Estimation de variance dans les enquêtes de l'Insee : le package R gustave [1]

Nicolas Paliod [1] & Martin Chevalier [2]

[1] *Insee, 88 avenue Verdier, 92120 Montrouge, nicolas.paliod@insee.fr*
[2] *Insee, martin.chevalier@insee.fr*

Résumé. Le *package* R gustave, développé par le Département des méthodes statistiques de l'Insee, est une nouvelle proposition visant à faciliter la mise en œuvre d'estimation de variance analytique pour tout type d'enquête. Il repose sur deux ensembles de fonctions : d'une part des fonctions méthodologiques, qui mettent en œuvre les estimateurs de variance classiques (Sen-Yates-Grundy, Deville-Tillé) ; d'autre part des fonctions techniques qui créent, à partir des fonctions méthodologiques, des outils d'estimation de variance autonomes et ergonomiques dont l'utilisation ne requiert aucune expertise méthodologique particulière. Le *package* gustave constitue ainsi une véritable « boîte à outils » pour le méthodologue qui systématise et simplifie la production de programmes d'estimation de variance tenant compte des spécificités de chaque enquête.

Mots-clés. Estimation de variance, qualité dans les enquêtes, techniques d'enquêtes avec le logiciel R

Abstract. The R *package* gustave, developed by Insee's statistical methods department, is a new proposal aiming at simplifying analytical variance computation for all kind of survey. It groups two sets of functions : on the one hand some methological functions compute usual variance estimators (Sen-Yates-Grundy, Deville-Tillé) ; on the other hand some technical functions which generate some self-contained and ergonomic tools of variance estimation, using the methological functions. Their use does not require any particular methological expertise. Therefore, the *package* gustave is a real "toolbox" for the methodologist. It systematizes and makes easier the production of variance estimation program taking account of each survey's specificities.

Key words. Variance estimation, survey quality, survey sampling with R

La mesure de la précision des résultats issus d'enquêtes statistiques constitue un enjeu de plus en plus important dans leur diffusion et leur analyse. Mesure de la qualité du processus de production statistique, les indicateurs de précision figurent dans les rapports communiqués périodiquement à Eurostat ainsi que dans certains règlements européens, en particulier le projet de règlement Integrated European Social Statistics (IESS) actuellement en discussion. Ils sont par ailleurs toujours plus mobilisés pour guider le commentaire

1. Gustave : a User-oriented Statistical Toolkit for Analytical Variance Estimation (`https://github.com/martinchevalier/gustave`)

des résultats d'une enquête, en particulier quand ceux-ci sont ventilés par domaine, secteur et tranche d'effectif dans les enquêtes auprès des entreprises ou par région dans les enquêtes auprès des ménages par exemple.

L'imprécision liée à un dispositif d'enquête renvoie à de nombreuses dimensions : celle-ci peut être liée à l'échantillonnage ou à la non-réponse, mais aussi à l'éventuelle imperfection de la base de sondage ou aux erreurs de mesure et d'observation (Biemer, 2010). Cette contribution présente les outils développés au sein du Département des méthodes statistiques de l'Insee pour estimer la variance associée à l'échantillonnage et à la non-réponse dans le cadre méthodologique défini par (Gros, Moussallam, 2015) et (Gros, Moussallam, 2016) et présenté lors des Journées de méthodologie statistiques de l'Insee 2015 (Chevalier, Gros, Moussallam, 2015).

Plusieurs outils ont été proposés pour faciliter la mise en œuvre d'estimation de variance dans ce type de contexte, en particulier la macro SAS Poulpe (Caron N., 1998, Petit J.-N., 1998). Ce type d'outil fait face à une gageure : il doit à la fois présenter un haut niveau de généralité, au sens de sa capacité à prendre en compte des plans de sondage et des redressements complexes, et en même temps rester suffisamment ergonomique pour être utilisé en pratique lors de l'exploitation de l'enquête.

Le *package* R gustave est une nouvelle proposition allant en ce sens. Son ambition est de fournir au méthodologue une « boîte à outils » souple et extensible qui facilite la création de programmes d'estimation de variance autonomes et simples d'utilisation. Ce *package* comporte deux ensembles de fonctions :

- d'une part, des fonctions méthodologiques qui facilitent la mise en œuvre de l'estimation de variance proprement dite : estimateurs de variance de Sen-Yates-Grundy (Sen, 1953, Yates-Grundy, 1953) et de Deville-Tillé (Deville-Tillé, 2005), prise en compte du calage sur marges (Deville, Särndal, 1992), fonctions de linéarisation usuelles (Caron, 1998), etc. ;
- d'autre part, des fonctions techniques qui créent, à partir des fonctions méthodologiques, des outils d'estimation de variance autonomes et ergonomiques (valeurs par défaut pour les paramètres techniques, contrôle et validation des données, évaluation non-standard [Wickham, 2014] notamment).

Ce faisant, le *package* R gustave simplifie la création d'outils d'estimation de variance sans imposer de contrainte *a priori* sur la forme ou la complexité de la fonction de variance utilisée, qui relève de l'expertise du méthodologue. Une fonction de variance « prête-à-estimer » pour les cas les plus simples (tirage à un degré stratifié, prise en compte de la non-réponse et du calage s'il y a lieu) est par ailleurs en cours de développement.

Cette présentation revient spécifiquement sur les principes méthodologiques qui ont guidé la construction du *package* gustave : choix d'implémentation pour les estimateurs de variance et les fonctions de linéarisation usuels, ergonomie générale du *package* et exemple d'utilisation sur un cas pratique.

Bibliographie

Biemer, P. (2010), "Total survey error : Design, implementation, and evaluation", *Public Opinion Quarterly*, Vol. 74, No. 5, 817–848

Caron, N. (1998), « Le logiciel Poulpe : Aspects méthodologiques », *Actes des journées de méthodologie statistique 1998*

Chevalier, M., Gros, E. et Moussallam, K. (2015), « Les calculs de précision dans Octopusse : Théorie et application à l'enquête Logement 2013 », *Actes des journées de méthodologie statistique 2015*

Deville, J.-C. et Särndal (1992), C.-E., "Calibration Estimators in Survey Sampling", *Journal of the American Statistical Association*, Vol. 87, 376-382

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

Gros, E. et Moussallam, K. (2015), *Les méthodes d'estimation de la précision pour les enquêtes ménages de l'Insee tirées dans Octopusse*, Document de travail de l'Insee M 2015/03

Gros, E. et Moussallam, K. (2016), *Les méthodes d'estimation de la précision de l'enquête Emploi en continu*, Document de travail de l'Insee M 2016/02

Petit, J.-N. (1998), « Le logiciel Poulpe : Modélisation informatique », *Actes des journées de méthodologie statistique 1998*

Sen, A. R. (1953). "On the estimate of the variance in sampling with varying probabilities", *Journal of the Indian Society of Agricultural Statistics*, 5, 119-127

Wickham, H. (2014), *Advanced R*, CRC press, 456 pp. (`http://adv-r.had.co.nz/`)

Yates, F. et Grundy, P. M. (1953), "Selection without replacement from within strata with probability proportional to size", *Journal of the Royal Statistical Society*, Series B, 15, 235-261