Introduction à la théorie des sondages
Cours 4

Martin Chevalier
martin.chevalier@insee.fr

INSEE, département des méthodes statistiques

20 février 2017

Plan de la séance

1. Rappel des épisodes précédents
2. Complément sur la stratification : tirage systématique
3. Correction du biais de non-réponse
   - Non-réponse : Définition et origine
   - Conséquences de la non-réponse
   - Correction d'une non-réponse MCAR
   - Correction d'une non-réponse MAR

Chapitre 1

Rappel des épisodes précédents

Sondage et cadre de l'estimation Horvitz-Thompson

Objectif Estimer une statistique (par exemple le total ou la moyenne d'une variable Y) à partir d'un **échantillon** s et non de l'ensemble de la population U.

Stratégie Maîtriser intégralement la manière dont les unités à interroger sont tirées au sort :

- notion de **plan de sondage** ;
- **probabilités d'inclusion** simples π_k et doubles π_{kl}.

Estimateur d'Horvitz-Thompson

$$\hat{T}_{HT}(Y) = \sum_{k \in s} \frac{y_k}{\pi_k} = \sum_{k \in s} w_k y_k \quad \text{avec } w_k \text{ le } \textbf{poids de sondage}.$$

- Dès lors que $\forall k \in \mathcal{U}$, $\hat{T}_{HT}(Y)$ est sans biais ;
- Variance calculable et estimable (sans biais) sur l'échantillon s.

Application : Enquête sur le patrimoine

Contexte On souhaite obtenir des informations sur le **patrimoine** d'une population U de $N = 10\ 000$ ménages : montant, composition, origine, transmission, etc.

Plan de sondage On réalise une enquête par sondage en tirant un échantillon s taille $n = 100$ ménages par **sondage aléatoire simple** :

$$\forall k \in \mathcal{U} \quad \pi_k = \frac{n}{N} = \frac{100}{10\ 000} = 0,01$$

Estimateur On estime alors sans biais la moyenne du patrimoine par :

$$\hat{\bar{Y}}_{HT} = \frac{1}{n}\sum_{k \in s} y_k = \bar{y}$$

Application : Enquête sur le patrimoine

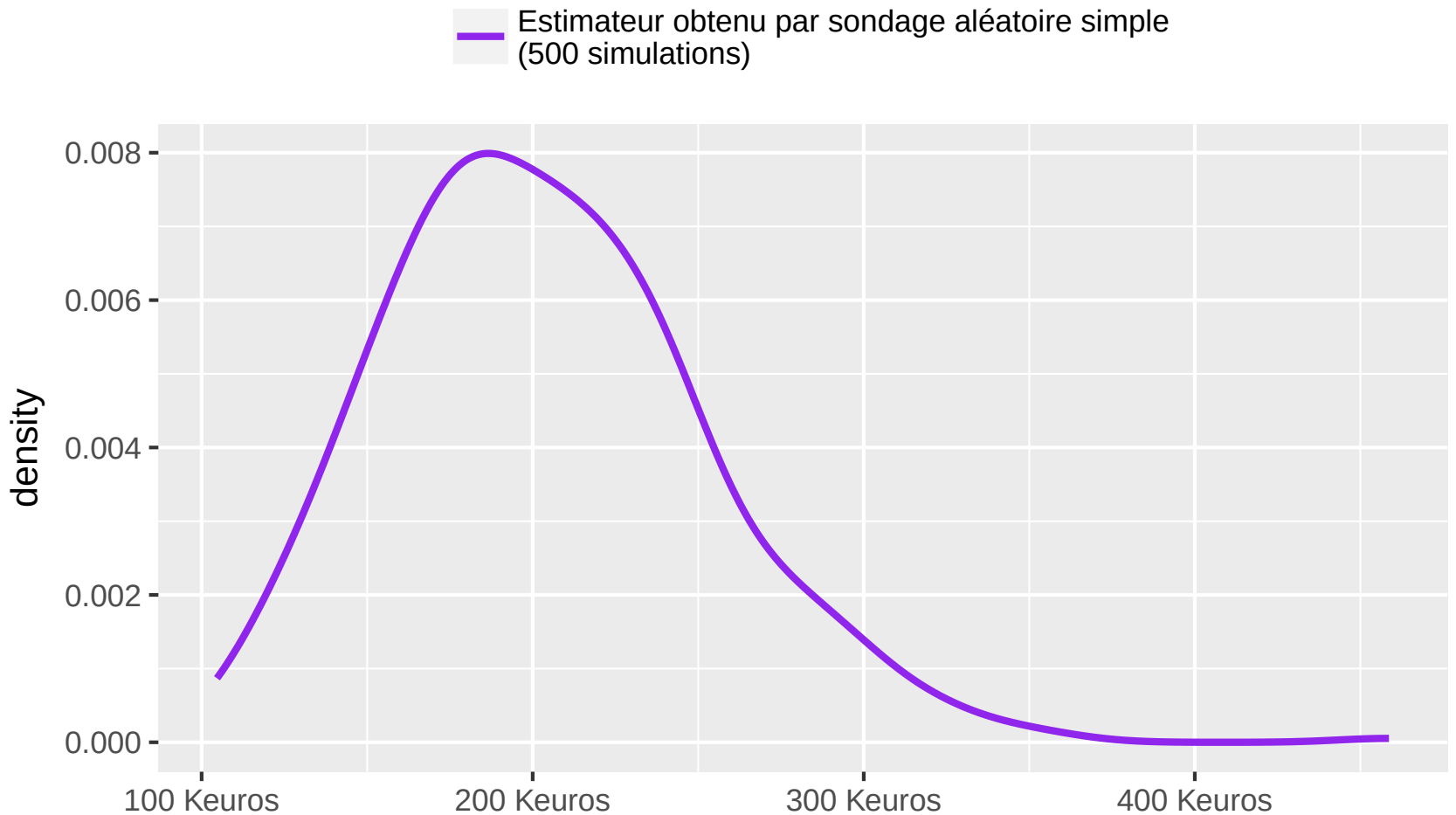
La variable de patrimoine est en fait **disponible dans la base de sondage** (source fiscale). On peut ainsi l'utiliser pour **vérifier que l'estimateur d'Horvitz-Thompson est bien sans biais**.

On procède par simulation :

1. On tire aléatoirement un échantillon s de taille 100 par sondage aléatoire simple ;
2. On estime la valeur de $\hat{\hat{Y}}_{HT}$ sur l'échantillon s.

On réitère les étapes 1 et 2 un **grand nombre de fois** (500 simulations) : on obtient ainsi 500 estimations dont on analyse la distribution.

Application : Enquête sur le patrimoine



Application : Enquête sur le patrimoine

Population La valeur moyenne du patrimoine **dans la population** est de 207 Keuros.

Échantillon L'estimateur d'Horvitz-Thompson à partir des **500 échantillons simulés** :

- prend des valeurs entre 105 Keuros et 459 Keuros ;
- a une moyenne empirique de 203 Keuros (≈ 207 : **absence de biais**) ;
- a un écart-type empirique de 48 Keuros.

Stratification : Exploiter l'information auxiliaire disponible

Bien souvent, la base de sondage comporte des informations auxiliaires susceptibles d'être liées à la variable d'intérêt.

Exemples

- enquête sur la formation professionnelle : position sur le marché du travail ;
- enquête sur l'investissement des entreprises : chiffre d'affaire ;
- enquête sur le logement : année de construction et nombre de pièces du logement.

Quand la ou les variables auxiliaires présentent des catégories, on peut les utiliser pour **définir des strates**.

Stratification : Exploiter l'information auxiliaire disponible

Principe de la stratification Mener le sondage indépendamment au sein des différentes strates, en sur-représentant éventuellement certaines.

Exemples Dans l'enquête sur la formation professionnelle, stratifier selon le fait d'être au chômage ou pas et sur-représenter les chômeurs.

Avantages

- contrôler *ex ante* le nombre d'unités de chaque strate dans l'échantillon ;
- allouer davantage d'unités de l'échantillon aux strates présentant la plus grande variabilité pour la variable d'intérêt Y.

Stratification : Exploiter l'information auxiliaire disponible

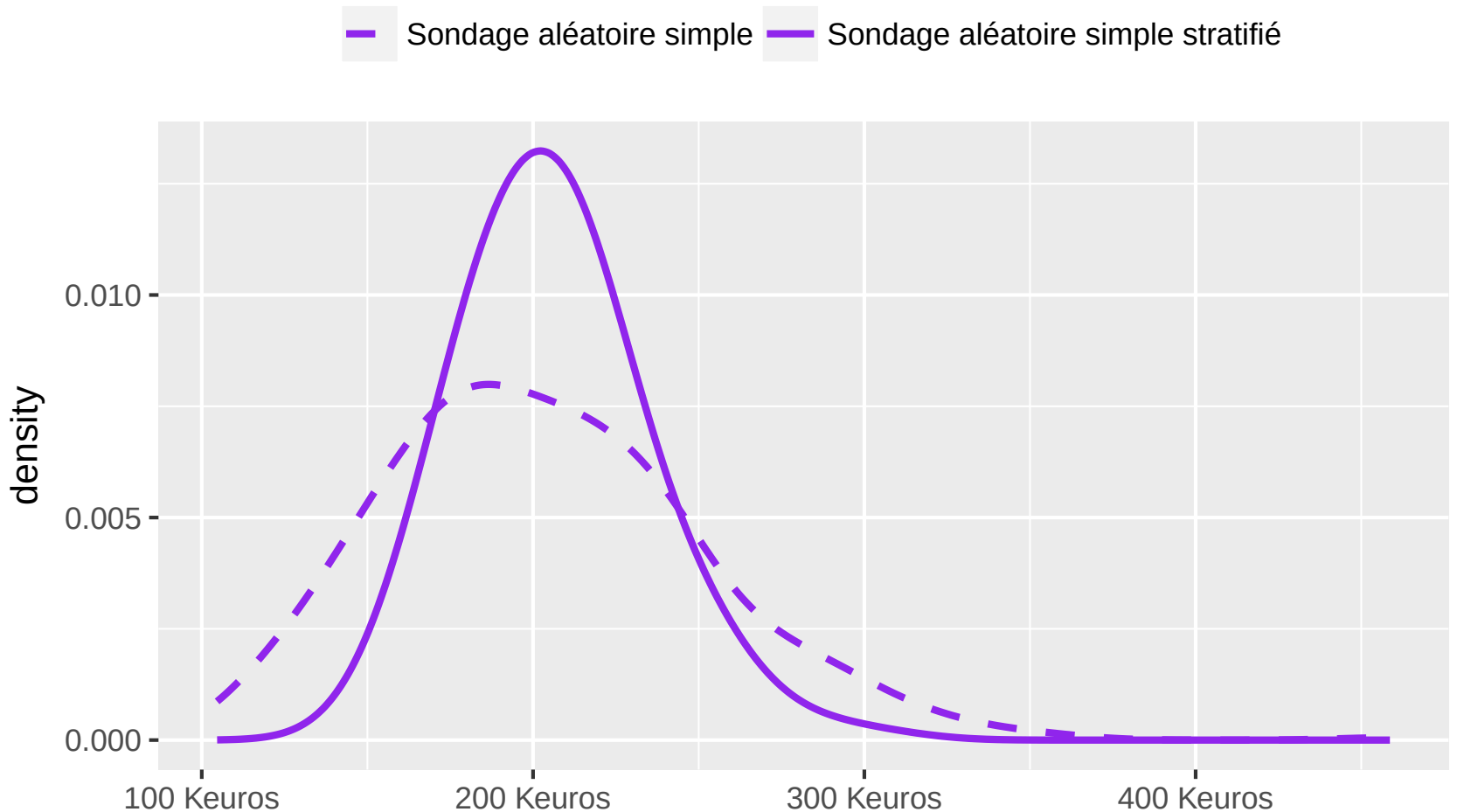
Estimateur d'Horvitz-Thompson Le plan de sondage stratifié **s'insère dans le cadre d'Horvitz-Thompson** : estimateur sans biais et variance calculable.

Sondage aléatoire simple dans chaque strate Quand un SAS est mené dans chaque strate, l'estimateur d'Horvitz-Thompson de la moyenne est :

$$\hat{\bar{Y}}_{SAS-str} = \frac{1}{N} \sum_{h=1}^{H} N_h \bar{y}_h$$

Application à l'enquête Patrimoine Stratification selon l'**assujettissement à l'Impôt de solidarité sur la fortune (ISF)** : patrimoine supérieur à 1,3 M d'euros ou non.

Stratification : Exploiter l'information auxiliaire disponible



Stratification : Exploiter l'information auxiliaire disponible

Population Moyenne : 207 Keuros

Sondage aléatoire simple

- Étendue : entre 105 Keuros et 459 Keuros ;
- Moyenne : 203 Keuros ;
- Écart-type : 48 Keuros.

SAS stratifié selon l'assujettissement à l'ISF

- Étendue : entre 145 Keuros et 308 Keuros ;
- Moyenne : 206 Keuros ;
- Écart-type : 27 Keuros.

Stratification : Choix des allocations

Dans un plan de sondage stratifié, le concepteur peut choisir les allocations n_h selon les objectifs de l'enquête :

- meilleure précision que le SAS pour toutes les variables : **allocation proportionnelle**

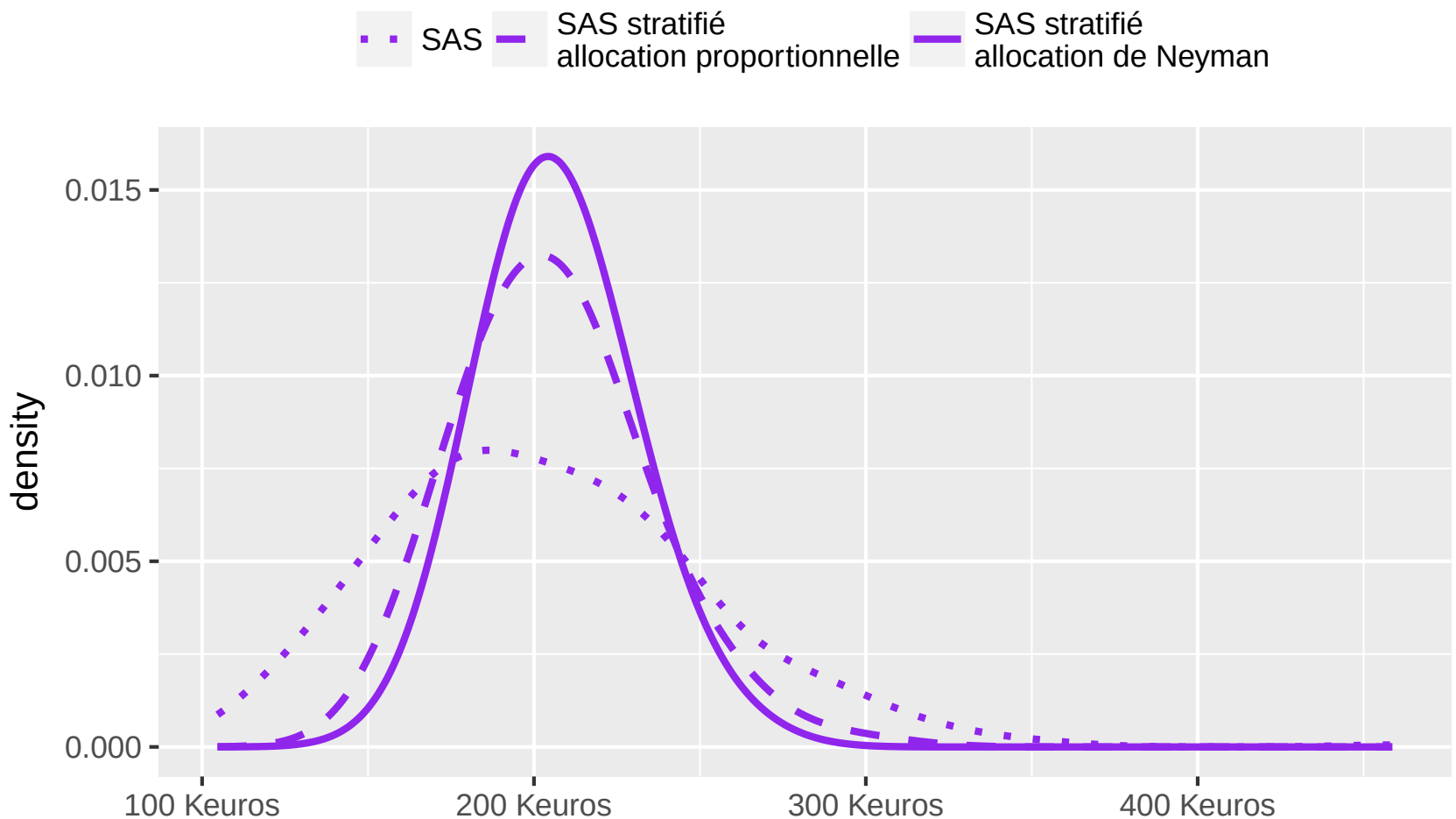
$$n_h = n \times \frac{N_h}{N}$$

- bien meilleure précision que le SAS pour une variable, mais moindre pour les autres : **allocation de Neyman**

$$n_h = n \times \frac{N_h S_h}{\sum_{h'=1}^{H} S_{h'} N_{h'}}$$

Remarque L'allocation de Neyman peut conduire à ce que certaines strates soient **exhaustives** (échantillonnées entièrement).

Stratification : Choix des allocations



Stratification : Choix des allocations

Population Moyenne : 207 Keuros

SAS stratifié avec allocation proportionnelle

- Étendue : entre 145 Keuros et 308 Keuros ;
- Moyenne : 206 Keuros ;
- Écart-type : 27 Keuros.

SAS stratifié avec allocation de Neyman

- Étendue : entre 151 Keuros et 275 Keuros ;
- Moyenne : 207 Keuros ;
- Écart-type : 20 Keuros.

Chapitre 2

Complément sur la stratification : tirage systématique

Chapitre 3

Correction du biais de non-réponse

Non-réponse : Définition et origine

Définition Incapacité d'obtenir des réponses utilisables, pour tout ou partie des variables d'intérêt.

On distingue deux types de non-réponse :

- **non-réponse totale** : non-réponse à l'ensemble des questions de l'enquête ;
- **non-réponse partielle** : non-réponse à certaines questions de l'enquête seulement.

Non-réponse : Définition et origine

Origine de la non-réponse totale

- Impossibilité de joindre l'unité (déménagement, absence) ;
- Incapacité à répondre ;
- Refus de répondre ;
- Abandon au tout début du questionnaire.

Origine de la non-réponse partielle

- Incompréhension des questions par l'enquêté ;
- Refus de répondre à certaines questions (jugées indiscrètes) ;
- Réponses incompréhensibles ;
- Abandon du questionnaire en cours d'enquête.

Conséquences de la non-réponse

En pratique, la non-réponse entraîne pour les estimateurs relatifs aux variables d'intérêt :

- l'introduction d'un biais
- une diminution de la précision.

Relation entre mécanisme de non-réponse et variable d'intérêt

- indépendance totale (*Missing completely at random* ou MCAR) ;
- indépendance conditionnelle à certaines variables auxiliaires (*Missing at random* ou MAR) ;
- dépendance même en contrôlant par les variables auxiliaires disponibles (*Missing not at random* ou MNAR).

Conséquences de la non-réponse

Application à l'enquête Patrimoine Vis-à-vis de l'**estimation du patrimoine** (total ou moyen) :

Non-réponse MCAR Le fait d'être répondant ou non est complètement indépendant du niveau de patrimoine.

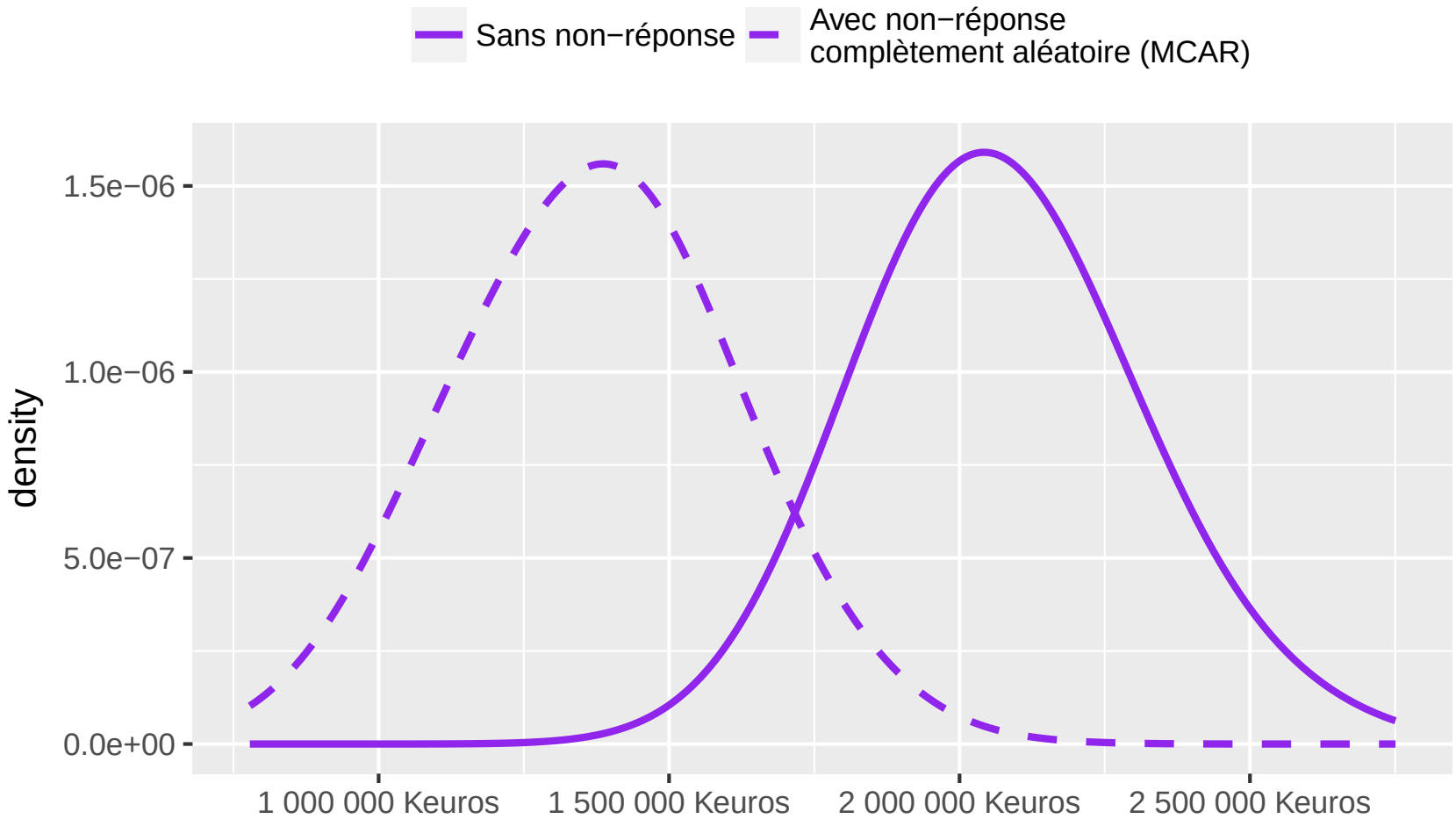
Non-réponse MAR Le fait d'être répondant ou non n'est pas indépendant du niveau de patrimoine, mais on dispose de variables auxiliaires pour le modéliser.

Non-réponse MNAR Le fait d'être répondant ou non n'est pas indépendant du niveau de patrimoine, et on ne dispose d'aucune variable auxiliaire pour le modéliser.

Remarque En pratique, la non-réponse est très rarement MCAR, souvent MAR et parfois MNAR.

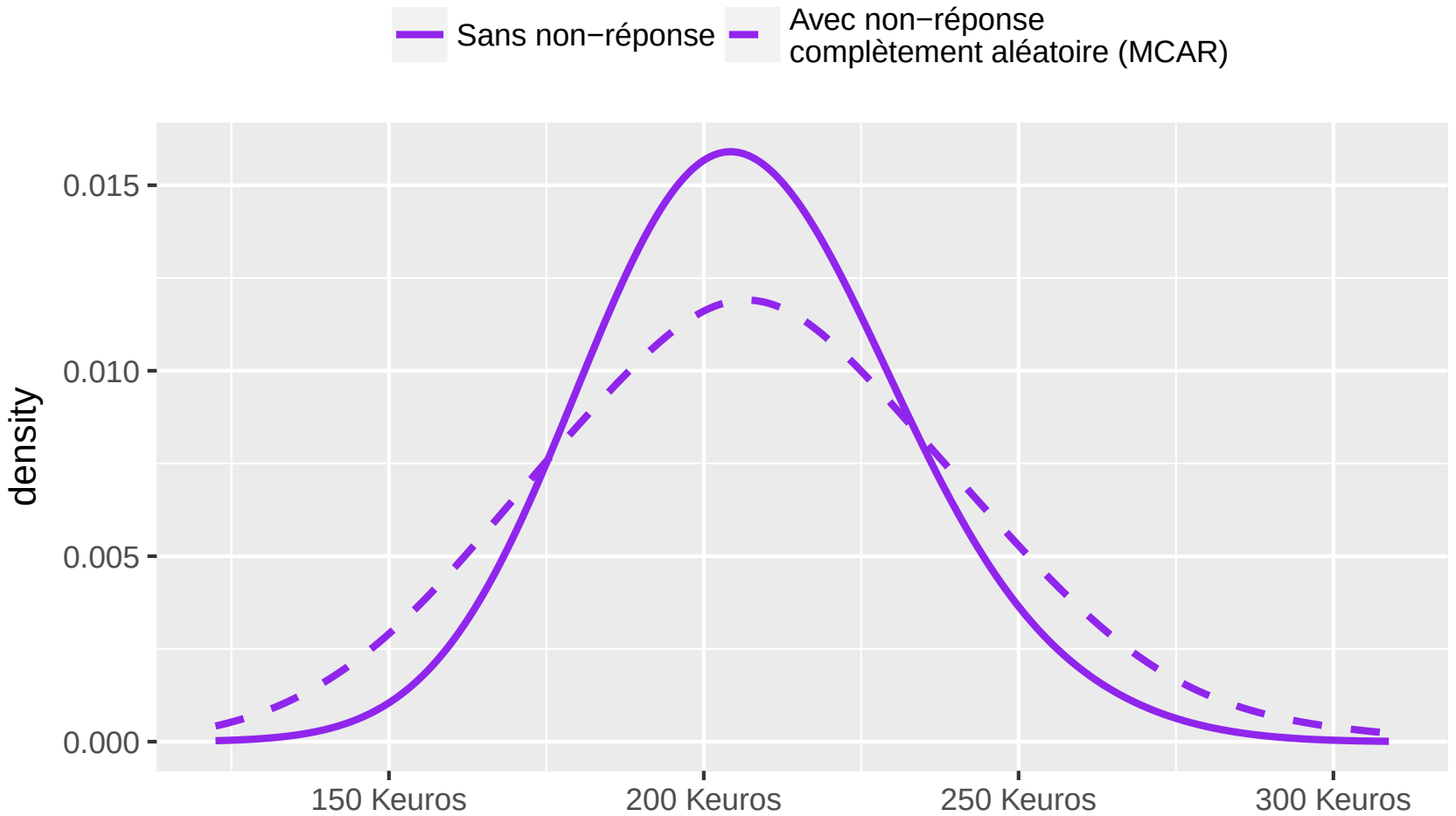
Conséquences de la non-réponse

Non-réponse MCAR et estimateur du total



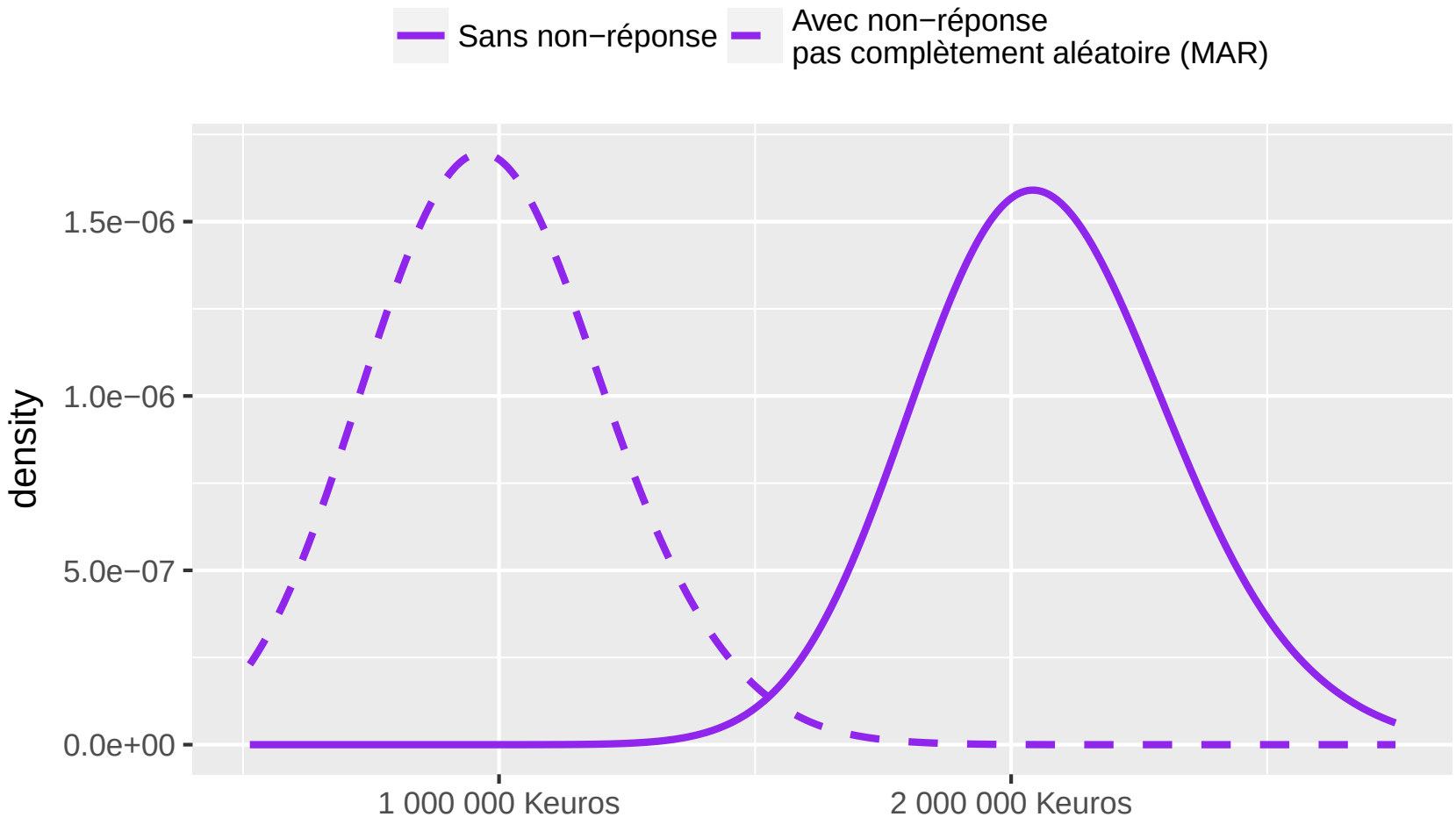
Conséquences de la non-réponse

Non-réponse MCAR et estimateur de la moyenne



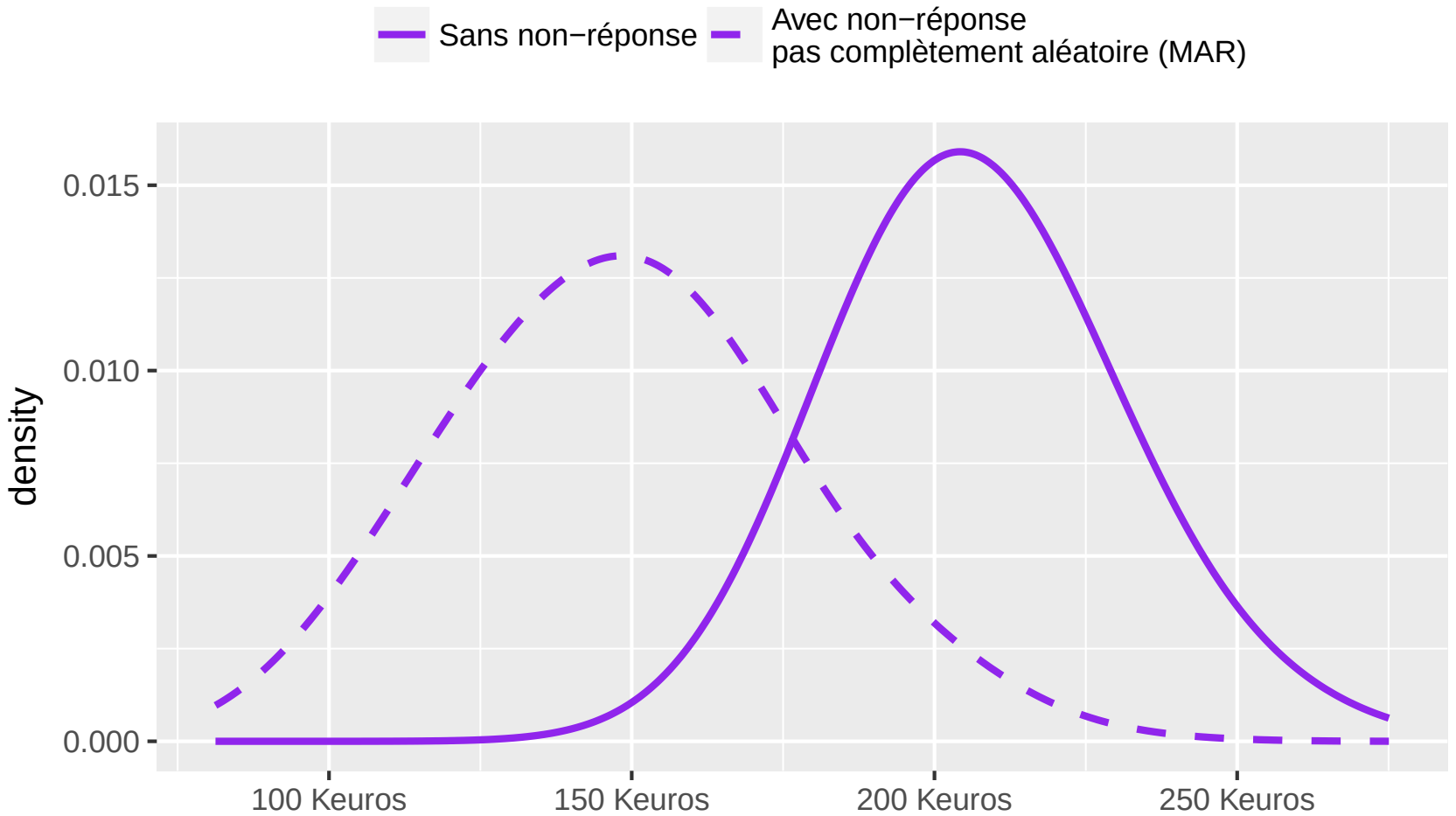
Conséquences de la non-réponse

Non-réponse MAR et estimateur du total



Conséquences de la non-réponse

Non-réponse MAR et estimateur de la moyenne



Conséquences de la non-réponse

Pour résumer Il est impératif de **minimiser la non-réponse en amont** et de la **corriger en aval** pour ne pas avoir d'estimateurs biaisés.

Il existe **deux grandes familles de méthodes de correction de la non-réponse** :

- **méthodes d'imputation** : on remplace les valeurs manquantes par des valeurs « plausibles » ;
- **méthodes de repondération** : on modifie le poids des unités répondantes en fonction de celui des unités non-répondantes.

Les méthodes à mettre en œuvre sont **analogues selon que la non-réponse est MCAR ou MAR**, mais appliquées au sein de **classes de correction de la non-réponse** dans le second cas.

Correction d'une non-réponse MCAR

Méthodes d'imputation déterministes

- Méthode déductive
- *Cold-deck*
- Moyenne, ratio, régression, tendance unitaire, etc.
- Plus proche voisin

Méthodes d'imputation aléatoires

- *Hot-deck* aléatoire
- Imputations avec résidus

Correction d'une non-réponse MCAR

Méthode de repondération

Utilisée en pratique uniquement pour corriger de la non-réponse totale.

Principe Inflater les poids des répondants pour conserver la somme totale des poids :

$$w_i^{CNR} = w_i \times \frac{\sum_{k \in s} w_k}{\sum_{k \in r} w_k}$$

Correction d'une non-réponse MCAR

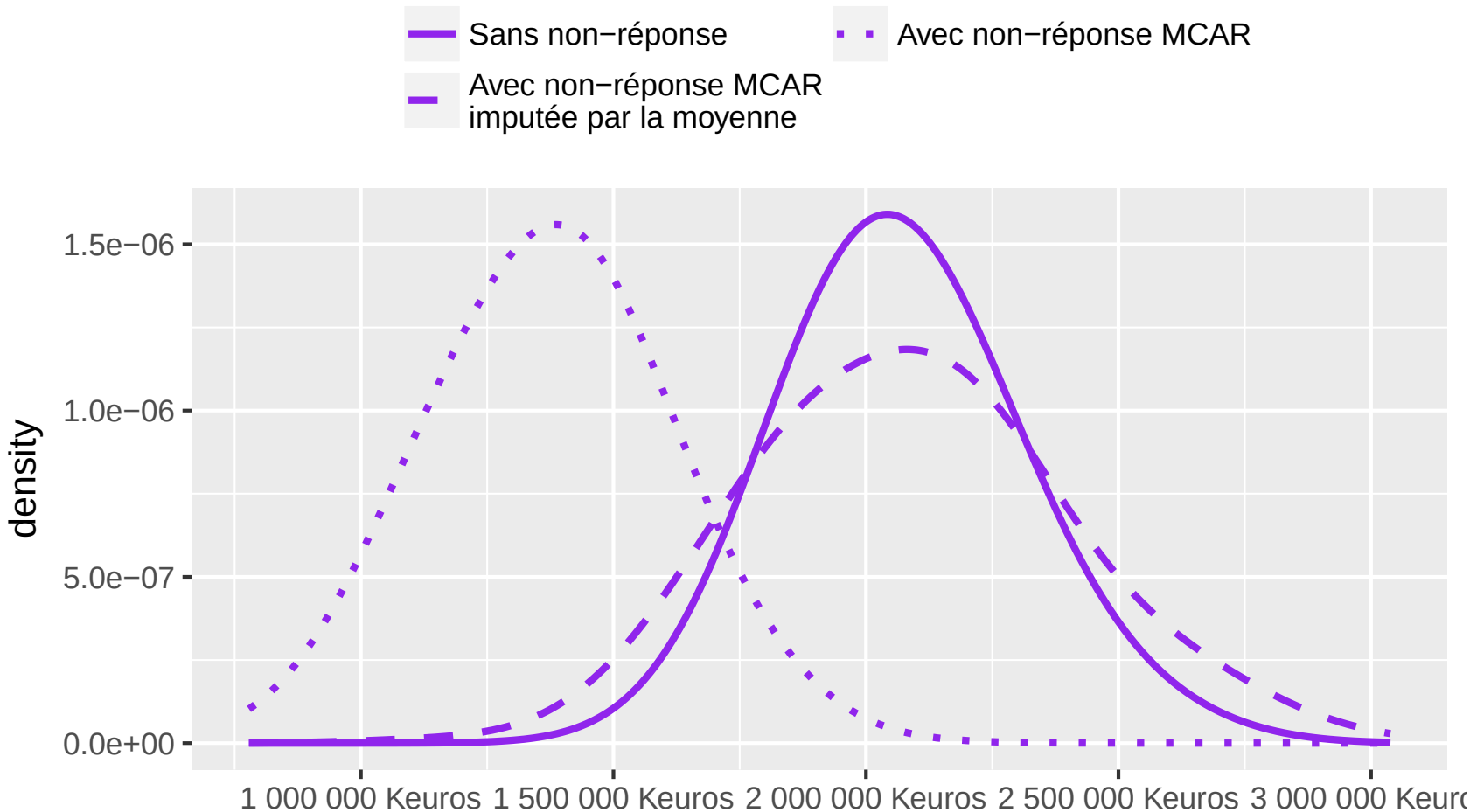
Application à l'enquête Patrimoine

Imputation par la moyenne On impute la variable de patrimoine en remplaçant toutes les non-réponses par la valeur moyenne dans l'échantillon.

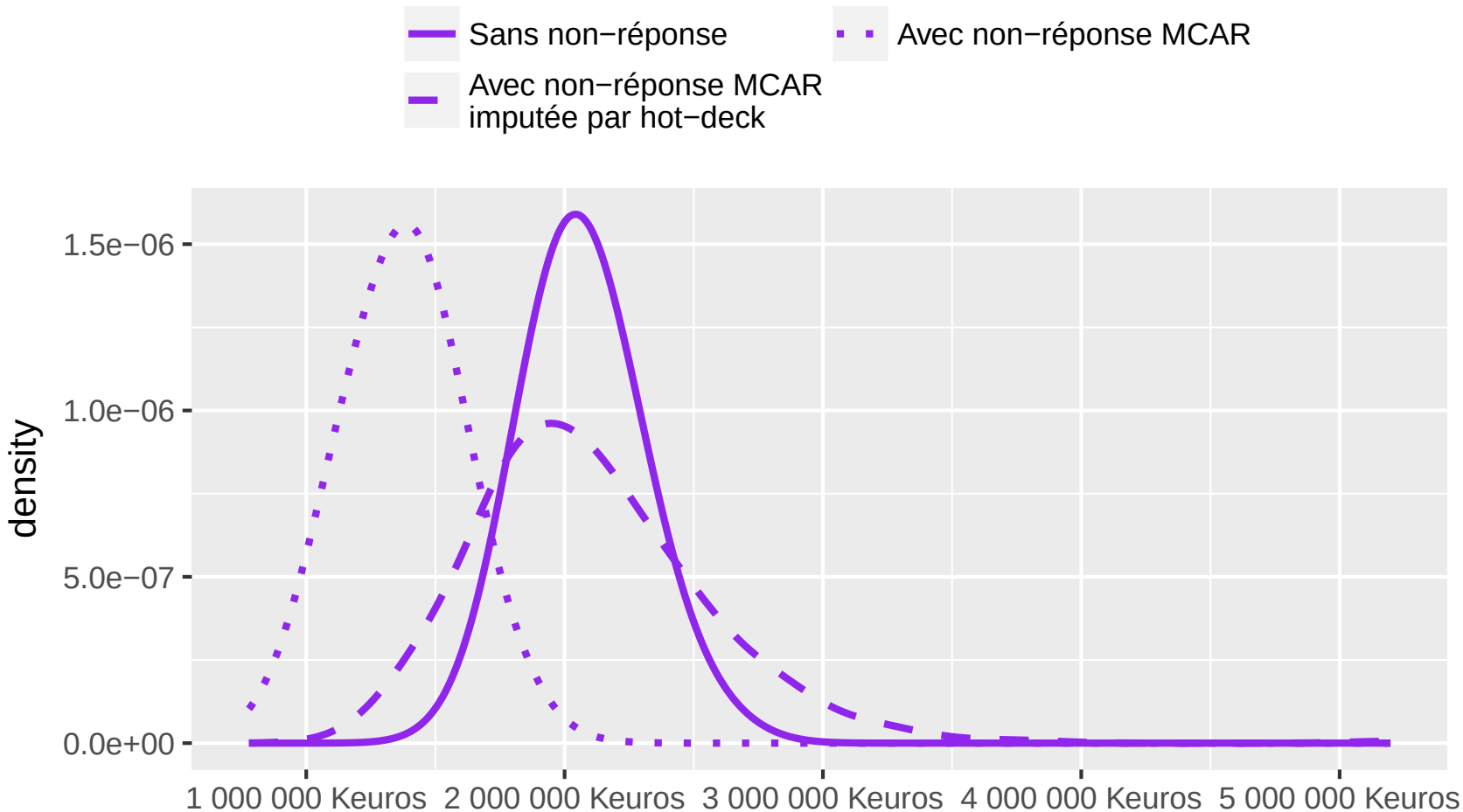
Imputation par *hot-deck* On impute la variable de patrimoine en assignant à chaque unité non-répondante le patrimoine d'une unité répondante tirée au sort.

Repondération On modifie les poids de sondage pour tenir compte de la non-réponse.

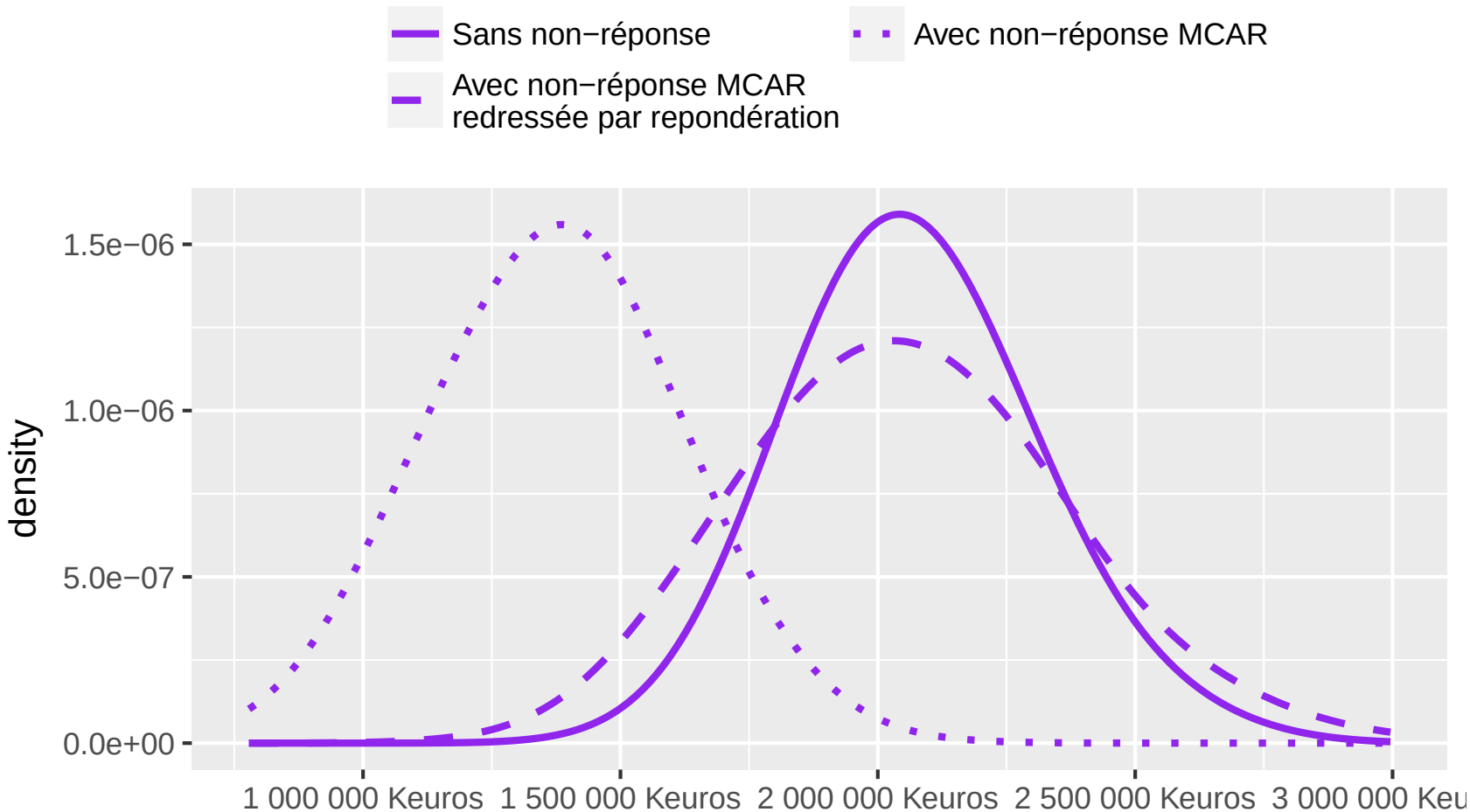
Correction d'une non-réponse MCAR



Correction d'une non-réponse MCAR



Correction d'une non-réponse MCAR



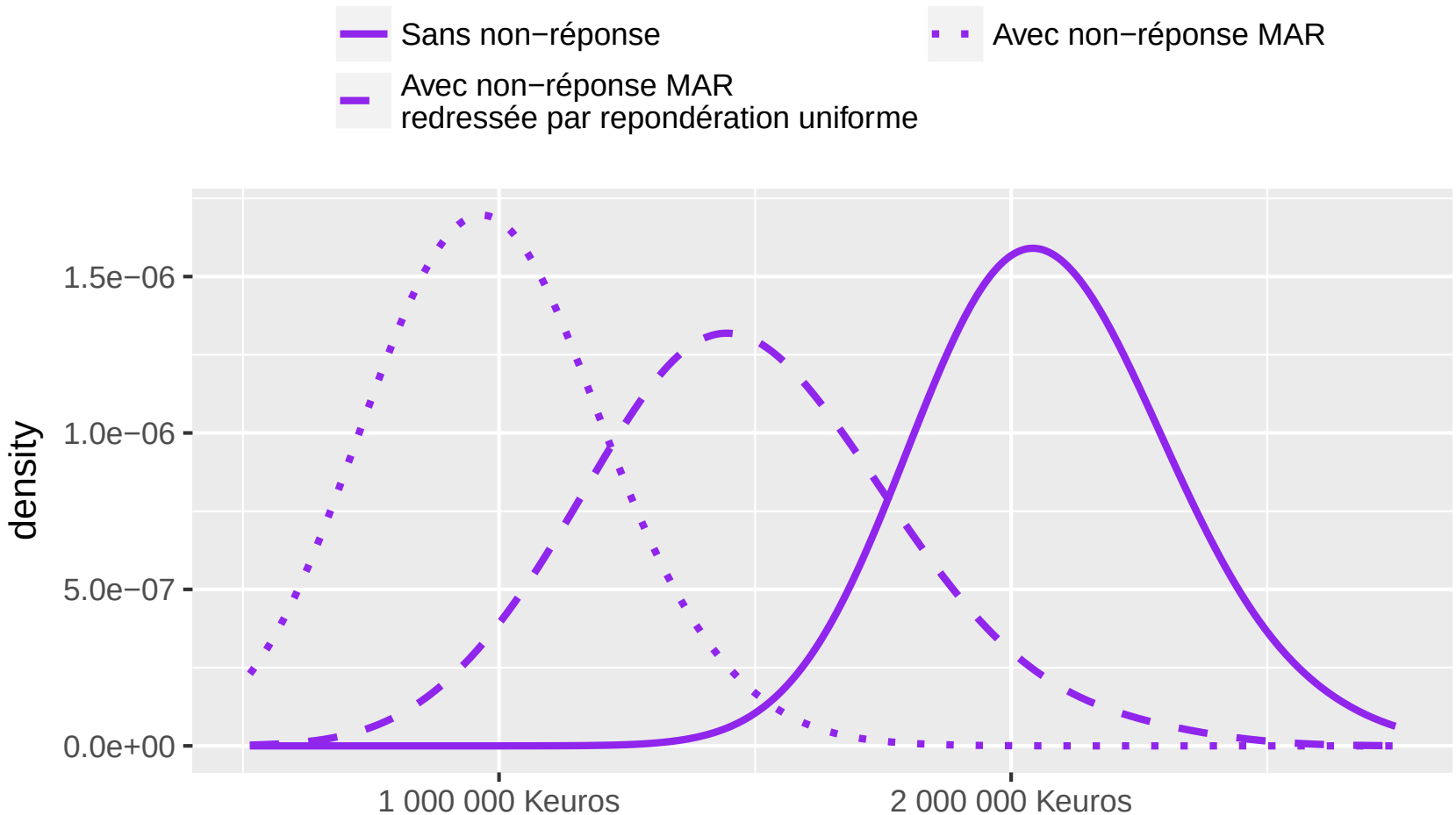
Correction d'une non-réponse MAR

La non-réponse n'est **pas indépendante de la variable d'intérêt** : il n'est **pas possible** d'utiliser à l'identique les méthodes applicables à une non-réponse MCAR pour corriger le biais de non-réponse.

Intuition Si les **hauts-patrimoines** ont tendance à **moins répondre**, repondérer en ajustant les poids de façon **uniforme** conduit à **sous-estimer le patrimoine moyen**.

Moralité Ne pas traiter la non-réponse (ou la corriger de façon uniforme), c'est considérer que les non-répondants et les répondants ont des **comportements identiques eu égard à la variable d'intérêt**, ce qui est en général faux.

Correction d'une non-réponse MAR



Correction d'une non-réponse MAR

Quand la non-réponse est MAR, cela signifie qu'**il existe des variables auxiliaires** telles que, conditionnellement à ces variables, le mécanisme de non-réponse est complètement aléatoire.

Méthode de correction de la non-réponse

1. Utiliser les variables auxiliaires pour constituer des classes de correction de la non-réponse.
2. Au sein de chaque classe, appliquer indépendamment une des méthodes de correction de la non-réponse mentionnées précédemment.

Remarque On parle en général de « classes d'imputation » pour les méthodes d'imputation et de « groupes de réponse homogènes » pour les méthodes de repondération.

Correction d'une non-réponse MAR

Méthodes de constitution des classes

- Croisement manuel de variables
- Algorithmes de segmentation : CHAID (*Chi-Square Assisted Interaction Detection*, CART (*Classification And Regression Tree*)
- Régression logistique + segmentation de la probabilité de réponse estimée :
  - quantiles ;
  - classification ascendante hiérarchique (CAH) ;
  - méthodes des centres mobiles (ou *k-means*) itérative (Haziza, Beamont, 2007).

Correction d'une non-réponse MAR

Application à l'enquête Patrimoine

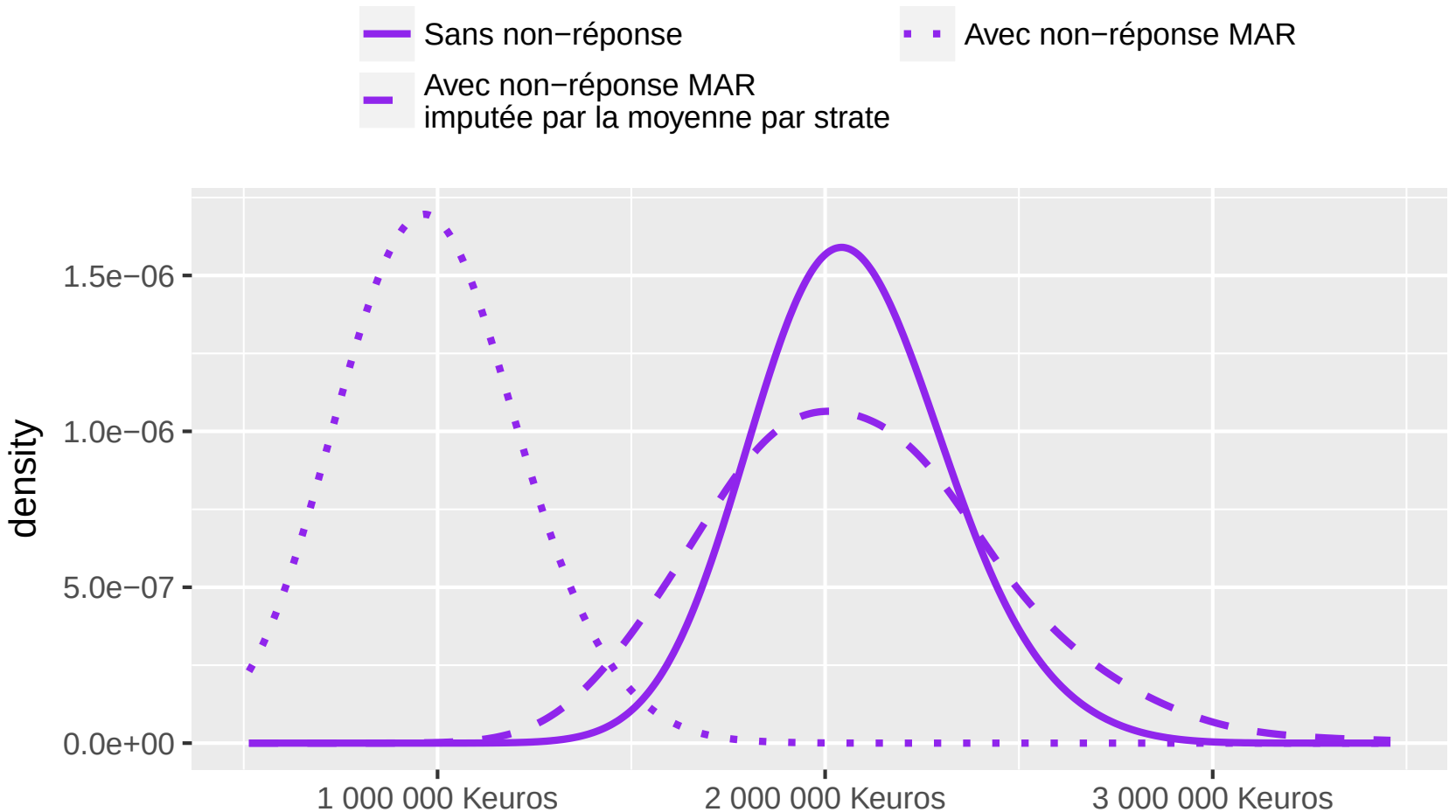
On fait l'hypothèse que le comportement de réponse est lié au niveau de patrimoine.

Assez naturellement, on **réutilise la variable de stratification**, à savoir l'assujettissement à l'impôt de solidarité sur la fortune (ISF).

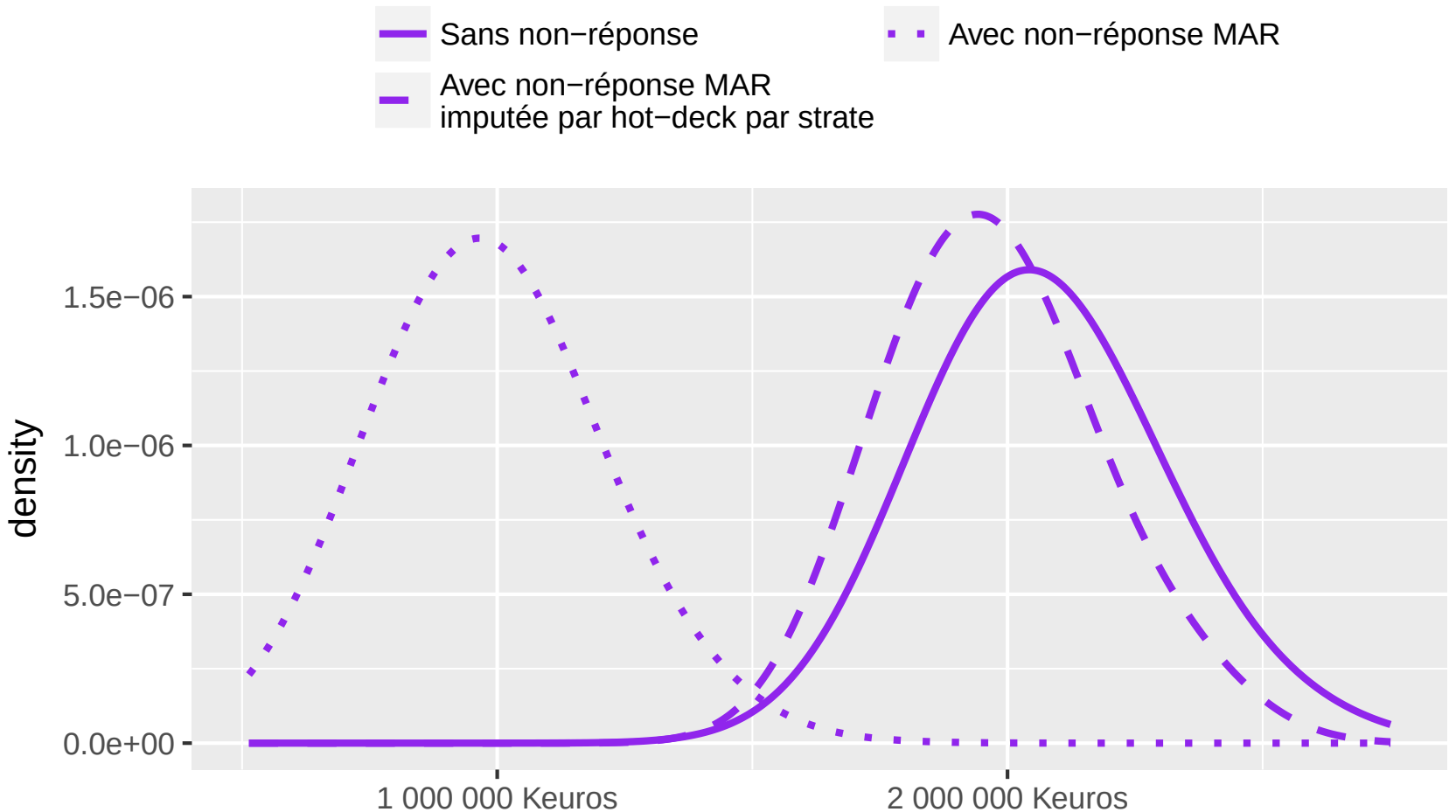
En pratique :

- on redresse de la non-réponse par les **trois méthodes utilisées précédemment** (imputation par la moyenne, impuation par *hot-deck*, repondération)...
- ... mais **séparément pour les assujettis à l'ISF et les autres**.

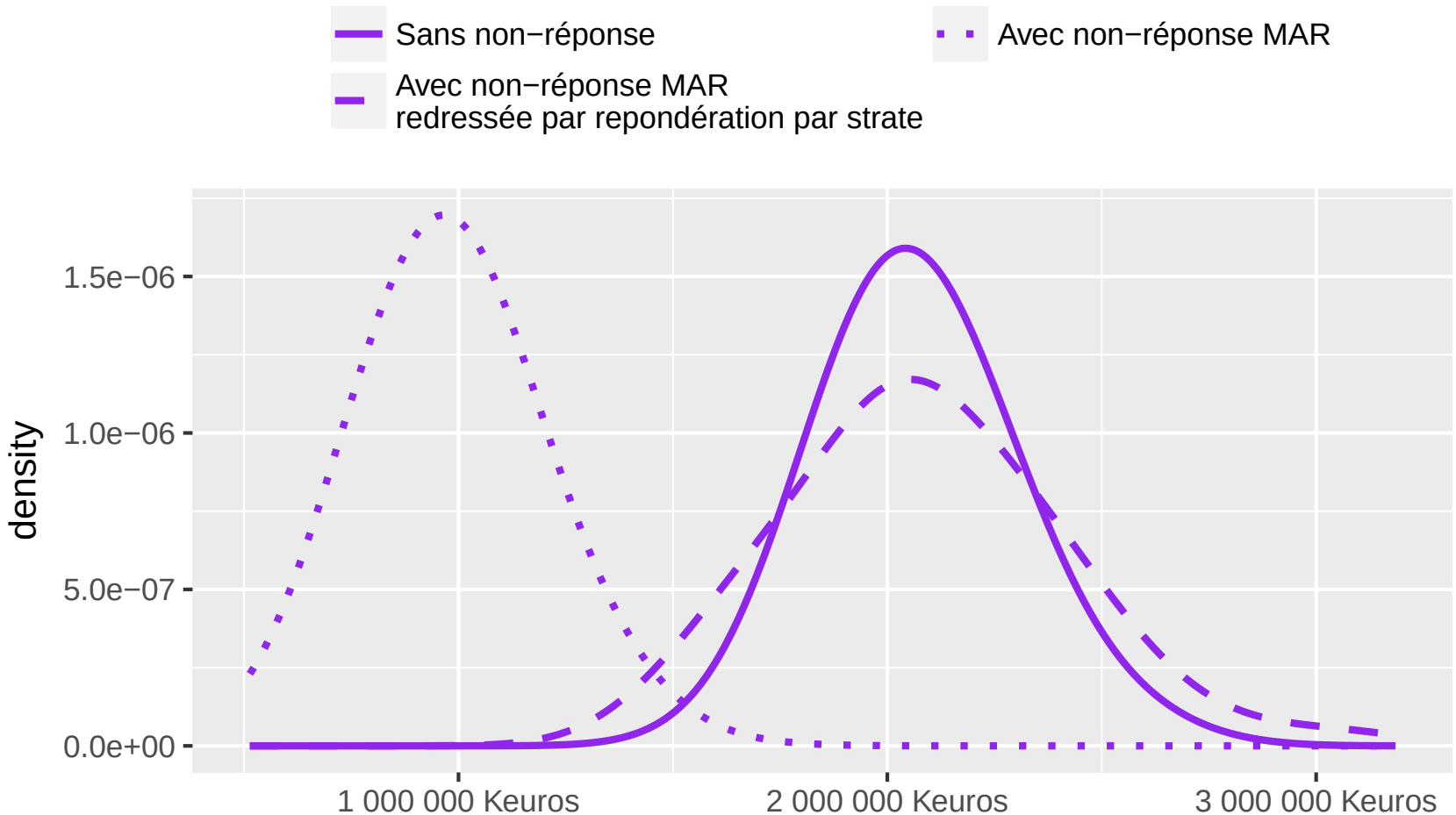
Correction d'une non-réponse MAR



Correction d'une non-réponse MAR



Correction d'une non-réponse MAR



En guise de conclusion

En pratique, les enquêtes par sondage font face à une **non-réponse de plus en plus importante**.

La mise en œuvre de méthodes de correction du biais au sein de **classes de corection de la non-réponse** est ainsi impératif :

- méthodes d'imputation ;
- méthodes de repondération.

Une fois l'estimateur corrigé du biais de non-réponse, d'autres méthodes d'estimation peuvent être utilisées pour **améliorer sa précision**.