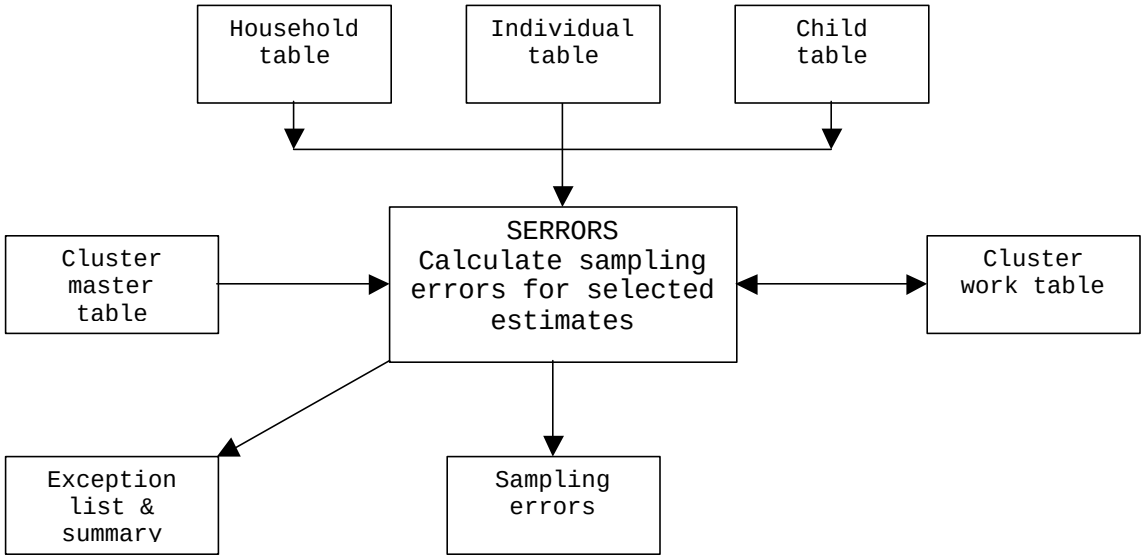
Program: SErrors2 - Calculate sampling errors using the clusters method

1. Schematic



2. Function

Calculate sampling errors for the principal estimates presented in the standard CWIQ tables at the national level and for urban, rural and regional subgroups. The list of estimates for which sampling errors are calculated appears in Appendix A.

3. Input

A. CWIQ questionnaire database.

The CWIQ questionnaire database consists of three tables: the household table; the individual table; and the child table. For each household surveyed there is one record in the household table (HHData), one record for each household member in the individual table (INData) and one record for each child under 5 in the child table (CHData).

4. Input/Output

A. Cluster work table (ClusterW)

The cluster work table contains intermediate results needed to calculate the sampling errors. There is an entry in the cluster work table for each cluster and each subgroup estimate. A detailed description of the cluster work table appears in Appendix B.

5. Output

A. Exception list and summary (EditList)

This table contains a summary list of the number of questionnaires processed and a list of any exceptional conditions encountered in the data.

B. Sampling errors for selected estimates (SErrors2)

This table contains the results of the sampling error calculations. These include the estimate name, the domain (subgroup) name, the estimate itself, the variance, the standard error, the relative standard error, the limits of a 95% confidence interval for the estimate and similar statistics as if the sample was drawn as a simple random sample (SRS). There is a table entry for each national and subgroup estimate. A complete description of the table appears in Appendix B.

6. Processing

This program uses the clusters method to calculate sampling errors for means, percentages and proportions. To use this method the clusters (primary sampling units) of the sample are organized into implicit strata each with two or more clusters. For maximum efficiency, the clusters should be relatively homogeneous and the number of clusters within a stratum should be kept to a minimum. Means, percentages and proportions can all be expressed as a ratio r = y / x, where y represents the weighted total sample value for variable y and x represents the weighted total number of cases in the group or subgroup under consideration. The variance of r is computed using the formula:

$$\text{var}(r) = \frac{1-f}{x^2} \sum_{h=1}^{H} \left[\frac{m_h}{m_h - 1} \left(\sum_{i=1}^{m_h} z_{hi}^2 - \frac{z_h^2}{m_h} \right) \right]$$

in which

$$z_{hi} = y_{hi} - r x_{hi} \quad \text{and} \quad z_h = y_h - r x_h$$

where

h represents the implicit stratum which varies from 1 to H

m_h is the total number of clusters in the h^{th} implicit stratum

y_{hi} is the sum of the values of variable y in cluster i in the h^{th} implicit stratum

x_{hi} is the number of cases in cluster i in the h^{th} implicit stratum

f is the overall sampling fraction, which is so small it is ignored.

The standard error of the estimate r is defined as the square root of the variance.

The program processes the database in cluster sequence and writes cluster totals of x and y (x_{hi} and y_{hi}) for each estimate to the cluster work table. The implicit stratum number, retrieved from the cluster master table, and the urban/rural and region codes are written to the cluster work table.

The program accumulates the cluster totals of x and y for each estimate and subgroup. After all clusters have been processed it calculates the overall ratio (r = y / x) for each estimate and subgroup.

Then the program processes the cluster work table in stratum and estimate sequence. It accumulates stratum totals of x and y for each estimate and subgroup from the cluster totals. It counts the number of clusters in the

stratum, calculates the cluster value of z ($z_{hi} = y_{hi} - rx_{hi}$) and accumulates z ($z_{hi}$) squared for each estimate to the stratum level.

When all the clusters in the stratum have been processed, the stratum component of the variance in the equation above is calculated and accumulated for each estimate and subgroup.

When all strata have been processed the variance for each estimate and subgroup is calculated. The following variables are derived from the variance:

Standard error = square root of the variance
Relative standard error = 100 * standard error / estimate
95% confidence interval minimum = estimate – 2 * standard error
95% confidence interval maximum = estimate + 2 * standard error

The program also calculates sampling errors as if the sample had been drawn as a simple random sample (SRS). The variance of a ratio r = y / x from a simple random sample is computed using the following formula:

$$\text{var}(r) = \frac{1-f}{n\bar{x}^2} \sum_{i=1}^{n} \frac{(y_i - rx_i)^2}{n-1}$$

where n is the unweighted number of observations

The sum of the squares above is calculated as:

$$\sum_{i=1}^{n} y_i^2 - 2r\sum_{i=1}^{n} x_i y_i + r^2 \sum_{i=1}^{n} x_i^2$$

The program accumulates the individual values of x, y, x^2, y^2 and xy as the database is being processed. When all the data have been processed, the SRS variance is calculated according to the formula above. SRS standard error and relative standard error are derived as described above.

Appendix A - sampling error specifications

Estimate	Type	Numerator	Denominator
Average household size	Mean	HhSize	All households
% of households owning land	Percentage	F3 > 0	All households
% of households with large livestock	Percentage	F8 > 0	All households
% of households with a radio	Percentage	F12e = 1	All households
% of households with a modern stove	Percentage	F12h = 1	All households
% of households with a bicycle	Percentage	F12i = 1	All households
% of households with a motorcycle	Percentage	F12j = 1	All households
% of households with a car	Percentage	F12k = 1	All households
% of households with safe water source	Percentage	G3 = 1-3	All households
% of households within 30 m. of water	Percentage	G7a = 1-3	All households
% of households within 30 m. of sec sch	Percentage	G7d = 1-3	All households
% of households within 30 m. of pri sch	Percentage	G7e = 1-3	All households
% of households within 30 m. of hlth f.	Percentage	G7f = 1-3	All households

Appendix B - ancillary table formats

SErrors2 - Sampling errors

Name	Description	Type
StatNo	Estimate number	Integer 2
SGVariable	Subgroup variable number	Byte 1
	0 National	
	1 UrbRur	
	2 Region	
SGNumber	Subgroup entry number (0, 1-2, 1-10)	Byte 1
SGName	Subgroup entry name	Text 16
EstimateName	Estimate name	Text 12
Estimate	Estimate value	Double 8
EstimateVar	Variance of the estimate	Double 8
EstimateSE	Standard error of the estimate	Double 8
RelSE	Relative standard error (%)	Double 8
CIMin	Lower bound of 95% confidence interval	Double 8
CIMax	Upper bound of 95% confidence interval	Double 8
H	Number of implicit strata	Long 4
N	Weighted number of observations	Double 8
Nunw	Unweighted number of observations	Long 4
Runw	Unweighted estimate value	Single 4
srsVar	Variance if simple random sample	Double 8
srsSE	Standard error if simple random sample	Single 4
srsRelSE	Relative standard error if simple random sample	Single 4

ClusterW - Cluster work table

Name	Description	Type
A1	Cluster number	Text 3
IStratum	Implicit stratum number	Integer 2
x	Total x value for the cluster	Double 8
y	Total y value for the cluster	Double 8
z	Ratio for the cluster: z = y / x	Double 8
UrbRur	Urban rural code	Text 1
Region	Region code	Text 2