Service statistics
Calculating Sampling Errors

CONTENTS

1. Introduction

Sampling error can be defined as the inaccuracy committed when estimating a characteristic of the study population (parameter) using the value obtained based on a part or sample of that population (statistic).

This error depends on many factors, including the sample extraction procedure, the number of units extracted, the estimate method, the nature of the characteristic to be estimated, etc. A generalised expression of the sampling error would be as follows:

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

Error de muestreo	Sampling error

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

The expression of this error will change according to the sampling technique used, which means that the more complicated the sample design is, the more complex it is to calculate. The majority of the Eustat surveys have a complex sample design that includes stratification, unequal selection probabilities, etc. These designs are applied in order to produce the best possible precise estimators, but in practice, they make it extremely complicate to estimate the sampling errors.

Various authors have put forward alternatives to the conventional methods for calculating sample errors. Among them, the replication [1] and linearlization techniques [5], [6] provide quick and simple estimates of the variance of any type of statistic (means, totals, proportions,…).

However, and for certain supposition, the Mean Quadratic Error (MQE) will need to be calculated. This takes into account not only the sample variance of the statistic but also any possible bias in the estimates due to factors outside the sample (e.g. use of ancillary information). This is the case of some Eustat economic surveys, which use the following expression to estimate the total error arising from population data interference [2]:

$$\text{ECM }(\hat{\theta}) = \text{Var}(\hat{\theta}) + \text{Bias}^2 \qquad \mathbf{(2)}$$

The Economic Survey of the Hotel and Catering Sector uses the above expression to estimate the sampling error. The error calculation and estimation system for the specific case of these surveys is considered below

2. Overview of the service survey

2.1 Definition

Population Coverage: The population that is the object of the study is made up of establishments whose exclusive or main activity is included in sections G, H, I, J, K, L, M, N, R and S of the National Classification of Economic Activities (CNAE-2009).

Population Sphere: The population studied consists of the establishments whose sole or main activity, according to CNAE93 (Spanish Classification of Business Activities codes), is Division 55 (Hotel and Catering Sector).

Time Scale: The survey base period is the year prior to the information being collected.

2.2 Sample Design

For 2012 statistics, a mixed method of information acquisition has been used which combines the direct collection of primary sample data and the acquisition of data from administrative information from the Business Registers in Álava, Bizkaia and Gipuzkoa and the Association of Registrars, as well as the Register of Cooperatives and the Register of Associations in the Basque Country attached to the Department of Justice, Employment and Social Security.

The establishments surveyed have been selected based on the Cube Method, where necessary [5]. The balancing variable that has been used is the number of establishments by territory, legal status and 3-digit CNAE-09 activity. The sample sizes for each strata of activity, legal status and territory were established based on the coverage of the administrative information available and the employment size:

- Stratum of 50 or more employees - censal
- Stratum of 20 to 49 employees - censal except for corporate entities
- Stratum of 10 to 19 employees - censal
- Stratum of 1 to 4 and 5 to 9 employees - the variability and coverage of the administrative registries were considered
- Stratum of employment without salaried workers (self-employed workers) - random sample proportional to the number of establishments

3. Estimate system and error calculation

3.1 Introduction

The Eustat economic surveys use different types of estimators when extrapolating the sample information to the population. On the one hand, direct estimators based on the sample design (Horvitz-Thompson estimator, a reason estimator,…) are used and on the other hand, model-assisted estimators that use ancillary information from other fields for areas where the sample is scarce. The latter have the advantage of reducing the sample error as the estimate is in a small area, but can also introduce a significant bias if the ancillary information in the different fields (or stratums) is not homogenous. Therefore, an optimum solution is to use estimators that on the one hand, offset the instability of the direct estimators and on the other hand, the bias of the indirect ones. See [2] and [3].

3.2 Composite estimators and their variance

The type of estimators referred to in the final part of the previous paragraph is the one used by the Services Statistics:

$$\hat{\theta}_{\text{COMPOSITE}} = \phi\,\hat{\theta}_{\text{DIRECT}} + (1 - \phi)\ \hat{\theta}_{\text{INDIRECT}} \qquad \text{with} \qquad 0 \leq \phi \leq 1 \qquad \textbf{(3)}$$

The expression of the Mean Quadratic Error for this type of estimator is not simple and an approximation is proposed as follows:

$$\text{ECM}(\hat{\theta}_{\text{COMPOSITE}}) = \phi^2\text{ECM}(\hat{\theta}_{\text{DIRECT}}) + (1-\phi)^2\text{ECM}(\hat{\theta}_{\text{INDIRECT}}) - 2\phi(1-\phi)[\text{ECM}(\hat{\theta}_{\text{DIRECT}}) - \hat{\theta}_{\text{INDIRECT}*}\text{Biais}] \quad \textbf{(4)}$$

Both the expression of the estimator and of its mean quadratic error are implemented in a SAS macro programmed for that purpose. Further details about the origin and calculation of the above expression can be consulted in reference [2] of the bibliography.

3.3 Estimate Tables and Variation Coefficients

The most relevant information provided by the Services Statistics refers to the main economic macro-magnitudes of the sectors of activity covered in the study and the profit and loss account of these sectors:

- Variation Coefficients for macro-magnitudes
- Variation Coefficients for the profit and loss account

The variation coefficient is a relative measure of the error that enables precisions between different groups or populations to be compared. It is an adimensional magnitude whose expression is:

$$CV(\hat{\theta}) = \frac{\sqrt{ECM\,(\hat{\theta})}}{\hat{\theta}} \qquad \textbf{(5)}$$

Another way of interpreting this information consists of calculating the **relative error at 95 % confidence**, which is obtained by multiplying the 1.96[1] percentile by the variation coefficient. This relative error allows us to talk about the estimate value in terms of percentage points.

In other words, at a 95% confidence level, the true value of the economic magnitude in the population can be said to be in the interval:

$(\hat{\theta}$ ± relative error $*\hat{\theta}$).= $(\hat{\theta} \pm 1{,}96 * \hat{\theta})$

Those estimates that exceed a certain percentage of the relative error at 95% should be highlighted so that the user can interpret the information given with the appropriate degree of caution. A reasonable threshold would be in those estimates that exceed 20% of the relative error (V.C. > approx. 10%), with special emphasis being placed on those errors where this error is greater that 30% (V.C. > approx. 15%).

[1] It is the normal distribution percentile (0.1) that corresponds to 95% probability.

Bibliografía

[1] EUSTAT (1998). "*El método de replicación para la estimación de errores de muestreo*". D. Morganstein, "Seminario Internacional de Estadística, 37". http://www.eustat.es/prodserv/vol37_c.html

[2] EUSTAT (2005). "*Cálculo de coeficientes de variación para diferentes estimadores directos e indirectos utilizados en las encuestas económicas de Eustat."* http://www.eustat.es/document/datos/Errores_c.pdf

[3] EUSTAT (2005). "*Estimación de Áreas Pequeñas en la Encuesta Industrial de la C.A. de Euskadi."* http://www.eustat.es/document/datos/ct_14_c.pdf

[4] EUSTAT (2007). Clasificaciones Sectoriales. http://www.eustat.es/document/datos/codigos/clasificacion_sectorial.xls

[5] EUSTAT (2010). "*Muestreo equilibrado y eficiente: el Método del Cubo".* Yves Tillé, "Seminario Internacional de Estadística, 52". *http://www.eustat.es/productosServicios/datos/Seminario_52.pdf*

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

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