Introduction à la théorie des sondages - Cours 5

Gaspar Massiot
gaspar.massiot@ined.fr

2024-2025

Chapitre 3

Compléments sur l'échantillonnage

Partie 1

Sondage à deux degrés

Introduction

On parle de sondage à deux degrés lorsque :

- On réalise un premier degré de tirage (par exemple : on sélectionne des logements) en suivant un plan de sondage spécifique.
- L'échantillon obtenu permet de constituer la base de sondage du second degré (ici, l'ensemble des occupants des logements)
- On définit alors un plan de sondage pour faire un tirage dans cette base (on sélectionne des individus)
- Le dernier plan de sondage peut dépendre de l'étape de tirage du premier degré.

Echantillon-maître

Cas classique :

- Le premier degré est constitué de communes ou de zones géographiques.
- L'échantillon de premier degré correspond à un territoire sur lequel l'enquête va se faire.
- On échantillonne ensuite des logements au sein de ce territoire.
- Remarque : on parle d'échantillon-maître lorsque ce tirage est fait pour plusieurs enquêtes (Octopusse 2009-2019).

Avantages et inconvenients

Avantages :

- Gestion : on sait où vont être les enquêtes.
- Coût : on n'a pas besoin de couvrir tout le territoire / tous les logements.

Inconvénients :

- Complexité : plus difficile à mettre en oeuvre qu'un sondage dispersé.
- Réinterrogation : plus de risques de ré-enquêter la même personne.
- Imprécision : on diminue la précision des estimations (voir après).

Auto-pondéré

Quel plan de sondage utiliser ?

- Notre principal objectif : tous les logements doivent avoir le même poids.
- Objectif secondaire : chaque zone a la même charge de collecte.
- Solution : on tire à probabilités inégales les zones puis uniformément les logements.

Sondage auto-pondéré

1. On échantillonne des zones à probabilités inégales selon X, le nombre de logements ;
2. On sélectionne 20 logements dans la zone.

Pourquoi ils ont la même probabilité ?

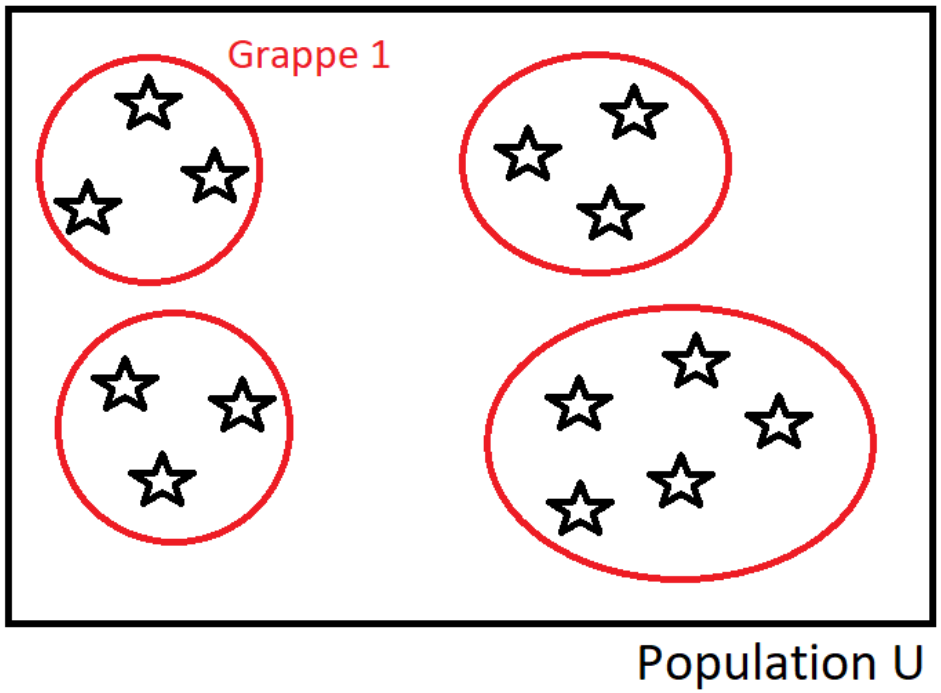
- Une zone de 2 000 logements a une chance sur 20 d'être sélectionnée. Ensuite, chaque logement a une chance sur 100 : au total, une chance sur 2 000.
- Une zone de 20 000 logements a une chance sur 2 d'être sélectionnée. Ensuite, chaque logement a une chance sur 1 000 : au total, une chance sur 2 000.

Partie 2

Sondage par grappes

Grappes

Le sondage par grappes repose sur l'idée que la population U est répartie entre groupes, qu'on appelle "grappes". On réalise ensuite un échantillon de grappes, et on va enquêter l'intégralité des occupants de la strate.



Grappes

Quelques exemples :

- On peut échantillonner des bâtiments pour enquêter tous les occupants (Recensement de la Population) ;
- On peut enquêter tous les occupants d'un ménage tiré au hasard

Attention :

- Ne pas confondre avec le sondage stratifié : on tire certaines grappes et non dans chaque strate.
- Ressemble au sondage à deux degrés

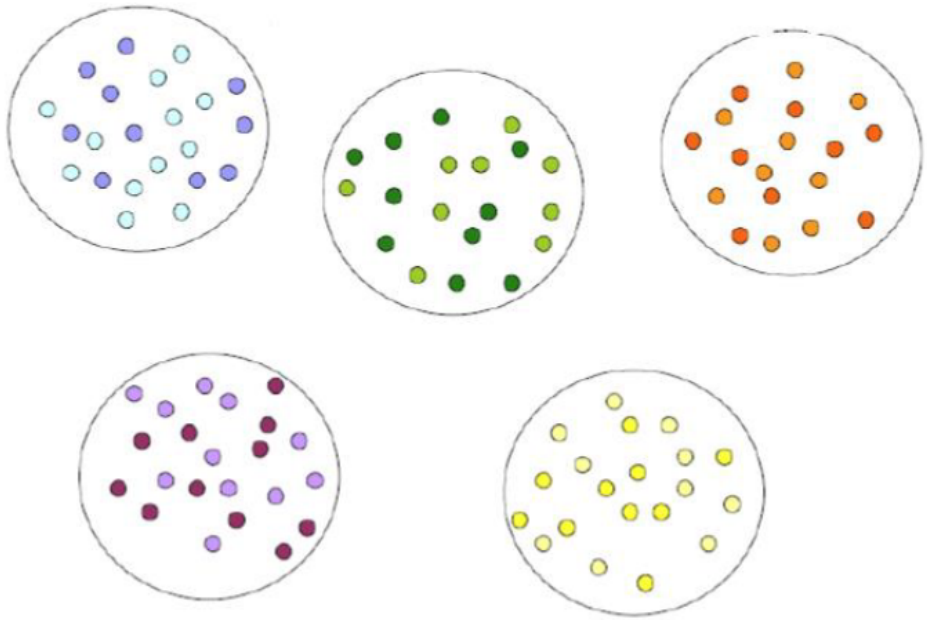
Effet de grappe

On appelle "effet de grappe" le phénomène de perte de précision due à la similarité des comportements entre individus d'une même grappe.
Lorsque toutes les grappes sont de même taille N_g, L'effet de grappe peut être mesuré par un coefficient appelé "coefficient de corrélation intra-grappe" noté ρ :

$$\rho = \frac{1}{N_g - 1}\left[N_g \frac{\sigma^2_{y,inter}}{\sigma^2_y} - 1\right]$$

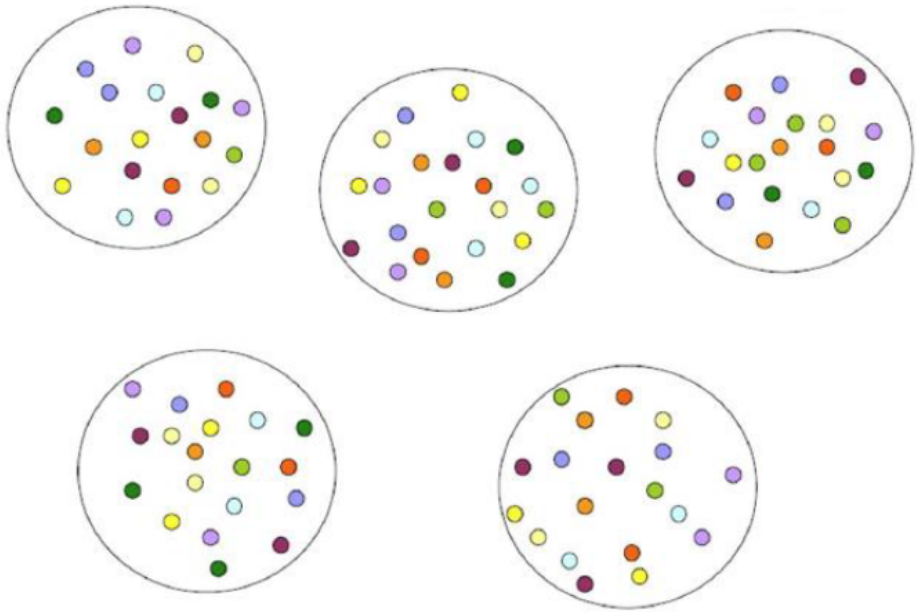
Effet de grappe

Quand ρ est grand, les grappes regroupent des individus qui se ressemblent :



Effet de grappe

Quand ρ est petit, à l'inverse, les grappes regroupent des individus différents entre eux :



Effet de grappe

Utiliser un sondage par grappes conduit à une précision dégradée ; on appelle *design effect* le ratio entre la précision obtenue et celle obtenue avec un SAS, et sa formule est :

$$Deff = \frac{Var_{Grappes}}{Var_{SAS}} = 1 + \rho(N_g - 1)$$

Ce ratio est d'autant plus grand que les grappes sont grandes (N_g) et homogènes (ρ). Malheureusement, souvent, les grappes sont des zones géographiques qui regroupent des individus qui se ressemblent (c'est le cas dans l'Enquête Emploi en Continu de l'Insee, par exemple).

Partie 3

Méthodes empiriques

Méthodes empiriques

- Méthode des Quotas
  - On fixe des quotas pour des catégories d'intérêt
  - Non probabiliste (pas de base de sondage)
- Méthode Responden Driven Sample (RDS), *boule de neige*
  - Un premier échantillon puis par recommendation
  - Utile pour cibler des populations rares
- Méthode des Itinéraires
  - Si on ne dispose pas de liste
  - On tire un point GPS puis logement suivant un itinéraire donné

Sommaire

Chapitre 1

Non-réponse aux enquêtes

Partie 1

Introduction

Définition

Qu'est-ce qu'un non répondant ? On aurait souhaité que ces individus nous donnent des informations car ils nous intéressent dans le cadre de l'enquête, mais ce n'est pas le cas, pour différentes raisons.

On peut distinguer deux types de non-réponse :

- Partielle : il manque une ou plusieurs réponses, mais pas toutes.
- Totale : il manque toutes ou quasiment toutes les réponses.

Causes

Il existe énormément de raisons de ne pas répondre à une enquête :

- Ne souhaite pas répondre ;
- N'est pas disponible pour répondre ;
- Ne peut pas répondre (langue, handicap) ;
- Considère que la question est intrusive ou inadaptée ;
- Considère qu'il n'y a pas de réponse adéquate ;
- L'enquêteur n'est jamais venu ;
- L'enquêteur n'a pas réussi à contacter la personne ;
- L'enquêteur n'a pas réussi à contacter la personne qui l'évitait ;
- L'enquêteur n'a pas réussi à rentrer dans l'immeuble.

Conséquences

Les conséquences de la non-réponse sont les suivantes :

- Baisse de la taille de l'échantillon exploitable ;
- Différences de structure entre échantillon et répondants ;
- Question, y-a-t-il des différences entre le comportement des répondants et des non-répondants ?

Ignorabilité

Trois façons de caractériser la non-réponse :

- Complètement ignorable : les répondants et les non-répondants sont en moyenne les mêmes (dit CMAR)
- Ignorable : les répondants et les non-répondants diffèrent en structure mais leur comportement de réponse est le même conditionnellement à leur âge, sexe...(dit MAR)
- Non ignorable : les non-répondants sont différents des répondants ; cette différence explique pourquoi ils n'ont pas répondu. (dit NMAR)

Le type de non-réponse vaut pour une variable Y et un paramètre précis (moyenne, ...).

Ignorabilité - Exemples

Exemples :

- Complètement ignorable : l'enquêteur n'a eu le temps de contacter que les fiches 1 à 10 et pas les suivantes.
- Ignorable : les jeunes répondent moins souvent à l'enquête que les personnes âgées.
- Non ignorable : on pose la question : "faites-vous confiance aux institutions de votre pays ?" $\rightarrow$ ne pas répondre à une enquête de l'INSEE semble un bon indicateur que non...

Ignorabilité - Exemples

- Les personnes ayant un plus fort patrimoine répondent moins volontiers à l'enquête HVP (Histoire de Vie et Patrimoine)
- Pour l'EEC, les personnes en emploi sont plus difficiles à joindre en journée que les personnes inactives.
- Un enquêteur peut ne pas souhaiter se rendre dans un quartier jugé difficile et exercer son droit de retrait.

Travail en amont

Remarque : pour un bon nombre de ces raisons, il est possible de réduire le risque de non-réponse en travaillant en amont sur le questionnaire et l'organisation de la collecte. Par exemple :

- *Ne peut pas répondre (langue, handicap) :* prévoir des traductions ou des interprètes.
- *Considère qu'il n'y a pas de réponse adéquate :* adapter les modalités.
- *L'enquêteur n'a pas réussi à contacter la personne :* organiser des contacts à différents jours et horaires.

Conséquences - retour

Que penser des conséquences ?

- *Baisse de la taille de l'échantillon exploitable :* correspond à une perte en précision ; on peut anticiper.
- *Différences de structure entre échantillon et répondants :* risque de biais, corrigeable.
- *Différences entre le comportement des répondants et des non-répondants :* risque de biais, corrigeable ?

Perte de précision

Anticiper la perte de précision (augmentation de la variance des estimateurs) :

- Non réponse $\Leftrightarrow$ Échantillon utilisable plus petit $\Leftrightarrow$ Variance d'estimation plus élevée
- Peut augmenter avec la dispersion des poids (voir après) ;
- Possible d'anticiper le taux de non-réponse et d'augmenter la taille de l'échantillon au préalable

Risque de biais

Corriger le biais des estimateurs :

- Plus problématique car difficile à anticiper ;
- Pas possible à calculer exactement si on ne connaît pas la vraie valeur ;
- Dépend de l'importance du taux de non-réponse et de son ignorabilité.

Cas CMAR (Completely Missing At Random)

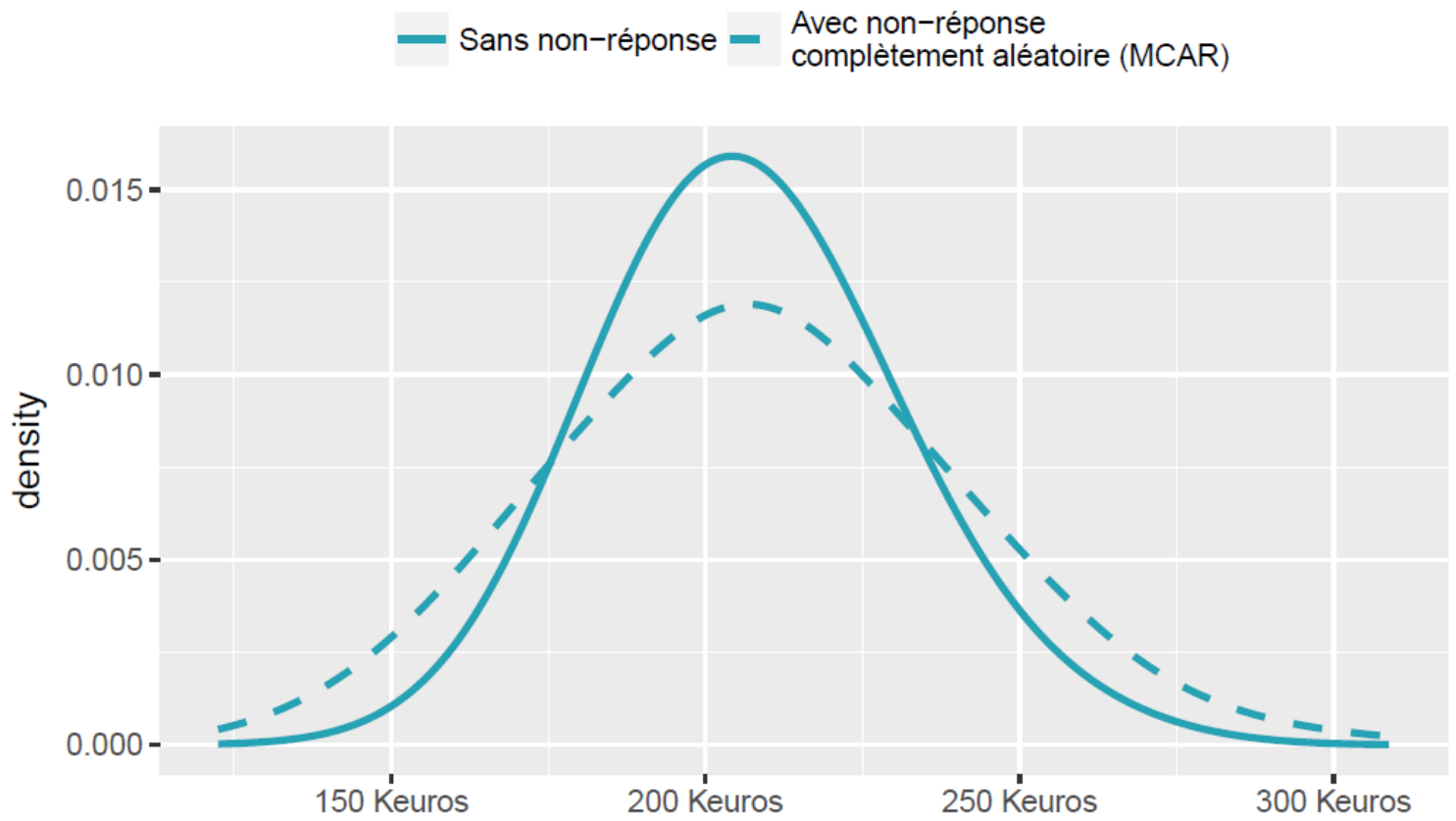
Dans le cas CMAR, la non-réponse est complètement ignorable :

- Il n'y a pas de biais sur l'estimateur de la moyenne : les répondants sont similaires aux non-répondants en moyenne ;
- On observe cependant un biais sur l'estimateur du total ;
- Le principal problème est la perte de précision.

Ce cas est extrêmement atypique à l'INSEE.

Cas CMAR - Exemple

Estimation de la moyenne :



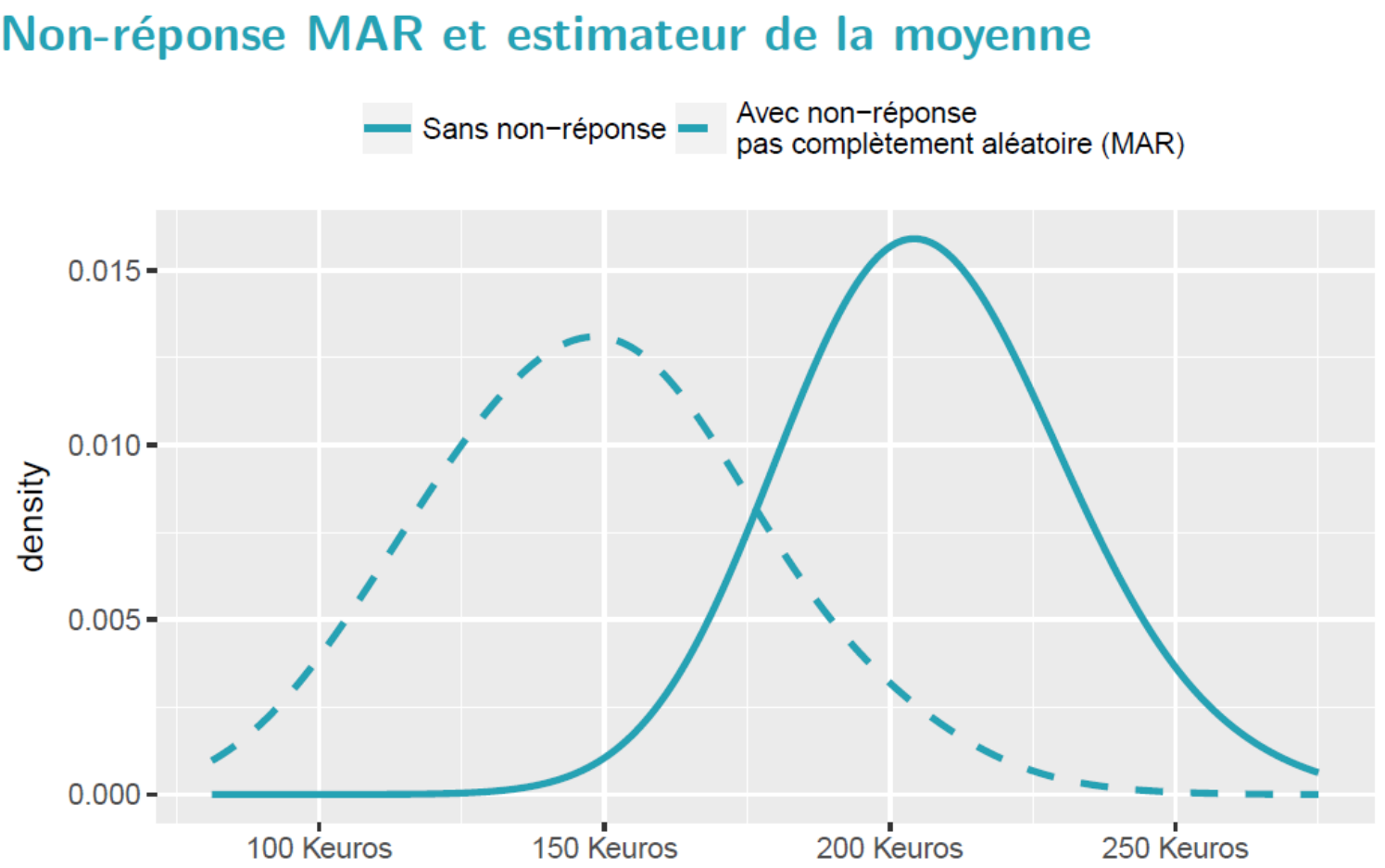
Cas MAR (Missing At Random) et NMAR (Not Missing At Random)

Dans les cas MAR et NMAR, la non-réponse entraîne une différence de structure :

- Il y a un biais sur les estimateurs du total et de la moyenne ;
- Les réponses des répondants ne sont pas similaires à celles qu'auraient donné les non-répondants en moyenne ;
- Le principal problème est d'essayer de corriger le biais.

Ce cas est le plus classique à l'INSEE : il peut être MAR ou NMAR.

Cas MAR et NMAR - Exemple



Exemple numérique

Id	Type de Logement	Loyer
A	Maison	1000
B	Maison	2000
C	Maison	1500
D	Maison	1200
E	Maison	1800
F	Maison	1500
G	Appartement	500
H	Appartement	400
I	Appartement	700
J	Appartement	600
K	Appartement	300
L	Appartement	500

Loyer moyen estimé : 1000

Exemple numérique - CMAR

Id	Type de Logement	Loyer
A	Maison	1000
B	Maison	/
C	Maison	1500
D	Maison	/
E	Maison	1800
F	Maison	/
G	Appartement	500
H	Appartement	/
I	Appartement	700
J	Appartement	/
K	Appartement	300
L	Appartement	/

Loyer moyen estimé : 967

Exemple numérique - MAR ou NMAR

Id	Type de Logement	Loyer
A	Maison	1000
B	Maison	2000
C	Maison	1500
D	Maison	1200
E	Maison	1800
F	Maison	1500
G	Appartement	500
H	Appartement	/
I	Appartement	/
J	Appartement	600
K	Appartement	/
L	Appartement	/

Loyer moyen estimé : 1267,5

Pistes de solution

Il existe deux visions de ce problème :

- Les poids de sondage ne sont plus égaux au nombre d'unités de la population que chaque répondant représente : il faut ajuster les poids pour que les poids des répondants absorbent les poids des non-répondants $\Rightarrow$ **Méthodes de repondération**.
- L'information dont nous avons besoin est manquante pour certaines observations de l'échantillon : nous devons boucher les trous $\Rightarrow$ **Méthodes d'imputation**.

Partie 2

Méthodes par repondération

Méthodes par repondération

On assimile la situation à un plan de sondage en deux phases :

1. PREMIÈRE PHASE : le plan de sondage initial détermine les probabilités d'inclusion π_i et les poids de sondage $1/\pi_i$
2. SECONDE PHASE : sélection aléatoire des répondants selon un plan de sondage inconnu $\Pi_{/\mathcal{S}}$ qui détermine des probabilités d'inclusion ρ_i

Méthodes par repondération

Si les probabilités d'inclusion de seconde phase étaient connues :

- l'estimateur HT $\hat{Y}_{HT} = \sum_{i \in \mathcal{R}} \frac{y_i}{\pi_i}$ est biaisé
- il serait remplacé par l'estimateur sans biais selon le plan de sondage $\hat{Y}_{NR} = \sum_{i \in \mathcal{R}} \frac{y_i}{\pi_i \, \rho_i}$

Cet estimateur est appelé **estimateur corrigé de la non-réponse**
$\Rightarrow$ objectif de la repondération : estimer ρ_i, parfois appelés **score de propension** à répondre

Méthodes par repondération

ρ_i sont inconnus $\Rightarrow$ ils doivent être remplacés par des estimateurs $\hat{\rho}_i$.

$\hat{Y}_{NR}$ est asymptotiquement sans biais (approximativement sans biais à distance finie) si et seulement si le modèle de non-réponse utilisé pour estimer les probabilités d'inclusion de seconde phase est correct, *i.e.* si

$$\hat{\rho}_i = \hat{\pi}_{i/\mathcal{S}} \xrightarrow[N,n\to+\infty]{} \rho_i$$

Méthodes par repondération

C'est difficile à vérifier avec les données : il est en général seulement possible de mobiliser le maximum d'information auxiliaire et de supposer que la majeure partie du biais de non-réponse a été éliminée.

Hypothèse et modèle

Nous supposons que :

- nous avons des variables auxiliaires Z disponibles pour les répondants et les non-répondants
- la non-réponse est ignorable : quand nous contrôlons Z, le fait de répondre ne dépend pas des caractéristiques des unités
- R_i, la variable égale à 1 pour les répondants et 0 sinon, a une distribution de Bernoulli de paramètre $\rho_i = f(Z_i)$

Modèle

On peut donc estimer ρ_i en utilisant un **modèle logistique** liant les R_i aux Z_i : $\hat{\rho}_i = \hat{f}(Z_i)$.

Cependant, on mobilise tellement d'information issue des répondants que le risque de **surapprentissage** est très grand.

Solution : Les méthodes dites de Groupes de Réponse Homogène (GRH) qui consistent à **diviser la population en sous-parties** dans lesquelles on corrigera plus "facilement" la non-réponse, permettent de limiter ce surapprentissage.

Partie 3

Méthodes par imputation

Définition

L'imputation consiste à remplacer une donnée manquante par une donnée ≪ plausible ≫ déduite ou calculée en fonction des renseignements obtenus pour l'unité défaillante et/ou pour les unités qui lui sont proches.
Les méthodes d'imputation ont pour but :

1. de réduire le biais de non-réponse ;
2. de produire des tableaux de données ≪ rectangulaires ≫ sans trous.

Définition

Les méthodes d'imputation peuvent être classées en deux groupes :

- les méthodes dites **déterministes** : méthodes qui fournissent une valeur fixe étant donné l'échantillon si le processus d'imputation est répété ;
- les méthodes dites **stochastiques** ou **aléatoires** : méthodes d'imputation ayant une composante aléatoire (et donc qui ne donnent pas nécessairement la même valeur imputée étant donné l'échantillon si le processus d'imputation est répété).

Méthodes déterministes

- **Méthode déductive** : Utilisée lorsque la valeur peut être déduite avec certitude (total avec toutes les valeurs données)
- **Cold-deck** : On déduit la valeur manquante à l'aide d'autres sources (données historiques, administratives, etc.)
- **Moyenne, Ratio, Régression** : On impute par la moyenne, le ratio (calculé avec une autre variable Y) ou une régression. Cela modifie la distribution. On peut le faire par classe. Les méthodes par ratio et régression sont dépendantes du lien entre les deux variables.
- **Tendance unitaire** : Lors d'enquêtes répétées, on peut utiliser la tendance observée sur les répétitions précédentes pour en déduire la nouvelle valeur.
- **Plus proche voisin** : On impute la valeur par celle d'un voisin le "plus proche" au sens d'une distance choisie.

Méthodes aléatoires

- **Hot-deck** : Imputation par une valeur existante dans la même enquête. La sélection peut se faire par l'aléatoire. Les plus proches voisins est un type de Hot-deck non aléatoire.
- **Imputation avec résidu** : Imputation déterministe bruitée par un résidu ε, le plus souvent de loi normale, $\varepsilon \sim \mathcal{N}(0, \sigma)$
  - Moyenne
  - Ratio
  - Régression

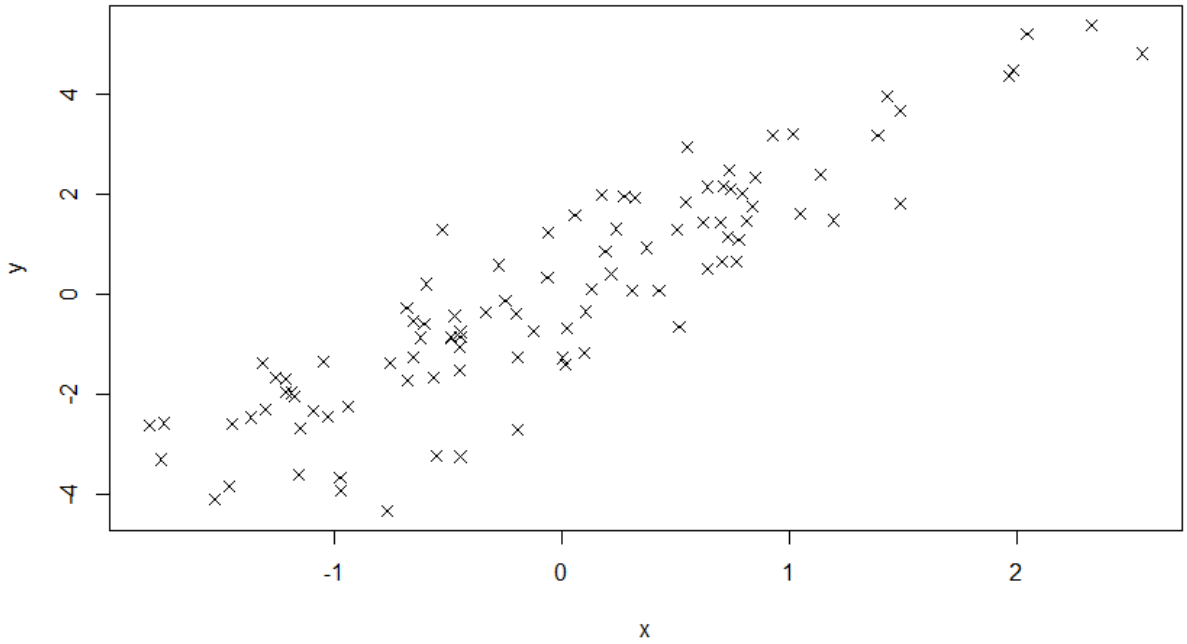
Définition

Une classification alternative peut être utilisée. On distingue alors les méthodes d'imputation par donneur de celles par valeur prédite :

- les méthodes par **donneur** : méthodes d'imputation qui utilisent les valeurs d'un autre individu pour remplacer les valeurs manquantes ;
  $\rightarrow$ hot-deck aléatoire et plus proche voisin
- les méthodes par **valeur prédite** : méthodes d'imputation qui utilisent diverses fonctions des valeurs des répondants pour obtenir les valeurs de remplacement (valeurs imputées)
  $\rightarrow$ méthode déductive, cold deck, [moyenne, ratio, régression et tendance unitaire] avec ou sans aléa...

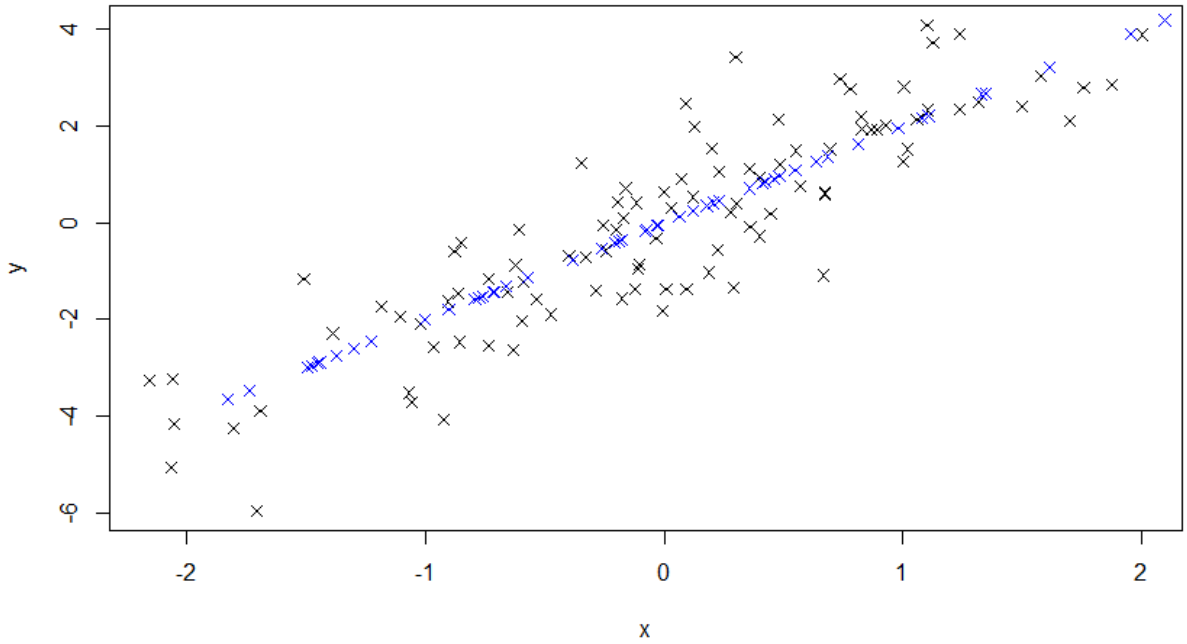
Exemple - Imputation par le ratio

Avant imputation :



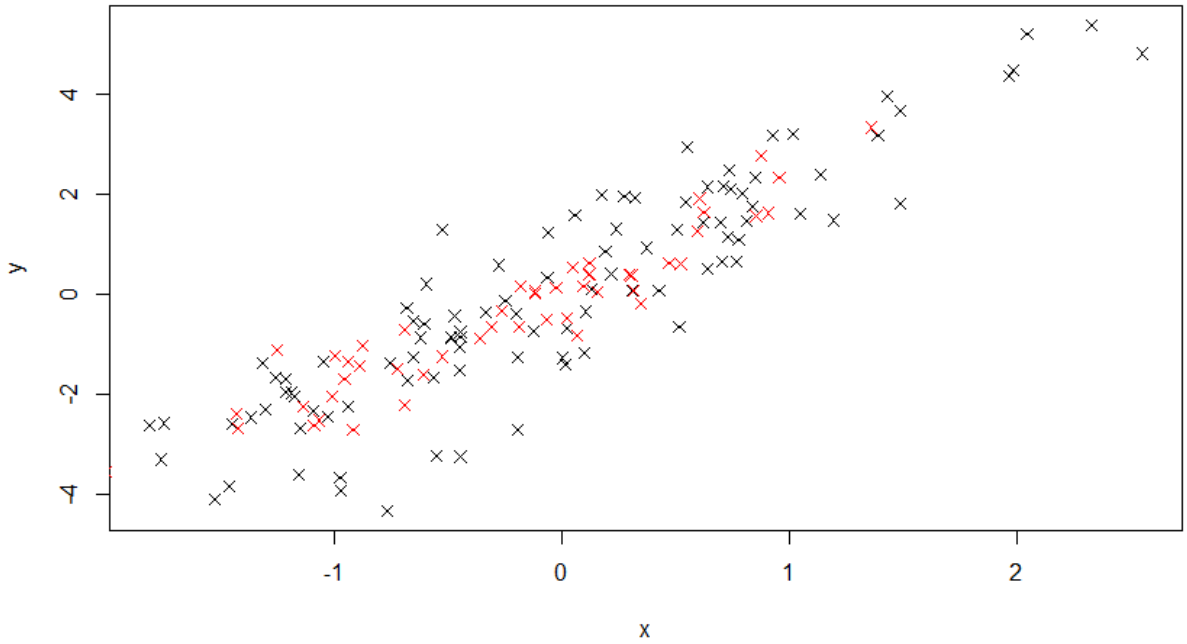
Exemple - Imputation par le ratio sans résidus

Après imputation par le ratio sans résidu :



Exemple - Imputation par le ratio avec résidus

Après imputation par le ratio avec résidus :



Partie 4

Quelle méthode choisir ?

Déterministe ou aléatoire ?

A propos de l'imputation :

- Les imputations **déterministes** faussent la distribution de la variable d'intérêt, donc **augmentent le biais**.
- Les imputations **aléatoires** préservent la distribution des variables d'intérêt, au prix d'une **augmentation de la variance** due à l'imputation.
  $\Rightarrow$ choix en fonction des objectifs principaux de l'enquête !

Remarque : les méthodes par donneur permettent d'imputer avec un donneur unique des ensembles de questions ; c'est très utile pour préserver les liens entre variables...

Repondération ou imputation ?

Repondération :

- Avantages :
  - **Aucune valeur n'est créée**
  - Permet l'utilisation d'un fichier complet mais...
- Inconvénients :
  - ...de taille plus petite
  - **Inadapté à la non-réponse partielle** : il faudrait créer un jeu de poids par variable

Repondération ou imputation ?

Imputation :

- Avantages :
  - Permet l'utilisation d'un poids de sondage unique
  - Permet l'utilisation d'un fichier complet
  - **Utilise toute l'information observée**, même un questionnaire partiellement rempli
- Inconvénients :
  - Peut donner une **fausse impression de complétude** et faire oublier la non-réponse ; penser à bien "flagger" les variables imputées
  - Peut **modifier distributions et corrélations entre variables**, ce qui est dangereux pour des exploitations de l'enquête (économétrie...)

Repondération ou imputation ?

- Pour la **non-réponse partielle**, uniquement **imputation** : en effet, repondération inadaptée et imputation permettent l'utilisation d'informations plus riches.
- Pour la **non-réponse totale**, imputation et repondération sont valables. Pas de supériorité théorique d'une méthode sur l'autre
- En général, on privilégie quand même la **repondération** :
  - perturbe moins les liaisons entre variables ;
  - ne donne pas l'illusion que l'on a affaire à des données complètes ;
  - calculs de variance beaucoup plus simples.

Repondération ou imputation ?

- Remarque : la séparation non-réponse totale / partielle n'est pas si claire que cela.
- Que dire d'un questionnaire où une seule question a été remplie ?
- Ou alors pour l'enquête Emploi, si il manque le module permettant de calculer le statut d'activité ?

Chapitre 2

Post-traitements et précision

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.

Application : Enquête sur la fréquentation des cinémas

Le distributeur d'un film souhaite connaître le **nombre d'entrées réalisées par un film sur 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 2020 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 20 au 27 février 2017.

L'estimateur d'Horvitz-Thompson obtenu est de **464 923** avec un **intervalle de confiance à 95 % de [243 061 ; 686 785]**.

Application : 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 5 061 fois ;
- à partir de l'échantillon, ce nombre est estimé à 3 333 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 $\frac{5061}{3333} = 1,52$.
- L'utilisation du nombre de projections comme information auxiliaire pourrait venir « **stabiliser** » l'estimateur.

Partie 1

Redressement par le ratio

Définition

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

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.

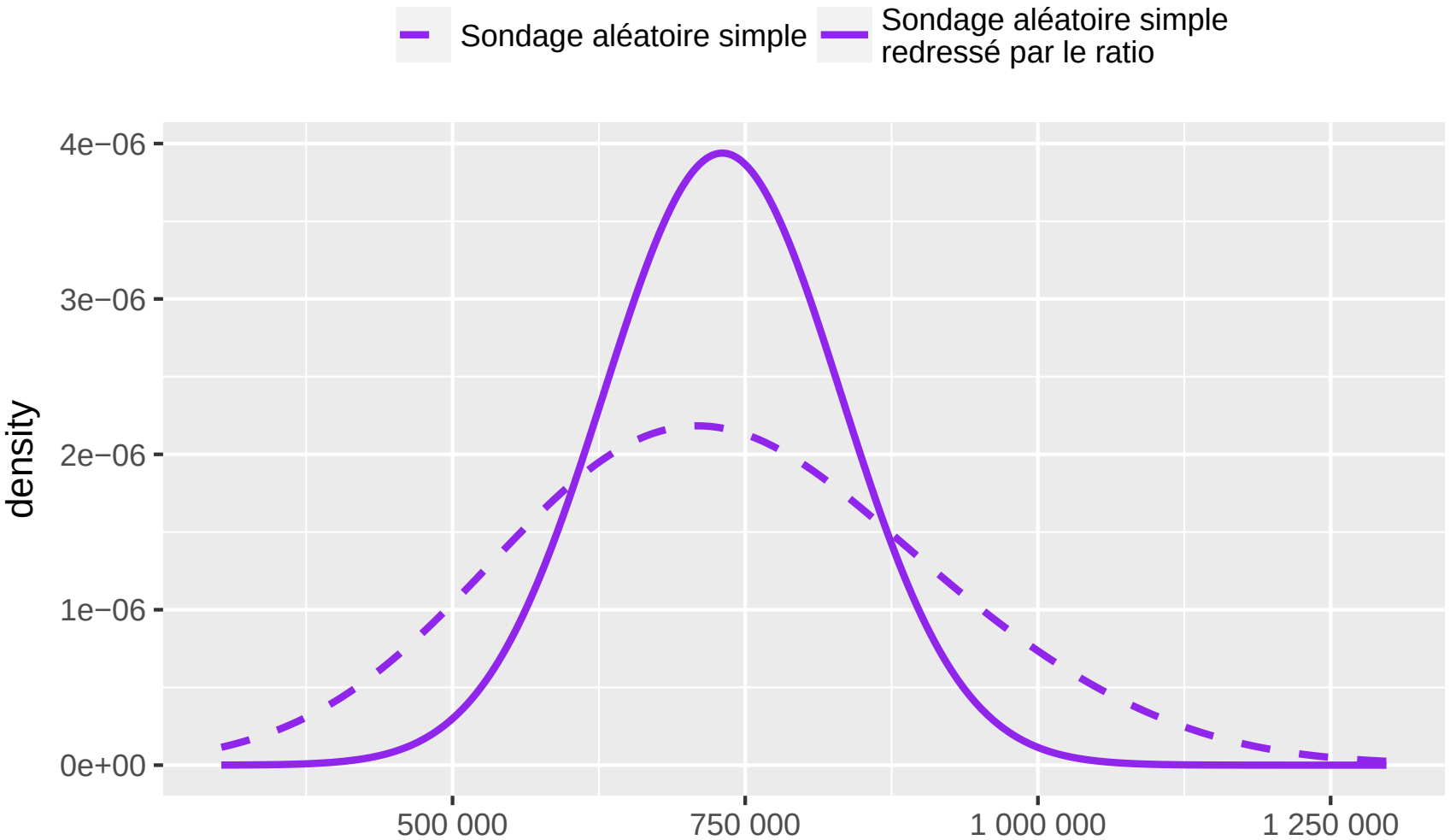
Application : Enquête sur la fréquentation des cinémas

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 :

$$\hat{T}_{ratio}(Y) = 464923 \times \frac{5061}{3333} = 705963$$

Application : Enquête sur la fréquentation des cinémas



Application : Enquête sur la fréquentation des cinémas

Valeur dans la population 731 892 entrées

Estimateur d'Horvitz-Thompson (1 000 simulations)

- moyenne empirique : 730 942
- écart-type empirique : 153 835

Estimateur redressé par le ratio (1 000 simulations)

- moyenne empirique : 730 299
- écart-type empirique : 16 039

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

Redressement par le ratio et repondération

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.

Partie 2

Redressement par post-stratification

Définition

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

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 $\hat{T}_{post}(Y) = \hat{T}_{HT}(Y)$.

Application : 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 20 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.

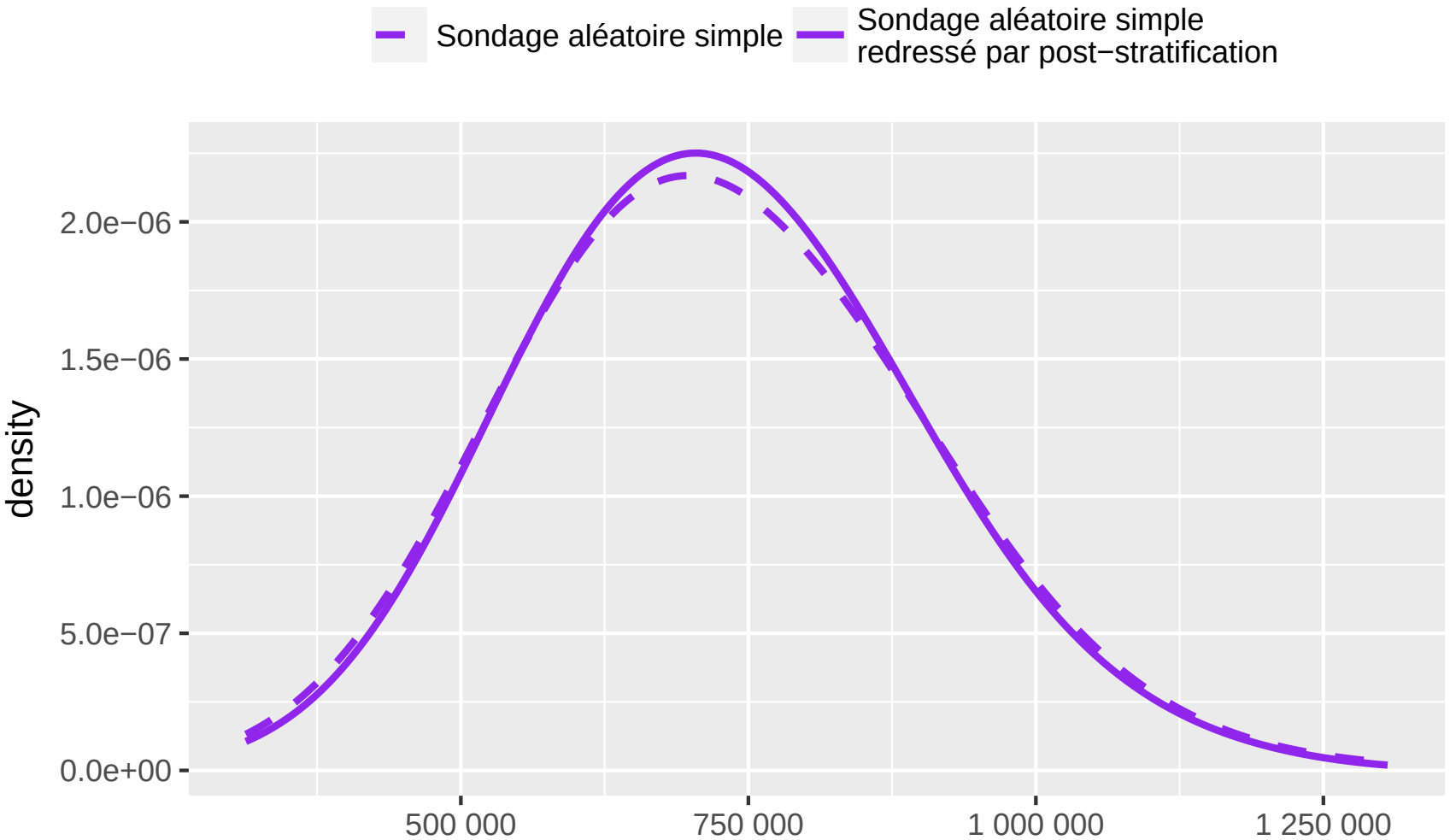
Application : Enquête sur la fréquentation des cinémas

Exemple À partir du tout premier échantillon, on estime :

- le nombre de cinéma dans une zone en vacances scolaires à 1 192 (contre 1 244 dans la population) ;
- le nombre de cinéma dans une zone non en vacances scolaires à 828 (contre 776 dans la population).

$$\hat{T}_{post}(Y) = 443915 \times \frac{1244}{1192} + 21008 \times \frac{776}{828} = 483042$$

Application : Enquête sur la fréquentation des cinémas



Application : Enquête sur la fréquentation des cinémas

Valeur dans la population 731 892 entrées

Estimateur d'Horvitz-Thompson (1 000 simulations)

- moyenne empirique : 725 513
- écart-type empirique : 153 752

Estimateur redressé par le ratio (1 000 simulations)

- moyenne empirique : 725 812
- écart-type empirique : 145 332

Partie 3

Généralisation : 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 $X = (x_1 \ x_2 \ \dots \ x_p)$ et renseignées pour chaque unité k de l'échantillon ;
- la valeur du total **dans la population** des p variables de calage : $T(X) = (T(x_1) \ T(x_2) \ \dots \ T(x_p))$

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.

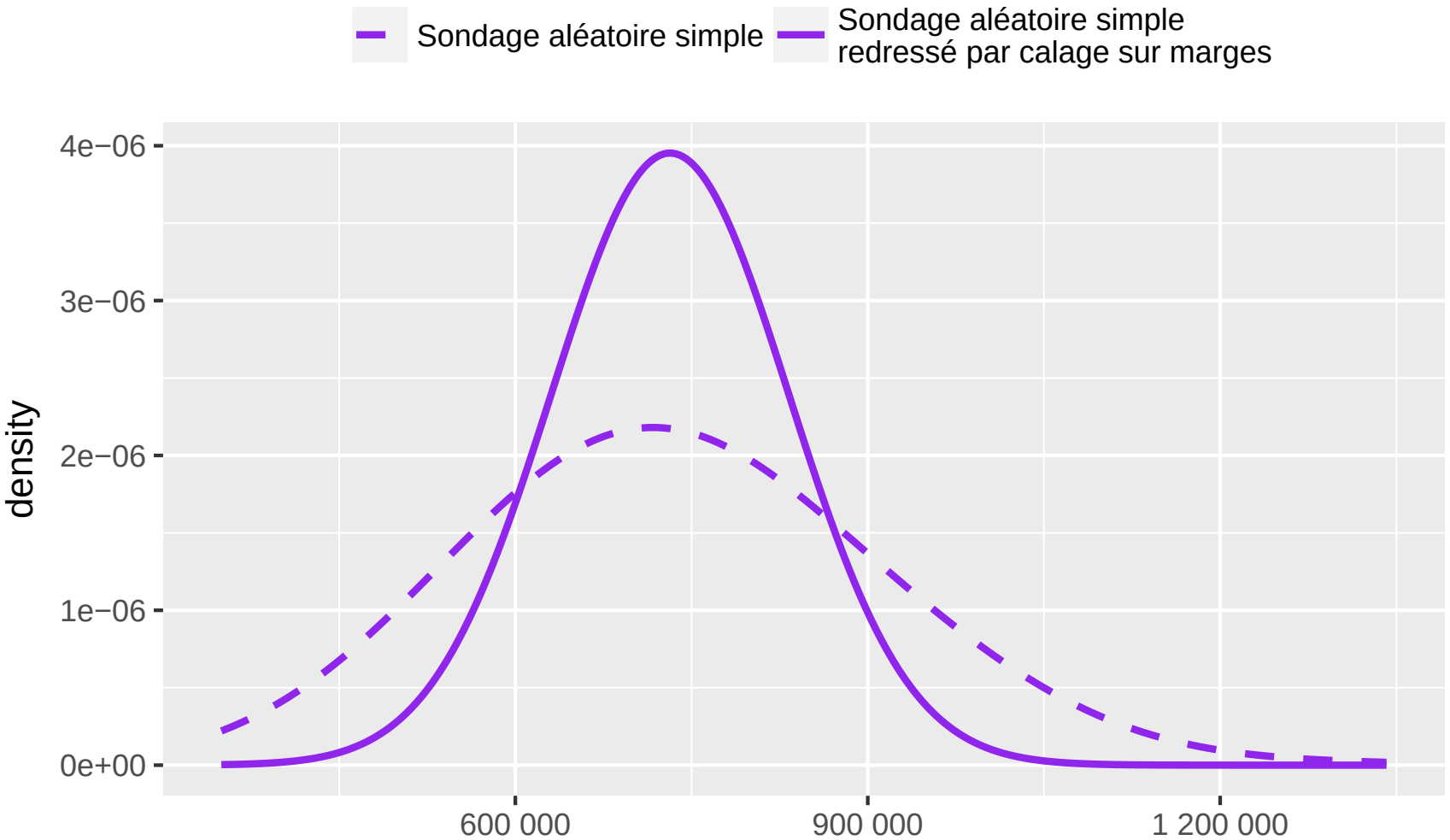
Application : Enquête sur la fréquentation des cinémas

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**.

Application : Enquête sur la fréquentation des cinémas



Application : Enquête sur la fréquentation des cinémas

Biais Moyenne empirique sur 1 000 simulations

- Valeur dans la population : 731 892 entrées
- Estimateur d'Horvitz-Thompson : 733 832
- Estimateur par le ratio : 730 299
- Estimateur par post-stratification : 725 812
- Estimateur par calage sur marges : 731 625

Précision Écart-type empirique sur 1 000 simulations

- Estimateur d'Horvitz-Thompson : 153 932
- Estimateur par le ratio : 16 039
- Estimateur par post-stratification : 145 332
- Estimateur par calage sur marges : 13 607

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**.

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.

Chapitre 3

Annexe

Méthode déductive

- La donnée manquante est remplacée selon une règle déterministe, utilisant les variables disponibles sur cette même unité.
- Cette méthode est souvent utilisée pour corriger des données jugées incohérentes ou invalides.
- Exemple : l'occupation d'un individu de 8 ans = écolier
- Méthode très coûteuse mais dont l'impact sur le biais et la précision est limité.

Cold deck

- On utilise une information extérieure à l'enquête relative à la même unité.
- Exemple : pour un panel, remplacer l'activité manquante par celle issue de l'enquête précédente, ou depuis la base de sondage.
- Avantages : si la source extérieure est de « bonne qualité », imputation par une donnée qui existe et sans doute proche de la vraie valeur.
- Inconvénients : disponibilité de la source extérieure ; qualité de ladite source ? (risque d'introduire des valeurs aberrantes...)

Imputation par la moyenne

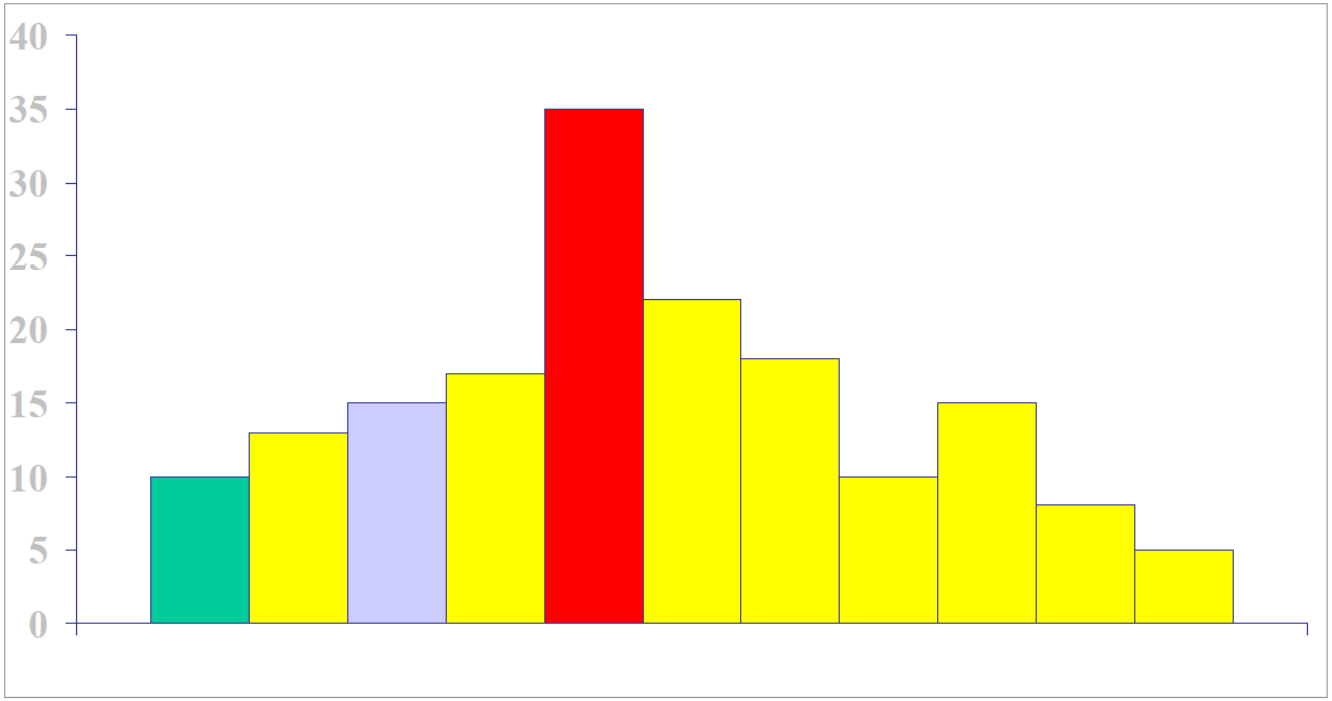
Les données manquantes sont remplacées par la moyenne des répondants.

- Avantages :
  - Simple à calculer ;
  - Ne nécessite aucune information auxiliaire ;
  - Si la population est homogène, imputation de bonne qualité
- Inconvénients :
  - Distord la distribution de la variable d'intérêt
  - Dispersion des données réelles après imputation sous-estimée

La méthode est plus efficace lorsqu'elle est appliquée sous-groupe homogène par sous-groupe de la population.

Imputation par la moyenne

Impact sur la distribution :



Imputation par le ratio

Les données manquantes sont remplacées par la valeur d'une variable auxiliaire ≪ actualisée ≫ par un ratio moyen :

$$\hat{\bar{Y}} = \frac{\sum_S w_k y_k}{\sum_S w_k}$$

$$\hat{\bar{X}} = \frac{\sum_S w_k x_k}{\sum_S w_k}$$

La valeur imputée vaut alors :

$$y_k^{\text{imp}} = \frac{\hat{\bar{Y}}}{\hat{\bar{X}}} x_k$$

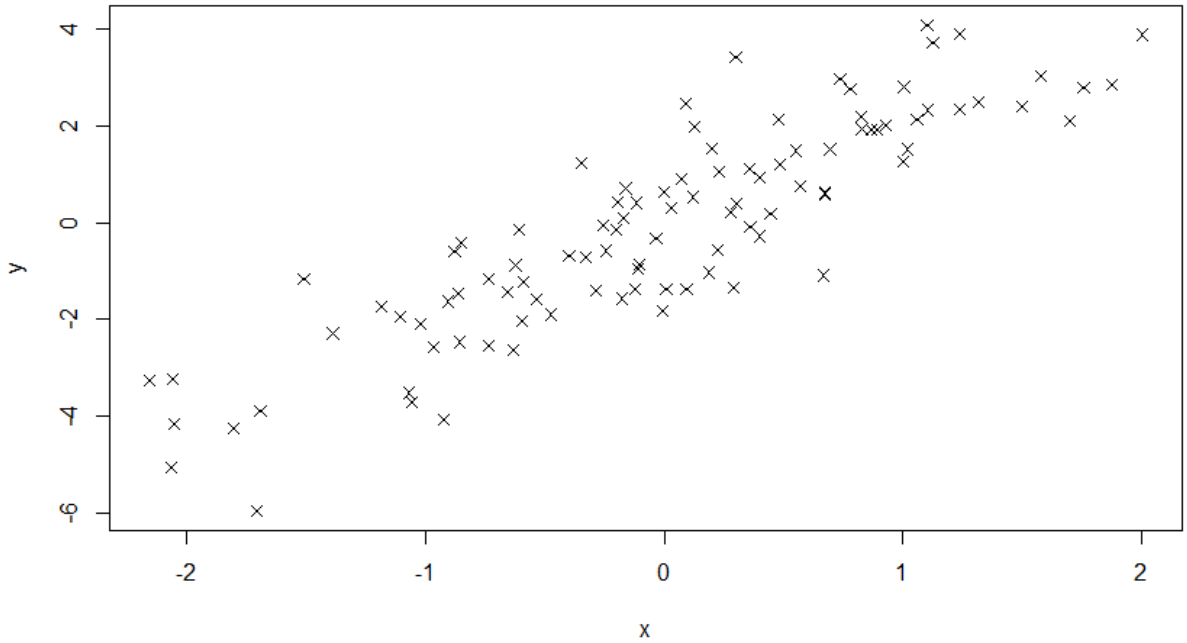
Comme pour l'imputation par la moyenne, cette méthode est généralement utilisée par classes.

Imputation par le ratio

- Avantages :
  - Simple à calculer ;
  - Si X et Y sont liés, meilleure qualité que la moyenne.
- Inconvénients :
  - Distord toujours la distribution de la variable d'intérêt ;
  - Nécessite une variable auxiliaire.
  - Le tableau final peut donner l'impression d'un fort lien entre Y et X.

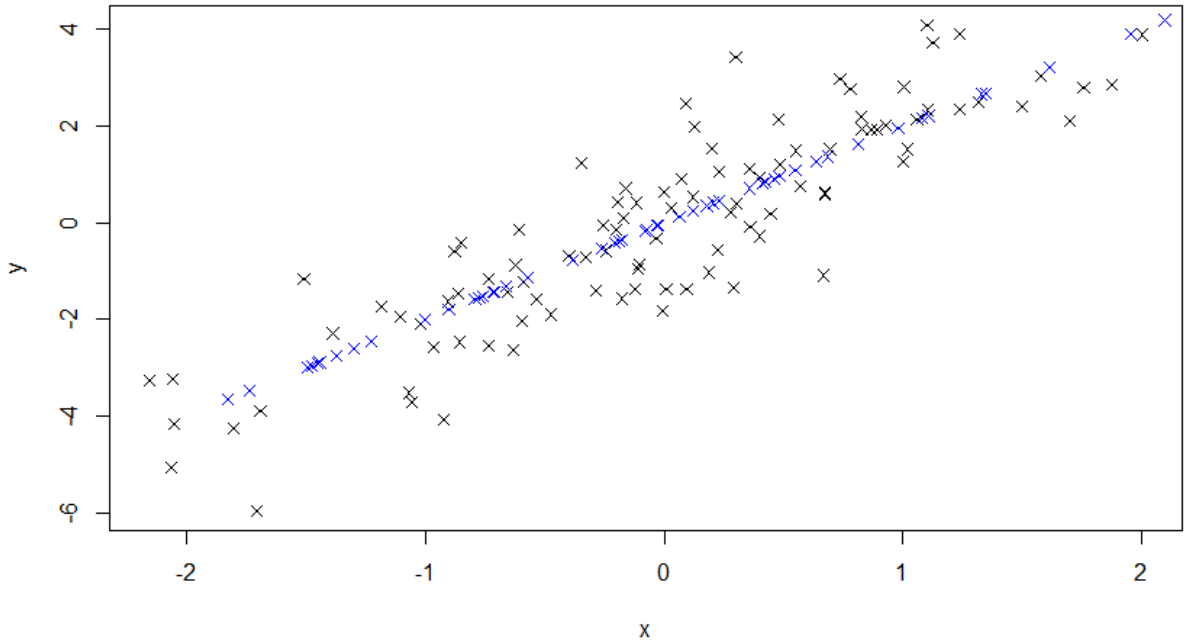
Imputation par le ratio

Avant imputation :



Imputation par le ratio

Après imputation :



Imputation par la tendance unitaire

Dans le cadre de panels ou d'enquêtes répétées : la valeur manquante est remplacée par la valeur déclarée lors d'une occasion précédente et modifiée selon la tendance de l'unité :

$$y_k^{\text{imp}} = \frac{x_{k,t}}{\hat{x}_{k,t-1}} y_{k,t-1}$$

- Avantages :
  - Simple à calculer ;
  - Fait intervenir de l'information à un niveau fin, donc a priori de bonne qualité.
- Inconvénients :
  - Nécessite une variable auxiliaire
  - Nécessite une édition précédente

Imputation par le plus proche voisin

La valeur manquante est remplacée par une valeur du ≪ plus proche ≫ répondant, selon une certaine distance par rapport à une ou plusieurs variables auxiliaires.

- Avantages :
  - Imputation par une valeur préexistante dans l'échantillon
  - Distord moins les distributions
- Inconvénients :
  - Nécessite une ou plusieurs variables auxiliaires
  - Dépend du choix de la distance (euclidienne, etc.)

Imputation par le plus proche voisin

Remarques

- Il est possible de sélectionner les k plus proches voisins et de réaliser la moyenne ou le vote majoritaire sur ces k individus.
- Un individu peut se retrouver donneur pour beaucoup d'autres si la non-réponse est très mal répartie.
- On appelle aussi cette méthode hot deck non aléatoire.

Hot deck

La valeur manquante est remplacée par une valeur choisie au hasard parmi les répondants.

- Avantages :
  - Imputation par une valeur préexistante dans l'échantillon
  - Distord moins les distributions
- Inconvénients :
  - Peut imputer par des valeurs atypiques
  - Augmente la variabilité de l'estimation

Amélioration : contrôler le hasard.

Hot deck

Comment contrôler le hasard ?

- Limiter le nombre de donneurs :
  - En se restreignant à une classe (croisement de variables, GRH, etc.)
  - En se limitant à k voisins
- Pondérer : les individus moins proches ont une probabilité plus faible de donner leur valeur de y pour l'imputation.
- Limiter le nombre de fois où un même individu peut être donneur.

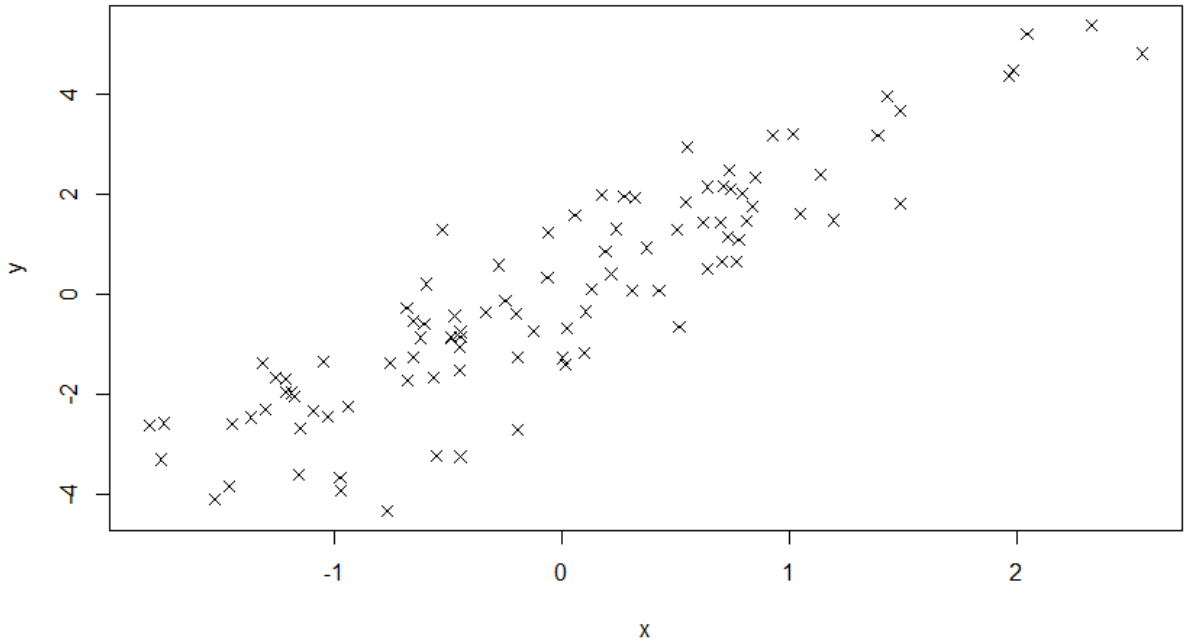
Imputation avec résidus

C'est une imputation déterministe (moyenne, ratio, régression, etc.) à laquelle on ajoute un résidu aléatoire ϵ_k. Il y a deux possibilité sur la génération de ce résidu aléatoire :

1. Générer les ϵ selon une loi normale $N(0, \sigma)$ où la variance est estimée à partir des répondants ;
2. Choisir les ϵ parmi les valeurs observés entre l'imputation déterministe et la vraie valeur pour les répondants de l'échantillon.

Imputation par le ratio avec résidus

Avant imputation :



Imputation par le ratio avec résidus

Après imputation :