Estimation de variance dans les enquêtes de l'Insee : le *package* R gustave et ses applications

Martin Chevalier (Insee, DMS)

13ème journées de méthodologie statistique
Session 13 : Calcul de précision

Paris, 13 juin 2018

Estimation de variance dans les enquêtes de l'Insee

L'estimation de variance est une opération qui gagne en importance dans le **processus de production d'une enquête** :

- outil pour évaluer la qualité de l'information collectée susceptible d'influer sur sa diffusion :
- indicateur utilisé dans les rapports qualité mais aussi dans le nouveau règlement européen IESS (*Integrated european social statistics*) en discussion.

Remarque Cette présentation porte sur une composante parmi d'autres de l'imprécision associée à un processus d'enquête.

Cette contribution présente :

- la stratégie mise en œuvre à l'Insee pour mener à bien l'estimation de variance sur des enquêtes complexes ;
- le *package* R gustave développé au sein du Département des méthodes statistiques (DMS).

Plan de la présentation

Objectif : rendre le calcul de précision (plus) simple

Exemple d'applications de gustave à l'Insee

Principe de fonctionnement du *package* gustave

Objectif : rendre le calcul de précision (plus) simple

Objectif : rendre le calcul de précision (plus) simple

Sources de complexité du calcul de précision

Plan de sondage

- algorithmes de tirage ;
- tirages à plusieurs degrés ;
- bases de sondage multiples.

Méthodes d'estimation

- correction de la non-réponse ;
- calage sur marges.

Estimateurs

- linéarisation : ratio, quantiles, indicateurs de pauvreté, etc. ;
- estimation sur des domaines.

Objectif : rendre le calcul de précision (plus) simple

Première solution : le logiciel Poulpe

Poulpe (**P**rogramme **O**ptimal et **U**niversel pour la **L**ivraison de la **P**récision des **E**nquêtes) est une macro SAS de calcul de précision présentée lors des $6^{\text{ème}}$ JMS (1998) :

- estimateurs de variance s'appuyant sur les probabilités d'inclusion simple ;
- modélisation générique du plan de sondage et des phases de redressement ;
- modules de linéarisation intégrés.

La présente contribution **s'appuie sur les travaux associés à Poulpe** avec quelques différences notables :

- aucune restriction sur le type d'estimateur de variance (échantillon-maître Octopusse) ;
- simplification de la mise en œuvre du calcul pour le non-spécialiste.

Objectif : rendre le calcul de précision (plus) simple

Nouvelle proposition (1) : procéder en deux étapes

Difficultés de l'estimation de variance en pratique :

- construire une modélisation méthodologiquement cohérente du plan de sondage et des redressements ;
- disposer des données nécessaires pour les prendre en compte.

Proposition organisationnelle : bien distinguer **deux étapes**

1. **Méthodologue** : analyse méthodologique, mobilisation de l'information auxiliaire, construction d'un **programme d'estimation de variance** raisonnablement exact ;
2. **Responsable d'enquête, chargé(e) d'étude** : utilisation du programme d'estimation dans le cadre d'études ou pour répondre à des obligations réglementaires.

Objectif : rendre le calcul de précision (plus) simple

Nouvelle proposition (2) : le *package* R gustave

Conséquence : les programmes d'estimation de variance doivent donc

- être autonomes et aussi simples d'utilisation que possible ;
- prendre en compte l'ensemble des éléments relatifs au calcul de précision (linéarisations, domaines, etc.) ;
- ne pas être trop complexes à développer ni à maintenir.

Proposition technique : *package* R **G**ustave : a **U**ser-oriented **S**tatistical **T**oolkit for **A**nalytical **V**ariance **E**stimation

- **Faciliter** la mise en œuvre du calcul de précision par tout un chacun...
- ...en fournissant au ou à la méthodologue des **outils dédiés**.

Exemple d'applications de gustave à l'Insee

Le *package* gustave à l'Insee

Utilisé pour l'estimation de variance des enquêtes ménages périodiques : Enquête emploi en continu (EEC), dispositif Statistique sur les revenus et les conditions de vie (SRCV), Cadre de vie et sécurité (CVS), Loyers et charges.

Exemple Enquête emploi en continu

- panel de logements initialisé en 2009, tirage équilibré ;
- correction de la non-réponse par calage en une étape ;
- indicateurs standards : ratios (taux de chômage, etc.) ventilés par domaine.

Nota bene *Les estimateurs ponctuels figurant sur les diapositives suivantes ne coïncident en général pas avec la diffusion officielle (champs de calcul différents, pas de désaisonnalisation, etc.).*

Les fichiers de calcul de précision

Le *package* gustave permet de produire, pour chaque millésime d'une enquête (chaque trimestre pour l'EEC) un **fichier de données R** qui contient :

- les micro-données de l'enquête (table `z` pour l'EEC) ;
- les programmes d'estimation de variance spécifiques à l'enquête (fonction `precisionEec()` pour l'EEC) ;
- toute l'information auxiliaire nécessaire.

Pour mettre en œuvre l'estimation de variance, **il suffit de charger ce fichier** (par exemple pour le T2 2014) :

```
load("precisionEec142.RData")
```

Remarque Ces fichiers de calcul de précision sont susceptibles de contenir des **informations auxiliaires réidentifiantes**.

Code : Précision du taux de chômage au T2 2014

Nombre total de chômeurs (acteu %in% 2)

```
precisionEec(z, total(acteu %in% 2))
```

```
##                         call     est   variance      std
## 1 total(y = acteu %in% 2) 3001046 2158830156 46463.21
##          cv   lower   upper
## 1 1.548234 2909980 3092112
```

Taux de chômage

```
precisionEec(z, ratio(acteu %in% 2, acteu %in% c(1, 2)))
```

```
##                                                    call
## 1 ratio(num = acteu %in% 2, denom = acteu %in% c(1, 2))
##         est   variance         std       cv     lower
## 1 0.1044647 2.5918e-06 0.001609907 1.541101 0.1013094
##       upper
## 1 0.1076201
```

Code : Précision du taux de chômage au T2 2014

Taux de chômage des 50 ans et plus

```
precisionEec(z,
  ratio(acteu %in% 2, acteu %in% c(1, 2)),
  where = age >= 50
)
```

```
##          est     variance          std       cv
## 1 0.07047492 5.402617e-06 0.002324353 3.298128
```

Taux de chômage par région

```
precisionEec(z,
  ratio(acteu %in% 2, acteu %in% c(1, 2)),
  by = reg
)
```

```
##   by       est     variance          std       cv
## 1 11 0.1003089 1.538408e-05 0.003922254 3.910175
## 2 21 0.1130015 1.068723e-04 0.010337904 9.148463
## 3 22 0.1220682 9.565600e-05 0.009780388 8.012235
```

Principe de fonctionnement du *package* gustave

Principe de fonctionnement du *package* gustave

« Emballer » (*wrap*) la complexité

L'objectif du *package* gustave est de **préserver l'utilisateur final de la complexité** du processus d'estimation de la variance.

Idée centrale « Emballer » la fonction d'estimation de variance complexe dans une autre fonction (appelée « *wrapper* ») plus simple d'utilisation :

- **fonction d'estimation de la variance** : fonction spécifique à chaque enquête développée par le ou la méthodologue ; → **complexité méthodologique**
- ***wrapper* d'estimation de variance** : fonction générique qui prend en charge des opérations systématiques (linéarisations, domaines), appelle la fonction de variance et affiche les résultats. → **complexité informatique**

Apports du *package* gustave

La production d'un programme d'estimation de variance avec le *package* gustave suppose en général **trois étapes pour le ou la méthodologue** :

1. Élaborer la fonction de variance spécifique à l'enquête
   $\rightarrow$ gustave propose des **fonctions optimisées** qui mettent en œuvre les estimateurs de variance standard.

2. Définir le *wrapper* de variance
   $\rightarrow$ gustave **simplifie la production** de *wrappers* de variance faciles à utiliser et intégrant toute l'information auxiliaire nécessaire.

3. Définir des linéarisations *ad hoc* si nécessaire
   $\rightarrow$ gustave permet l'**interaction** entre *wrappers* de variance et fonctions de linéarisation.

Diffusion et perspectives

- Version 0.3.0 en ligne sur le CRAN
- Code source accessible sur github.com : https://github.com/martinchevalier/gustave
- Maintenance assurée par la division Sondages de l'Insee
- Fonctionnalités en développement :
  - création d'une fonction « prête-à-estimer » pour les cas les plus simples (SAS stratifié, repondération dans des GRH, calage) similaire à la macro SAS %everest ;
  - intégration dans le *package* de linéarisations plus complexes.

En guise de conclusion

Le Département des méthodes statistiques a mis en place une organisation pour **industrialiser l'estimation de variance** :

- processus systématisé et documenté ;
- programmes simples d'utilisation et faciles à diffuser ;
- fichiers d'estimation de variance résilients.

Le développement du *package* gustave constitue un **investissement important** :

- présenté en *workshop* européen ;
- utilisé pour vérifier le respect des objectifs de précision prévus par le réglement IESS ;
- qui inscrit avec d'autres les travaux du DMS et de l'Insee dans l'univers du logiciel libre.

Merci de votre attention !

Martin Chevalier
martin.chevalier@insee.fr
https://github.com/martinchevalier/gustave