Stratification and Computation of Allocations in Business Surveys

Ronan Le Gleut

***Résumé* — The aim of this methodological note is to provide a brief description of the methods usually used for stratification and the computation of allocations in business surveys.**

I. The Sampling Frame

INSEE statisticians make use of the Sirus ("Système d'Identification au Répertoire des Unités Statistiques") statistical business register in order to produce the sampling frame for business surveys. Sirus lists French businesses, legal units and establishments, together with some of their characteristics : geographical location, principal activity, number of employees and annual turnover reported to the authorities, probability of existence, etc.

This statistical register presents a few differences when compared with the National Enterprise and Establishment Register Database (Sirene - for "Système Informatisé du Répertoire national des ENtreprises et des Établissements") which has long formed the basis for the sampling frames for business surveys carried out at INSEE. The Sirius register also lists businesses that have an economic direction, whereas Sirene identifies legal units that have a legal direction.

II. Stratification

The population U is said to be stratified when the units can be partitioned into H disjointed sub-populations U_1,...,U_H known as strata (see diagram in Figure 1). It is therefore essential to have auxiliary information for the entire population.

The sampling design is said to be stratified when independent samples are selected in each stratum. A sample S_h of size n_h is therefore drawn from each stratum U_h of size N_h. Where simple random samples are selected from each stratum, this is known as stratified simple random sampling.

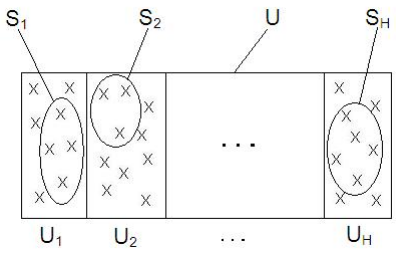

Fig. 1. The population U is said to be stratified when the units can be partitioned into H disjointed sub-populations $U_1, U_2, ..., U_H$ known as strata.

A. What are the Stratification Criteria ?

Samples for business surveys are drawn in accordance with stratified simple random sampling designs [1]. Most of time, the population of businesses is stratified by combining two criteria [2] :

- an activity criterion using more or less refined levels of the French Classification of Activities (NAF)
- a size criterion (using bands of salaried workers and/or bands of turnover).

For example (see [3]), the survey on information and communication technologies (ICT) is drawn by stratifying in accordance with :

- the activity sector with very different levels of aggregation (of the class or grouping of sections of the NAF) ;
- the business workforce size category (10-19, 20-49, 50-249, 250-499, 500+) ;
- turnover ;

The companies with the largest workforce bands (500+ employees) and the greatest turnover (last stratification criterion) are systematically included in the sample (exhaustive stratum).

B. How are the Strata Defined ?

The question of how many strata need to be constructed arises, which here means choosing a level of detail for our two criteria (activity sector and workforce band, for example).

First of all, it should be recalled that the exhaustive units belong to a separate stratum (called the "exhaustive stratum") in which all of the units are surveyed. In order to define these strata, exhaustivity thresholds (in terms of workforce or turnover) are often defined in order to force the largest units into the sample [3]. "Cut-off sampling" methods [2] allow all of the largest units to be automatically included in the sample, which allows a certain rate (of turnover, for example) of the population to be covered.

In business surveys, it is often the case that half of the sample involves these completeness thresholds.

Secondly, the aim of stratification is first and foremost to define the strata within which the behaviour is homogeneous in the sense of the variable of interest. In order to achieve this, the variable used for stratification must be related to the variable of interest.

In other words, the aim is to minimise the intra-strata dispersion of the variable of interest (i.e. the dispersion within the strata), or to maximise the inter-strata dispersion (i.e. the dispersion between the strata). The variance of a variable of

1. The new sample coordination method implemented by INSEE [1] results in the need to select samples by stratified simple random sampling.

2. Geographical location is sometimes used as a third criterion for drawing samples, but it is generally not taken into account when optimising the sampling design.

3. It is also possible to force other units into exhaustivity where these are known to behave atypically (e.g. restructuring, atypical units, etc.)

interest y can be broken down and written as follows :

$$\begin{aligned} S_y^2 &= \frac{1}{N-1}\sum_{k\in U}(y_k-\mu_y)^2 \\ &= \underbrace{\sum_{h=1}^{H}\frac{N_h-1}{N-1}S_{yh}^2}_{S_{y,intra}^2} + \underbrace{\sum_{h=1}^{H}\frac{N_h}{N-1}(\mu_{yh}-\mu_y)^2}_{S_{y,inter}^2} \end{aligned}$$

where $S_{yh}^2 = \frac{1}{N_h-1}\sum_{k\in U_h}(y_k-\mu_{yh})^2$, $\mu_y = \frac{1}{N}\sum_{k\in U} y_k = \sum_{h=1}^{H}\frac{N_h}{N}\mu_{yh}$ and $\mu_{yh} = \frac{1}{N_h}\sum_{k\in U_h} y_k$.

There are methods that can be used to ensure optimal division in order to define stratification boundaries that serve to minimise variance (of the stratification variable [4]). By way of a few examples, there is the Dalenius method [5], the geometric method proposed by Gunning and Horgan [6] or even the Lavallée-Hidiroglou method [7], where the strata are defined by the values of a quantitative variable that is well correlated with the variable of interest. The latter method also makes it possible to define an optimal threshold, based on which all of the units can be considered as exhaustive.

In practice, the strata are often defined from an expert's opinion, according to the levels at which the results are to be published (the dissemination fields, e.g. NAF section or division level, groupings of workforce bands). The fine sampling strata (defined in Section II-A)must therefore be included in the aggregated "optimisation" strata (e.g. aggregated NAF level combined with the workforce bands), which are themselves included in the dissemination fields.

Two levels of stratification are therefore generally used. With the exception of a few adjustments, the first level consists of combinations of the fields in which the results are to be disseminated. As we will see in the following section, this level is often used to calculate sampling rates t_h ensuring a certain degree of accuracy in each field in which they are to be disseminated. The fact that these optimisation strata are relatively well aggregated allows robust estimates of dispersions (and therefore of advance precision computations) to be made, since they are based on a large number of units.

The second level, which is used for sampling, is more refined than the first. More precisely, each sampling stratum t is included in an optimisation stratum h .The number of units to be drawn n_t is calculated in sampling stratum t by applying sampling rate t_h to the corresponding optimisation stratum :

$$n_t = t_h \times N_t$$

This procedure, which is based on the properties of the allocations that are proportional to the number of units (see Section III-A), allows the precision of future estimates to be improved if the stratification criteria are linked to the parameters that are to be measured [5].

The sampling strata therefore correspond to the most refined level of detail possible [6] given the scope of the survey and the sample size envisaged. The hope is that, by proceeding in this manner, estimates will be obtained that are at least as precise as if the samples had been taken from the optimisation strata.

Where there is a large number of sampling strata, rounding methods are used so as not to deviate too far from the sample size initially intended. This is performed using the τ-argus software (initially used to anonymise data, see[8]) or the Cox method [9].

III. Allocation computation

It is assumed that the overall sample size n is fixed, and that the strata have been defined. The size $n_1, n_2, ..., n_H$ of the sub-samples that are to be drawn from each stratum must be chosen.

Where there is a single, exhaustive stratum h, the allocation n_h is equal to the size of the stratum N_h (all of the units are automatically selected from the sample).

A. Proportional Allocations

Where the allocation is proportional **to the number of units**, the sampling rate is the same in each stratum :

$$f_h = \frac{n_h}{N_h} = \frac{n}{N} = f$$

This can be re-written as follows :

$$n_h = n\frac{N_h}{N}$$

In other words, the larger the stratum, the larger the sample selected from within it.

Each unit within the population has the same probability of inclusion $\pi_k = n/N$.This allocation therefore leads to a sampling design that is *self-weighted* where each individual unit has the same weighting $d_k = N/n$. This ensures excellent robustness of the results when analysing several variables simultaneously, particularly with categorical variables .

The variance of the stratified estimator for the total of a variable of interest y with proportional allocation is given by :

$$\begin{aligned} \mathbb{V}_p[\hat{t}_{y_\pi}] &= \sum_{h=1}^{H}\mathbb{V}_p[\hat{t}_{y_{h\pi}}] = \sum_{h=1}^{H}N_h^2\frac{1-f_h}{n_h}S_{yh}^2 \\ &= \frac{1-f}{n}N^2\sum_{h=1}^{H}\frac{N_h}{N}S_{yh}^2 \approx N^2\frac{1-f}{n}S_{y,intra}^2 \end{aligned}$$

We can therefore see that stratified simple random sampling performed with proportional allocation is (almost) always more efficient than simple random sampling (for

4. There are also methods that take account of the discrepancies between the stratification variables and the variable of interest (see [4]).

5. Even in cases where the stratification criteria turn out to not be linked to the parameters that are to be measured, this procedure does not have any adverse effect on future estimates (except in some very specific cases).

6. In practice, in order to facilitate post-survey processing (and precision computations in particular), survey managers generally wish to impose a minimum number of units to be drawn from each sampling stratum.

which the variance formula is identical, replacing $S^2_{y,intra}$ with S^2_y).Once again, the stratification must be selected such that the dispersion within strata is minimised (see Section II-B).

Other allocations, which are proportional **to an auxiliary variable** x are possible, and lead to better results than the allocation proportional to the number of units if the auxiliary variable x is positively correlated with the variable of interest y. In order to achieve this, the totals for x must be known for each stratum :

$$n_h = n\frac{t_{xh}}{t_x} = n\frac{\sum_{k\in U_h} x_k}{\sum_{k\in U} x_k}$$

Allocation proportional to the number of units is used if $x_k = 1\ \forall k \in U$.

For example, within the scope of a survey aiming to question the same number of employees within each establishment, if the aim is for there to be little dispersion in the sampling weight at the second level, it may be useful to proportionally allocate the employees of the establishments to each stratum at the first degree of sampling.

The proportional allocations to the economic quantities x provide approximations to Neyman's allocations (see Section Section III-B), assuming that the empirical coefficient of variation of the variable x (S_{xh}/μ_{xh}) is the same within each stratum.

B. Neyman's allocation

Neyman's allocation according to a variable of interest, which is widely documented in survey theory and regularly used by INSEE in the sampling designs for business surveys, optimises the precision of the estimator of the total **of this variable of interest at the level of the population as a whole**.

We are therefore seeking to resolve a (variance) minimisation issue under constraints (total fixed sample size of n) :

$$\begin{cases} \min\limits_{n_1,\dots,n_H} \mathbb{V}_p[\hat{t}_{y\pi}] = \sum\limits_{h=1}^{H} N_h^2 \dfrac{1-f_h}{n_h} S^2_{y,h} \\ s.c.\ \sum\limits_{h=1}^{H} n_h = n \end{cases}$$

The output allocation for this minimisation programme is as follows :

$$n_h = n\frac{N_h S_{yh}}{\sum_{j=1}^{H} N_j S_{yj}}$$

Neyman's allocation indicates that a larger sample should be selected :

— within the large strata ;
— within the strata with high dispersion.

The allocation is optimal for the variable of interest y and near-optimal for the variables positively correlated to y. Nevertheless, for variables that are negatively correlated or not at all correlated to the variable of interest, it may lead to results that are less precise than those achieved with proportional allocation (or even with simple random sampling).

Neyman's allocation may result in sample sizes being larger than strata sizes where these present high dispersion and/or are large in size. In this case :

— a census is conducted within the strata in question (we set $n_h = N_h$) ;
— the allocation is recalculated within the other strata.

An other way to set the allocation is to deal with the optimisation of precision under a total fixed cost constraint C :

$$C_0 + \sum_{h=1}^{H} C_h n_h = C$$

where C_0 indicates the fixed cost of the survey, and C_h the cost associated with the collection of a unit of U_h.

It is also possible to incorporate anticipated response rates when calculating Neyman's allocation. It is therefore standard practice to gather the response rates by stratum for a previous edition of the survey or a survey performed in the same field.

Neyman's allocation, in its "traditional" form, generally only partially meets the objectives of a survey since, as we have seen in Section II-B, the totals of the variable are not just published at the level of the population as a whole, but also at intermediate levels known as dissemination fields and corresponding to sub-sections of the population (only certain activities and certain sizes of business, etc.).

There can be no assurance that Neyman's allocation will be successful in these sub-sections. In particular, businesses in sectors with small (or more homogeneous) amounts relative to others are likely to be less well represented within the sample and the precision of the estimates limited to these businesses may not be sufficient.

The sampling rates corresponding to the business surveys conducted by INSEE are therefore increasingly based on a variation of Neyman's allocation, which introduces **local constraints on precision** [10]. This variant, which was proposed by Koubi and Mathern (2009) , optimises the precision of the estimator of the total variable of interest at the level of the population as a whole by guaranteeing a minimum level of precision in each dissemination field. Cases where the strata are saturated ($n_h > N_h$) are also handled by this algorithm.

The minimisation programme solved by the algorithm can therefore be written as follows :

$$\begin{cases} \min\limits_{n_1,\dots,n_H} \mathbb{V}_p[\hat{t}_{y\pi}] = \sum\limits_{h=1}^{H} N_h^2 \dfrac{1-f_h}{n_h} S^2_{y,h} \\ s.c.\ \sum\limits_{h=1}^{H} n_h = n,\ n_h \le N_h \\ s.c.\ \max\limits_{d\in D} CV_d \le CV_{loc} \end{cases}$$

where D represents all of the dissemination fields and CV_{loc} the maximum expected coefficient of variation.

In order to use this method, INSEE's statisticians must estimate the dispersions of the variable on which the constrained Neyman's allocation will be based. If the survey relates to a new topic, the most common practice is to optimise the allocation using a known variable within the sampling frame (turnover or number of employees) that is assumed to be linked to the variables of interest within the survey. Where there are previous editions of the survey, the results of those surveys are generally used to estimate dispersions.

C. Mixed Allocations

Surveys often have several distinct objectives. These two objectives usually have a good level of precision for a variable of interest, but a limited weight dispersion to ensure good quality estimates for other survey variables. One solution, in this case, is to take the mathematical mean of two allocations, i.e. :

$$n_{mixed} = \frac{1}{2}n_1 + \frac{1}{2}n_2$$

This allocation, which allows the benefits of the two low-cost methods to be combined is presented in the Cochran Handbook [11].

Taking the ICT survey as an example (see [3]), the decision was made to use mixed allocation corresponding to the average of :

— an allocation proportional to the number of units by ensuring, for each activity, that the half-length of the confidence interval will not exceed 10 points while also imposing a minimum of 10 units drawn from each stratum ;

— an allocation proportional to the number of persons employed (by imposing a minimum number of units to be drawn from each stratum).

The use of allocation that is proportional to the number of units is intended to meet the objective of precision for proportion-type variables. This is a specific case of Neyman's allocation under local constraints, described in Section III-B. Neyman's allocation is calculated using a variable indicator for which the dispersion (or empirical standard deviation S_y) is estimated at 0.5 in each stratum [7], and the local constraints correspond to a confidence interval half-length of 10 points, per activity (dissemination field), for the estimate of the proportion corresponding to this variable.

The proportional allocation to the number of persons employed aims to meet an objective of precision relating to the amount-type variables (by favouring strata containing large enterprises).

However, the choice of a factor of $1/2$ for the mean allocation is questionable. The paper from Merly-Alpa et Rebecq [12] specifically aims at investigating a method based on a minimisation program involving the dispersion of weights and the distance from Neyman's allocation. This program aims at selecting a parameter α which weights the two allocations, proportional (n_{prop}) and Neyman's (n_{Neyman}), in an optimal way such as :

$$n_{mixte}^{opt} = \alpha n_{prop} + (1-\alpha) n_{Neyman}$$

7. This represents an increase in the dispersion of an indicator variable y : $S_y = \sqrt{\frac{N}{N-1}P(1-P)} \approx \sqrt{P(1-P)}$ where $P = \frac{1}{N}\sum_{k\in U} y_k = 0.5$ (with no a priori on the value of the proportion to be estimated).

References

[1] Gros, E., Merly-Alpa, T. (2016). La coordination d'échantillons. *Note de méthodologie du DMS - Insee*.

[2] Särndal, C. E., Swensson, B., Wretman, J. (2003). Model assisted survey sampling. *Springer Science & Business Media*, pp. 531-533.

[3] Demoly, E., Fizzala, A., Gros, E. (2014). Méthodes et pratiques des enquêtes entreprises à l'Insee. *Journal de la Société Française de Statistique*, vol. 155, No 4, pp. 134-159.

[4] Rivest, L.P. (2002). Une généralisation de l'algorithme de Lavallée et Hidiroglou pour la stratification dans les enquêtes auprès des entreprises. *Techniques d'enquête*, vol. 28, No 2, pp. 207-214.

[5] Dalenius, T., Hodges Jr, J.L. (1959). Minimum variance stratification. *Journal of the American Statistical Association*, vol. 54, No 285, pp. 88-101.

[6] Gunning, P., Horgan, J.M. (2004). A new algorithm for the construction of stratum boundaries in skewed populations. *Survey Methodology*, vol. 30, No 2, pp. 159-166.

[7] Lavallee, P., Hidiroglou, M.A. (1988). On the stratification of skewed populations. *Survey Methodology*, vol. 14, No 1, pp. 33-43.

[8] De Wolf, P.P., Hundepool, A., Giessing, S., Salazar, J.J., Castro, J. (2014). τ-argus User's manual. *Argus Open Source-project*, pp. 28-30.

[9] Cox, L.H. (1987). A constructive procedure for unbiased controlled rounding. *Journal of the American Statistical Association*, vol. 82, No 398, pp. 520-524.

[10] Koubi, M., Mathern, S. (2009). Résolution d'une des limites de l'allocation de Neyman. *Journées de Méthodologie Statistique, Paris*.

[11] Cochran, W.G. (1977). *Sampling Techniques, third edition*, pp. 119-120.

[12] Merly-Alpa, T., Rebecq, A. (2016). Optimisation d'une allocation mixte. *9ème colloque francophone sur les Sondages, Gatineau*.

Département des méthodes statistiques
Version n° 1, diffusée le 11 septembre 2017.