Introduction aux traitements statistiques
d'enquêtes sociologiques

Cours 8 : Panorama des méthodes de sondages
Correction de la non-réponse et redressements

Damien Cartron et Martin Chevalier

Année universitaire 2024-2025

Le travail ne s'arrête pas à l'échantillon

Les méthodes d'échantillonnage permettent d'obtenir des estimateurs **sans biais** sous le plan de sondage et dont la **variance est calculable** (Horvitz-Thompson).

Plusieurs méthodes, dont la stratification, **exploitent l'information auxiliaire** de la base de sondage pour **améliorer la précision du sondage** à taille d'échantillon donnée.

Plusieurs opérations doivent néanmoins être opérées après la collecte pour **garantir la qualité des résultats de l'enquête** :

- en général tous les individus ne répondent pas à l'enquête : la **non-réponse** est susceptible de **biaiser** les estimateurs ;
- de l'information auxiliaire est souvent disponible : les **redressements** permettent d'en tirer parti pour **améliorer la précision de l'enquête**.

Objectifs de la séance

1. Avoir conscience de l'impact de la non-réponse sur les résultats d'une enquête
2. Connaître les méthodes classiques de correction de la non-réponse et savoir les appliquer à bon escient
3. Comprendre en quoi l'information auxiliaire peut être utilisée pour améliorer la précision des estimateurs
4. Connaître les principales méthodes de redressement, notamment le calage sur marges

Non-réponse : Définition et impacts

Définition

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.

En pratique, la frontière entre non-réponse totale et non-réponse partielle est **floue** : quand il y a trop de non-réponse partielle, on a plutôt intérêt à considérer le questionnaire en non-réponse totale.

Exemple Dans l'enquête Emploi en continu, on considère qu'un individu est non-répondant dès lors que le module A (servant au calcul du chômage BIT) n'est pas exploitable.

Origine de la non-réponse

Origine de la non-réponse totale

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

Origine de la non-réponse partielle

- Refus de répondre à certaines questions car jugées indiscrètes (sexualité, violences, mais aussi salaires par exemple) ;
- Incompréhension des questions ;
- Réponses incompréhensibles ;
- Abandon du questionnaire en cours d'enquête.

Mode de collecte et non-réponse

Le mode de collecte **conditionne fortement** l'ampleur et la nature de la non-réponse :

- **internet** : forte non-réponse totale (mail de contact considéré comme spam), abandon tout au long du questionnaire s'il est trop long, risque de non-réponse partielle déguisée (*satisficing*) ;
- **téléphone** : très forte non-réponse totale (appel assimilé au démarchage téléphonique), questionnaire nécessairement court sinon abandon ;
- **dépôt-retrait** : non-réponse totale moins forte (contact lors du dépôt avec l'enquêté), risque d'incompréhension des questions ;
- **face-à-face** : non-réponse totale moins forte si les moyens sont mis (relances, rendez-vous, etc.), possibilité d'aider à la compréhension des questions (mais pas trop), de motiver pour finir le questionnaire.

La plupart des enquêtes de la statistique publique sont assorties d'une **obligation de réponse** (accordée par le comité du label du conseil national de l'information statistique [CNIS]).

En théorie donc, les individus ou les entreprises figurant dans l'échantillon et ne répondant pas après plusieurs relances peuvent **faire l'objet de poursuites**.

Ces dispositions sont néanmoins appliquées **avec discernement** :

- l'Insee ne va au contentieux que quand une **entreprise importante au sein d'un secteur d'activité** refuse de répondre ;
- pour les plus petites entreprises ou les individus, en général **aucune poursuite n'est engagée**.

Exemple Recensement de la population : « Votre participation est essentielle. Elle est rendue obligatoire par la loi, mais c'est avant tout un devoir civique, simple et utile à tous. »

Méthodes pour limiter la non-réponse totale

L'importance de la **lettre-avis** :

- signifier le **caractère officiel** : références juridiques, Marianne en couleur ;
- rassurer quant à l'**utilisation des données** de l'enquête : à des fins statistiques uniquement (pas de contrôle), anonymisation, pas de transmission à des tiers ;
- rassurer quant à la **personne qui va se présenter** (face-à-face ou dépôt-retrait) : donner son identité, préciser qu'elle sera porteuse d'une carte professionnelle ;
- présenter l'**objectif de l'enquête** pour motiver les enquêtés : brochure avec les résultats de l'enquête précédente, etc.

Tout faire pour **faciliter la réponse** :

- bien **préparer une collecte** en face-à-face ou en dépôt retrait : repérage des adresses à l'avance, récupération des digicodes ;
- appels à des horaires et des jours **variés** pour une enquête par téléphone ;
- **bonne conception** du questionnaire internet : pas trop chargé ni trop long, utilisable sur smartphone, accessible.

Méthodes pour limiter la non-réponse partielle

Bien poser les questions :

- questions **claires et compréhensibles** (importance des tests de questionnaire) ;
- **indications** quand c'est nécessaire mais pas trop ;
- ordre des questions **logique** ;
- (papier) pas de **filtres complexes**.

Bien organiser le questionnaire :

- (papier) bien **aérer et structurer** le document ;
- (papier) éviter les **décrochages de mise en page** (*cf.* Sumer 2010, p.2) ;
- (internet) **informer** sur la progression ;
- (internet) utiliser les fonctions d'**auto-complétion** pour les nomenclatures.

Impacts de la non-réponse : simulations

Simulation de la non-réponse sur des **données analogues à celles collectées par l'enquête Patrimoine**.

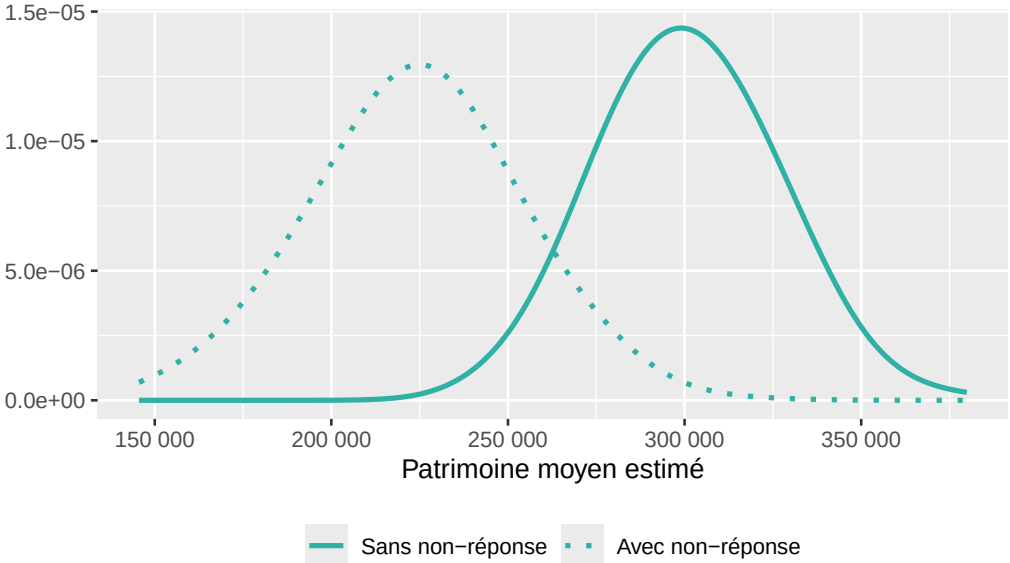
Rappel du plan de sondage Sondage aléatoire simple stratifié selon l'assujettissement à l'ISF, surreprésentation des « hauts patrimoines ».

Cadre de simulation Echantillon de taille 100 parmi 10 000, 500 simulations, patrimoine moyen dans la population 300 000.

Tous les individus ne répondent pas :

- taux de réponse compris entre 63 % et 86 % ;
- taux de réponse moyen de 75 %.

Impacts de la non-réponse : simulations



Impacts de la non-réponse

La présence d'individus non-répondants semble induire **deux phénomènes** :

- d'une part, l'estimateur d'Horvitz-Thompson n'est **plus sans biais** : la courbe de densité n'est pas centrée autour de la vraie valeur dans la population ;
- d'autre part, sa **variance augmente légèrement** : la courbe de densité est légèrement moins « piquée » autour de l'espérance.

Si malgré tous les efforts déployés la non-réponse à l'enquête devait être importante, il apparaît ainsi **indispensable d'en neutraliser l'impact sur les estimateurs**.

Plusieurs méthodes existent pour chercher à **corriger ce biais de non-réponse** : **imputation** et **repondération**.

Néanmoins, au-delà de la méthode de correction mise en œuvre , c'est la manière d'**exploiter l'information auxiliaire** pour procéder à la correction qui est déterminante.

Non-réponse : correction du biais de non-réponse

Méthodes d'imputation

Les méthodes d'imputation désignent l'ensemble des opérations visant à **affecter une valeur en lieu et place des données manquantes**.

Ces méthodes peuvent être utilisées en présence de **non-réponse totale ou de non-réponse partielle**.

Quand des méthodes d'imputation sont utilisés pour corriger de la non-réponse totale, le fichier de l'enquête comporte *in fine* **autant d'observations que d'unités échantillonnées**.

Quand plusieurs variables doivent faire l'objet d'une imputation, l'**ordre est en général déterminant** : certaines imputations peuvent en effet mobiliser des variables ayant elles-mêmes fait l'objet d'une imputation auparavant.

Principales méthodes d'imputation

Méthode déductive Déduire des variables renseignées la valeur de la variable manquante.

Modélisation Estimer un **modèle** sur les observations renseignées puis l'utiliser pour **prédire la valeur** pour les individus non-répondants : imputation par la moyenne, imputation par la régression.

Recherche d'un « donneur » Imputer la ou les valeurs manquantes d'un individu en les remplaçant par celles d'un **autre individu jugé comparable** :

- *cold-deck* : valeur de la variable pour le même individu dans un autre fichier ;
- plus proche voisin : recherche du donneur qui **ressemble le plus** à l'individu non-répondant en termes de variables renseignées ;
- *hot-deck* : tirage aléatoire d'un donneur au sein du fichier de l'enquête.

Méthodes de repondération

Les méthodes de repondération désignent l'ensemble des opérations visant à corriger la non-réponse en **modifiant les poids de sondage de l'enquête**.

Cette méthode est utilisée uniquement pour corriger de la **non-réponse totale**.

Principe Inflater les poids des répondants pour retrouver la somme des poids de l'ensemble des individus tirées (répondants ou non) :

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

Quand des méthodes de repondération sont utilisées pour corriger de la non-réponse totale, le fichier de l'enquête comporte *in fine* **autant d'observations que d'unités répondantes**.

Bien exploiter l'information auxiliaire pour corriger la non-réponse

Bien plus que la méthode choisie, l'enjeu essentiel de la correction de la non-réponse consiste à **bien exploiter l'information auxiliaire pour neutraliser effectivement le biais**.

La non-réponse n'est en général **pas répartie de façon uniforme dans l'échantillon** : certains individus ont une **probabilité supérieure** de ne pas répondre que les autres.

Quand cette probabilité de réponse est **(anti-)corrélée à la variable d'intérêt**, ne pas en tenir compte conduit en pratique à une **correction imparfaite du biais de non-réponse**.

Simulations : correction dans l'ensemble de l'échantillon

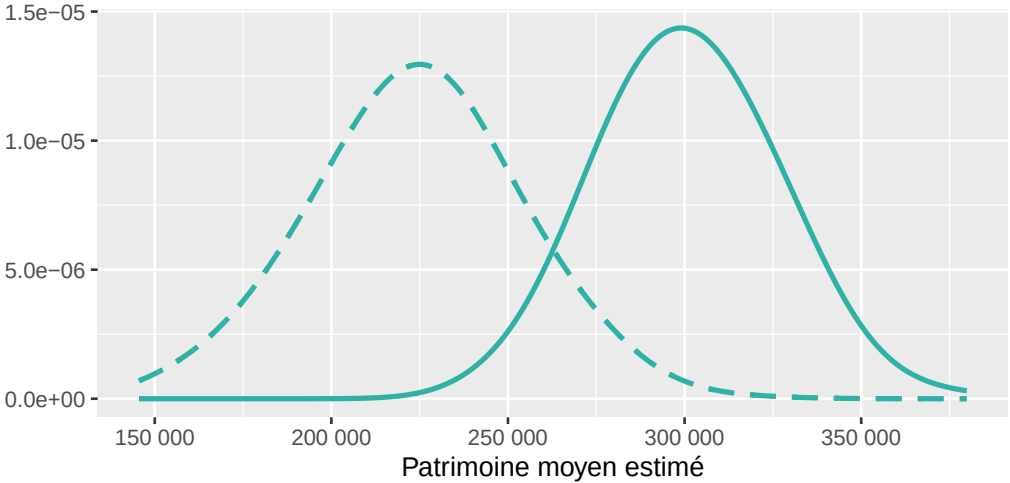
Dans ces simulations, on compare trois méthodes de correction de la non-réponse :

- imputation par la moyenne ;
- imputation par *hot-deck* ;
- repondération.

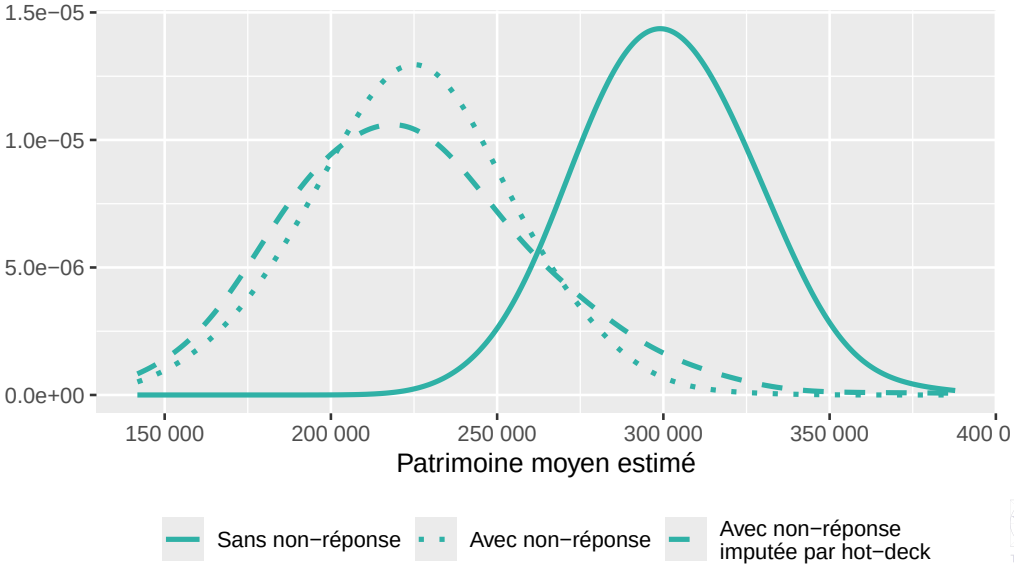
Dans les trois cas, on opère la correction de façon homogène dans l'**ensemble de l'échantillon**, *i.e.* **sans tenir compte de l'information auxiliaire**.

On compare à chaque fois à la distribution de l'estimateur du patrimoine moyen en l'absence de non-réponse.

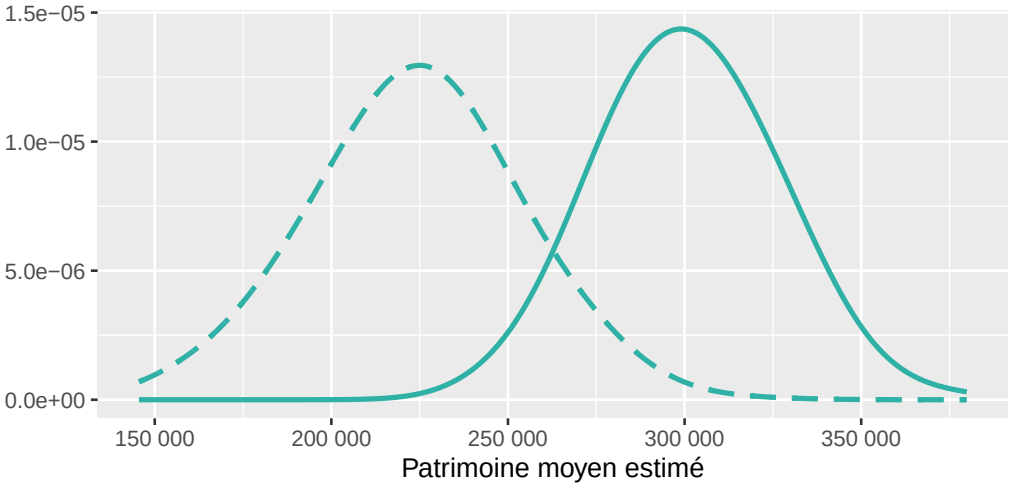
Simulations : imputation par la moyenne



Simulations : imputation par *hot-deck*



Simulations : repondération



Quand la correction de la non-réponse est effectuée dans l'ensemble de l'échantillon, quelle que soit la méthode **elle échoue à neutraliser le biais de non-réponse**.

Quand elles sont appliquées de façon homogène dans l'**ensemble de l'échantillon**, les méthodes de correction de la non-réponse font l'**hypothèse** que les éventuels différences dans la probabilité de réponse ne sont **pas liées à la variable d'intérêt**.

Or ici c'est **vraisemblablement faux**, le niveau de patrimoine est susceptible d'affecter la probabilité de réponse :

- plus le patrimoine est élevé, moins on peut avoir souhaiter vouloir en parler ;
- plus le patrimoine est élevé, plus on peut anticiper que l'enquête sera longue.

En pratique, il est possible de tenir compte de ce lien dans la correction de la non-réponse en exploitant l'**information auxiliaire disponible pour les répondants et les non-répondants**, par exemple celle de la base de sondage.

Imputer en exploitant l'information auxiliaire

- **Modélisation** : intégrer dans le modèle des variables auxiliaires sensément corrélées à la non-réponse et à la variable d'intérêt.

  Exemple Imputer le patrimoine par la moyenne du patrimoine chez les « haut patrimoine » et chez les « bas patrimoine ».

- **Recherche d'un donneur** : limiter la recherche du donneur aux unités avec des caractéristiques proches. On constitue en pratique des **classes d'imputation**.

  Exemple Imputer le patrimoine par *hot-deck* dans deux classes d'imputation différentes, l'une constituée des « haut patrimoine », l'autre constituée des « bas patrimoine ».

De la sorte, on remplace la réponse des individus non-répondants par celles d'individus **de même profil** et on corrige ainsi une éventuelle **déformation liée à la non-réponse**.

Repondérer en exploitant l'information auxiliaire

En présence de non-réponse, l'estimateur par expansion $\hat{\hat{Y}}^{exp}$ est **sans biais** :

$$\hat{\hat{Y}}^{exp} = \frac{1}{N} \sum_{i \in r} \frac{y_i}{\pi_i \times p_i}$$

en notant r les individus répondants, π_i la probabilité d'inclusion simple de l'unité i et p_i sa probabilité de réponse.

En pratique cependant, p_i est **inconnue et doit être estimée**.

On recourt en général pour ce faire à des **groupes de réponse homogène** :

1. On constitue une **partition de l'échantillon** : croisement de variables, méthodes de classification supervisée (CHAID, CART), régression logistique puis constitution de groupes à partir de la probabilité prédite ;
2. Au sein de chaque groupe, on assimile la probabilité de réponse des unités au **taux de réponse du groupe**.

Simulations : correction en exploitant l'information auxiliaire

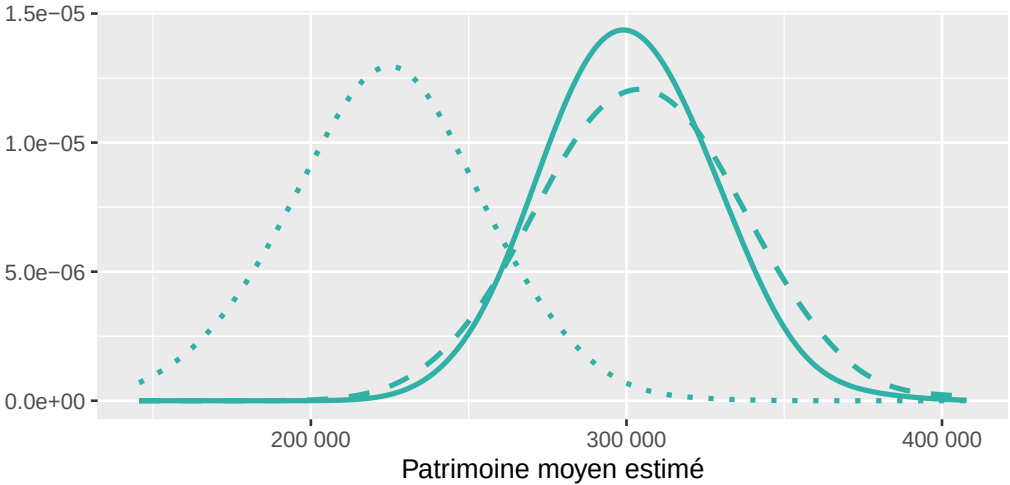
Dans ces nouvelles simulations, la correction est opérée en **distinguant les « haut patrimoine » d'une part et les « bas patrimoine » d'autre part**.

Les méthodes sont les mêmes que précédemment, mais **appliquées au sein de ces deux classes d'imputation** :

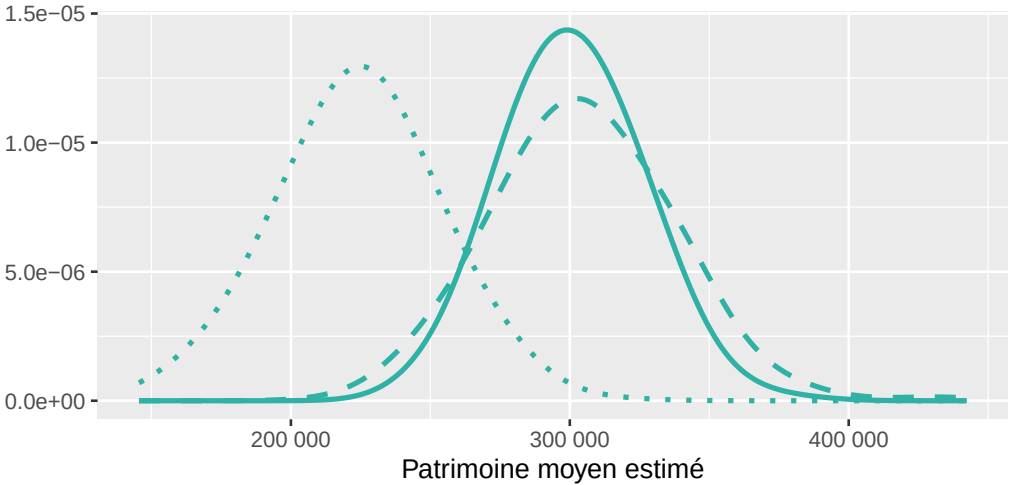
- imputation par la moyenne ;
- imputation par *hot-deck* ;
- repondération.

On espère ce faisant capter un **éventuel comportement spécifique des « haut patrimoine »**, qui seraient susceptibles de **moins répondre à l'enquête**.

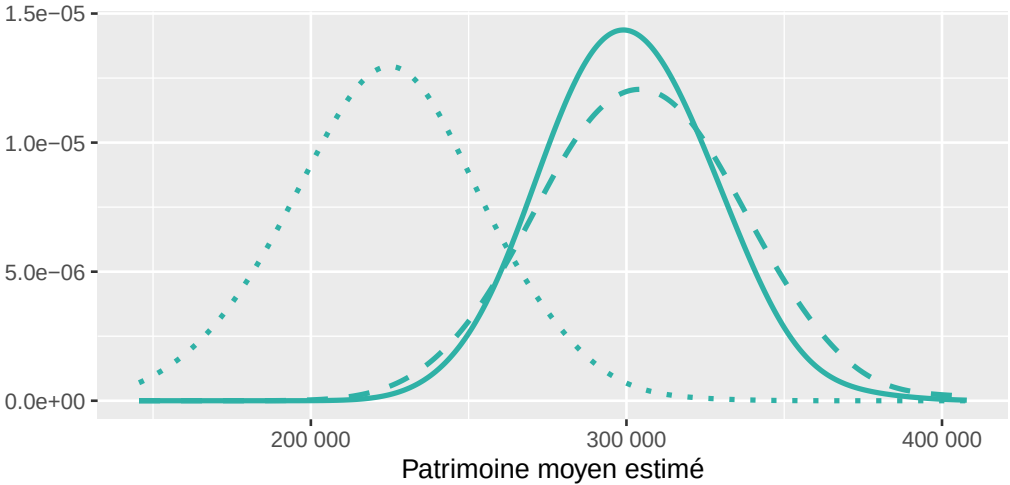
Simulations : imputation par la moyenne par classe d'imputation



Simulations : imputation par *hot-deck* par classe d'imputation



Simulations : repondération par groupes de réponse homogène



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 correction de la non-réponse** est ainsi impératif :

- méthodes d'imputation ;
- méthodes de repondération.

La non-réponse induit également une **augmentation de la variance** due à la diminution de l'échantillon utile.

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

Redressements : Principe et estimateur par le ratio

Objectifs des méthodes de redressement

1. Exploiter l'information auxiliaire qui n'a pas pu l'être au moment du tirage pour **améliorer la précision de l'estimateur**.
2. **Assurer la cohérence** entre les estimations produites par l'enquête et une ou plusieurs sources de référence.

En pratique Ajustement de l'estimateur d'Horvitz-Thompson...

- ...pour garantir une **estimation parfaite** de certaines variables...
- ...et ainsi **diminuer sa variance**...
- ...tout en gardant le caractère **sans biais**.

Remarque Dans l'ensemble de cette partie, le plan de sondage est un sondage aléatoire simple et il n'y a pas de non-réponse.

Exemple : Enquête sur la fréquentation des cinémas

Le distributeur d'un film souhaite connaître le **nombre d'entrées réalisées une semaine donnée**.

Habituellement des remontées sont effectuées tous les mois, mais il souhaite avoir une **information plus rapidement** pour ajuster sa campagne promotionnelle.

Pour ce faire, il interroge un **échantillon de 100 cinémas** (parmi les 2 020 exploitants en activité) tiré par sondage aléatoire simple.

La variable d'intérêt est le **nombre d'entrées réalisées par le film** pour la semaine du 21 au 27 février 2022.

L'estimateur d'Horvitz-Thompson obtenu est de 862 944 avec un **intervalle de confiance à 95 %** de [538 825 ; 1 187 063].

Exemple : Enquête sur la fréquentation des cinémas

Le distributeur n'est **pas très satisfait** de cette fourchette extrêmement large.

Il envisage d'exploiter une information disponible quelques jours après l'enquête, le **nombre de projections du film** :

- sur l'ensemble de la France, le film a été projeté 5 413 fois ;
- à partir de l'échantillon, ce nombre est estimé à 6 080 fois.

Intuition

- Nombre de projections et nombre d'entrées étant **corrélées**, le distributeur pourrait être tenté de **redresser** l'estimateur du nombre d'entrées en le multipliant par $\frac{5413}{6080} = 0,89$.
- L'utilisation du nombre de projections comme information auxiliaire pourrait venir « **stabiliser** » l'estimateur.

Estimateur par le ratio : définition

L'estimateur par le ratio est utilisé quand la variable auxiliaire X est **quantitative**.

Exemple Nombre de projections pour estimer le nombre d'entrées réalisées par un film.

Sachant que le total de la variable auxiliaire $T(X)$ est connu, on définit l'**estimateur par le ratio du total de la variable** Y par :

$$\hat{T}_{ratio}(Y) = \hat{T}_{HT}(Y) \times \frac{T(X)}{\hat{T}_{HT}(X)}$$

Intuition Si $T(X) > \hat{T}_{HT}(X)$, l'estimateur par le ratio de Y est supérieur à l'estimateur d'Horvitz-Thompson.

Estimateur par le ratio : Propriétés

1. Asymptotiquement sans biais :

$$B(\hat{T}_{ratio}(Y)) \xrightarrow[n\to+\infty]{} 0$$

2. Variance d'autant plus faible que Y est corrélée à X :

$$V(\hat{T}_{ratio}(Y)) \approx N^2 \left(1 - \frac{n}{N}\right) \frac{S_Y^2 + R^2 S_X^2 - 2RS_{X,Y}}{n}$$

avec $R = \dfrac{T(Y)}{T(X)}$

3. **Propriété de calage** T(X) est estimé parfaitement :

$$\hat{T}_{ratio}(X) = \hat{T}_{HT}(X) \times \frac{T(X)}{\hat{T}_{HT}(X)} = T(X)$$

Exemple : Enquête sur la fréquentation des cinémas

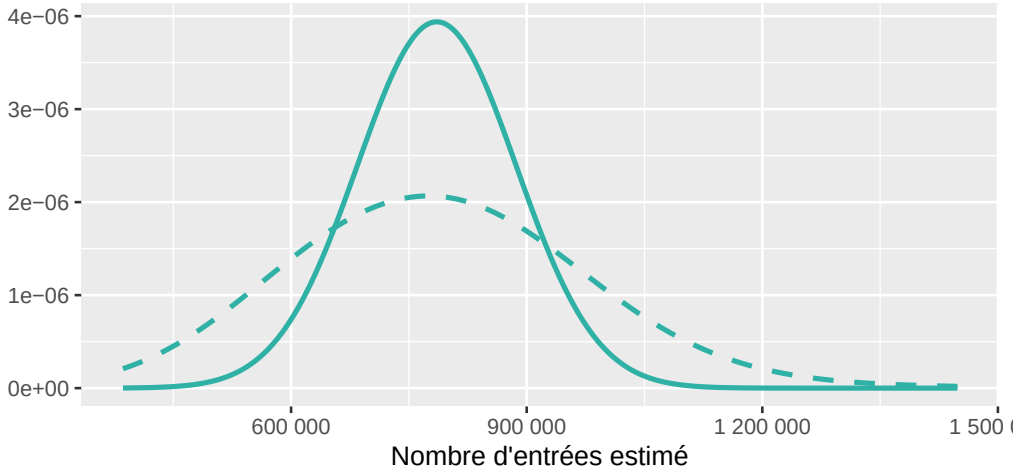
Une fois les informations complètes sur le film remontées, le distributeur **évalue la pertinence d'un redressement par le ratio** en utilisant le nombre de projections comme variable auxiliaire.

Il tire 1 000 échantillons de taille 100 et calcule pour chacun la valeur de l'estimateur d'Horvitz-Thompson et celle de l'estimateur redressé par le ratio.

Exemple L'estimateur par le ratio associé au premier échantillon tiré donne :

$$\hat{T}_{ratio}(Y) = 862944 \times \frac{5413}{6080} = 768250$$

Exemple : Enquête sur la fréquentation des cinémas



Exemple : Enquête sur la fréquentation des cinémas

Valeur dans la population 785 824 entrées

Estimateur d'Horvitz-Thompson (1 000 simulations)

- moyenne empirique : 786 446
- écart-type empirique : 164 687

Estimateur redressé par le ratio (1 000 simulations)

- moyenne empirique : 785 339
- écart-type empirique : 15 981

Redressement par le ratio et repondération

L'estimation par le ratio peut être vue comme une repondération. En notant $d_k = \frac{1}{\pi_k}$ le poids de sondage de l'unité k, l'estimateur d'Horvitz-Thompson s'écrit en effet :

$$\hat{T}_{HT}(Y) = \sum_{k \in s} d_k y_k$$

Dès lors, on peut réécrire l'estimation par le ratio :

$$\hat{T}_{ratio}(Y) = \sum_{k \in s} d_k y_k \times \frac{T(X)}{\hat{T}_{HT}(X)} = \sum_{k \in s} \left(d_k \times \frac{T(X)}{\hat{T}_{HT}(X)} \right) \times y_k = \sum_{k \in s} w_k y_k$$

avec $\forall k \in s \quad w_k = d_k \times \frac{T(X)}{\hat{T}_{HT}(X)}$

En pratique Les redressements sont effectués **une fois pour toutes** au moment de la production d'une enquête. Un vecteur de **poids redressés** est ainsi produit et a vocation à être utilisé à la place des poids de sondage.

Redressements : Post-stratification

Définition

L'estimateur post-stratifié est utilisé quand la variable auxiliaire X est **qualitative** (ou recodée en tranches).

Exemple Le fait pour une salle de cinéma d'être située dans une zone en période de vacances scolaires pour la semaine de référence ou non.

On peut alors définir H groupes d'unités (les **post-strates**) selon les modalités de cette variables et calculer l'estimateur post-stratifié :

$$\hat{T}_{post}(Y) = \sum_{h=1}^{H} \hat{T}_{h,HT}(Y) \frac{N_h}{\hat{N}_{h,HT}}$$

où N_h est le nombre d'unités de la population dans la post-strate h et $\hat{N}_{h,HT}$ son estimateur à partir de l'échantillon.

Remarque Quand le plan de sondage est stratifié selon X, $\hat{N}_{h,HT} = N_h$ et donc $\hat{T}_{post}(Y) = \hat{T}_{HT}(Y)$.

Propriétés

1. Sans biais si tous les N_h **sont entiers**.

2. Variance supérieure à celle d'un SAS stratifié avec allocation proportionnelle ;

$$V(\hat{T}_{post}(Y)) \approx \underbrace{N^2 \frac{1-f}{n} \sum_{h=1}^{H} \frac{N_h}{N} S_h^2}_{\text{Variance d'un SAS stratfié avec alloc. proportionnelle}} + \underbrace{N^2 \frac{1-f}{n^2} \sum_{h=1}^{H} \frac{N-N_h}{N} S_h^2}_{\text{Variance supplémentaire due à la post-stratification}}$$

3. **Propriété de calage** La taille des H post-strates est estimée parfaitement :

$$\forall h = 1, \ldots, H \quad \hat{N}_{h,post} = N_h$$

Exemple : Enquête sur la fréquentation des cinémas

Le distributeur envisage également d'utiliser comme variable auxiliaire le fait que la zone dans laquelle sont situés les cinémas a été en **vacances scolaires** du 21 au 27 février.

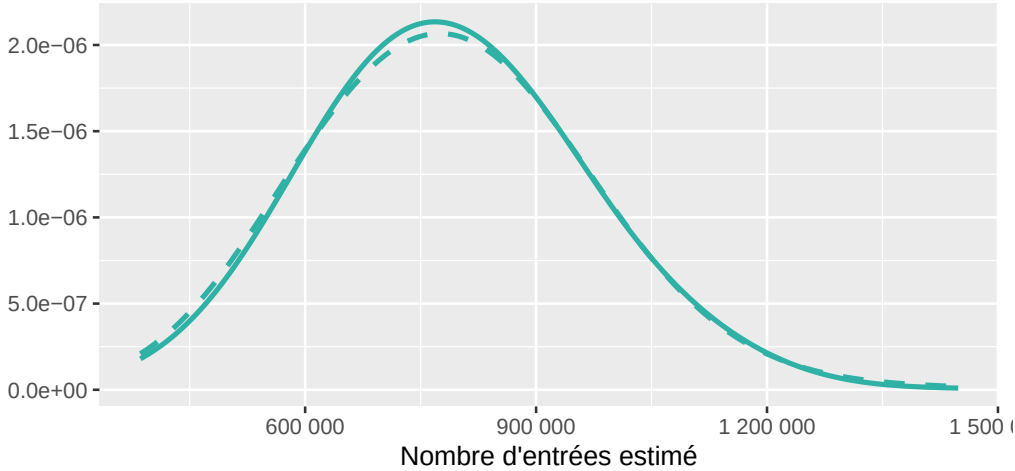
Il constitue donc **deux post-strates** et les utilise pour redresser l'estimateur d'Horvitz-Thompson. À nouveau l'évaluation de la performance de ce redressement est effectuée sur 1 000 simulations.

Exemple À partir du tout premier échantillon, on estime :

- le nombre de cinéma dans une zone en vacances scolaires à 1 252 (contre 1 278 dans la population) ;
- le nombre de cinéma dans une zone non en vacances scolaires à 768 (contre 742 dans la population).

$$\hat{T}_{post}(Y) = 836947 \times \frac{1278}{1252} + 25997 \times \frac{742}{768} = 879185$$

Exemple : Enquête sur la fréquentation des cinémas



Exemple : Enquête sur la fréquentation des cinémas

Valeur dans la population 785 824 entrées

Estimateur d'Horvitz-Thompson (1 000 simulations)

- moyenne empirique : 786 446
- écart-type empirique : 164 687

Estimateur redressé par le ratio (1 000 simulations)

- moyenne empirique : 788 283
- écart-type empirique : 157 773

Post-stratification et repondération

Comme pour l'estimation par le ratio, il est possible de réécrire l'estimateur post-stratifié sous la forme d'une répondération :

$$\hat{T}_{post}(Y) = \sum_{h=1}^{H} \sum_{k \in s_h} d_k y_k \frac{N_h}{\hat{N}_{h,HT}} = \sum_{k \in s} \left(d_k \times \underbrace{\frac{N_h}{\hat{N}_{h,HT}}}_{h|k \in s_h} \right) y_k = \sum_{k \in s} w_k y_k$$

avec $\forall k \in s \quad w_k = d_k \times \underbrace{\frac{N_h}{\hat{N}_{h,HT}}}_{h|k \in s_h}$

En pratique À nouveau, cette propriété permet de simplifier la mise en œuvre des redressements en calculant au moment de la production de l'enquête un **vecteur de poids redressés** à utiliser à la place des poids de sondage.

Redressements : Calage sur marges

Redresser sur plusieurs variables simultanément

Le redressement par le ratio ou la post-stratification sont des méthodes simples et classiques pour utiliser de l'information auxiliaire au moment de l'estimation.

Néanmoins, elles présentent l'une et l'autre une limite principale : **elles ne peuvent intégrer l'information auxiliaire que d'une seule variable**.

Exemple On ne peut pas utiliser conjointement dans les redressements l'information sur le nombre de projections et les vacances scolaires.

Remarque Dans le cas de la post-stratification, une possibilité consiste à croiser les modalités de toutes les variables (qualitatives) que l'on souhaite utiliser, mais cela suppose d'avoir une **information auxiliaire sur leur distribution jointe**.

Calage sur marges : intuition et principe

Au moment de l'estimation on dispose des éléments suivants :

- pour chaque unité k de l'échantillon, un poids de sondage d_k ;
- p **variables de calage** formant la matrice $X = (x_1\ x_2\ \ldots\ x_p)$ et renseignées pour chaque unité k de l'échantillon ;
- la valeur du total **dans la population** des p variables de calage : $T(X) = (T(x_1)\ T(x_2)\ \ldots\ T(x_p))$

Intuition

- Utiliser les poids de sondage d_k **garantit une estimation sans biais**...
- ... mais les modifier de façon à obtenir une estimation parfaite des marges de calage **améliore la précision des estimateurs**.

Principe du calage sur marges Trouver le **vecteur de poids calés** w_k qui conduise à **estimer parfaitement les marges de calage** et qui soit **le plus proche possible de** d_k.

Calage sur marges : formulation du problème

D'un point de vue mathématique, ce problème se formule de la façon suivante :

$$\begin{cases} \min\limits_{w_k} \sum\limits_{k \in s} d_k G\left(\dfrac{w_k}{d_k}\right) \\ \text{sous la contrainte } \sum\limits_{k \in s} w_k X_k = T(X) \end{cases}$$

où G est une certaine **fonction de distance** entre les poids initiaux d_k et les poids finaux w_k :

- $G(1) = 0$;
- $G\left(\dfrac{w_k}{d_k}\right)$ est d'autant plus grand que $\dfrac{w_k}{d_k}$ est différent de 1.

Calage sur marges : fonction de distance et résolution

La résolution de ce problème fait intervenir la **fonction réciproque de la dérivée** de la fonction G, notée en général F.

La forme de la fonction F identifie la **méthode de calage** mise en œuvre, dont les propriétés diffèrent :

- méthode linéaire : $F(x) = 1 + x$
- méthode exponentielle (ou *raking ratio*) : $F(x) = exp(x)$
- méthode logistique : $F(x) = \frac{L(U-1) + U(L-1)exp(Au)}{U - 1 + (1-L)exp(Au)}$ avec L et U des bornes pour $\frac{w_k}{d_k}$ et A une constante
- méthode linéaire tronquée : $F(x) = 1 + x$ pour $x \in [L; U]$

Dans tous les cas, la résolution s'appuie sur un **algorithme itératif**.

Illustration : *Raking ratio* sur deux variables dichotomiques

Identifiant	Sexe	Île-de-France	Poids de sondage d_k
A	H	Oui	10
B	H	Non	10
C	H	Non	10
D	F	Oui	10
E	F	Oui	10
F	F	Non	10

Marges dans la population

- $T(\text{Sexe} = \text{H}) = 20$
- $T(\text{Sexe} = \text{F}) = 40$
- $T(\text{Île-de-France} = \text{Oui}) = 40$
- $T(\text{Île-de-France} = \text{Non}) = 20$

Illustration : *Raking ratio* sur deux variables dichotomiques

Étape 1

Sexe \ IdF	Oui	Non	Marge	
H	10	20	30 (20)	$\times$ 20/30
F	20	10	30 (40)	$\times$ 40/30
Marge	30 (40)	30 (20)	60 (60)	

Étape 2

Sexe \ IdF	Oui	Non	Marge
H	6,67	13,33	20 (20)
F	26,67	13,33	40 (40)
Marge	33,34 (40)	26,67 (20)	60 (60)
	$\times$ 40/33,34	$\times$ 20/26,67	

Illustration : *Raking ratio* sur deux variables dichotomiques

Étape 3

Sexe \ IdF	Oui	Non	Marge	
H	8	10	18 (20)	× 20/18
F	32	10	42 (40)	× 40/42
Marge	40 (40)	20 (20)	60 (60)	

Étape 4

Sexe \ IdF	Oui	Non	Marge
H	8,88	11,11	20 (20)
F	30,48	9,52	40 (40)
Marge	39,36 (40)	20,63 (20)	60 (60)
	× 40/39,36	× 20/20,63	

Illustration : *Raking ratio* sur deux variables dichotomiques

Étape 5

9,03	10,77	19,80 (20)
30,97	9,23	40,20 (40)
40 (40)	20 (20)	60 (60)

Étape 6

9,12	10,88	20 (20)
30,81	9,19	40 (40)
39,93 (40)	20,07 (20)	60 (60)

Étape 7

9,14	10,84	19,98 (20)
30,86	9,16	40,02 (40)
40 (40)	20 (20)	60 (60)

Étape 8

9,15	10,85	20 (20)
30,85	9,15	40 (40)
40 (40)	20 (20)	60 (60)

Détermination des poids finaux Multiplication du poids initial d_k par le rapport entre les totaux de chaque cellule après/avant l'algorithme de calage.

Exemple $d_B = 10$, $sexe_B = \text{H}$ et $idf_B = \text{Non}$

- total final/initial de la cellule : 10,85/20
- poids final $w_b = 10 \times 10{,}85/20 = 5{,}425$

Illustration : *Raking ratio* sur deux variables dichotomiques

Identifiant	Sexe	Île-de-France	d_k	Poids calé w_k
A	H	Oui	10	9,150
B	H	Non	10	5,425
C	H	Non	10	5,425
D	F	Oui	10	15,425
E	F	Oui	10	15,425
F	F	Non	10	9,150

Vérification des contraintes de calage

- $\hat{T}(\text{Sexe} = \text{H}) = 9{,}150 + 5{,}425 + 5{,}425 = 20 = T(\text{Sexe} = \text{H})$
- $\hat{T}(\text{Sexe} = \text{F}) = 15{,}425 + 15{,}425 + 9{,}150 = 40 = T(\text{Sexe} = \text{F})$
- $\hat{T}(\text{Idf} = \text{Oui}) = 9{,}150 + 15{,}425 + 15{,}425 = 40 = T(\text{Idf} = \text{Oui})$
- $\hat{T}(\text{Idf} = \text{Non}) = 5{,}425 + 5{,}425 + 9{,}150 = 20 = T(\text{Idf} = \text{Non})$

Propriétés de l'estimateur obtenu par calage

1. Quelle que soit la méthode, asymptotiquement sans biais :

$$B(\hat{T}_{calage}(Y)) \xrightarrow[n\to+\infty]{} 0$$

2. Quelle que soit la méthode, variance **approximativement égale** et qui s'exprime en fonction d'un résidu :

$$V(\hat{T}_{calage}(Y)) \approx V(\hat{T}_{calage}(\varepsilon))$$

où ε est le **résidu de la régression (linéaire) de Y sur les variables de calage**.

Moralité Plus les variables de calage X sont corrélées à Y, plus le résidu de la régression de Y sur X est faible et **plus la variance de l'estimateur du total de Y est elle-même faible**.

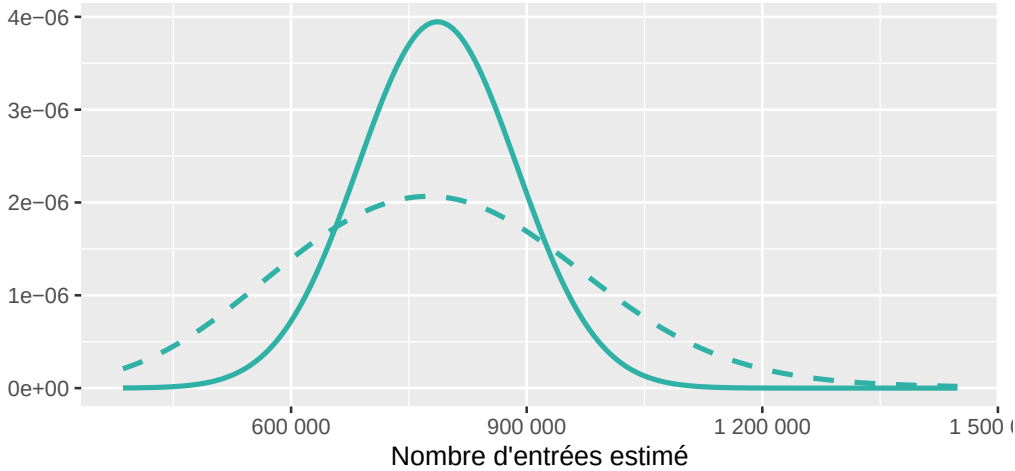
Exemple : Enquête sur la fréquentation des cinémas

Le distributeur souhaite **exploiter conjointement** l'information auxiliaire sur le nombre de projections et les périodes de vacances scolaires.

Pour ce faire, il introduit ces deux variables dans un calage sur marges par la **méthode exponentielle** (ou méthode du *raking ratio*).

À nouveau, il évalue les propriétés de l'estimateur en **répliquant 1 000 fois** l'ensemble des opérations (tirage puis redressement) et en représentant la **distribution des estimations ainsi obtenues**.

Exemple : Enquête sur la fréquentation des cinémas



Exemple : Enquête sur la fréquentation des cinémas

Biais Moyenne empirique sur 1 000 simulations

- Valeur dans la population : 785 824 entrées
- Estimateur d'Horvitz-Thompson : 786 446
- Estimateur par le ratio : 785 339
- Estimateur par post-stratification : 788 283
- Estimateur par calage sur marges : 786 292

Précision Écart-type empirique sur 1 000 simulations

- Estimateur d'Horvitz-Thompson : 164 687
- Estimateur par le ratio : 15 981
- Estimateur par post-stratification : 157 773
- Estimateur par calage sur marges : 14 679

Le calage sur marges en pratique

La plupart des enquêtes par sondage font l'objet d'un calage sur marges sur les **grandes structures de la population**, la pyramide des âges notamment.

En effet, une telle opération **ne peut qu'améliorer la précision** et garantit la **cohérence avec des sources extérieures à l'enquête**.

Est ainsi diffusé dans le fichier de l'enquête non pas le poids de sondage mais le **poids calé** sur de nombreuses marges.

En pratique, le calage sur marges est implémenté dans de nombreux logiciels :

- SAS : macro *%calmar* ;
- R : *packages* `sampling` et `icarus`.

En guise de conclusion

Les méthodes de redressement cherchent à **exploiter l'information auxiliaire disponible** au moment de l'estimation pour **améliorer la précision**.

Les estimateurs par le **ratio** et **post-stratifié** présentent une **variance plus faible** que l'estimateur d'Horvitz-Thompson pour autant que la variable d'intérêt soit **bien corrélée** à la variable explicative utilisée.

La méthode du **calage sur marges** généralise ce principe et permet de tirer parti de plusieurs variables auxiliaires simultanément.