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

Nicolas Paliod, Martin Chevalier

10ème Colloque francophone sur les sondages
Session Estimation de variance et Robustesse

26 octobre 2018

Pourquoi calculer la variance d'un indicateur ?

- L'estimation de variance est importante pour le chargé d'études :
  - Permet de disposer d'un intervalle de confiance
  - Permet de commenter la significativité des variations de l'indicateur
- L'estimation de variance permet de mesurer la qualité des indicateurs produits
  - Utilisée pour les rapports qualité transmis à Eurostat
  - Requise pour divers indicateurs dans différentes enquêtes par le nouveau règlement européen IESS (*Integrated european social statistics*) en discussion
- L'estimation de variance est donc une opération qui gagne en importance dans le processus de production d'une enquête

Les sources d'erreurs dans une enquête

Chaque maillon de la chaîne de production d'une enquête peut être source d'erreur :

- Base de sondage
- Plan de sondage
- Taille de l'échantillon
- Estimateur retenu
- Spécification des questions
- Réponse aux questions
- Non-réponse
- Chaînes de traitement des données

Ces erreurs recouvrent biais et incertitude.

Les parties de la chaîne de production prises en compte par les estimations de variance à l'Insee

- Eléments liés au plan de sondage :
  - Algorithmes de tirage
  - Degrés de tirage
  - Bases de sondage multiples
- Eléments liés aux méthodes d'estimation
  - Correction de la non-réponse
  - Calage sur marges

Des solutions existantes mais imparfaites (1/2)

- macro SAS %calker : sondage aléatoire simple stratifié, correction de la non-réponse par repondération ou par imputation au sein des strates de tirage
  => pas adapté à des groupes de réponse homogènes

- macro SAS %calker_grh : sondage aléatoire simple stratifié, correction de la non-réponse par repondération au sein de groupes de réponse homogènes quelconques
  => pas adapté à des plans de sondage complexes

Des solutions existantes mais imparfaites (2/2)

- macro SAS %everest : sondage aléatoire simple stratifié, correction de la non-réponse par imputation au sein des strates de tirage ou par repondération au sein de groupes de réponse homogènes quelconques, calage sur marges
  => pas adapté à des plans de sondage complexes

- macro SAS %Poulpe : 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
  => pas adapté à l'utilisation en production par un chargé d'études

Plan

1 Objectifs du package

2 Fonctionnement du package

3 Principe et outils de développement

Trois grands objectifs

package R Gustave : a User-oriented Statistical Toolkit for Analytical Variance Estimation

3 objectifs axés autour de :

- user
- toolkit
- analytical variance estimation

Un *package* de calcul de variance analytique

- Systématiser le calcul de variance en s'abstrayant des éléments communs à tous les calculs de variance, à toutes les enquêtes

- Le *package* permet de prendre en compte simplement :
  - le calage sur marges (fonction *res_cal*)
  - l'estimation sur un domaine (arguments *by*, *where*)
  - la linéarisation (fonctions de linéarisation *mean*, *ratio*, *diff_of_ratio*, *ratio_of_ratio*)

Un *package* orienté utilisateur

- Simplifier le calcul de variance en limitant le travail du méthodologue au codage de la fonction d'estimation de variance analytique adapté à l'enquête

- Standardiser la mise en forme des fonctions de calcul de variance, avec des fonctions déjà présentes dans le package, pour permettre au chargé d'études de récupérer une fonction d'estimation de variance simple d'utilisation

Une boîte à outils

Un *package* qui permet au méthodologue :

- d'intégrer dans la fonction d'estimation de variance n'importe quelle étape de l'enquête pourvu qu'il existe un calcul analytique qui puisse être codé
- de disposer de fonctions déjà codées

Faire interagir les différents acteurs du processus de production

Le *package* permet à chaque acteur du processus de production de limiter son travail à son champ d'action :

- le chargé d'études pour la production d'estimations de variance d'indicateurs
- le méthodologue pour la production de la fonction d'estimation de variance adaptée à l'enquête
- le développeur pour améliorer l'ergonomie des fonctions de variance produites par le package, pour intégrer de nouvelles fonctionnalités

Plan

1 Objectifs du package

2 Fonctionnement du package

3 Principe et outils de développement

Le *wrapper* de variance

Parmi les objectifs :

- intégrer des fonctionnalités communes pour toutes les enquêtes comme l'estimation sur un domaine
- avoir une mise en forme standard des fonctions de variance pour simplifier leur utilisation

Solution mise en œuvre : le *wrapper* de variance

- define_variance_wrapper() : fonction générique qui prend en charge des opérations systématiques (statistiques de linéarisation, domaines), appelle la fonction d'estimation de variance et affiche les résultats

Des fonctions de variance déjà codées dans le *package*

Le *package* contient un certain nombre de fonctions utiles dans différentes enquêtes :

- des fonctions d'estimation de variance analytique
  - Variance de Sen-Yates-Grundy
  - Variance de Deville-Tillé (Deville, Tillé, 2005)
- des statistiques de linéarisation (*mean*, *ratio*, *diff_of_ratio*, *ratio_of_ratio*)
- une fonction *res_cal* pour prendre en compte le calage

L'utilisation du 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.)

Exemple de calcul de variance à l'Insee (1/3)

- Première phase : préparation des données pour qu'elles contiennent toutes les variables ensuite utilisées par la fonction de variance codée en deuxième phase

- Deuxième phase : codage de la fonction de variance

```
varEec <- function(y, up, log, ind){

  variance <- list()

  # Etape 0 : Agrégation par logement
  y <- sum_by(y, by = ind$idlog)

  # Etape 1 : Prise en compte du calage
  y <- add_zero(y, log$id[log$cal])
  y <- res_cal(y, precalc = log$res_cal_precalc)
```

Exemple de calcul de variance à l'Insee (2/3)

```
  (...)

  # Etape 2 : Prise en compte de la non-réponse
  variance[["nr"]] <- colSums(
    (1/log$pilog[log$cal]^2 - log$qlog[log$cal]) *
      (1 - log$pinr[log$cal]) * (y/log$pinr[log$cal])^2
  )
  y <- add_zero(y / log$pinr[log$cal], log$id)

  # Etape 3 : Sélection des logements dans les up
  variance[["log"]] <- varDT(
    y, w = 1/(log$piup^2) - log$qup,
    precalc = log$varDT_precalc
  )

  # Etape 4 : Sélection des up
  y <- sum_by(y, by = log$idup, w = 1/log$pilog_up)
  y <- add_zero(y, up$id)
  variance[["up"]] <- varDT(y, precalc = up$precalc)

  colSums(do.call(rbind, variance))

}
```

Exemple de calcul d'estimation de variance à l'Insee (3/3)

- Troisième phase : à partir de la fonction de variance et de l'information auxiliaire nécessaire, la fonction define_variance_wrapper() crée un *wrapper* de variance simple d'utilisation

```
# Création du wrapper de variance avec define_variance_wrapper()
precisionEec <- define_variance_wrapper(
  variance_function = varEec,
  technical_data = list(up = up, log = log, ind = ind),
  reference_id = technical_data$ind$id,
  reference_weight = technical_data$ind$w,
  default_id = quote(paste0(ident, noi))
)

# Utilisation du wrapper de variance (données du T4 2014)
precisionEec(z, acteu %in% 2)
```

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

- Remarque pour la diffusion de la fonction de variance produite :

  Le *wrapper* de variance est une fonction complètement autonome : toute l'information auxiliaire spécifiée au paramètre technical_data est intégrée dans la fonction (il s'agit d'une *closure*)

La fonction qvar()

- La fonction qvar() : « une fonction prête-à-estimer » pour des plans de sondage simples et des redressements standards
- La fonction qvar() combine les autres fonctions du package gustave
- La fonction qvar() s'applique dans un cadre similaire à la macro SAS %everest :
  - sondage aléatoire simple stratifié
  - correction de la non-réponse par repondération dans des groupes de réponse homogènes
  - calage sur marges

Un *package* pensé pour être extensible (1/3)

La fonction define_variance_wrapper() accepte n'importe quelle fonction de variance en entrée :

- autant d'information auxiliaire que nécessaire
- utilisation des fonctions d'autres *packages* pour coder la fonction de variance (utiliser require() dans la fonction de variance)

Large éventail de méthodologies couvert à ce jour :

- échantillons tirés dans l'échantillon-maître Octopusse (formule spécifique dérivée de la formule de Sen-Yates-Grundy (Chauvet, 2011 et Gros, Moussallam, 2015))
- degrés multiples (CVS)
- partage des poids complexes (SRCV)

Un *package* pensé pour être extensible (2/3)

La fonction de variance peut exporter, en plus des variances estimées, des résultats intermédiaires de l'estimation de variance.

Cette fonctionnalité facilite la création de surcouches à partir des *wrappers* de variance produits par le *package* gustave.

Exemple : Dans l'EEC, l'estimation de variance pour des indicateurs faisant intervenir plusieurs trimestres (évolution d'un trimestre à l'autre, moyennes annuelles, etc.) s'appuie sur la récupération des résultats intermédiaires des *wrappers* de variance de chaque trimestre concerné (estimation des covariances trimestrielles).

Un *package* pensé pour être extensible (3/3)

Il est également possible de définir de nouvelles fonctions pour estimer la précision de statistiques complexes grâce à la fonction define_statistic_wrapper() :

```
# Définition du coefficient de gini à partir du package vardpoor
gini <- define_statistic_wrapper(
  statistic_function = function(y, weight){
    require(vardpoor)
    result <- lingini(Y = y, weight = weight)
    list(point = result$value$Gini, lin = result$lin$lin_gini)
  },
  arg_type = list(data = "y", weight = "weight", param = NULL)
)

# Utilisation pour calculer la précision dans l'enquête SRCV en 2014
precisionSrcv(r, gini(HX090))
```

```
##              call      est  variance       std      cv    lower    upper
## 1 gini(y = HX090) 29.21328 0.1013441 0.3183458 1.08973 28.58933 29.83722
```

Plan

1 Objectifs du package

2 Fonctionnement du package

3 Principe et outils de développement

Diffusion sur le Cran, tests unitaires et intégration en continu

- *package* disponible sur le Cran, dernière version : août 2018
- Le développement sous la forme de *packages* favorise également le développement de tests unitaires
  - à chaque fonctionnalité du *package* est associé un test qui vérifie son bon fonctionnement
  - gustave comporte plus de 180 tests unitaires
  - à chaque nouvelle version du *package*, des tests sont automatiquement réalisés : intégration en continu

Suivi de versions

Les évolutions du *package* sont suivies en version depuis l'été 2017 :

- code source librement accessible sur plusieurs plateformes de développement, notamment github.com
- conservation de toutes les versions (plus de 300 *commits* à ce jour) avec leurs métadonnées : une description est associée à chaque ensemble cohérent de modifications
- travail collaboratif facilité, y compris de façon concomittante : création de branches pour des développements particuliers, gestion des conflits
- possibilité pour des utilisateurs externes de proposer efficacement des modifications (remontée de *bugs*, demandes spécifiques, *pull requests*)

Conclusion

Le Département des méthodes statistiques de l'Insee a mis en place un *package* R pour systématiser l'estimation de variance :

- une solution qui permet à chaque participant de la chaîne de production à ne se préoccuper que de la partie qu'il a en charge
- des fonctions d'estimation de variance simples d'utilisation
- un *package* documenté

Un *package* déjà utilisé pour les enquêtes ménages de l'Insee :

- pour produire les estimations de variance jointes aux rapports qualité
- pour vérifier le respect des objectifs de précision prévus par le règlement IESS