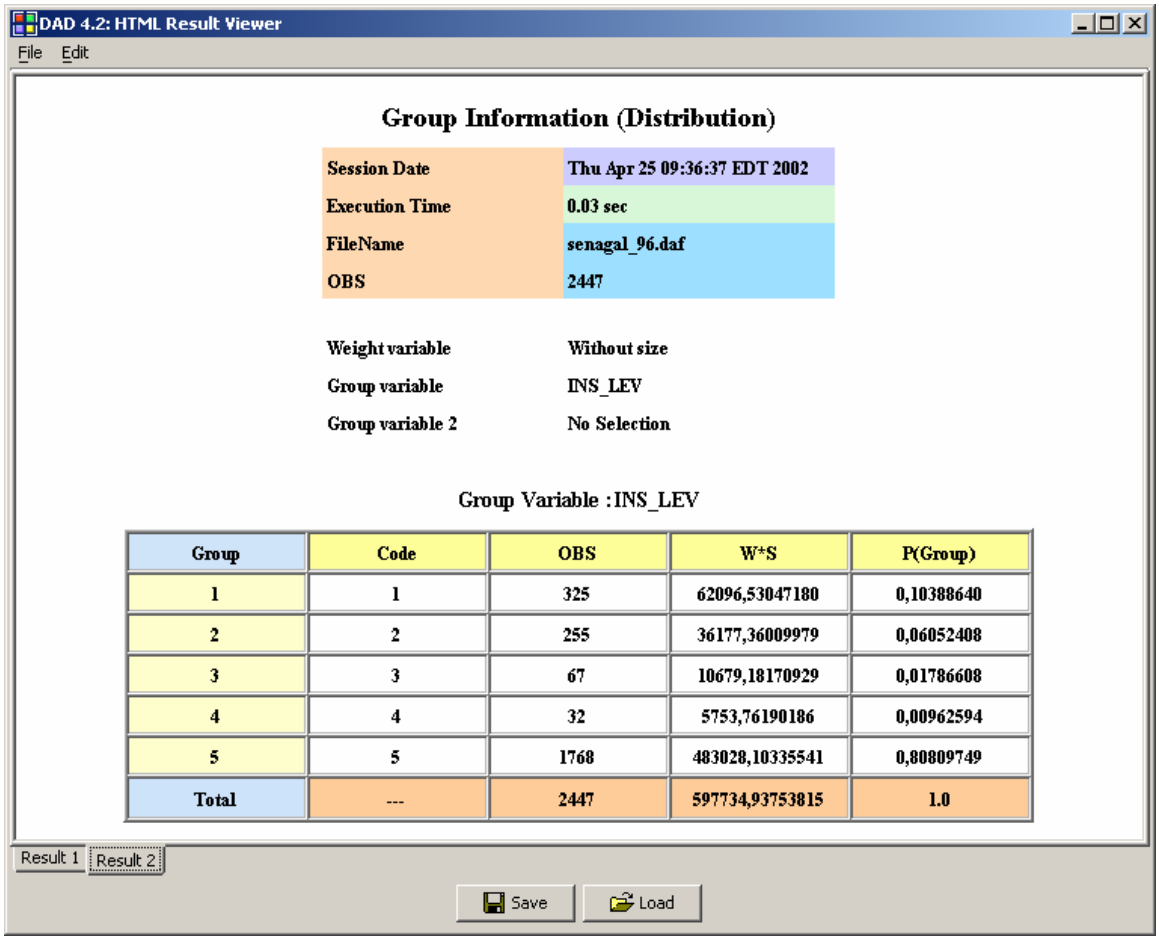
GROUP INFORMATION

This application estimates the cross-group composition of a population. The group details are provided by the user through either or both of two *Group* variables.

To reach this application:

- From the main menu, choose: "Distribution ⇒ Group Information".
- Choose the first group variable.
- Choose the size variable if the observations must be weighted by size.
- Choose the second group variable if you would like cross-group (or cross-tabulation) information to be provided across two groups.

Example 1:



This example uses only one group variable “INS-LEV” (level of instruction of the household head), categorized as

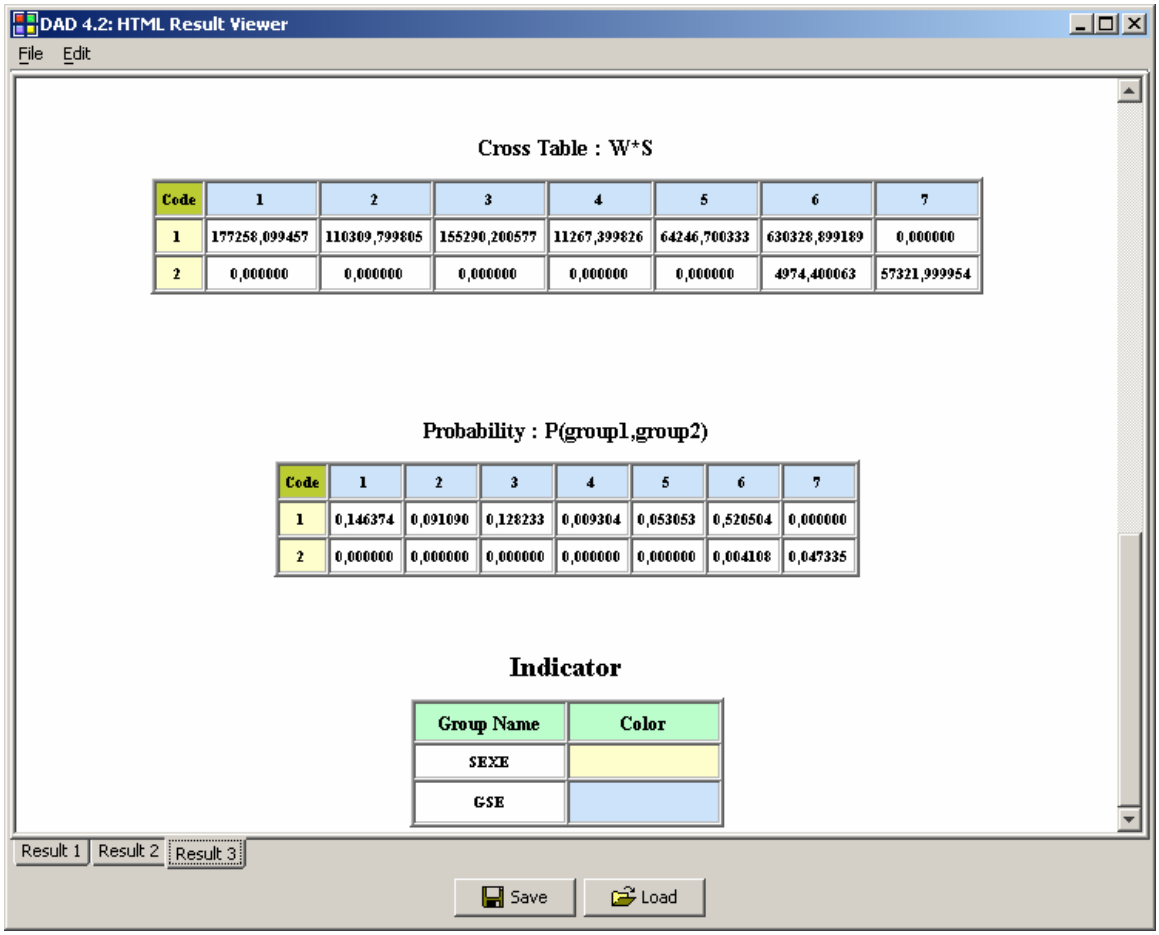
1. Primary
2. Secondary
3. Superior
4. Not available
5. None

The output shows:

Code	The exact code of the group
Group	The group number: (1,2,3,…)
OBS	The number of observations in the group
W*S	The sum of the products of Sampling Weight times Size
P(Group)	The estimated proportion of population found in that group

The use of two group variables shows the following information.

Example 2:



The *"Cross Table"* table shows the sum of the products of Sampling Weight times Size for those observations belonging to the two groups simultaneously. The second table, *"Probability"*, shows the estimated proportion of the population who belong to both of the groups.