

Report on the Calculation of Sample Errors

Information Society Survey
(ESI – Families)



CONTENTS

1. Introduction.....	3
2. Taylor Expansion Method.....	3
3. Error Calculation E.S.I. - Families.	4
3.1 Sample Design	4
3.2 Calculation procedure.....	5
3.3 Statistics and domains for error calculation in the E.S.I.....	5
3.4 Results and Interpretation.....	7
Bibliography.....	9

1. Introduction

Sample error can be defined as the inaccuracy that occurs when a characteristic of the study population (parameter) is estimated by means of the value obtained from a part or sample of that population (statistic).

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

$$\text{Error de muestreo} = \sqrt{\text{Var}(\hat{\theta})} \quad (1)$$

Where $\hat{\theta}$; is the statistic of interest (mean, total, ratio,..). This statistic will take on different values depending on the extracted sample. The viability of the statistic in the sample will determine the sample error.

The expression of this error will change depending on the sample technique used. It becomes more complicated to calculate as the sample design gets more complex. Furthermore, incidences when collecting the information, adjustment to specific 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 puts forward several alternatives to conventional sample error calculation methods. These heuristic techniques provide a good estimate of the sample error from the final weights and characteristics of the sample design [3], [5].

These methods and their specific application in the case of the Information Society Survey since 2000 are set out below.

2. Taylor expansion method [3], [5].

This method enables the calculation of sample 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 sample error.

The expression to calculate the estimate variance for the mean population is as follows:

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

where:

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

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

and

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

Notation:

$h = 1, 2, \dots, H$ indicates the stratum with a total of H strata.

$i = 1, 2, \dots, n_h$ indicates the number of clusters in stratum h , with a total of n_h clusters.

$j = 1, 2, \dots, m_{hi}$ indicates the unit number 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 weighting of observation j in cluster i of stratum h

$y_{hij} = (y_{hij}(1), y_{hij}(2), \dots, y_{hij}(P))$ are the observed values of variable Y in observation j of cluster i of stratum h . (categorical and numerical variables).

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

3. ESI-Families error calculations

3.1 Sample Design [1].

The Information Society Survey of Families (ESI – Families) is a sample survey of the population of the Basque Country aged 6 and over. The panel of 5,088 family households selected for the Population in Relation to Activity survey (PRA) for the same reference quarter is used as the basis of the sample. Since 2004, this sample has consisted of 5,088 households extracted at random from the Household Directory and stratified by province. Within each strata, households are shown systematically (with the same probability) [2]

The selection of the first person within each household is performed randomly using a Kish table and when there are working or student members, one of each is also selected using the same

procedure. Since 2003, the sample has been completed with all the minors aged 6 to 14 until a sample of nearly 7,500 individuals is reached.

The survey that is exploited annually (1st quarter of the benchmark year) and the results refer to both individual and families, with a special emphasis on Internet users.

This sample design adapts perfectly to the specifications of the heuristic method described in the previous section. Only the parameters required by the SAS procedure to estimate the variance correctly will have to be indicated.

3.2 Calculation procedure

The basic syntax of the SAS procedure implemented to calculate the errors of this survey is as follows [4]:

```
PROC SURVEYMEANS < file_name > < output options >;  
  BY variables ; /*error calculation by independent subpopulations*/  
  CLASS variables ; /*error calculation by qualitative variables*/  
  CLUSTER variables ; /*variable that indicates the cluster in sample by conglomerates*/  
  DOMAIN variables ; /*variables that demarcate the domain/cross for which the errors are  
  calculated*/  
  RATIO variable/variable ; /*variables ratio for which the sample error is to be calculated*/  
  STRATA variables < / option > ; /*variable that indicates the stratum in the stratified sample*/  
  VAR variables ; /* quantitative and qualitative variables for which sample errors are to be  
  calculated*/  
  WEIGHT variable ; /* pre-calculated weight variable (optional)*/
```

The general parameters of this syntax used for the specific case of the ESI - Families will be as follows:

CLUSTER = Household identifier.

STRATA = Province.

WEIGHT = Annual elevator of persons / Annual elevator of families.

VAR = Equipment variables and use of Information Technologies.

DOMAIN = Crosses by socio-demographic and economic variables.

3.3 Statistics and domains for error calculation in the ESI - Families

Sample errors will be estimated for the following crosses and statistics:

Families

- Families of the Basque Country by ICT equipment, according to province (%) 2011.
Sample errors.
- Families of the Basque Country by ICT equipment, according to type of family (%) 2011.
Sample errors.

- Families of the Basque Country by television sets in the home, according to province (%) 2011. Sample errors.
- Families of the Basque Country by television sets in the home, according to type of family (%) 2011. Sample errors.

Population

- Population aged 15 and over of the Basque Country by ICT equipment and television sets in the home, according to province (%) 2011. Sample errors.
- Population aged 15 and over of the Basque Country with a computer in the home by sex and age, according to Province (%) 2011. Sample errors
- Population aged 15 and over of the Basque Country with a computer in the home, by level of education and activity status, according to Province (%) 2011. Sample errors.
- Population aged 15 and over of the Basque Country with internet in the home by sex and age, according to Province (%) 2011. Sample errors
- Population aged 15 and over of the Basque Country with internet in the home, by level of education and activity status, according to Province (%) 2011. Sample errors.
- Population aged 15 and over of the Basque Country according to Province and socio-demographic characteristics, according to household ICT equipment (%) 2011. Sample errors.
- Population aged 15 and over of the Basque Country by possibility of Internet access and province (%) 2011. Sample errors.

Use of Internet

- Population aged 15 and over of the Basque Country who are internet users by sex and age, according to Province (%) 2011. Sample errors
- Population aged 15 and over of the Basque Country who are internet users, by level of education and activity status, according to Province (%) 2011. Sample errors.
- Population aged 15 and over of the Basque Country who are Internet users by services used and average length of the last connection, according to Province (%) 2011. Sample errors.
- Population aged 15 and over of the Basque Country who are Internet users by place of access and languages used, according to Province (%) 2011. Sample errors.
- Population aged 15 and over of the Basque Country who have shopped online by goods purchased, payment method and opinion regarding payment security, according to province (%) 2011. Sample errors.
- erESIF17. Population aged 15 and over of the Basque Country by end of connection, access frequency, length of the connection and province (%) 2011. Sample errors.
- Population aged 15 and over of the Basque Country by services used, types of websites visited and province (%) 2011. Sample errors

The above can be summarised in the following tables according to statistics and cross variable:

Statistic	ICT equipment	Province	Sex	Age	Level of Education	Relation to activity	Family type	Relation with Internet	E-commerce
Population aged 15 and over percentage	X	X	X	X	X	X	X		
Total population aged 15 and over (thousands)	X	X	X	X	X	X	X		
Percentage of population aged 15 and over who are Internet users		X	X	X	X	X		X	X
Total population aged 15 and over who are internet users (thousands)		X	X	X	X	X	X	X	X
Family percentage (%)	X	X					X		
Total families (in thousands)	X	X					X		

3.4 Results and Interpretation

Apart from calculating the sample error (2), SAS provides other error measurements that are useful and help to interpret the survey. The most interesting are:

- The **Variation Coefficient**. It is a relative measurement of the error that enables accuracies to be compared between different groups or populations. **It is a dimensionless quantity commonly used to measure sample error and its expression is:**

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

- **Confidence Interval** at 95%. This confidence interval is based on the distribution in the sample of the statistic (ratio, mean, rate,...). According to the Central Limit Theorem, a Normal¹ distribution can be assumed for the most common statistics and this interval will therefore be constructed according to the following expression:

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

The value 1.96 is the percentile of a Normal distribution with mean 0 and standard deviation 1 which entail a probability of 95%. It can therefore be stated that the interval calculated for statistic $\hat{\theta}$; contains the true value of the population parameter in 95% of the cases (possible samples).

The information provided by SAS is used to build the final error tables that will contain the estimate of the statistic, the lower and upper limit of the confidence interval at 95% and the variation coefficient as a percentage.

An error dissemination table model is included below:

Families of the Basque Country by ICT equipment in the home (%) 2011. Sample errors.

¹ A sufficiently large sample size ($n > 30$) is assumed. When that is not the case, the confidence interval will be calculated using the relevant percentile at 95% of the distribution t-Student with $n-1$ degrees of freedom.

	Total (thousands)	Computer	Internet	Mobile phone	E-mail
Basque Country					
Estimate	846,3	62,4	57,0	90,1	55,1
Lower limit 95%	842,8	61,0	55,5	89,2	53,7
Upper limit 95%	849,8	63,8	58,4	90,9	56,6
VC (%)	0,2	1,1	1,3	0,5	1,3

Source: EUSTAT. Information Society Survey - ESIF

Another way to interpret this information consists of calculating the **relative error** at confidence of 95%, which is obtained by multiplying the 1.96 percentile by the Variation Coefficient. This relative error means that we can talk about the value of the estimate in terms of percentage points.

For the above table, the relative error at 95% for the percentage of families of the Basque Country with a computer at home is 2.2 % ($1.96 \cdot 1.1$). Or in other words, at a confidence level of 95%, we can confirm that the real value of the percentage of families of the Basque Country with a computer at home ranges in an interval of ± 2.2 % of the given estimate. That is:

$$(62.4 \pm 0.02156 \cdot 62.4) = \text{between } 61.0\% \text{ and } 63.8\%$$

Estimates that exceed a certain percentage of the relative error at 95% should be pointed out so that user can take the necessary precautions when interpreting the given information. A responsible threshold would be for estimates with over 20% relative error (V.C. approx. > 10%) and those fields when this error is greater than 30% (V.C. approx > 15%) should be specifically pinpointed.

Bibliography

[1] EUSTAT (2005), "*Information Society Survey-ESI-Families. Methodological File.*".
http://www.eustat.es/document/esi_c.html

[2] EUSTAT (2005), "*Population Related to Activity Survey.. Methodological Note. 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® 9.1 Guía de Usuario*". 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.