Proceedings of Q2006
European Conference on Quality in Survey Statistics

Dissemination of Complex Survey Sampling Errors

Ayestaran M., Goni E., Mas M., Prado C.[1]

1. Introduction

1.1 Objective and parts of the paper

The aim of this paper is to describe the process followed in the last years by Eustat, Basque Statistical Office, to disseminate the sampling errors, mainly for household surveys.
There are three parts in this paper. Firstly, there is a short introduction to Eustat; then a description of the process to disseminate the sampling errors in household surveys in the last 90s are given: which was the initial situation, the reviews and the proposals regarding the issue. After that, we will deal with the current situation and, finally, with the outlook for the future.

1.2 Brief presentation of Eustat

Eustat is a regional statistical office. It is responsible for official statistics in the Basque Country, which is one of the autonomous regions situated in the north of Spain, with a population of over 2 million inhabitants.

Twenty years ago it was set up as an Official Statistics Institute under the Basque Statistics Law. Here are some indicators of its activity:
- A workforce of 99 persons
- The most recent Statistics Plan provides for 180 surveys, 104 of which are under the responsibility of Eustat
- 138 Annual press releases with 1,605 media hits
- 600,200 web users
- 3.2 million accesses to the web
- Annual budget of approximately 10 million Euros

1.3 The household surveys in Eustat

An important part of Eustat statistical production correspons to household statistics, where the sampling design is traditionally carried out in two or sometimes three stages :

[1] Marina Ayestaran, Marina_Ayestaran@Eustat.es; Elena Goni, Elena_Goni@Eustat.es; Marta Mas, Marta_Mas@Eustat.es; Cristina Prado, Cristina_Prado@eustat.es; all of them of the Basque Statistics Office, EUSTAT, Methodology Department, Donostia-San Sebastian 1, 01013, Vitoria-Gasteiz, Basque Country, Spain,

- a first stage to select the census sections
- a second stage to select the dwellings
- and, occasionally, a third stage to select individuals in the dwelling.

Examples of surveys using this design are:

- Labour Force Survey (PRA),
- Living Conditions Survey (ECV),
- Time Use Survey (EPT)
- The Information Society Survey (ESIF)

At present, in some of the most important surveys such as the Labour Force Survey Survey, the design is being changed for a simple stratified random design.
In any of these surveys there is great interest in measuring the accuracy and in disseminating this information.

Therefore, the aim of this contribution is to present the development of this work since the beginning of Eustat and especially to present the dissemination policy which is being applied at this moment.

Although these quality indicators are classic and have always been calculated by statistical offices, it is not so common for them to form part of standard dissemination.

2. Process to publish sampling errors in household surveys

2.1 Until the late 90s.

Going back in time, to the end of the 90s, one of the most important surveys was the Labour Force Survey (PRA). Great interest was placed in estimating accuracy, especially accuracy related to the unemployment rate, due to high disaggregation by provinces.

Between 1987 and 1992 the coefficients of variation of this survey were published in short term bulletins by Eustat and computed on the basis of ad-hoc programmes. In 1993, after some methodological changes in the survey, sampling errors were no longer published until 1997. At that time the calculation method changed and the WESVAR, which is a programme based on sample replication methods, was used.
Between 1997 y 1999, the coefficients of variation were estimated in this way. The statistical bulletins published estimates together with information on the coefficient of variation. Cells were coloured in different tones according to whether the error was between 10 and 15% or greater than 15%.

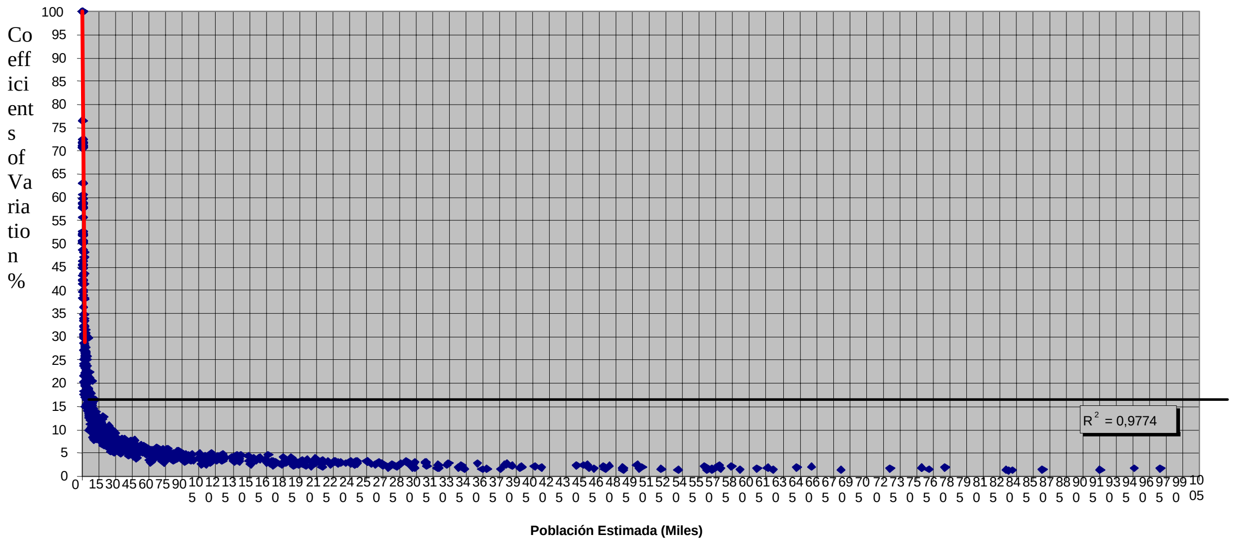
TABLE 1. EMPLOYED AND UNEMPLOYED OVER 16 POPULATION (O.I.T.) BY INDUSTRY, PROVINCE AND SEX
In thousands, 4th quarter 1997

	CAE			Alava			Bizkaia			Gipuzkoa		
	Total	Males	Females	Total	Males	Females	Total	Males	Females	Total	Males	Females
EMPLOYED												
TOTAL	754,8	485,9	268,9	116,7	73,5	43,3	382,0	247,1	134,9	256,1	165,4	90,7
Agriculture	23,8	18,3	5,4	5,0	3,6	1,4	10,8	8,9	2,0	7,9	5,8	2,1
Manufacturing	216,5	182,9	33,6	41,5	34,2	7,3	89,7	75,3	14,3	85,3	73,3	12,0
Construction	63,9	61,5	2,4	7,2	6,3	0,9	36,0	34,9	1,1	20,8	20,4	0,4
Services	450,7	223,2	227,5	63,0	29,3	33,7	245,5	128,0	117,6	142,2	65,9	76,2
UNEMPLOYED												
TOTAL	191,0	80,4	110,7	21,4	8,1	13,3	117,0	50,2	66,9	52,6	22,1	30,5
Agriculture	1,7	1,3	0,4	0,4	0,4	-	0,5	0,3	0,1	0,8	0,5	0,3
Manufacturing	28,6	15,8	12,7	5,3	2,6	2,7	13,5	7,7	5,8	9,8	5,5	4,3
Construction	14,6	13,9	0,7	0,6	0,5	0,1	10,6	10,4	0,2	3,4	2,9	0,5
Services	91,5	28,4	63,1	10,1	2,4	7,7	57,1	18,1	39,0	24,3	7,9	16,4
Seek 1st employment	54,7	20,9	33,7	5,0	2,1	2,9	35,3	13,5	21,8	14,3	5,3	9,1

🟩 - Relative Sampling Error between 10% and 15%
▨ - Relative Sampling Error over 15%
Source: EUSTAT

The tables were also accompanied by graphs, showing the estimates and theirs coefficients of variation. That is to say, the graph relates the size of the cells and the estimated coefficient of variation. The data come from tables made using WesVar and by the graph we know that an estimate of around 15,000 people have a 10-15% CV, approximately.

Graph 1. Coefficients of Variation and Estimated Population



From 1999, the statistical bulletin in paper disappeared and information on errors was no longer published. They were still calculated for internal dissemination, to be used by statistical producers or to be released on request.

At this time references to errors in tables published on the web and in the Data Bank began to appear. Eatimates with errors of over 30% were referred in the table footnotes.

2.2 Background and revision of other agencies' webs

The publication of errors was sporadic in the 90s and was carried out mainly for the Labour Force Survey (PRA). It was in 2004 when a more systematic policy for the publication of sampling errors, together with the documents on the computation methods, was considered. This was approacched for three main reasons:

- To fulfill the statistical quality standards set out by EUROSTAT (9 quality factors) which recommend the publication of metadata (methodology documents) to provide better information for the user.
- To follow the documentation process of all statistical surveys followed by Eustat through technical projects, establishing quality indicators and their assessment. Errors are the most classic quality indicators.
- For the interest in publishing the main sampling error tables already calculated for the surveys.

In order to carry out a proposal in this sense, we started to study what was being done by other statistical offices, to draw conclusions and define lines of action that were discussed in Eustat in the corresponding Committees, and were later applied.

We studied the treatment given to the publication of errors by the following offices:
INE (Spain), EUROSTAT, ONS (UK), Bureau Census (USA), CBS (Netherlands),
Statistics Canada and ISTAT (Italy)

From this study based on statistical offices very different models and types of publication were found. Some of them were part of quality reports, many were specific error tables, others were in the data bank and others were accessed by payment… .

As a conclusion, there were two common points for most of these offices :
- publication of tables with the coefficient of variation or standard error.
- publication of documents with methodology explanations on the computing of errors.

2.3 Proposals to disseminate sampling errors

After this study the following proposal was made:
- What to publish: Tables of standard errors or coefficients of variation of the main tables for all the sampling surveys and their methodological documents.
- How: Estimate, upper and lower limit and coefficient of variation
- Periodicity: The same as the general data tables of the survey

Other proposals made for the dissemination of these errors:
- tables with footnotes, or shaded for errors above a threshold value.

- to publish the tables in quality reports, together with other non sampling errors in specific Quality sections on the website. This will be the aim of a further project

3. Current situation

3.1 Errors dissemination policy

Now, as a result of all the above, the policy of sampling error dissemination in Eustat is based on the following points:

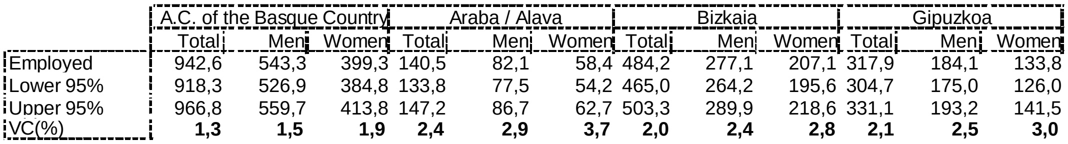
- All the sampling surveys have to publish sampling error tables
- For this reason, a proposal is made jointly with the statistic producer. Proposals are based on the results published in press releases, tables on the web, etc. The trend now is to publish errors on all the statistical tables that are published on the web.
- This policy includes producing a document on the methods used to estimate or compute sampling errors

The sampling error tables must contain the estimate, lower and upper limits above 95% of the Level of Confidence and the coefficient of variation (standard error/estimate) as a percentaje. This proposal tries to combine maximum information with clarity and possibility of space.

The sampling errors are published on Eustat's Web, in the survey Methodology section. An example of this is the PRA (the survey on labour force). In this section the Report on the Sampling Errors Estimate and sampling error tables are available. There are tables for quarterly and annual results. In both cases, confidence intervals are presented as well as coefficients of variation of tables mainly on activity and unemployment rates, working and unemployed population for the main tables with sex, age and province. The table which is shown below as an example is specifically on the working population, by province and sex. Estimates, lower and upper limits, and the coefficients of variation (in percentages) for the estimates are shown.

Table 2. Variation coefficients (%) and confidence intervals for the employed population aged 16 and over by province and sex (thousands). I-2006

Source: EUSTAT. Survey on the Population in Relation to Activity.



	A.C. of the Basque Country			Araba / Alava			Bizkaia			Gipuzkoa		
	Total	Men	Women	Total	Men	Women	Total	Men	Women	Total	Men	Women
Employed	942,6	543,3	399,3	140,5	82,1	58,4	484,2	277,1	207,1	317,9	184,1	133,8
Lower 95%	918,3	526,9	384,8	133,8	77,5	54,2	465,0	264,2	195,6	304,7	175,0	126,0
Upper 95%	966,8	559,7	413,8	147,2	86,7	62,7	503,3	289,9	218,6	331,1	193,2	141,5
VC(%)	**1,3**	**1,5**	**1,9**	**2,4**	**2,9**	**3,7**	**2,0**	**2,4**	**2,8**	**2,1**	**2,5**	**3,0**

Another way to access the information on sampling errors is through the methodological note, where the general characteristics of the survey are presented (objectives, sampling design, etc.).

3.2 Surveys with information on sampling error and methods of error estimation available

There are sampling errors and documentation available for some household surveys. One of them is already mentioned, the Labour Force Survey. The sampling design of this survey changed in 2005, from two stage sampling to stratified sampling on housing.

The error estimates are carried out in SAS, with the PROC SURVEMEANS, within the Enterprise Guide environment. This procedure implements the Taylor expansion method. SAS tables are generated in html format and by means of an Excel macro, the specific web dissemination format (according to the rules) is applied.

Other surveys for which this kind of information is available are the Survey on Living Conditions (ECV), the Time Use Survey (EPT) and the Information Society Survey on - Families (ESIF).
These surveys have a design with, at least, two stages, as one person is selected in the dwelling to answer the survey. For this sampling with more than one stage, WesVar PC is used. This involves sub-sample replication methods to obtain the variance of the estimator. In this case, to be more precise, the method is the Jackknife n, for stratified designs with two or more Primary Sampling Units per stratum.

4. Outlook for the future

- It has been also considered whether in the future information on sampling errors should be in the data bank. This would involve more flexibility for users, who could obtain the coefficients of variation for the specified tables in an interactive way. This would also mean a continuous updating of information on the sampling errors (series maintenance). At this first stage, our aim is to give just a punctual or precise information.
- Availability of sampling errors is promoted on other surveys run by the Departments of the Basque Government (Basque Statistical Organization). For surveys such as the Survey on Health or the Census on the Labour Market, widely disseminated, Eustat will have to recommend the computation of sampling errors and their dissemination. At this moment, Eustat does not carry out a systematic publication of sampling errors coming from other sources.
- Another important aspect is the creation of a quality section on the website, containing quality reports relating to sampling and non-sampling errors and other documents related to quality assessment.

No less important is the effort made to publish errors in economic surveys, as in the Industrial and Construction Survey. In this case a macro programmed in a SAS environment has been used. The macro computes the variance of the estimator, in this case the indirect ratio estimator, with auxiliary information on employment. The mean square error or MSE is estimated as the sum of variance and bias squared. For the fact that it is an indirect estimator, not unbiased, the coefficient of variation is estimated in reference to the MSE. The coefficient of variation is the square root of the MSE between the estimate.

The computation of errors in other surveys based on economic establishments is planned (Survey on Technological Innovation, Survey on the Information Society - Companies,…), with the SAS Proc Surveymeans.

References

Eurostat (2002). Quality in the European Statistical System. The way forward. 2002 Edition.

Eustat (1998) The replication method for the estimation of sampling errors”. D. Morganstein, “International Statistics Seminar, 37“. Eustat. 1998.
http://www.eustat.es/prodserv/vol37_c.html

Eustat (2004) “Publication and treatment of sampling errors. National and international statistical sources and proposals for Eustat” unpublished report, Vitoria-Gasteiz, Spain, Eustat.

EUSTAT (2005), Labour Force Survey (P.R.A.). Methodological note.
http://www.eustat.es/document/poblact_i.html

EUSTAT (2005), Report on the Calculation of Sampling Errors. Labour Force Survey (PRA) http://www.eustat.es/document/datos/Calculo_errores_PRA_i.pdf

EUSTAT (2005), Labour Force Survey (PRA), Sampling error tables 1st quarter 2006 and annual average 2005
http://www.eustat.es/estad/temalista.asp?idioma=i&tema=37&opt=0&tipo=7&mas=&otro=

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

Särndal, C.E., Hidiroglou, M.A. (1989). Small Domain Estimation: A Conditional Analysis. Journal of the American Statistical Association, 84, 266-275.

WESTAT (2002) *WesVar 4.2 Users' Manual*” Copyright 2002. WESTAT, Inc.

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

INE www.ine.es
ONS www.statistics.gov.uk
Eurostat, www.europa.eu.int/comm/eurostat/
Bureau of the Census (USA), American Community Sample,
www.census.gov/acs/www/index.html

CBS - Central Bureau of Statistics (Netherlands), www.cbs.nl/en-GB/default.htm
Statistics Canada, www.statcan.ca/start.html
ISTAT www.istat.it