Report on the Calculation of Sampling Errors

Survey on the Population in Relation to Activity
(PRA)

CONTENTS

1. Introduction

We may define sampling error as the imprecision that occurs when a characteristic of the study is estimated (parameter) through the value obtained from a part or sample of this population (statistic).

This error depends on many factors, including the procedure to extract this part of the population (sampling design), the number of units to be extracted (size of the sample), the nature of the characteristic to be estimated, etc. A generalised expression of the sampling error would be as follows:

$$\text{Sampling error} = \sqrt{Var\ (\hat{\theta})} \qquad (1)$$

$\hat{\theta}$ being the statistic of interest (mean, total, proportion,..). This statistic will take on different values depending on the extracted sample. The variability of the statistic in the sampling will determine the sampling error.

The expression of this error will change depending on the sampling technique used, the calculation becoming more complex as the sampling design gets more complex. Furthermore, incidences produced during the collection of information, adjustment to determined characteristics of the population (post-stratification) and other factors during the development of a survey, imply variations in the calculation of the elevators or final weights.

The literature suggests several alternatives to conventional sampling error calculation methods. These heuristic techniques provide a good estimate of the sampling error from the final weights and characteristics of the sampling design [3], [5].

These methods and specific application are introduced below in the case of the Survey on the Population in Relation to Activity (hereafter referred to as PRA) from 2005 on.

2. Taylor Expansion Method [3], [5].

This method enables the calculation of sampling error estimates for totals, means and ratios in samples with stratification, clusters and unequal probabilities, as is the case of many EUSTAT statistical operations. The method obtains linear approximations of the estimator and calculates the variance by using it as an estimate of the sampling error.

The expression for the calculation of the estimated variance for the mean population is as follows:

$$\hat{V}(\hat{\bar{Y}}) = \sum_{h=1}^{H} \frac{n_h(1-f_h)}{n_h-1} \sum_{i=1}^{n_h} (e_{hi.} - \bar{e}_{h..})^2 \qquad (2)$$

Where:

$$e_{hi.} = \frac{\sum_{j=1}^{m_{hi}} w_{hij}(y_{hij} - \hat{\bar{Y}})}{w...}$$

$$\bar{e}_h = \frac{\sum_{j=1}^{n_h} e_{hi.}}{n_h}$$

y

$$w... = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij}$$

Note:

$h = 1, 2, \ldots, H$ indicates the stratum with a total H strata.
$i = 1, 2, \ldots, n_h$ indicates the number of clusters in stratum h, with a total of n_h clusters.
$j = 1, 2, \ldots, m_{hi}$ indicates the number of unit within cluster i of stratum h, with a total of m_{hi} units
$n = \sum_{h=1}^{H} \sum_{i=1}^{n_h} m_{hi}$ is the total number of observations in the sample.

w_{hij} indicates the elevator of observation j in cluster i of stratum h
$y_{hij} = (y_{hij}(1), y_{hij}(2), \ldots, y_{hij}(P))$ are the values observed for the variable and in the observation j of cluster i of stratum h. (numeric and categorical variables).

The procedure PROC SURVEYMEANS in the statistical package SAS [4], implements this method to estimate sampling errors and will be the tool used for the calculation of sampling errors in the statistical operation in question.

3. Calculation of PRA errors

3.1 Sampling Design [1]

In 2005, the sample of the PRA was changed and resulted in not only an increase in units with regard to previous periods, but also a lower level of complexity in design [2]. The current sample is based on a probabilistic continuous sample, i.e. a panel of households that are continually revised. The sample has a size of approximately 5,000 households per quarter (an approximate total of 13,500 individuals) and a rotation of an eighth from one quarter to another, such that each household remains on the sample for two years.

The sample of households is extracted at random from the Housing Directory and stratified at Provincial level. Within each stratum households are sampled systematically (with the same

probability). Through an informer the information is collected from all the individuals in the household by which, for the purpose of the design, the households behave as small conglomerates. The selected units are households and the results refer to individuals.

This design adapts perfectly to the specifications of the heuristic method described in the previous section. It should only be indicated that the parameters are required by the SAS procedure for the correct estimate of the variance.

3.2 Calculation procedure.

The basic syntaxis of the SAS procedure implemented for the error calculation is as follows [4]:

PROC SURVEYMEANS < *nombre_fichero* > < *opciones de salida* >*;*
BY *variables* ; /*calculation of errors by independent subpopulations */
CLASS *variables* ; /* calculation of errors by qualitative variables */
CLUSTER *variables* ; */variable that indicates the cluster in the sampling by conglomerates*/
DOMAIN *variables* ; /*variables that demarcate the dominium/cross for which the errors are calculated*/
RATIO *variable/variable* ; /*variables ratio for which sampling error calculations are desired*/
STRATA *variables* < / *option* > ; /*variable that indicates the stratum in the stratified sampling*/
VAR *variables* ; /* quantitative and qualitative variables for which sampling errors are calculated*/
WEIGHT *variable* ; /* pre-calculated weight variable (optional)*/

The general parameters of this syntaxis for the specific case of the new PRA will be the following:

CLUSTER = Household identifier.
STRATA = Province.
WEIGHT = Quarterly elevator of persons /Quarterly elevator of families.
RATIO = Unemployment rates, activity and occupation.
VAR = Total Unemployed population, employed, active,…
DOMAIN = Crossed by socio-demographic and economic variables. (See section 3.3)

3.3 Statistics and domaims for the calculation of errors in PRA

Sampling errors are estimated for the following crossed variables and statistics:

Quarterly

- Activity and unemployment rate of the population aged 16 and over according to province (%).
- Activity rate of the population aged 16 and over according to sex and age (%).
- Unemployment rate of the population aged 16 and over according to sex and age (%).
- Employment rate of the population aged 16 to 64 according to sex, age and province (%).
- Population aged 16 and over according to province and sex (thousand).

- Active population aged 16 and over according to province and sex (thousand).
- Working population aged 16 and over according to province and sex (thousand).
- Working population aged 16 and over by province according to the economic sector (thousand).
- Unemployed population aged 16 and over according to province and sex (thousand).

Annual

- Population aged 16 and over by the relation with activity according to province and sex (thousand).
- Working population aged 16 and over by professional situation and type of contract (thousand).
- Families by the relation with the activity of members according to province (thousand).

We can sum up the above in the following tables according to crossed statistics and variables:

Quarterly errors

Crossed Statistic\Variable	Variable	Sex	Age (3 groups)	Economic sector
Activity rate	X	X	X	
Unemployment rate	X	X	X	
Employment rate	X	X	X	
Active population	X	X		
Working population	X	X		X
Unemployed population	X	X		
Population aged 16 and over	X	X		

Note: Crossed data are shaded and are given simultaneously

Annual errors

Crossed Statistic\Variable	Variable	Sex	Relation to activity	Professional Situation	Type of contract
Population aged 16 and over	X	X	X		
Working population				X	X
Families	X		X		

Note: Crossed data are shaded and are given simultaneously

3.4 Results and Interpretation.

Apart from the estimate of the sampling error (2), SAS provides other error measurements that are useful and help with interpretation. Among these, the most interesting are:

- The **Variation Coefficient,** a measurement relating to error that enables precisions to be compared between different groups or populations. It is an adimensional figure which is often used as a measurement of sampling error and is expressed thus:

$$CV = \frac{\sqrt{Var(\hat{\theta})}}{\hat{\theta}} \qquad (3)$$

- **Confidence Interval** at 95%. This confidence interval is based on the distribution in the statistical sampling (proportion, mean, rate). By the Central Limit Theorem, most times we can assume a Normal[1] law for the most common statistics, by which the construction of this interval will be given by the following expression:

$$\left[\hat{\theta} - 1{,}96\sqrt{Var(\hat{\theta})},\quad \hat{\theta} + 1{,}96\sqrt{Var(\hat{\theta})}\right] \qquad (4)$$

  The value 1.96 is the percentile of a Normal distribution with mean 0 and typical deviation 1 which entail a probablity of 95%. This makes it possible to state the calculated interval for the statistic $\hat{\theta}$ contains the true value for the population parameter in 95% of cases (possible samples).

With the information provided by SAS, definitive error tables are constructed that contain the estimation of the statistic, the lower and upper limit of the confidence limit at 95% and the variation coefficient as a percentage. Below you can see a model of error difusion table:

T.1 Coeficientes de variación e intervalos de confianza para la tasa de actividad y paro de la población de 16 y más años según el territorio histórico (%). IV-2004

Fuente: EUSTAT. Encuesta de Población en Relación con la Actividad.

	C.A. de Euskadi		Araba / Alava		Bizkaia		Gipuzkoa	
	Tasa de actividad	Tasa de paro	Tasa de actividad	Tasa de paro	Tasa de actividad	Tasa de paro	Tasa de actividad	Tasa de paro
Estimación	55,0	7,0	57,7	5,1	53,3	7,7	56,5	6,7
L. Inferior 95%	53,8	6,2	55,1	3,7	51,6	6,4	54,7	5,5
L. Superior 95%	56,2	7,8	60,2	6,4	55,1	9,0	58,4	8,0
CV(%)	**1,1**	**5,9**	**2,3**	**13,5**	**1,7**	**8,4**	**1,7**	**9,3**

Another way to interpret this information consists of calculating the **relative error** to 95% confidence, which is obtained by multiplying the percentile 1.96 by the Variation Coefficient. This relative error enables us to speak of the value of the estimation in terms of percentage points.

For the above table, the relative error at 95% of the Activity Rate of the A.C. of the Basque Country is 2.1% (1.96*1.1). This is the same as saying that, at a confidence level of 95%, we can state that the true value of the Activity Rate of the A.C. of the Basque Country varies between an interval of ±2.1% of the given estimation:

$$(55{,}0 \pm 0{,}021*55{,}0) = (53{,}8,\ 56{,}2)$$

It is important to point out estimations that surpass a certain percentage of relative error at 95%, so that the user may take the necessary precautions when interpreting the given information. A reasonible threshold would be for estimates with over 20% relative error (V.C. > 10% approx.), showing up especially those fields where this error is higher than 30% (V.C. > 15% approx.).

Bibliography

[1] EUSTAT (2005), *"Encuesta de Población en Relación con la Actividad. Ficha metodológica."* http://www.eustat.es/document/poblact_c.html

[2] EUSTAT (2005), *"Encuesta de Población en Relación con la Actividad. Nota metodológica.2005."* http://www.eustat.es/document/datos/notamet_nuevaPRA_c.pdf

[3] Fuller, W. A. (1975), *"Regression Analysis for Sample Survey,"* Sankhy , 37, Series C, Pt. 3, 117 - 132.

[4] Sas Institute Inc. (2004), *"SAS/STAT® 9. "User's Guide".* Copyright © 2004, Cary, NC, USA. ISBN 1-59047-243-8

[5] Woodruff, R. S. (1971), "*A Simple Method for Approximating the Variance of a Complicated Estimate"* Journal of the American Statistical Association, 66, 411 -414.