Survey Research Methods (2007)
Vol.1 , No.2 , pp. 75-83
ISSN 1864-3361
http://www.surveymethods.org

Cross-sectional variance estimation for the French “Labour Force Survey”

Pascal Ardilly and Guillaume Osier
Institut National de la Statistique et des Études Économiques (INSEE)

This paper describes the method that was implemented by the Statistical Office of France (INSEE) to calculate cross-sectional variance estimates for the French “Labour Force Survey” (LFS). A home-made SAS application, named POULPE, was used for the calculations. After outlining the LFS sample design in the first part, the paper presents the software POULPE, particularly its technical capabilities and the theoretical principles underlying the variance estimation methods it implements. The third part develops how POULPE was utilized to compute variance estimates for the LFS, focusing on the statistical problems which were met and how they were solved. Finally, we provide estimated standard errors for totals and ratios and we make comments about those values.

Keywords: Multi-stage sampling, software POULPE, standard error, confidence interval, design effect

1 The French “Labour Force Survey”

1.1 Scope and objectives of the survey

The “Labour Force Survey” (LFS) is a quarterly survey carried out by the INSEE on a sample of around 54,000 private dwellings. Its purpose is to study the French labour market every quarter (in particular, estimate the number of unemployed people and the unemployment rate) and measure quarterly employment variations. All the persons aged 15 and over living in Continental France are eligible for inclusion in the sample.

1.2 A multi-stage sample

The LFS is based on a multi-stage selection of dwellings. The Primary Sampling Units (PSU) are either municipalities or blocks. They were stratified according to NUTS2 region and degree of urbanisation. At the first stage, PSUs were selected with probability proportional to the number of dwellings. One sector (a sector is a contiguous area containing between 120 and 240 dwellings) was then chosen in each PSU with probability proportional to the number of dwellings, and one area of around 20 dwellings was drawn from each sector by simple random sampling. Finally, all the dwellings enumerated by the 1999 Census within the selected areas were surveyed.

The dwellings which came into being after the Census (the new dwellings) were sampled according to a specific design. Let X denote the total number of new dwellings in an area (value collected during the fieldwork):

- $X \leq 10$: all the new dwellings are surveyed
- $11 \leq X \leq 40$: 10 dwellings are selected by simple random sampling (SRS)
- $41 \leq X$: 1/4 of the dwellings are selected by SRS

2 The variance estimation software POULPE

2.1 Introduction and key features

POULPE is a SAS macro-based application which was developed by the INSEE for variance estimation in complex designs. The sampling plans POULPE can deal with are:

1. The one-phase multi-stage plans with at each stage one of the following
   (a) Simple random sampling without replacement
   (b) Balanced simple random sampling
   (c) Sampling with unequal inclusion probabilities (probability proportional-to-size sampling)
   (d) Systematic sampling with equal inclusion probabilities
2. The two-phase multi-stage plans where the second phase is carried out by either Poisson sampling or post-stratified sampling.
3. The three-phase multi-stage plans where the second phase is carried out by post-stratified sampling and the third one by Poisson sampling.

Contact information: Institut National de la Statistique et des Études Économiques (INSEE), 18 Boulevard Adolphe Pinard, 75675 PARIS Cedex 14, France (pascal.ardilly@insee.fr)

In particular, the impact of unit non-response on variance estimates can be included in the calculations by viewing a sample of respondents as the outcome of an additional phase of selection. In addition, POULPE can take into account the impact of weight adjustments to external data sources (calibration procedure, cf. Deville and Särndal 1992).

For a given set of estimators, POULPE will estimate:

- Their variances and standard errors
- The lower and upper bounds of a 95% confidence interval. The estimators are assumed to follow a normal distribution, provided the sample size is large enough. Thereby, a 95% confidence interval for a parameter θ is given by:

$$C\hat{I}\left(\theta, 95\%\right) = \left[\hat{\theta} - 2\sqrt{\hat{V}\left(\hat{\theta}\right)}\ \ ,\ \ \hat{\theta} + 2\sqrt{\hat{V}\left(\hat{\theta}\right)}\right] \quad (1)$$

- The design effect *Deff*. This is the ratio of the actual variance, under the sampling plan P actually used, to the variance that would be obtained under a simple random sampling without replacement and of same size:

$$Deff = \frac{V_P\left(\hat{Y}\right)}{V_{SRS}\left(N \cdot \bar{y}\right)} \quad (2)$$

N is the target population size and $\bar{y}$ the sample mean of the variable y. $N\,\bar{y}$ is the standard unbiased estimator of the population total of y under simple random sampling. Basically, a design effect greater than one indicates that the design has increased the variance, while a value less than one indicates that the design actually decreased the variance of the estimate. The *Deff* estimation formula is given in the appendix.

2.2 Theoretical principles underlying the variance calculations

POULPE is based on analytic variance formulas, i.e. explicit formulas reflecting the peculiarities of a sample design, and Taylor linearisation. Detailed documentation on this topic can be found in Caron (1998).

2.2.1 The general formula

Every multi-stage sampling is splitted in "elementary" samplings. The variances contributed by each "elementary" sampling are estimated and then combined so as to form an estimate for the overall variance. The underlying formula is due to Raj (1968). Consider a two-stage sampling design where the second-stage sampling is assumed to have the properties of independence and invariance and let $\hat{t}$ denote the Horvitz-Thompson estimator of a total t. An unbiased variance estimator for $\hat{t}$ is given by:

$$\hat{V}\left(\hat{t}\right) = f\left(\hat{T}\right) + \sum_{i\in s} \omega_{is} \cdot \hat{V}_i \quad (3)$$

- $f\left(\hat{T}\right)$ is the estimated variance contributed by the first stage. $\hat{T}$ is the vector of the estimated totals for the PSUs.
- $\hat{V}_i$ is an unbiased estimator for the second-stage variance in the PSU$_i$.
- ω_{is} is the sampling weight of i.

It is easy to see that the formula (3) can be extended to multi-stage designs. Thus, the variance of a multi-stage design can be expressed as a sum of variance terms representing the contribution of each sampling stage.

2.2.2 Variance formulas for element sampling designs

Let

$$\hat{t} = \sum_{k\in s} \frac{y_k}{\pi_k}$$

denote the Horvitz-Thompson estimator of a population total

$$t = \sum_{k\in U} y_k$$

.

Simple random sampling without replacement of size n

$$\hat{V}\left(\hat{t}\right) = N^2 \cdot \left(\frac{1}{n} - \frac{1}{N}\right) \cdot s^2 \quad (4)$$

Where n is the sample size, N the population size and s^2 the sample variance of the target variable y.

Balanced simple random sampling of size n

$$\hat{V}\left(\hat{t}\right) = N^2 \cdot \left(\frac{1}{n} - \frac{1}{N}\right) \cdot e^2 \quad (5)$$

Where e^2 is the sample variance of the linear regression residuals of the target variables on the balance variables (Deville and Tillé 2005).

Sampling with unequal inclusion probabilities

Sampling schemes with unequal inclusion probabilities are difficult to handle because:

- The double inclusion probabilities π_{ij} (i.e. the probability that both i and j be selected) generally cannot be calculated.
- The "classical" Horvitz-Thompson variance estimator (Särndal et al. 1992) is expressed as a quadratic form. The number of terms in the sum is prohibitive.

POULPE relies on the following variance formula:

$$\hat{V}\left(\hat{t}\right) = \frac{n}{n-1} \sum_{k\in s} \left(1 - \pi_k\right) \cdot \left[\frac{y_k}{\pi_k} - D\left(\frac{y}{\pi}\right)\right]^2 \quad (6)$$

Where

$$D\left(\frac{y}{\pi}\right) = \frac{\sum_{i\in s}(1-\pi_i)\,y_i}{\sum_{i\in s}(1-\pi_i)\,\pi_i}$$

The formula (6) is an approximation which is valid under fixed-sized sample designs with large entropy (randomness). It only requires knowing the simple inclusion probabilities $\{\pi_i, i \in s\}$. Besides, the sum has only n terms, where n is the effective sample size, which makes (6) tractable.

Systematic sampling of size n with equal inclusion probabilities

$$\hat{V}(\hat{t}) = N^2\left(\frac{1}{n} - \frac{1}{N}\right)\cdot\delta^2 \qquad (7)$$

Where

$$\delta^2 = \frac{1}{2(n-1)}\sum_{i=1}^{n-1}(y_i - y_{i+1})^2$$

If the population is listed in random order, (7) is similar to (4). For information on variance estimation under systematic sampling, see Wolter (1985).

2.2.3 Multi-phase sampling

Variance estimation under multi-phase sampling is a long-established theory (Särndal et al. 1992). It can be regarded as an extension of variance estimation under multi-stage sampling (cf. 2.2.1): the variance of a multi-phase design can be expressed as a sum of variance terms representing the contribution of each sampling phase. POULPE tackles multi-phase sampling designs by estimating the variances contributed by each phase of selection and then combining them in order to obtain an estimate for the overall variance.

2.2.4 Treatment of non-linear statistics

POULPE can deal with non-linear statistics expressed as ratios or products of linear estimators by linearising them. Linearising a non-linear statistic $\hat{\theta}$ consists of deriving a linear statistic which has the same asymptotic variance:

$$Var\left(\hat{\theta}\right) \approx Var\left(\sum_{i\in s}\omega_{is}\times z_i\right) \qquad (8)$$

For instance, consider the ratio $\hat{\theta}$ of two linear estimators $\hat{X}$ and $\hat{Y}$ for the totals X and Y of two variables x and y. Then, a "linearised" variable at k for $\hat{\theta}$ is:

$$z_k = \frac{1}{X}\left(y_k - \frac{Y}{X}x_k\right) \qquad (9)$$

An extensive literature about the linearisation technique is available, e.g. Woodruff (1971), Binder and Patak (1994), Deville (1999).

2.3 *Running POULPE*

As an "analytic" variance estimation software, POULPE has strong theoretical foundations. The direct consequence of this is much information is needed to make the software run. That information is conveyed through three datasets which must be created before running POULPE.

2.3.1 The design dataset

The design dataset will describe a multi-stage sampling plan according to a stage-by-stage hierarchy. The following information will be given for each sampling stage:

- The description of the selected units
- The description of the "aggregation" units, from which the selected units are drawn
- The type of sampling that is implemented

2.3.2 The survey dataset

The survey dataset has one record per unit that is drawn, including out-of-scope and non-responding units. The dataset also contains identifying codes for the units which are selected at each sampling stage. Finally, additional variables must be recorded depending on the type of sampling design, e.g.:

- An identifying variable for the sampling phases (in case of multi-phase design)
- Estimated response probabilities
- Calibration variables
- The variables used to arrange the sampling frame (in case of systematic sampling)...

2.3.3 The geographical dataset

It contains relevant auxiliary numerical information about the sampling units, basically:

- Population sizes (in case of simple random sampling or systematic sampling)
- The sample distribution of the "size" variable (in case of probability proportional-to-size sampling)

On the basis of the information contained in the geographical and the survey datasets, POULPE will manage to calculate the inclusion probabilities for each sampling stage. This information is essential for variance estimation.

2.3.4 The interactive environment

The construction of the design, the survey and the geographical datasets is the most difficult step in using POULPE. It is as difficult as the sample selection is complex. Once those three preliminary datasets have been created, running the software will be pretty easy. POULPE carries out variance estimation in four stages. At each stage, a macro is executed and some work is done. Thanks to a "push-button" environment, assigning values to the macro parameters is interactive.

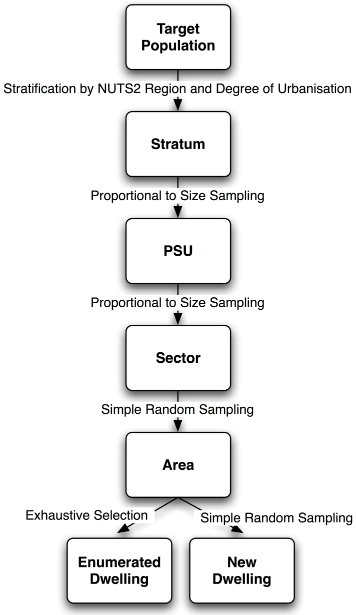

Figure 1. Sampling "Tree" of the LFS

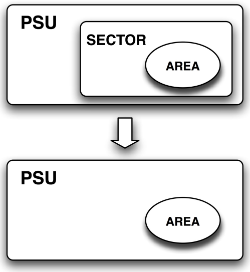

Figure 2. Elimination of the Sector Stage

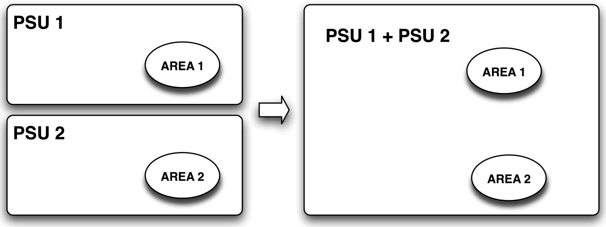

Figure 3. Merging Primary Sampling Units

3 Implementation of POULPE to the French "Labour Force Survey"

This section deals with the implementation of POULPE to produce variance estimates in the particular situation of the French "Labour Force Survey". The emphasis is put on the creation of the three preliminary datasets, the problems which were encountered during this stage and the solutions which were found.

3.1 Creation of the design dataset

As stated in 2.3.1, the design dataset aims to describe a multi-stage sampling design according to a stage-by-stage hierarchy. In the case of the LFS, on the basis of the information given in 1.2, that hierarchy is:

(0): Stratification by NUTS2 region and degree of urbanisation
(1): Selection of PSUs with probability proportional to the number of dwellings
(2): Selection of one sector with probability proportional to the number of dwellings
(3): Selection of one area by simple random sampling
(4): Exhaustive selection of the enumerated dwellings; simple random selection of new dwellings

A sampling "tree" is a good way to represent multi-stage sampling designs, as shown in Figure 1.

However, a major statistical problem happens. It is due to samples of size 1: within a PSU, one sector is selected and one area is then selected within a sector. Samples of size 1 do not allow unbiased variance estimation. The solution to this problem consists in using a proxy design for which unbiased variance estimation can be achieved, i.e. with no selection of samples of size 1. This entails some approximation as a proxy design never exactly reflects the exact design. In the case of the LFS, two actions were taken in order to deal with samples of size 1:

- Elimination of the sector stage
- Merging Primary Sampling Units

3.1.1 Elimination of the sector stage

We assume that one area is directly selected from a PSU by simple random sampling, as shown in Figure 2. This assumption does not affect the accuracy. It can be shown the variances calculated under the actual design (with the sector stage) and under the proxy design (after eliminating the sector stage) are equal.

3.1.2 Merging Primary Sampling Units

Despite the elimination of the sector stage, a problem remains: only one area is selected within a PSU so we still have to deal with samples of size 1. A possible solution to this problem consists in forming groups of 2 or 3 PSUs. We then assume that a sample of those groups has been selected with probability proportional to the number of dwellings. Two or three areas have been then selected within each group by simple random sampling, as shown in Figure 3.

Obviously, forming groups of PSUs will make variance estimates biased and one must strive to make groups so as to control that bias as much as possible.

Consider two Primary Sampling Units PSU_1 and PSU_2 of sizes[1] N_1 and N_2. One area is selected within both of

[1] Size=number of "areas"

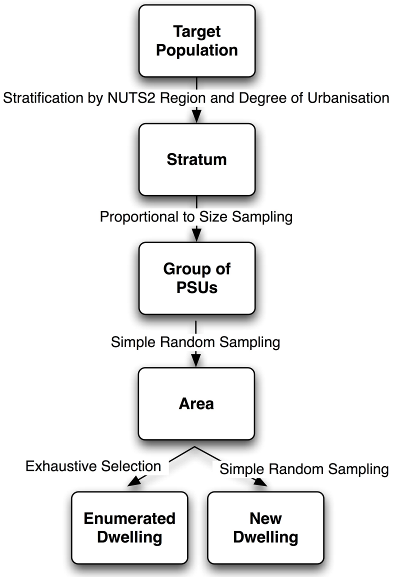

Figure 4. Final Sampling ”Tree” of the LFS

them by simple random sampling. Let V_P denote the variance under this design (denoted P). Suppose now that those two areas are selected by simple random sampling within the group PSU$_1$ + PSU$_2$. Let Y denote a variable defined at the area level. A standard unbiased variance estimator for the sample mean of Y is:

$$\hat{V}\left(\hat{\bar{Y}}\right) = (1-f)\cdot\frac{s^2}{2} \qquad (10)$$

Where s^2 is the sample variance of Y and f the sampling rate.

Assuming $N_1 \approx N_2$ and both N_1 and N_2 are large enough, the expectation of (10) with respect to P is:

$$E_P\left(\hat{V}\right) = V_P + f\cdot V_P + (1-f)\cdot\left(\frac{\bar{Y}_1-\bar{Y}_2}{2}\right)^2 \qquad (11)$$

Where $\bar{Y}_1$ and $\bar{Y}_2$ are the PSU$_1$ and PSU$_2$ means of Y.

As a conclusion, merging PSUs turns out to overestimate the variance, which is a conservative estimator. Moreover, the closer the means $\bar{Y}_1$ and $\bar{Y}_2$ are, the lower that increase of variance is.

Thus, we sought to merge PSUs of same sizes and of same characteristics on some target survey variables.

3.1.3 Final sampling ”tree”

We used the sampling tree that is represented in Figure 4.

3.2 *Creation of the geographical dataset*

The geographical dataset contains all the auxiliary numerical information that POULPE needs to calculate the inclusion probabilities of the sampling units. For the LFS, two difficulties happened.

3.2.1 The total numbers of areas in the groups of PSUs

Under the proxy design introduced in the previous section, the total number of areas in a group of PSUs is needed in order to compute the inclusion probabilities. For a given area, which is picked up by simple random sampling within a group of PSUs, the inclusion probability is equal to n/N, where n is the total number of selected areas and N the total number of areas in the group. POULPE is able to determine the numerator n by counting the records in the survey dataset. However, concerning the denominator N, the value is not available as it actually does not exist (only a sector was divided into areas and not a whole PSU).

As an area contains about 20 dwellings, the total number of areas in a group of PSUs was estimated by dividing its size (in number of dwellings) by 20. For instance, the total numbers of areas in a group of 1200 dwellings is 1200 / 20 = 60.

3.2.2 The total numbers of new dwellings in the areas

Those values are necessary to calculate the inclusion probabilities for the new dwellings. In practice, the total number of new dwellings in an area was not available. Nevertheless, we managed to derive the information for most of the areas using the number of new dwellings that had been surveyed in each of them.

The idea is to ”inverse” the sampling design presented in 1.2. Let n denote the number of new dwellings selected within an area and N the total number of those dwellings. According to 1.2, we have:

- $N \leq 10$: all the new dwellings are surveyed
- $11 \leq N \leq 40$: 10 dwellings are selected by simple random sampling
- $41 \leq N$: 1/4 of the dwellings are selected by SRS

It is easy to deduce N from n:

- $n < 10$: N = n
- $n > 10$: N = $4 \times n$
- $n = 10$: N is unknown. We only know that N∈ $[10, 40]$. We chose N=40, which is the most conservative solution.

The number of areas with n=10 is very small (8 areas for the first quarter 2003), so the uncertainty as to the value of N when n=10 should not be very problematic.

3.3 *Creation of the survey dataset*

All the selected dwellings, including the out-of-scope and the non-responding ones, were recorded in the survey dataset. Identifying variables for the sampling units, i.e. the groups of PSUs and the areas were added. No serious difficulty happened at this stage.

Table 1: Estimated standard errors for totals (first quarter 2003)

			Confidence interval at 95%			
Indicator	**Value**	**Standard error**	**Lower bound**	**Upper bound**	**CV*(%)**	**Design effect**
		Number of unemployed people by age, group and gender				
Total	2 684 701	56 992	2 572 996	2 796 406	2.1	1.85
Male	1 289 136	35 458	1 219 638	1 358 634	2.7	1.63
Female	1 395 565	34 244	1 328 447	1 462 683	2.4	1.37
15-29 years	918 009	30 190	858 836	977 182	3.3	1.66
Male	472 643	19 663	434 104	511 182	4.2	1.37
Female	445 366	18 157	409 778	480 954	4.1	1.29
30-49 years	1 305 894	33 650	1 239 940	1 371 848	2.6	1.41
Male	576 036	20 735	535 395	616 677	3.6	1.31
Female	729 858	24 002	682 815	776 901	3.3	1.29
50 years and over	460 798	21 785	418 099	503 497	4.7	1.56
Male	240 457	15 082	210 897	270 017	6.3	1.54
Female	220 341	12 959	194 941	245 741	5.9	1.16
		Number of unemployed people by NUTS2 region				
Ile De France	541 076	34 361	473 727	608 425	6.3	3.28
Rhône-Alpes	221 330	14 382	193 141	249 519	6.5	1.26
Auvergne	43 763	6 308	31 398	56 127	14.4	1.29
Nord - Pas de Calais	219 867	15 418	189 648	250 086	7.0	1.44

*Coefficient of variation = Standard error/value

4 Numerical results[2]

Estimated standard errors for totals (number of unemployed people) and ratios (unemployment rate) have been calculated using POULPE and are set out in Table 1 and Table 2.

The coefficients of variation for the subpopulation indicators (the numbers of unemployed people or the unemployment rates by age group and gender, by NUTS2 region) appear to be higher than for the population indicators. This is a basic result, which can be explained by the smaller sample size at subpopulation level. On the other hand, domain estimators seem to have lower design effects. This could be figured out by considering the impact of intra-class correlation. For simplicity, all the dwellings are assumed to have the same composition: one father, one mother, one son and one daughter. The design effect for the total number of unemployed people assuming a simple random selection of dwellings is (Cochran 1977; Ardilly 2006):

$$Deff \approx 1 + \rho \cdot (\bar{n} - 1) \tag{12}$$

Where ρ is the intra-class correlation coefficient of the 0/1 variable "unemployed or not" and $\bar{n}$ is the average household size ($\bar{n} = 4$). Let us consider now the total number of unemployed people in the male population m. The expression (12) is still valid, but each of its terms has to be calculated at the male population level:

$$Deff_m \approx 1 + \rho_m \cdot (\bar{n}_m - 1) \tag{13}$$

As $\bar{n}_m = \frac{\bar{n}}{2}$ and $\rho_m \approx \rho, Deff_m$ ought to be lower than *Deff*.

Besides, it is worth mentioning the design effects for the number of unemployed people and the unemployment rate in the Ile De France region (3.28 and 2.65), and for the unemployment rate in the Nord-Pas de Calais region (1.44 and 0.78). The two first values are likely to result from strongly positive intra-class correlation in the region. On the contrary, the value for the Nord-Pas de Calais region might be caused by small intra-class correlation, i.e. the dwelling population in this region is similar to the national dwelling population regarding unemployment characteristics.

Another advantage of POULPE is it can measure the impact on the accuracy of weight adjustments to external sources (calibration procedure). The LFS sample was calibrated to many external sources, basically census data at household level (number of rooms, household tenure status. . .) and updated data at individual level (population counts by age group and gender. . .). POULPE was run assuming no weight calibration in order to measure the impact the LFS calibration model had on the estimated standard errors for the main indicators. The results are set out in Table 3.

In our calculations, calibration always makes the accuracy better[3], but the impact varies depending on the indicator. In general, the better the calibration model is, the stronger the impact of the procedure. That impact appears to be weak for the subpopulation indicators. Actually, the LFS calibration model is a good fit for unemployment. At subpopulation level, unemployment patterns can become more complex and

[2] Source: Osier (2003)

[3] The negative relative difference (-0.06) for the number of unemployed people in the Rhône-Alpes region must not be significant.

Table 2: Estimated standard errors for ratios (first quarter 2003)

			Confidence interval at 95%			
Indicator	**Value**	**Standard error**	**Lower bound**	**Upper bound**	**CV*(%)**	**Design effect**
		Unemployment rate by age, group and gender (%)				
Total	9.9	0.22	9.5	10.3	2.2	1.96
Male	8.8	0.26	8.3	9.3	2.9	1.68
Female	11.2	0.28	10.6	11.7	2.5	1.46
15-29 years	16.9	0.57	15.8	18.0	3.4	1.64
Male	15.9	0.68	14.5	17.2	4.3	1.34
Female	18.2	0.76	16.7	19.7	4.2	1.37
30-49 years	8.6	0.24	8.2	9.1	2.8	1.47
Male	7.1	0.28	6.6	7.7	3.9	1.33
Female	10.4	0.35	9.7	11.0	3.7	1.35
50 years and over	7.1	0.33	6.4	7.7	4.6	1.58
Male	6.8	0.43	5.9	7.6	6.3	1.57
Female	7.4	0.42	6.6	8.2	5.7	1.14
		Unemployment rate by NUTS2 region (%)				
Ile De France	10.0	0.59	8.8	11.1	5.9	2.65
Rhône-Alpes	12.6	0.92	10.8	14.4	7.3	1.70
Auvergne	8.6	0.52	7.6	9.7	6.0	1.22
Nord – Pas de Calais	7.6	0.91	5.9	9.4	12.0	0.78

*Coefficient of variation = Standard error/value

Table 3: Impact of calibration on the estimated standard errors

Indicator	**Value**	**Standard error before calibration**	**Standard error after calibration**	**Relative difference (%)**
		Number of unemployed people by age, group and gender		
Total	2 684 701	67559	56 992	15,6
Male	1 289 136	39328	35 458	9.8
Female	1 395 565	40531	34 244	15.5
		Number of unemployed people by NUTS2 region		
Ile De France	541 076	34512	34 361	0.44
Rhône-Alpes	221 330	14374	14 382	-0.06
		Unemployment rate by age, group and gender (%)		
Total	9.9	0.23	0.22	4.35
Male	8.8	0.26	0.26	0.00
Female	11.2	0.30	0.28	6.67
		Unemployment rate by NUTS2 region (%)		
Ile De France	10.0	0.59	0.59	0.00
Rhône-Alpes	12.6	0.92	0.92	0.00

then the quality of the model may decrease, which may explain the loss of efficiency at domain level. For a general discussion on this point, see Ardilly (2006).

5 Conclusion

In our opinion, the variance estimation method that was presented in this document has three advantages:

- It has strong theoretical foundations.
- Contrary to re-sampling methods (Bootstrap, Jackknife…), it is not computer-intensive.
- It is easily reproducible once the design, the survey and the geographical datasets have been created.

However, although POULPE has been developed as a universal variance estimation tool, it cannot easily deal with ”highly” complex sample designs. For instance, each quarter, 1/6 of the LFS sample rotates out. The difference between unemployment rates estimated at two consecutive quarters is affected by covariance effects between those quarters. Variance estimation in POULPE taking this aspect into account is conceptually much more difficult to handle.

References

Ardilly, P. (2006). *Les techniques de sondage* (2 ed.). Paris: Technip.

Binder, D., & Patak, Z. (1994). Use of estimating functions for estimation from complex surveys. *Journal of the American Statistical Association*, *89*, 1035-1043.

Caron, N. (1998). Le logiciel poulpe: aspects méthodologiques. *Proceedings of the Journées de Méthodologie Statistique 1998*, 173-200.

Cochran, W. G. (1977). *Sampling techniques*. New York: Wiley.

Deville, J. C. (1999). Variance estimation for complex statistics and estimators: linearization and residual techniques. *Survey Methodology*, *25*, 219-230.

Deville, J. C., & Särndal, C. E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, *87*, 376-382.

Deville, J. C., & Tillé, Y. (2005). Variance approximation under balanced sampling. *Journal of Statistical Planning and Inference*, *128*, 569-591.

Osier, G. (2003). *Utilisation du logiciel poulpe pour des calculs de précision sur l'enquête emploi en continu* (Internal Report). INSEE.

Raj, D. (1968). *Sampling theory*. New York: Mc Graw-Hill.

Särndal, C. E., Swensson, B., & Wretman, J. (1992). *Model assisted survey sampling*. New York: Springer.

Wolter, K. M. (1985). *Introduction to variance estimation*. New York: Springer.

Woodruff, R. (1971). A simple method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association*, *66*(334), 411-414.

Appendix: Estimation of the Design Effect in POULPE

1 Definition

Let $\hat{Y}$ denote the linear estimator for the total Y of a variable y with respect to a sample design P. Let $\tilde{n}$ be the (expected) sample size. Let $\hat{Y}_{SRS}$ be the linear estimator that would be obtained from a simple random sampling (SRS) without replacement and of size $\tilde{n}$. Then, the Design Effect is:

$$\mathit{Deff} = \frac{V_P\left(\hat{Y}\right)}{V_{SRS}\left(\hat{Y}_{SRS}\right)}$$

2 Estimation in case of a one-phase sampling

Preliminary notations:

U = *target population, of size N*
s = *sample of size n, drawn from U according to a sample design P*
y_k = *value of a target variable y on k*
π_k = *inclusion probability of k*

Each of the two variances above is estimated separately. An estimate of $V_P\left(\hat{Y}\right)$ is the result of a POULPE session. What remains is estimating the variance $V_{SRS}\left(\hat{Y}_{SRS}\right)$.

1. We have:

$$V_{SRS}\left(\hat{Y}_{SRS}\right) = N^2 \cdot \left(1 - \frac{n}{N}\right) \cdot \frac{S^2}{n}$$

Where:

$$S^2 = \frac{1}{N-1} \sum_{k \in U} \left(y_k - \bar{Y}\right)^2 = \frac{N}{N-1} \cdot \left(\frac{1}{N} \sum_{k \in U} y_k^2 - \bar{Y}^2\right)$$

2. Hence:

$$\begin{aligned} V_{SRS}\left(\hat{Y}_{SRS}\right) &= N^2 \cdot \left(1 - \frac{n}{N}\right) \cdot \frac{1}{n} \cdot \frac{N}{N-1} \cdot \left(\frac{1}{N} \sum_{k \in U} y_k^2 - \bar{Y}^2\right) \\ &= \frac{1}{n} \cdot \left(1 - \frac{n-1}{N-1}\right) \cdot \left[N \cdot \sum_{k \in U} y_k^2 - \left(\sum_{k \in U} y_k\right)^2\right] \end{aligned}$$

3. With respect to the sample design P, we have:

$$\sum_{k \in U} y_k^2 \quad \text{estimated by} \quad \sum_{k \in s} \frac{y_k^2}{\pi_k}$$

$$\left(\sum_{k \in U} y_k\right)^2 \quad \text{estimated by} \quad \left(\sum_{k \in s} \frac{y_k}{\pi_k}\right)^2 - \hat{V}_P\left(\hat{Y}\right)$$

Indeed,

$$\begin{aligned} E_P\left(\sum_{k \in s} \frac{y_k}{\pi_k}\right)^2 &= V_P\left(\sum_{k \in s} \frac{y_k}{\pi_k}\right) + E_P^2\left(\sum_{k \in s} \frac{y_k}{\pi_k}\right) \\ &= V_P\left(\sum_{k \in s} \frac{y_k}{\pi_k}\right) + \left(\sum_{k \in U} y_k\right)^2 \end{aligned}$$

4. Then, we have:

$$\hat{V}_{SRS}\left(\hat{Y}_{SRS}\right) = \frac{1}{n} \cdot \left(1 - \frac{n-1}{N-1}\right) \cdot \left[N \cdot \sum_{k \in s} \frac{y_k^2}{\pi_k} - \left(\sum_{k \in s} \frac{y_k}{\pi_k}\right)^2 + \hat{V}_p\left(\hat{Y}\right)\right]$$

5. Replacing $\hat{V}_p\left(\hat{Y}\right)$ with $D\hat{e}\mathit{ff} \cdot \hat{V}_{SRS}\left(\hat{Y}_{SRS}\right)$ and rearranging the expression, we get:

$$\hat{V}_{SRS}\left(\hat{Y}_{SRS}\right)\cdot\left[1-\frac{D\hat{e}ff}{n}\cdot\left(1-\frac{n-1}{N-1}\right)\right]$$
$$=\frac{1}{n}\cdot\left(1-\frac{n-1}{N-1}\right)\cdot\left[N\cdot\sum_{k\in s}\frac{y_k^2}{\pi_k}-\left(\sum_{k\in s}\frac{y_k}{\pi_k}\right)^2\right]$$

6. Finally, by replacing N with $\hat{N}=\sum_{k\in s}\frac{1}{\pi_k}$ and $\left[1-\frac{D\hat{e}ff}{n}\cdot\left(1-\frac{n-1}{N-1}\right)\right]$ with 1, we obtain the following approximation formula for the Deff:

$$D\hat{e}ff=\frac{\hat{V}_p\left(\hat{Y}\right)}{\frac{1}{n}\cdot\left(1-\frac{n-1}{\hat{N}-1}\right)\cdot\hat{N}\cdot\sum_{k\in s}\frac{1}{\pi_k}\left(y_k-\bar{y}\right)^2}$$

Where $\bar{y}=\frac{\sum_{k\in s}\frac{y_k}{\pi_k}}{\sum_{k\in s}\frac{1}{\pi_k}}$

3 Estimation in case of a multi-phase sampling

We have:

$$\hat{V}_{SRS}\left(\hat{Y}_{SRS}\right) = \frac{1}{r}\cdot\left(1-\frac{r-1}{\hat{N}-1}\right)\cdot\hat{N}\cdot\sum_{k\in s}\frac{1}{\pi_k}\left(y_k-\bar{y}\right)^2$$

Where r denotes the effective sample size and π_k the inclusion probability of k.