Données Manquantes dans les Enquêtes

Guillaume Chauvet

École Nationale de la Statistique et de l'Analyse de l'Information

22/01/2025

Objectifs du cours

- Donner un aperçu plus réaliste du déroulement d'une enquête.
- Expliquer le phénomène de non-réponse, et ses conséquences sur l'estimation.
- Décrire les méthodes de correction de la non-réponse totale dans les enquêtes.
- Décrire les méthodes de correction de la non-réponse partielle dans les enquêtes.

1. Compléments sur les enquêtes
   - Cadre du cours de théorie des sondages
   - Sources d'erreur dans l'estimation
   - Les types de non-réponse

2. Echantillonnage en population finie
   - Rappels
   - Echantillonnage en deux phases

3. Traitement de la non-réponse totale
   - Identification des non-répondants
   - Modélisation du mécanisme de non-réponse
   - Estimation des probabilités de réponse
   - Estimation ponctuelle et estimation de précision
   - Cas des groupes homogènes de réponse
   - Application

4. Traitement de la non-réponse partielle
   - Contexte
   - Choix d'un modèle d'imputation
   - Choix d'un mécanisme d'imputation
   - Estimation de paramètres après imputation

Compléments sur les enquêtes

Cadre du cours de théorie des sondages

Cadre du cours de Théorie des Sondages

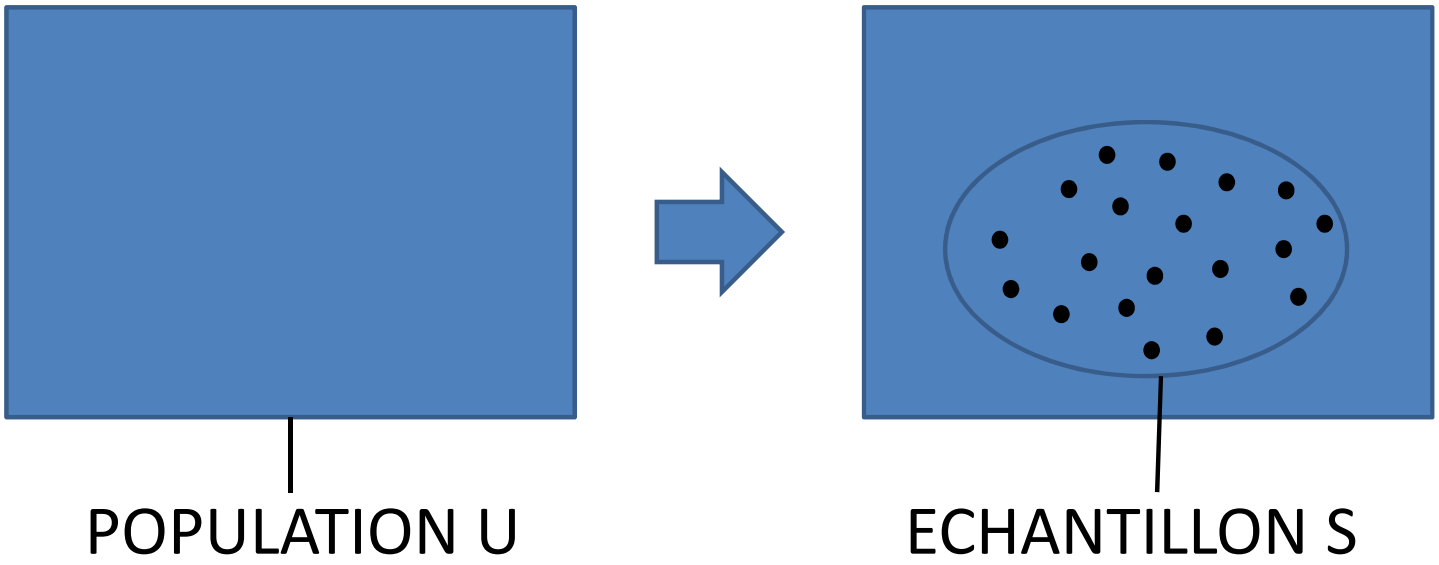
Nous nous intéressons à une **population-cible** notée $U = \{1, \ldots, N\}$. Nous supposons disposer d'une **base de sondage** listant de façon exacte les unités de la population. Nous nous intéressons à l'estimation du total $t_y = \sum_{k \in U} y_k$ d'une variable d'intérêt y_k.

Nous sélectionnons dans U un échantillon S au moyen d'un plan de sondage $p(\cdot)$. Les probabilités d'inclusion π_k sont supposées **connues et strictement positives**. L'estimateur de Horvitz-Thompson

$$\hat{t}_{y\pi} = \sum_{k \in S} \frac{y_k}{\pi_k} = \sum_{k \in U} \frac{y_k}{\pi_k} I_k \tag{1}$$

est **sans biais sous le plan de sondage** pour le total t_y.

Cadre du cours de Théorie des Sondages



Quelques définitions utiles

Base de sondage : liste des unités d'échantillonnage.

Population-cible : ensemble des unités statistiques pour lesquelles nous souhaitons estimer des paramètres. Si nécessaire, il est important de bien spécifier les sous-groupes qui en sont exclus (restriction du champ de l'enquête).

Champ de l'enquête : définition des caractéristiques des unités statistiques qui appartiennent à la population-cible.

Unité statistique : unité d'observation pour laquelle des données sont recueillies ou calculées.

Questions préliminaires avant estimation

1. Quelle est la population d'intérêt, et en particulier quel est le champ de l'enquête ?
2. Quelle est l'unité d'observation, et quelle est l'unité d'échantillonnage ?
3. Quel est le paramètre d'intérêt, et quelle est la variable d'intérêt ?
4. Quel est le plan de sondage ?

Les étapes d'une enquête (Haziza, 2011)

En pratique, l'échantillonnage n'est qu'une des nombreuses étapes d'une enquête. Produire des estimations précises dépend de la bonne réalisation de chacune de ces étapes :

1. Planification : objectifs, concepts, champ de l'enquête
2. Constitution de la base de sondage
3. Conception du questionnaire
4. Conception du plan de sondage et tirage de l'échantillon
5. Collecte des données
6. Traitement des données
7. Estimation ponctuelle et estimation de variance

Les étapes d'une enquête (Haziza, 2011)

En pratique, l'échantillonnage n'est qu'une des nombreuses étapes d'une enquête. Produire des estimations précises dépend de la bonne réalisation de chacune de ces étapes :

1. Planification : objectifs, concepts, champ de l'enquête
2. Constitution de la base de sondage
3. Conception du questionnaire
4. **Conception du plan de sondage et tirage de l'échantillon**
5. Collecte des données
6. **Traitement des données**
7. Estimation ponctuelle et estimation de variance

Les enquêtes Insee

Enquêtes auprès des entreprises

Base de sondage d'entreprises régulièrement mise à jour (SIRUS : Système d'Immatriculation au Répertoire des Unités Statistiques).

Echantillonnage direct par sondage aléatoire simple stratifié selon :

- un critère d'activité utilisant un niveau plus ou moins fin de la Nomenclature d'Activités Française (NAF)
- un critère de taille (tranches d'effectifs salariés et/ou tranches de CA).

Difficultés liées :

- à la coordination des échantillons : négative pour favoriser la sélection d'entreprises non encore enquêtées, et réduire la charge de réponse ; positive pour panéliser une partie de l'échantillon pour produire des mesures d'évolution,
- à la gestion des unités influentes.

Le Gleut, R. (2017), *Stratification et calcul d'allocations dans les enquêtes auprès des entreprises*

Gros, E., et Merly-Alpa, T., *La coordination d'échantillons des enquêtes auprès des entreprises*.

Lien vers les fiches méthodologiques Insee

Les enquêtes Insee

Enquêtes auprès des ménages

Base de sondage constituée à partir de sources fiscales. Le répertoire statistique des individus et des logements (RESIL) devrait remplacer fin 2025 l'ancienne base FIDELI.

Dans les enquêtes auprès des ménages/individus :

- Tirage **à plusieurs degrés** : tirage de zones (obtenues par agrégation ou découpage de communes), de logements dans ces zones. L'enquête est réalisée auprès de tous les individus du ménage ou d'un représentant tiré aléatoirement (**individu Kish**).
- Difficultés : différents niveaux d'unité statistique (individu, ménage, logement) à prendre en compte dans la pondération, le traitement de la non-réponse, le calcul de variance.
- L'Insee réfléchit à un tirage direct des individus.

Faivre, S. (2017). *L'échantillonnage des enquêtes auprès des ménages dans la source fiscale*. Lien vers les fiches méthodologiques Insee.

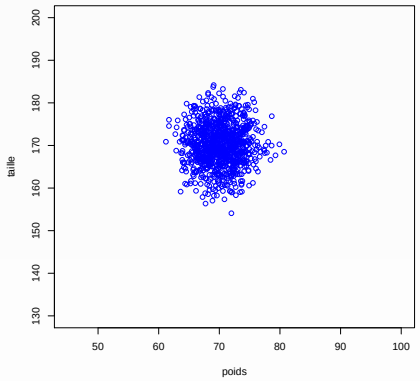
Chaput, H., Gros, E., Merly-Alpa, T. (2025). *Comment choisir entre échantillonner des individus ou des logements dans un contexte multimode*. Journées de méthodologie statistique, Paris, 25-27 novembre 2025.

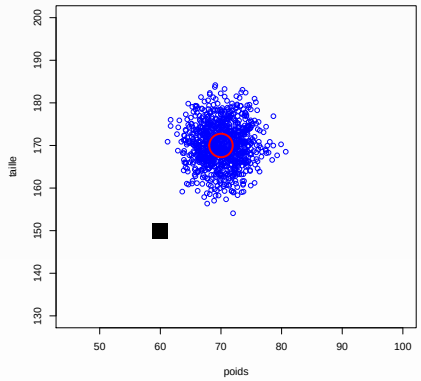
Sources d'erreur dans l'estimation

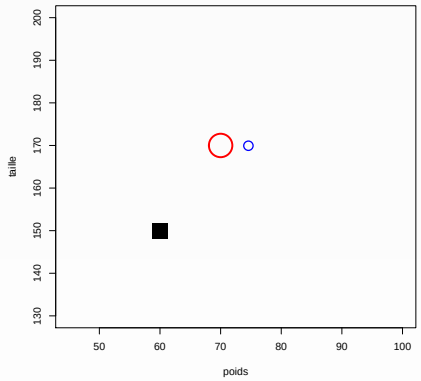
Erreur associée à l'estimateur

Soit $\hat{\theta}$ l'estimateur d'un paramètre θ. La précision de cet estimateur peut être mesurée par :

- son biais : $B(\hat{\theta}) = E(\hat{\theta} - \theta)$,
- sa variance : $V(\hat{\theta}) = E(\hat{\theta} - E\ \hat{\theta})^2$
- son EQM : $EQM(\hat{\theta}) = B(\hat{\theta})^2 + V(\hat{\theta})$.







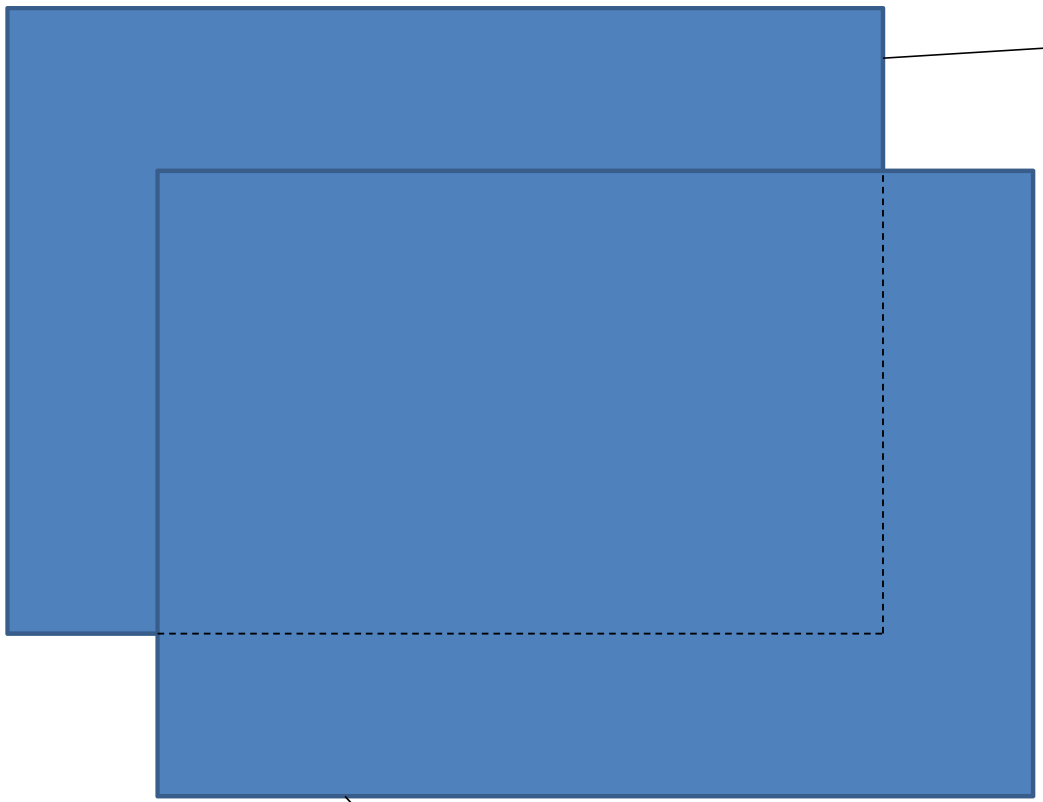
Sources d'erreur

L'erreur totale de l'estimateur $\hat{\theta} - \theta$ dépend des erreurs réalisées à toutes les étapes de l'enquêtes.

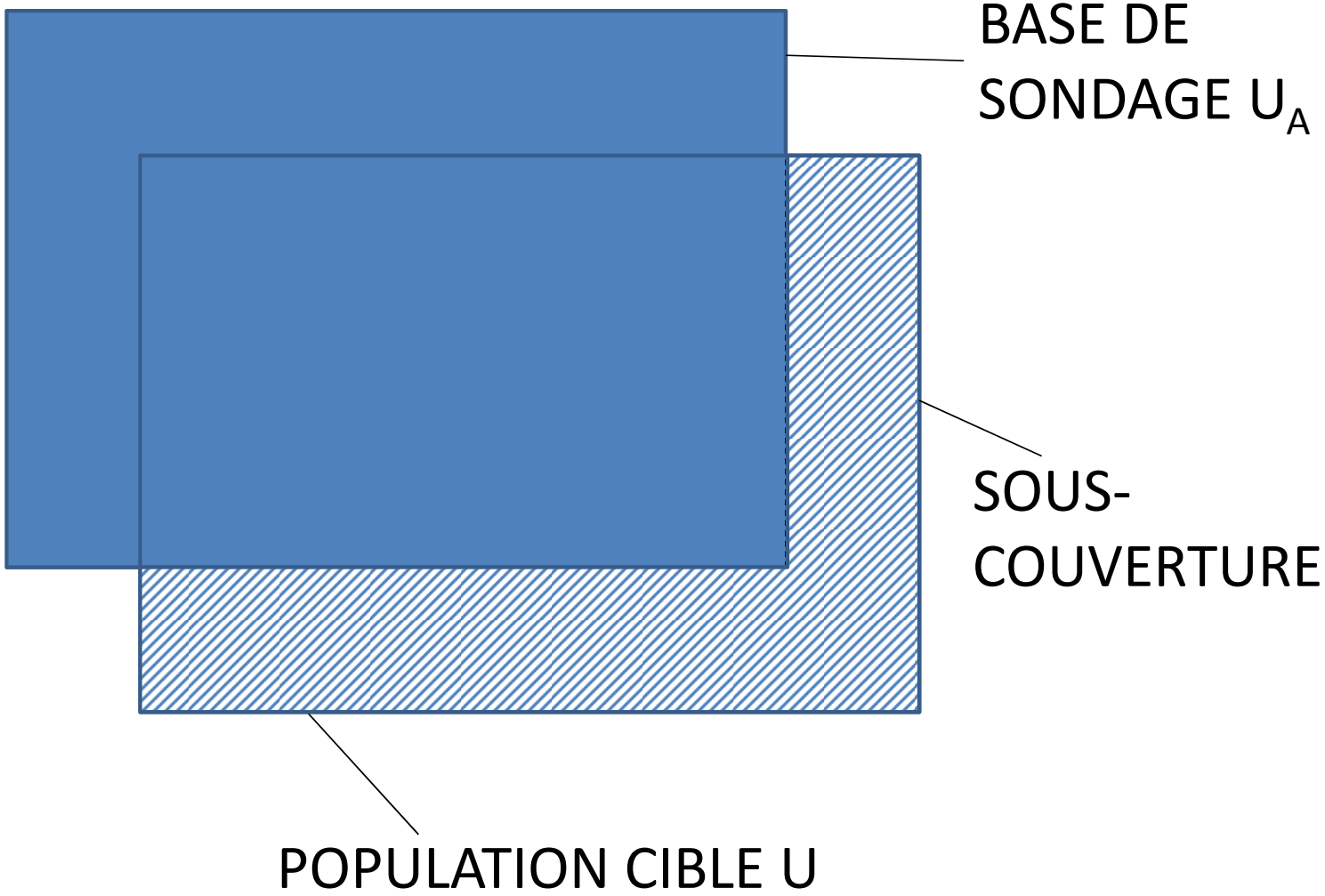
Ceci inclut :

- les erreurs de couverture,
- l'erreur d'échantillonnage,
- les erreurs dues à la non-réponse,
- les erreurs de mesure.

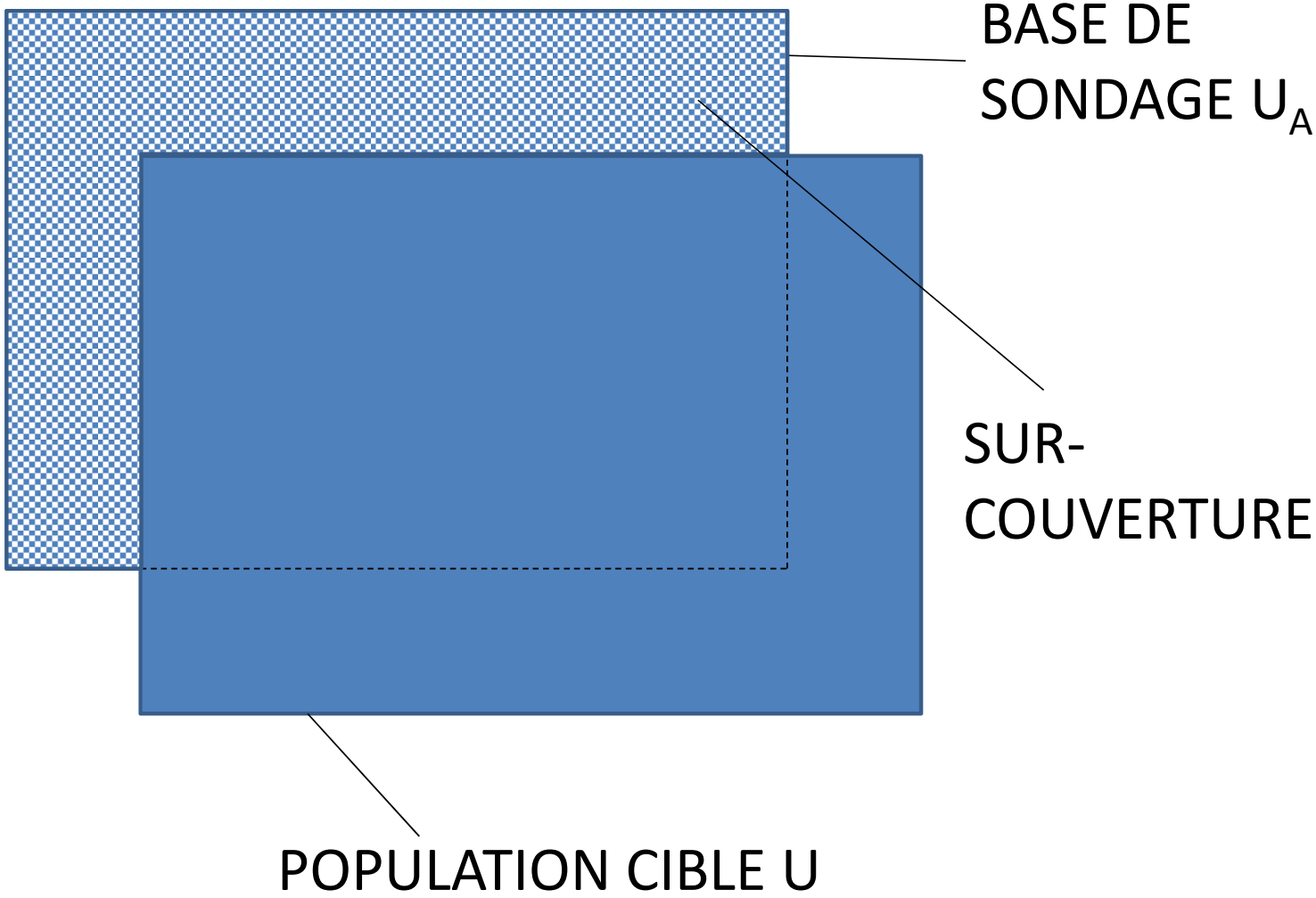
Erreurs de couverture



Sous-couverture



Sur-couverture



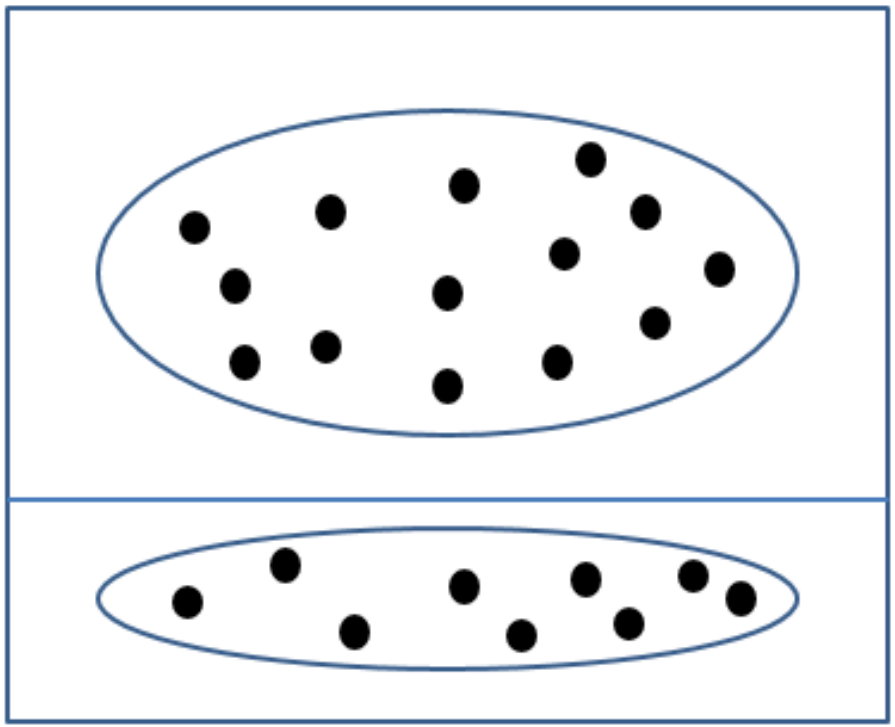
Erreurs de couverture

Les erreurs de couverture proviennent du fait que la base de sondage et la population-cible ne coincident pas. Nous distinguons :

- la sous-couverture (des individus de la population-cible sont absents de la base de sondage) :
  - nouvelles entreprises pas encore inscrites dans le répertoire SIRUS,
  - enquête auprès des ménages avec réponse par internet (15 % d'illectronistes en 2021, source : TIC ménages)
- la sur-couverture (la base de sondage contient des individus qui ne sont pas dans la population-cible) :
  - échantillonnage de logements, dont le statut (résidence principale/secondaire, logement occasionnel/vacant) n'est pas connu au moment du tirage, en vue d'une enquête en résidence principale.

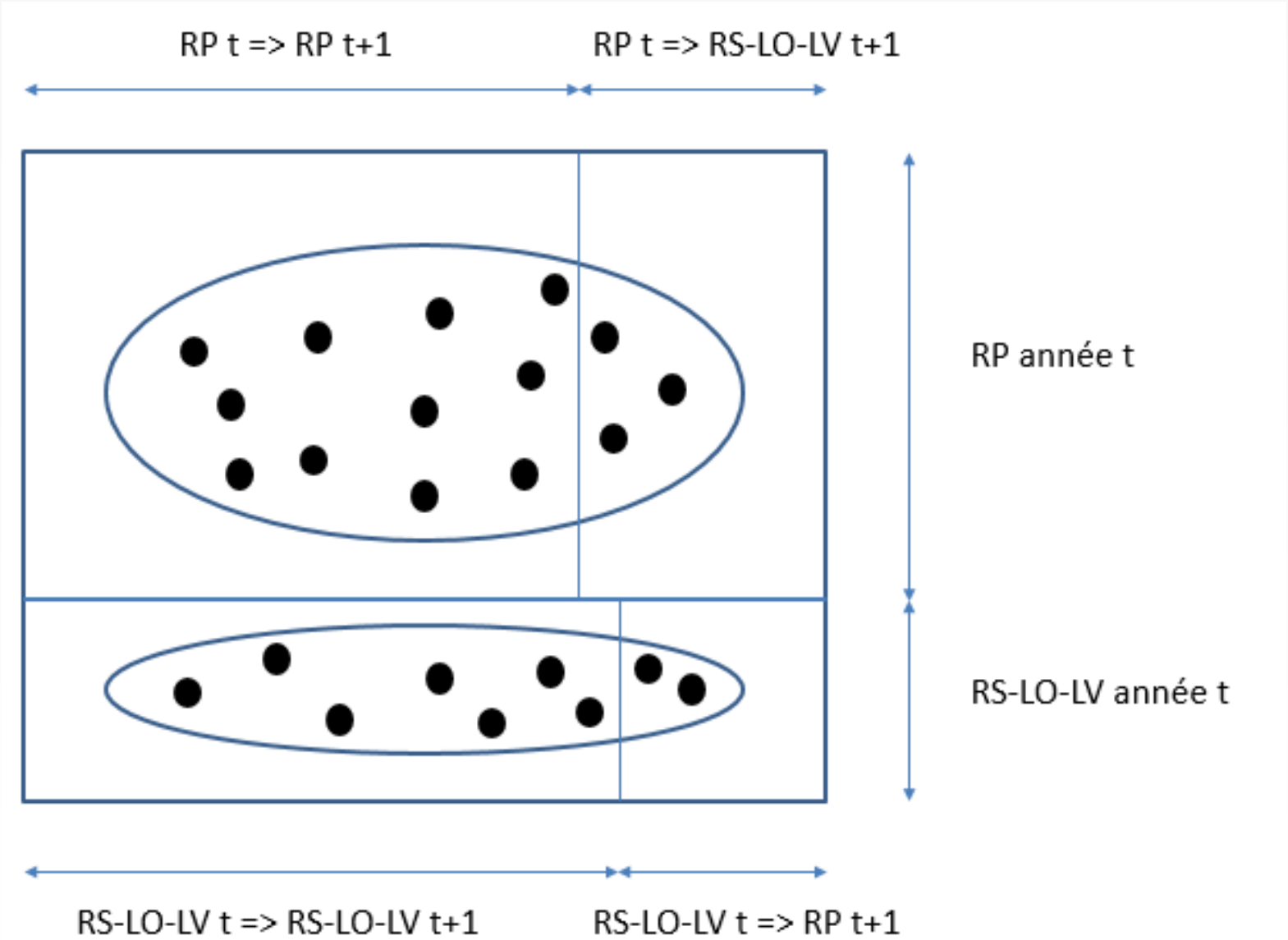
Sélection d'un échantillon de logements

Tirage année t



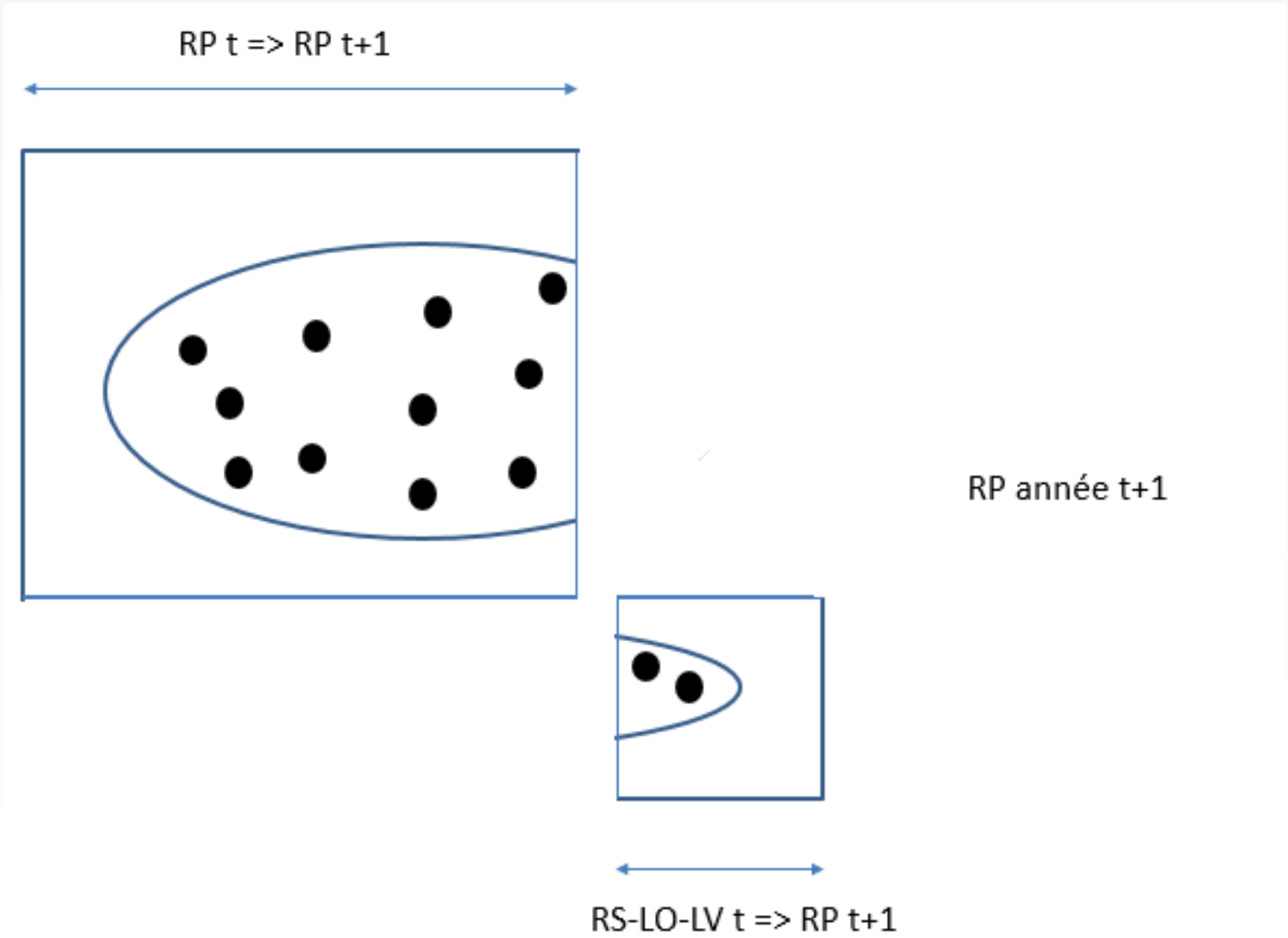
Sélection d'un échantillon de logements

Enquête année t+1



Sélection d'un échantillon de logements

Estimation année t+1 avec les logements dans le champ



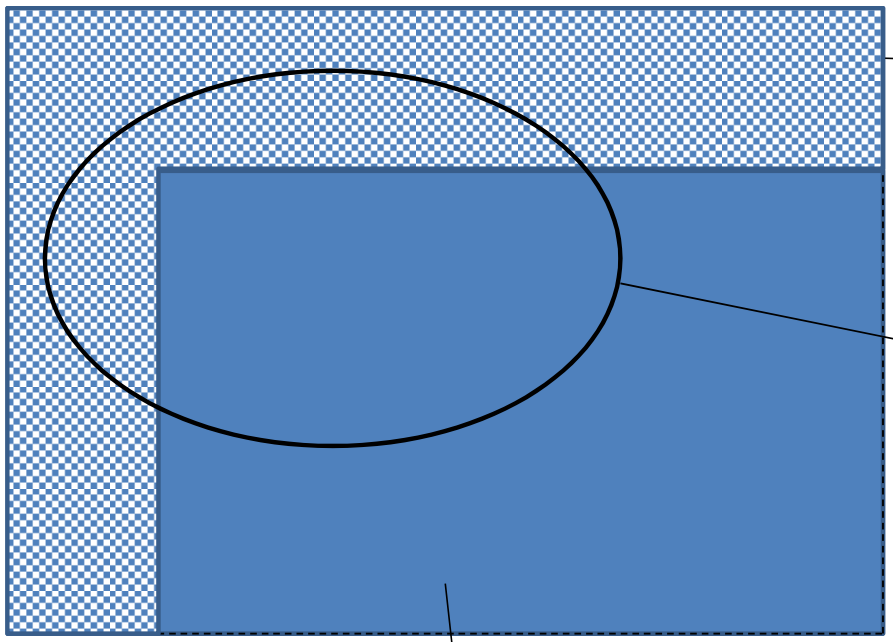
Situation de sous-couverture

C'est la situation la moins favorable des deux. Nous avons $\pi_k = 0$ pour toutes les unités de la population-cible qui ne sont pas dans la base de sondage, ce qui occasionne un **biais de couverture**.

Le biais peut être important si la sous-couverture est forte et/ou si les unités non couvertes contribuent de façon importante au total t_y. Il peut alors être nécessaire de redéfinir le champ de l'enquête, ou d'inclure une correction de la pondération pour tenir compte de la sous-couverture.

Dans la suite, nous supposerons simplement que la base de sondage recouvre la population-cible.

Situation de sur-couverture



Situation de sur-couverture

C'est la situation la plus simple :

- la population-cible est un **domaine** dans la base de sondage,
- l'appartenance au domaine peut être identifiée pendant l'enquête pour les unités de S.

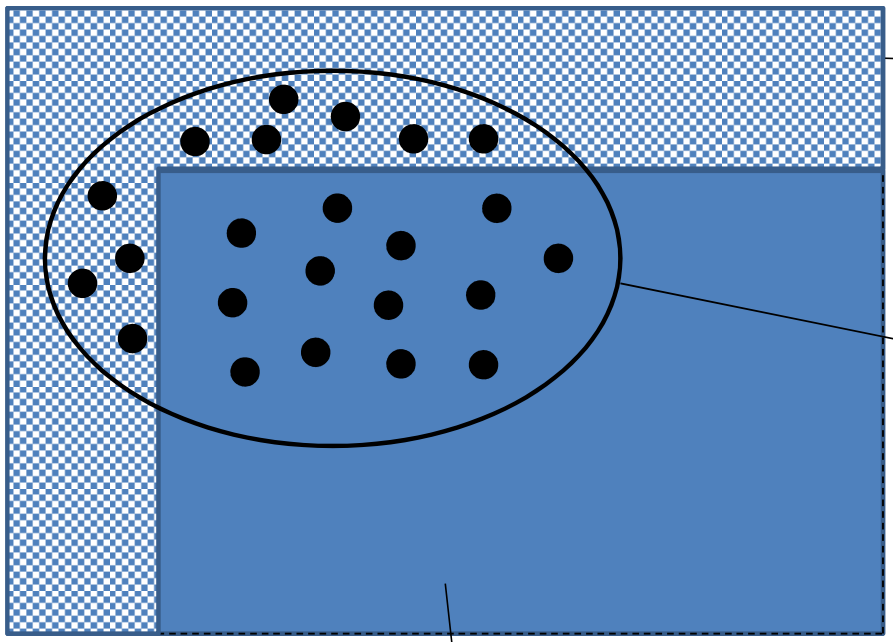
L'estimateur sur domaine du total $t_y = \sum_{k \in U} y_k$ est simplement

$$\hat{t}_y = \sum_{k \in S \cap U} d_k y_k \quad \text{et} \quad E_p(\hat{t}_y) = t_y.$$

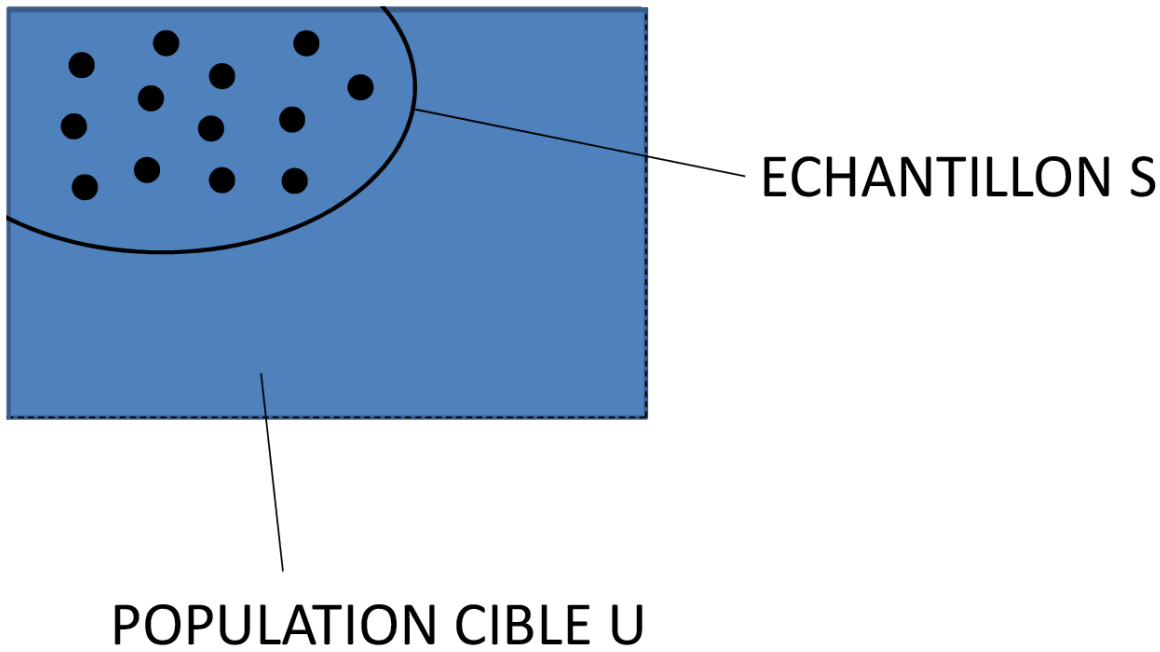
Cet estimateur est sans biais pour le total, mais il souffre d'une perte d'efficacité due à l'utilisation d'une partie de l'échantillon seulement.

Connaître l'appartenance au domaine peut poser problème pour les individus non-répondants.

Situation de sur-couverture



Situation de sur-couverture



Erreurs d'échantillonnage et de non-réponse

L'erreur d'échantillonnage provient du fait que l'information n'est collectée que sur une partie de la population : cette erreur est **volontaire et planifiée**.

L'erreur de non-réponse provient du fait que l'information n'est observée que sur une partie de l'échantillon : cette erreur est **subie et non maîtrisée**.

La non-réponse a des conséquences

- sur le biais des estimateurs : les individus répondant peuvent présenter un profil particulier par rapport aux variables de l'enquête (**biais de NR**),
- sur la variance des estimateurs : la taille effective de l'échantillon diminue (**variance de NR**). De plus, une imputation aléatoire peut introduire une variabilité additionnelle (**variance d'imputation**).

Erreurs de mesure

Les erreurs de mesure proviennent du fait que les valeurs obtenues sont différentes des vraies valeurs de la variable d'intérêt.

Parmi les causes des erreurs de mesure :

- questionnaire mal conçu,
- problème d'enquêteur,
- appel à la mémoire des enquêtés,
- erreur de codage.

Un effet de mesure peut également provenir d'un dispositif d'enquête multi-mode, e.g. bimode séquentiel internet-téléphone.

Dans ce qui suit, nous supposerons que les erreurs de mesure peuvent être négligées. Nous nous focalisons sur l'erreur due à l'échantillonnage et sur l'erreur due à la non-réponse.

En résumé

Nous négligeons dans ce cours l'erreur de sous-couverture et les erreurs de mesure.

Nous supposons disposer d'une base de sondage U_A qui recouvre la population-cible U. Nous sélectionnons un échantillon S dans U_A selon le plan de sondage choisi.

L'échantillon exploitable est constitué des individus de l'échantillon qui sont situés dans la population-cible. On dit encore que ces individus appartiennent au champ de l'enquête.

Les estimateurs produits doivent prendre en compte l'échantillonnage et la non-réponse.

Les types de non-réponse

Type de non-réponse

Dans le contexte des enquêtes, nous distinguerons schématiquement :

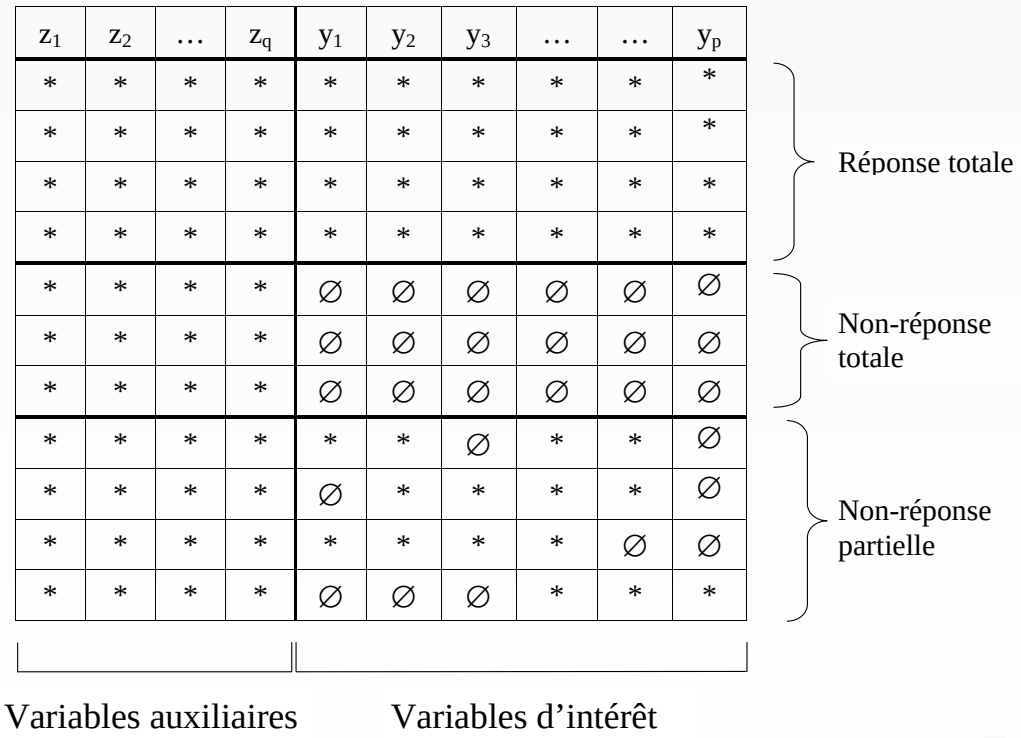
- la **non-réponse totale** ("unit non-response") : aucune information n'est relevée pour une unité,
- la **non-réponse partielle** ("item non-response") : une partie seulement de l'information est relevée pour une unité.

y_1	y_2	y_3	y_4	…	…	…	…	…	y_p	
*	*	*	*	*	*	*	*	*	*	Réponse totale
*	*	*	*	*	*	*	*	*	*	
*	*	*	*	*	*	*	*	*	*	
*	*	*	*	*	*	*	*	*	*	
ϕ	ϕ	ϕ	ϕ	ϕ	ϕ	ϕ	ϕ	ϕ	ϕ	Non-réponse totale
ϕ	ϕ	ϕ	ϕ	ϕ	ϕ	ϕ	ϕ	ϕ	ϕ	
ϕ	ϕ	ϕ	ϕ	ϕ	ϕ	ϕ	ϕ	ϕ	ϕ	
*	*	ϕ	*	ϕ	*	ϕ	*	*	ϕ	Non-réponse partielle
ϕ	*	*	*	ϕ	*	ϕ	*	*	ϕ	
*	*	*	*	*	*	*	*	ϕ	ϕ	
ϕ	ϕ	ϕ	*	*	ϕ	*	*	*	*	

Type de non-réponse

La correction de la non-réponse passe par la connaissance d'**information auxiliaire** connue sur l'ensemble de l'échantillon S, et qui soit

- explicative de la probabilité de répondre,
- et/ou explicative de la variable d'intérêt.



Traitement de la non-réponse dans les enquêtes

La non-réponse totale est habituellement traitée par une **méthode de repondération** :

- les non-répondants totaux sont supprimés du fichier,
- les poids des répondants sont augmentés pour compenser de la non-réponse totale.

La non-réponse partielle est habituellement traitée par **imputation** : une valeur manquante est remplacée par une valeur plausible.

L'objectif prioritaire est de **réduire autant que possible le biais de non-réponse** : cela passe par une recherche des facteurs explicatifs de la non-réponse.

Quelques facteurs de non-réponse totale

- Impossibilité de joindre l'individu (personnes activés, horaires décalés, résidences multiples).
- Type d'enquête (obligatoire ou volontaire). En France, la labellisation peut être attribuée par le Comité du Label de la Statistique Publique.
- Fardeau de réponse : le temps, la complexité du questionnaire et la fréquence des sollicitations augmentent la non-réponse et l'abandon.
- Méthode de collecte : face à face $>$ téléphone $>$ web en termes de taux de réponse, toutes choses égales par ailleurs.
- Durée de collecte : l'allongement de la période de collecte permet de réduire la non-réponse par non-contact, par suivi (et relance) des non-répondants.
- Formation des enquêteurs.

Echantillonnage en population finie

Rappels

Plan de sondage

Nous nous plaçons dans le cadre d'une population finie d'individus, notée U. Nous nous intéressons à une **variable d'intérêt** y qui prend la valeur y_k sur l'individu k de U.

Les valeurs prises par la variable y sont collectées sur un échantillon S. L'objet de la théorie des sondages est d'utiliser cette information afin d'estimer des paramètres définis sur la population entière.

L'échantillon S est sélectionné dans U au moyen d'un **plan de sondage** $p(\cdot)$, i.e. d'une loi de probabilité sur l'ensemble des parties de U.

Plan de sondage

Nous supposons connues les probabilités d'appartenance à l'échantillon de chaque unité k :

$$\pi_k = \Pr(k \in S).$$

Si toutes les π_k sont > 0, le total $t_y = \sum_{k \in U} y_k$ est estimé sans biais par l'**estimateur de Horvitz-Thompson**

$$\hat{t}_{y\pi} = \sum_{k \in S} d_k y_k = \sum_{k \in U} \frac{I_k}{\pi_k} y_k \tag{2}$$

avec $d_k = 1/\pi_k$ le **poids de sondage** de l'unité k, et $I_k = 1(k \in S)$ l'indicatrice de sélection.

Les mêmes poids peuvent être utilisés pour toutes les variables d'intérêt.

Plan de sondage

La forme générale de variance est donnée par la formule de Horvitz-Thompson :

$$V_p\left(\hat{t}_{y\pi}\right) = \sum_{k,l\in U} \frac{y_k}{\pi_k}\frac{y_l}{\pi_l}\Delta_{kl} \tag{3}$$

avec $\Delta_{kl} = \pi_{kl} - \pi_k\pi_l$ et $\pi_{kl} = Pr(k, l \in S)$ la probabilité d'inclusion jointe des unités k et l dans l'échantillon.

Nous pouvons estimer sans biais cette variance par

$$\widehat{V}_{HT}\left(\hat{t}_{y\pi}\right) = \sum_{k,l\in S} \frac{y_k}{\pi_k}\frac{y_l}{\pi_l}\frac{\Delta_{kl}}{\pi_{kl}} \tag{4}$$

si tous les π_{kl} sont > 0.

Le tirage de Poisson

Chaque individu k est tiré dans l'échantillon avec une probabilité π_k, indépendamment des autres individus. Les indicatrices de sélection $I_k,\ k \in U$ sont donc indépendantes. Pour $k \neq l \in U$, nous avons :

$$\pi_{kl} = \pi_k \pi_l \quad \text{et} \quad \Delta_{kl} = 0.$$

L'estimateur de Horvitz-Thompson $\hat{t}_{y\pi}$ a pour variance

$$V_p\left(\hat{t}_{y\pi}\right) = \sum_{k \in U} \left(\frac{y_k}{\pi_k}\right)^2 \pi_k (1 - \pi_k). \tag{5}$$

Elle est estimée sans biais par

$$\hat{V}_{HT}\left(\hat{t}_{y\pi}\right) = \sum_{k \in S} \left(\frac{y_k}{\pi_k}\right)^2 (1 - \pi_k). \tag{6}$$

Le tirage de Poisson nous sera utile en particulier pour modéliser la non-réponse totale.

Sondage aléatoire simple

Sondage aléatoire simple (SRS) de taille n : plan de taille fixe, où tous les échantillons de taille n ont la même probabilité d'être sélectionnés. L'estimateur de Horvitz-Thompson se réécrit

$$\hat{t}_{y\pi} = \frac{N}{n}\sum_{k\in S} y_k = N\bar{y}.$$

Sa variance est donnée par

$$V_p\left(\hat{t}_{y\pi}\right) = N^2(1-f)\frac{S_y^2}{n}.$$

Nous pouvons l'estimer sans biais par

$$\hat{V}_{HT}\left(\hat{t}_{y\pi}\right) = N^2(1-f)\frac{s_y^2}{n}. \quad (7)$$

Sondage aléatoire simple stratifié (STSRS)

La population est partitionnée en H strates $U_1, \ldots, U_H$. Nous effectuons un SRS(n_h) indépendamment dans chaque strate. L'estimateur de Horvitz-Thompson se réécrit

$$\hat{t}_{y\pi} = \sum_{h=1}^{H} N_h \bar{y}_h.$$

Sa variance est donnée par

$$V_p\left(\hat{t}_{y\pi}\right) = \sum_{h=1}^{H} N_h^2 (1 - f_h) \frac{S_{yh}^2}{n_h},$$

et nous l'estimons sans biais par

$$\hat{V}_{HT}\left(\hat{t}_{y\pi}\right) = \sum_{h=1}^{H} N_h^2 (1 - f_h) \frac{s_{yh}^2}{n_h}.$$

Echantillonnage en deux phases

Principe

Dans le cadre d'une enquête, nous pouvons être amenés à sélectionner l'échantillon en deux temps :

- Nous sélectionnons tout d'abord un gros sur-échantillon S selon un plan de sondage $p(\cdot)$.
- Nous tirons ensuite dans S un sous-échantillon S_2 selon un plan de sondage $q(\cdot|S)$.

C'est une méthode d'**échantillonnage en deux phases**. Elle est par exemple utilisée pour cibler une population spécifique.

Exemple : Enquête Capacités, Aides et REssources des seniors (CARE-Ménages), réalisée en 2016. Echantillon de 15000 personnes nées avant le 02/05/1955, résidant en logements "ordinaires" en France métropolitaine et ayant répondu à l'enquête Vie Quotidienne et Santé 2014.

Estimation

Nous notons :

- I_k l'indicatrice d'appartenance à l'échant. de 1ère phase S, et π_k (>0) la probabilité de sélection associée,
- r_k l'indicatrice d'appartenance à l'échant. de 2nde phase S_2, et $p_{k|S} \equiv p_k$ (>0) la probabilité de sélection (cond. à S) associée.

L'**estimateur par expansion** est défini par

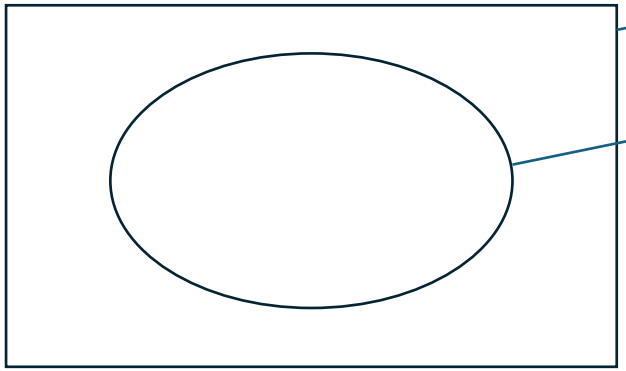
$$\hat{t}_{ye} = \sum_{k \in S_2} \frac{y_k}{\pi_k \, p_k} = \sum_{k \in U} \frac{I_k}{\pi_k} \frac{r_k}{p_k} y_k. \tag{8}$$

Cet estimateur contient une double pondération :

- le facteur $1/\pi_k$ compense pour la sélection de l'échantillon S,
- le facteur $1/p_k$ compense pour la sélection du sous-échantillon S_2,

qui conduit aux poids $d_{ek} = \dfrac{1}{\pi_k p_k}$.

Exemple d'un tirage en deux phases



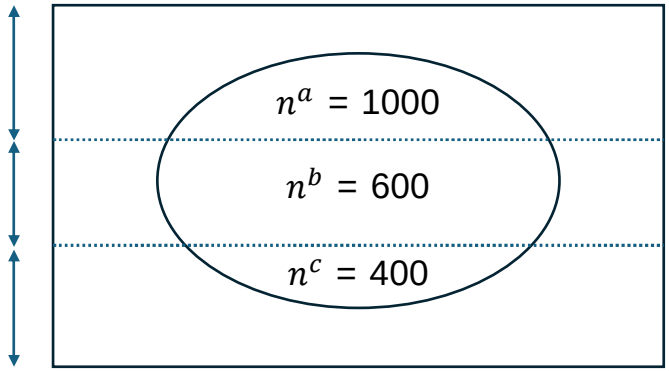
U $(N = 10\,000)$
$S \sim Poisson(\pi_k = 0.2)$
$d_k = 5$

$S_2|S \sim STSRS\ (n_2^a = 200, n_2^b = 300, n_2^c = 320,)$

U^a : sans handicap

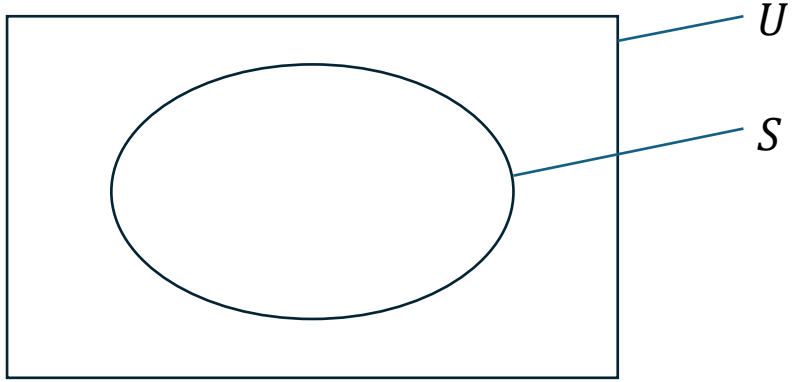
U^b : Handicap léger

U^c: Handicap lourd



$$p_k = \begin{cases} \frac{200}{1000} \text{ pour } k \in S^a, \\ \frac{300}{600} \text{ pour } k \in S^b, \\ \frac{320}{400} \text{ pour } k \in S^c. \end{cases}$$

Exemple d'un tirage en deux phases

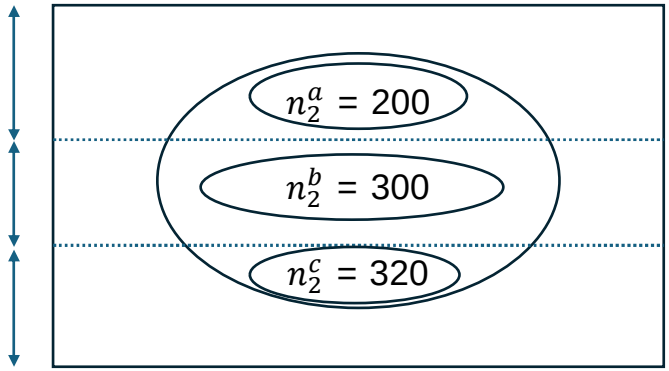


U $(N = 10\,000)$
$S \sim Poisson(\pi_k = 0.2)$
$d_k = 5$

U^a : sans handicap

U^b : Handicap léger

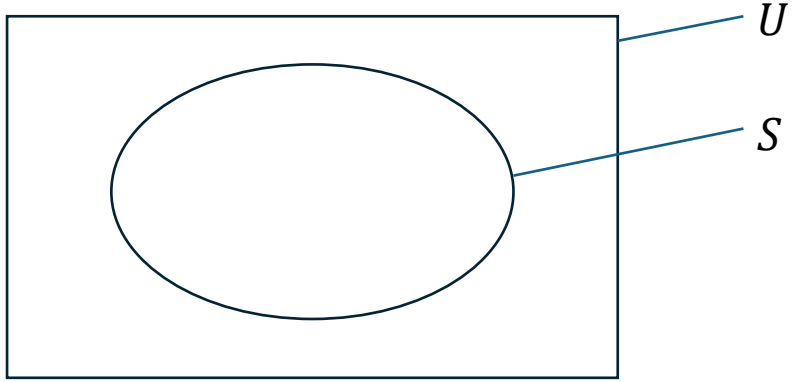
U^c: Handicap lourd



$$d_{ek} = \begin{cases} 25 \text{ pour } k \in S_2^a, \\ 10 \text{ pour } k \in S_2^b, \\ 6.25 \text{ pour } k \in S_2^c. \end{cases}$$

Champ : population U entière

Exemple d'un tirage en deux phases

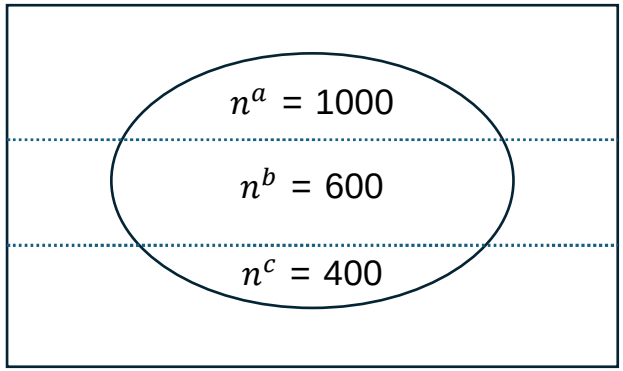


U $(N = 10\,000)$
$S \sim Poisson(\pi_k = 0.2)$
$d_k = 5$

U^a : sans handicap

U^b : Handicap léger

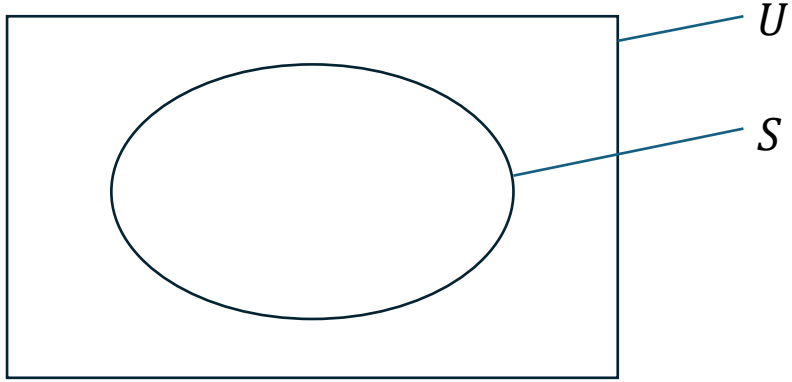
U^c: Handicap lourd



$$S_2|S \sim STSRS(n_2^b = 300, n_2^C = 320)$$

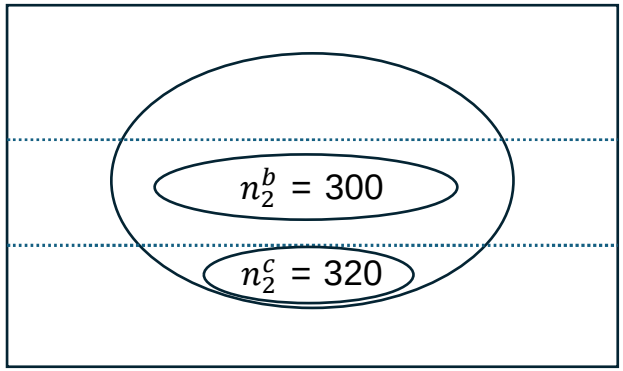
$$p_k = \begin{cases} \frac{300}{600} \text{ pour } k \in S^b, \\ \frac{320}{400} \text{ pour } k \in S^c. \end{cases}$$

Exemple d'un tirage en deux phases



U $(N = 10\,000)$
$S \sim Poisson(\pi_k = 0.2)$
$d_k = 5$

U^b : Handicap léger
U^c: Handicap lourd



$$d_{ek} = \begin{cases} 10 \text{ pour } k \in S_2^b, \\ 6.25 \text{ pour } k \in S_2^c. \end{cases}$$

Champ : population en situation de handicap

Estimation (2)

Nous avons

$$E\left(\hat{t}_{ye}\right) = EE\left(\hat{t}_{ye}|S\right)$$
$$= EE\left(\sum_{k\in S}\frac{y_k}{\pi_k}\frac{r_k}{p_k}\middle| S\right) = E\left(\sum_{k\in S}\frac{y_k}{\pi_k}\right) = t_y. \quad (9)$$

L'estimateur par expansion est donc sans biais sous les hypothèses précédentes : $\pi_k > 0$ pour tout $k \in U$ et $p_k > 0$ pour tout $k \in S$ (pas de biais de sélection en première ou en seconde phase).

$$V\left(\hat{t}_{ye}\right) = VE\left(\hat{t}_{ye}|S\right) + EV\left(\hat{t}_{ye}|S\right)$$
$$= \underbrace{V\left(\sum_{k\in S}\frac{y_k}{\pi_k}\right)}_{\text{Variance 1ère phase}} + \underbrace{EV\left(\hat{t}_{ye}|S\right)}_{\text{Variance 2nde phase}}. \quad (10)$$

Estimation de la variance de 1ère phase

La variance de 1ère phase peut être estimée sans biais par

$$\hat{V}_1\left(\hat{t}_{ye}\right) = \sum_{k,l\in S_2} \frac{\pi_{kl}-\pi_k\pi_l}{\pi_{kl}p_{kl}} \frac{y_k}{\pi_k}\frac{y_l}{\pi_l}$$

si tous les $\pi_{kl} > 0,\ k,l \in U$ et tous les $p_{kl} > 0,\ k,l \in S$.

Preuve : $E\left\{\hat{V}_1\left(\hat{t}_{ye}\right)\right\} = EE\left\{\sum_{k,l\in S} \frac{\pi_{kl}-\pi_k\pi_l}{\pi_{kl}} \frac{y_k}{\pi_k}\frac{y_l}{\pi_l}\frac{r_k r_l}{p_{kl}}\middle| S\right\}$

$$= E\left\{\sum_{k,l\in S} \frac{\pi_{kl}-\pi_k\pi_l}{\pi_{kl}} \frac{y_k}{\pi_k}\frac{y_l}{\pi_l}\right\}$$

$$= E\left\{\hat{V}_{HT}\left(\hat{t}_{y\pi}\right)\right\} = V\left(\sum_{k\in S}\frac{y_k}{\pi_k}\right).$$

Estimation de la variance de 2nde phase

Nous nous limitons au cas où l'échantillon S_2 est sélectionné dans S selon un plan de Poisson de probabilités d'inclusion p_k.
La variance de seconde phase vaut alors

$$EV\left(\hat{t}_{ye}|S\right) = EV\left(\sum_{k\in S}\frac{y_k}{\pi_k}\frac{r_k}{p_k}\middle| S\right) = E\left\{\sum_{k\in S}\left(\frac{y_k}{\pi_k}\right)^2\frac{1-p_k}{p_k}\right\}.$$

Elle peut être estimée sans biais par

$$\hat{V}_2\left(\hat{t}_{ye}\right) = \sum_{k\in S_2}\left(\frac{y_k}{\pi_k}\right)^2\frac{1-p_k}{(p_k)^2}.$$

Preuve : $$E\left\{\hat{V}_2\left(\hat{t}_{ye}\right)\right\} = EE\left\{\sum_{k\in S}\left(\frac{y_k}{\pi_k}\right)^2\frac{(1-p_k)r_k}{(p_k)^2}\middle| S\right\}$$

$$= E\left\{\sum_{k\in S}\left(\frac{y_k}{\pi_k}\right)^2\frac{1-p_k}{p_k}\right\}.$$

Traitement de la non-réponse totale

Le problème

La non-réponse totale ("unit non-response") survient lorsqu'aucune information (autre que celle de la base de sondage) n'est relevée pour une unité.

Nous allons traiter ce problème par **repondération** : nous faisons porter aux répondants le poids des non-répondants. Cette repondération se justifie sous une modélisation du mécanisme de non-réponse.

Cette modélisation permet d'estimer les probabilités de réponse à l'enquête, pour obtenir les poids corrigés de la non-réponse totale.

Les étapes du traitement de la non-réponse totale

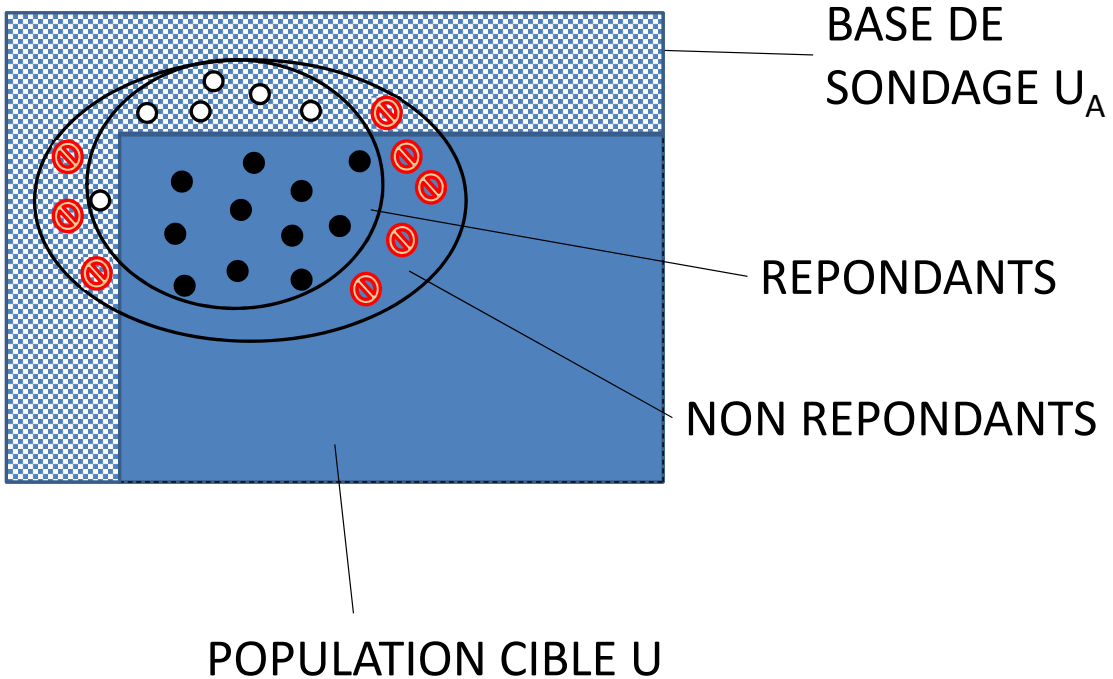
1. Identification des non-répondants,
2. Modélisation du mécanisme de non-réponse (recherche des facteurs explicatifs),
3. Estimation des probabilités de réponse,
4. Calcul des poids corrigés de la non-réponse totale,
5. Estimation ponctuelle et estimation de précision.

Etape 1
Identification des non-répondants

Identification des non-répondants

La distinction entre individus **hors-champ** et individus non-répondants peut être difficile. Les individus **non-répondants** font partie du champ de l'enquête, mais leur réponse n'est pas observée et doit être compensée.

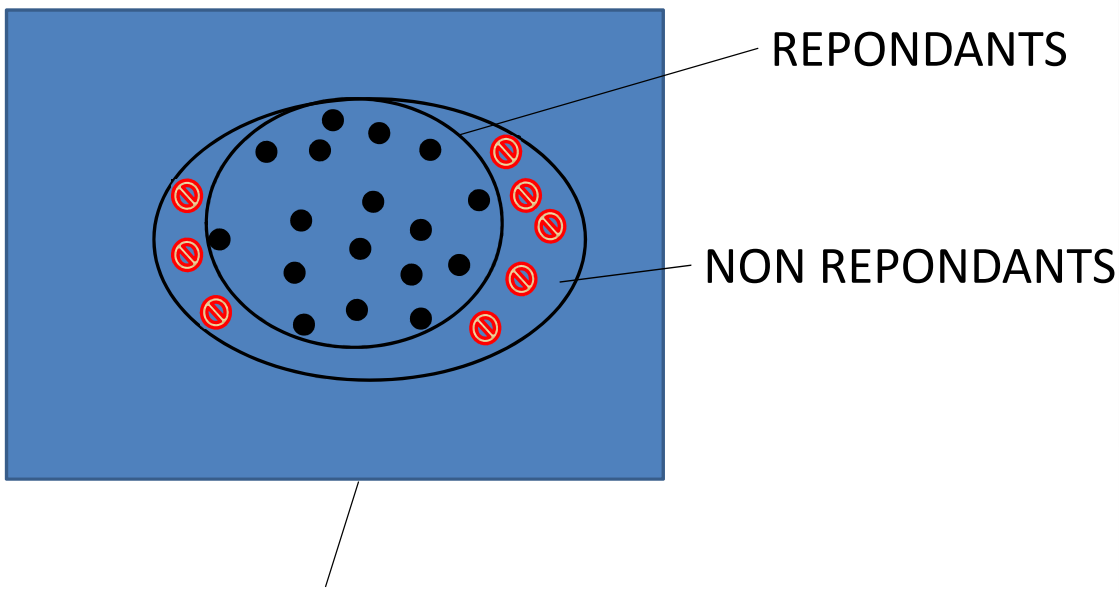
Le statut (champ/hors champ) est connu pour les répondants de l'enquête, mais il peut être difficile à obtenir pour les non-répondants ce qui peut occasionner une surestimation.



Identification des non-répondants

Il est important d'essayer d'obtenir de l'information (par exemple, auprès du voisinage) permettant de classer une unité en hors-champ ou en non-répondant.

Pour simplifier, nous supposerons dans la suite que la base de sondage U_A coincide exactement avec la population-cible U : pas de problème d'identification des non-répondants.



Etape 2
Modélisation du mécanisme de non-réponse

Modélisation du mécanisme de non-réponse

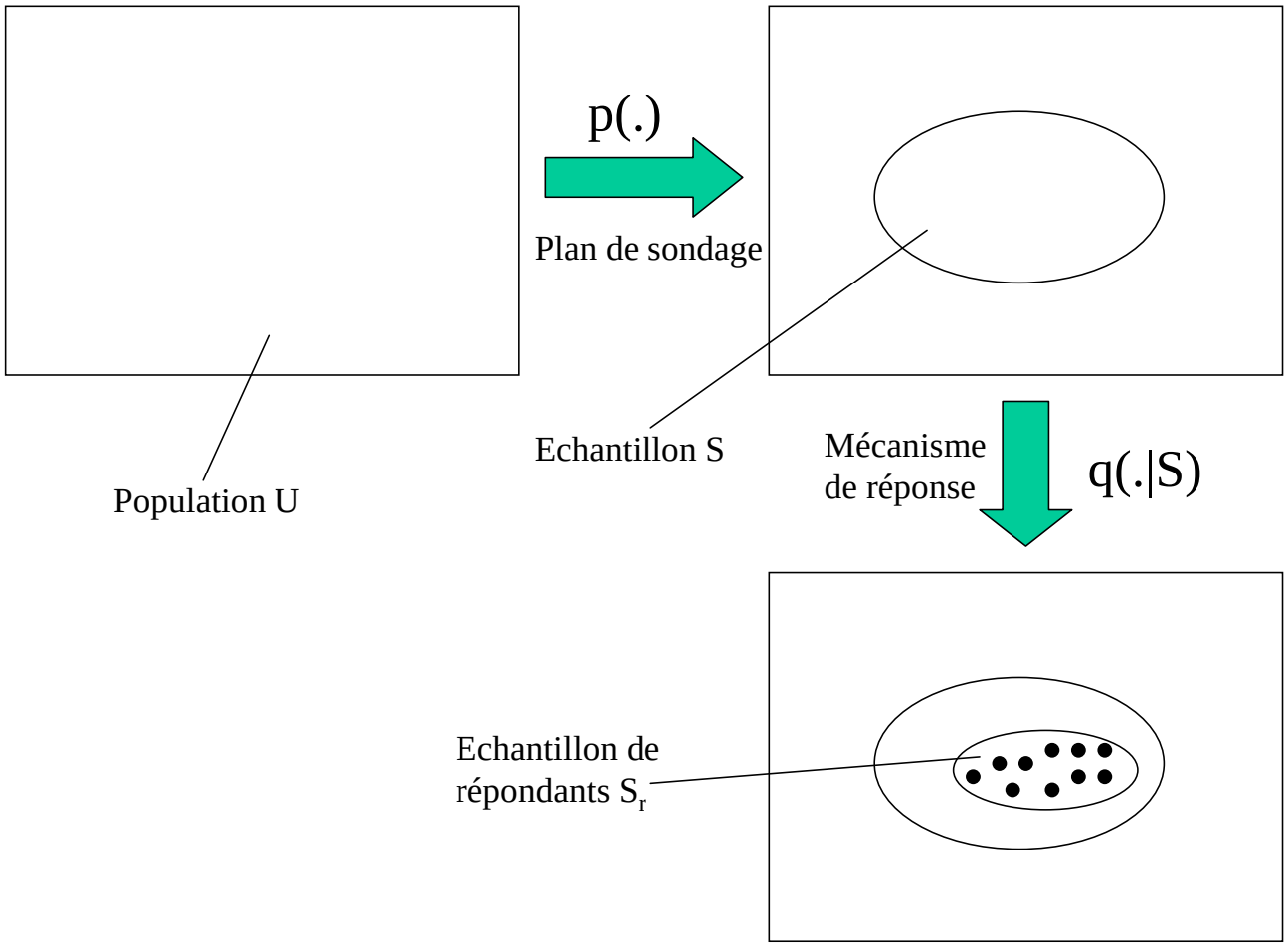
En situation de non-réponse totale :

- le mécanisme de sélection de l'échantillon S est connu,
- le mécanisme de réponse qui conduit au sous-échantillon de répondants S_r est en revanche inconnu.

Nous avons recours à une modélisation du mécanisme aléatoire conduisant à S_r sous la forme d'un **échantillonnage en deux phases** :

- la 1ère phase correspond à la sélection de l'échantillon S,
- la 2nde phase correspond à la "sélection" du sous-échantillon de répondants S_r
  $\Rightarrow$ mécanisme de réponse

Mécanisme de réponse



Modélisation du mécanisme de réponse

Nous notons r_k la variable indicatrice de réponse pour l'individu k, valant 1 si l'individu a répondu à l'enquête et 0 sinon.
Nous notons $p_{k|S} \equiv p_k$ la probabilité de réponse pour l'unité k :

$$p_k = \Pr(r_k = 1|S).$$

Nous faisons les hypothèses suivantes :

1. Pas de non-répondants irréductibles : toutes les probabilités de réponse vérifient $0 < p_k \leq 1$.
2. Mécanisme de réponse de Poisson : les individus répondent indépendamment les uns des autres, de sorte que :

$$\Pr(k, l \in S_r|S) \equiv p_{kl} \quad = \quad p_k p_l.$$

L'enquête doit être réalisée de façon à ce que l'hypothèse 1 soit plausible. L'hypothèse 2 permet de simplifier l'étude des estimateurs corrigés de la non-réponse. Elle peut être affaiblie (Haziza et Rao, 2003 ; Skinner et D'Arrigo, 2011).

Types de mécanisme

Nous distinguerons trois types de mécanisme de non-réponse :

- uniforme (ou MCAR),
- ignorable (ou MAR),
- non-ignorable (ou NMAR).

Le mécanisme est dit uniforme (ou Missing Completely At Random) quand $p_k = p$, i.e. quand tous les individus ont la même probabilité de réponse. C'est une hypothèse généralement peu réaliste.

Exemple : non-réponse provenant de la perte de questionnaires.

Mécanisme MAR

Le mécanisme de non-réponse est dit ignorable (ou Missing At Random) quand les probabilités de réponse peuvent être expliquées à l'aide de l'information auxiliaire disponible :

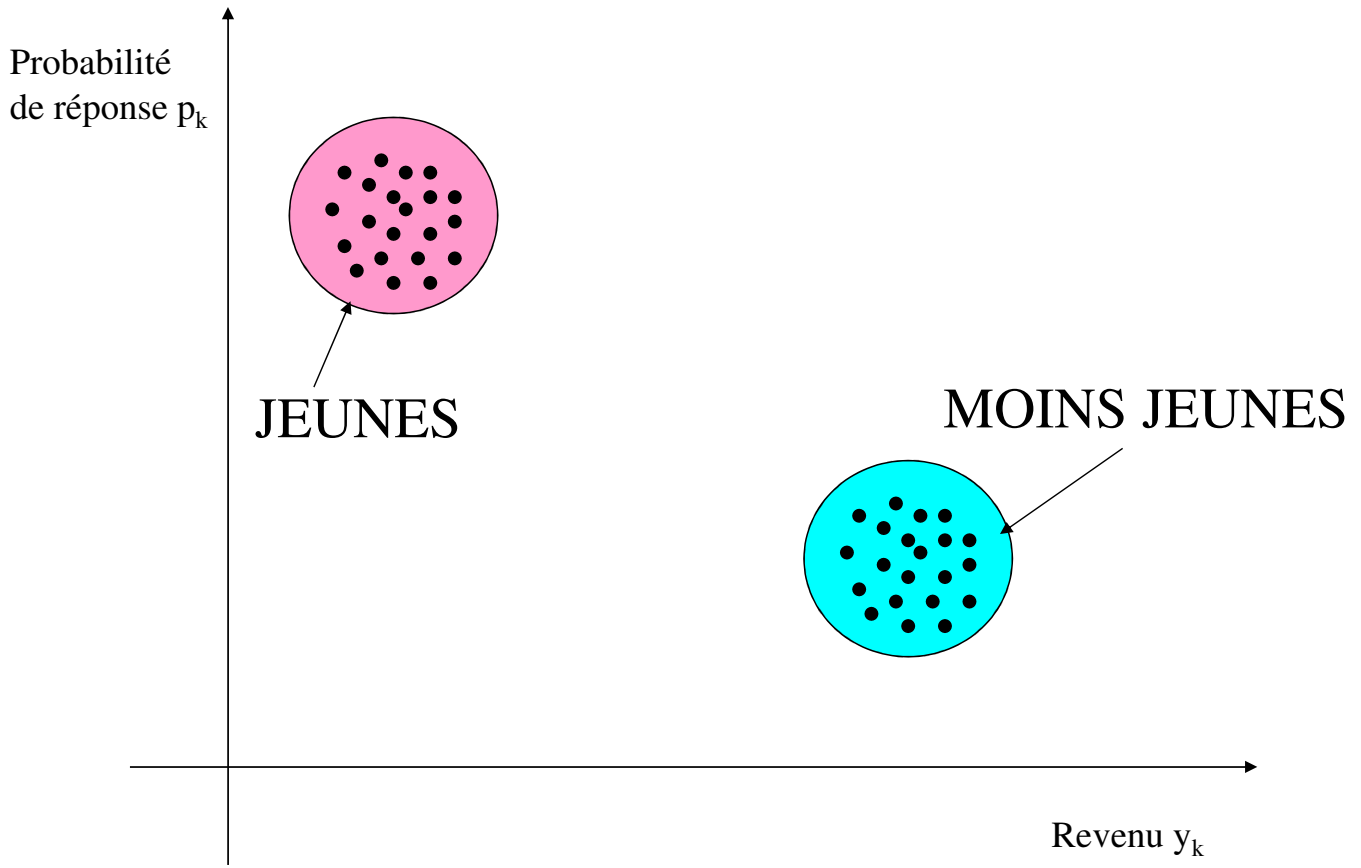
$$\Pr(r_k = 1|y_k, \mathbf{z}_k) \quad = \quad \Pr(r_k = 1|\mathbf{z}_k),$$

avec

- y_k la variable d'intérêt,
- $\mathbf{z}_k$ le vecteur des valeurs prises par un vecteur $\mathbf{z}$ de variables auxiliaires pour l'individu k de S.

Exemple : enquête sur le revenu + non-réponse expliquée par l'âge des individus.

Un exemple de non-réponse MAR



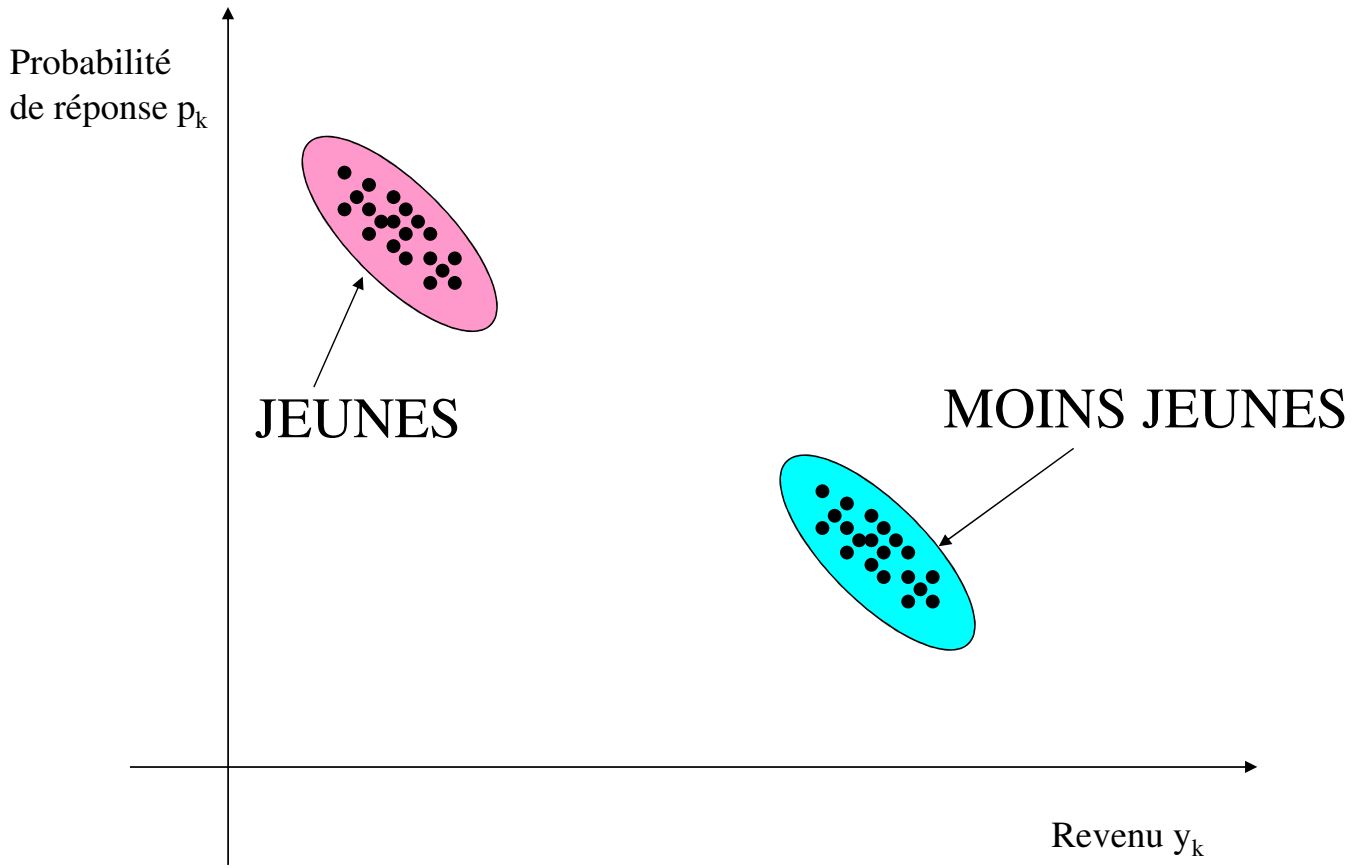
Mécanisme NMAR

Un mécanisme de non-réponse qui n'est pas ignorable est dit non-ignorable (ou Non Missing At Random). Cela signifie que la non-réponse dépend de la variable d'intérêt, même une fois que l'on a pris en compte les variables auxiliaires.

Il est difficile de corriger de la non-réponse non ignorable, ou même de la détecter. **Dans la suite, nous supposerons être dans le cas d'un mécanisme MAR**.

Exemple : enquête sur le revenu + non-réponse expliquée par le croisement âge $\times$ revenu.

Un exemple de non-réponse NMAR



Exemple sur données simulées

Nous considérons une population artificielle contenant $N_1 = 250$ jeunes (sous-pop. U_1) et $N_2 = 250$ moins jeunes (sous-pop. U_2), et une variable d'intérêt y (revenu) générée selon le modèle

$$y_k = \begin{cases} 50 + 10\ \epsilon_k & \text{pour les jeunes,} \\ 100 + 10\ \epsilon_k & \text{pour les moins jeunes,} \end{cases}$$

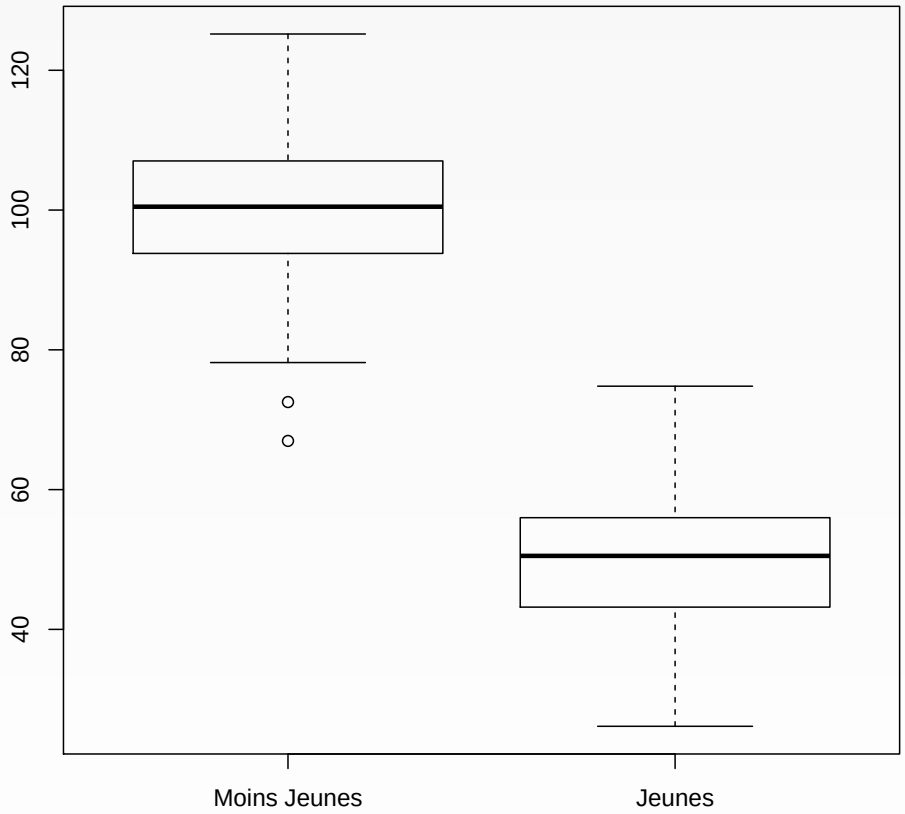
avec les ϵ_k générés selon une loi Normale$(0, 1)$.

Nous nous plaçons dans le cas d'un recensement (tous les individus de U sont théoriquement enquêtés), avec de la non-réponse totale. Nous considérons deux jeux de probabilités de réponse :

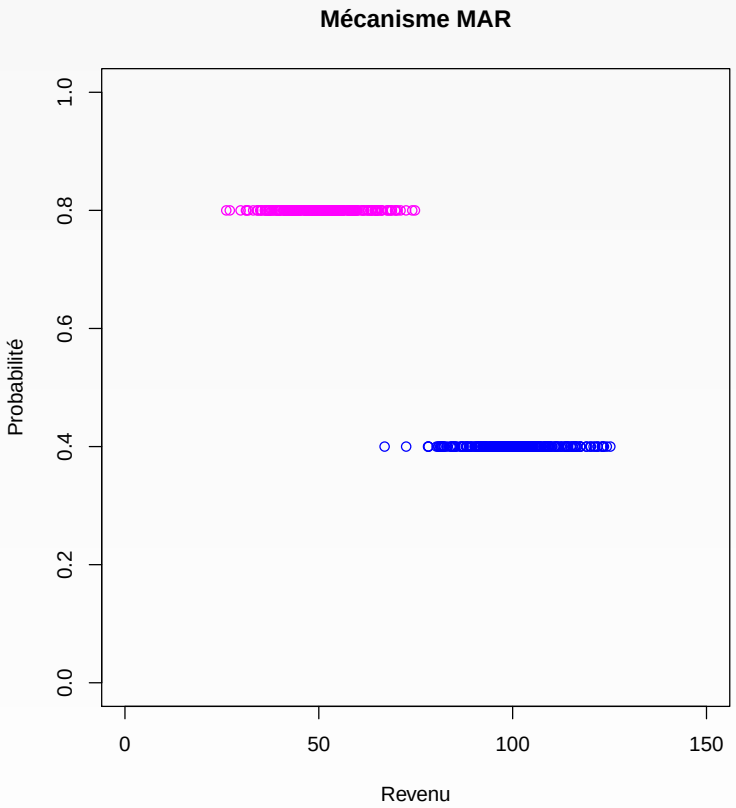
- mécanisme MAR q_1 : $p_{1k} = 0.8$ pour les jeunes et $p_{1k} = 0.4$ pour les moins jeunes,
- mécanisme NMAR q_2 : $p_{2k} = \dfrac{\exp^{8.5-0.1\times y}}{1 + \exp^{8.5-0.1\times y}}$.

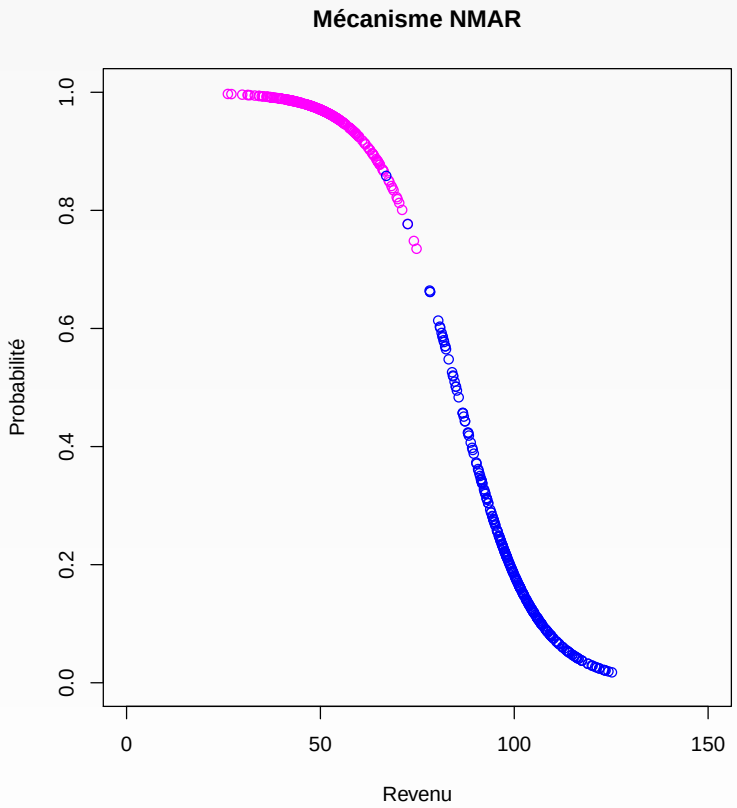
Probabilité de réponse moyenne de 0.60 environ dans chaque cas.

Distribution des revenus par classe d'âge



Mécanismes de réponse





Estimation naïve du revenu moyen

Soit S_r le sous-ensemble de répondants dans U. Une solution est d'utiliser la moyenne simple des répondants

$$\bar{y}_r = \frac{1}{n_r} \sum_{k \in S_r} y_k.$$

Nous allons montrer que cet estimateur est biaisé. Nous notons

$$\bar{y}_r = \frac{\sum_{k \in U} r_k y_k}{\sum_{k \in U} r_k} = f(\hat{\tau}_{\mathbf{y}}) \text{ avec } \left\{ \begin{array}{rcl} \hat{\tau}_{\mathbf{y}} & = & \sum_{k \in U} r_k \mathbf{y}_k \\ \mathbf{y}_k & = & (1, y_k)^\top, \end{array} \right.$$

et avec $f(x, y) = x/y$ la fonction ratio. Nous avons

$$E_q\left(\hat{\tau}_{\mathbf{y}}\right) = E_q\left(\sum_{k \in U} r_k \mathbf{y}_k\right) = \sum_{k \in U} p_k \mathbf{y}_k \equiv \tau_{\mathbf{y}}.$$

Estimation naïve du revenu moyen (2)

Alors par linéarisation

$$f(\hat{\tau}_{\mathbf{y}}) - f(\tau_{\mathbf{y}}) = f'(\tau_{\mathbf{y}}) \times (\hat{\tau}_{\mathbf{y}} - \tau_{\mathbf{y}}) + o_p(n^{-1/2})$$

$$\Rightarrow \bar{y}_r - \frac{\sum_{k\in U} p_k y_k}{\sum_{k\in U} p_k} \simeq f'(\tau_{\mathbf{y}}) \times (\hat{\tau}_{\mathbf{y}} - \tau_{\mathbf{y}})$$

$$\Rightarrow E_q(\bar{y}_r) \simeq \frac{\sum_{k\in U} p_k y_k}{\sum_{k\in U} p_k}.$$

Après quelques calculs, nous obtenons :

$$E_q\left(\bar{y}_r - \mu_y\right) \simeq \frac{\frac{1}{N}\sum_{k\in U}(p_k - \mu_p)(y_k - \mu_y)}{\frac{1}{N}\sum_{k\in U} p_k} \equiv \frac{S_{py}}{\mu_p}.$$

L'estimateur $\bar{y}_r$ est biaisé si la variable d'intérêt est corrélée avec la probabilité de réponse. Dans le cas traité, la corrélation est négative (les jeunes ont de plus fortes probabilités de réponse, et des revenus plus faibles) donc l'estimateur $\bar{y}_r$ est biaisé négativement.

Estimation redressé de la non-réponse : cas MAR

Le mécanisme de réponse MAR déséquilibre l'échantillon par rapport à l'âge (les jeunes sont sur-représentés parmi les répondants), et donc indirectement par rapport au revenu.

Le mécanisme de réponse est entièrement expliqué par l'âge, et la structure par âge de la population est connue. Il est possible d'estimer les probabilités de réponse dans chaque groupe d'âge par

$$\hat{p}_h = \frac{n_{rh}}{N_h} = \frac{\sum_{k \in U_h} r_k}{N_h} \text{ avec } E_{q_1}(\hat{p}_h) = \frac{E_{q_1}\left(\sum_{k \in U_h} r_k\right)}{N_h} = \frac{N_h p_h}{N_h} = p_h.$$

Nous pouvons alors utiliser l'estimateur redressé

$$\bar{y}_{red} = \frac{\sum_{k \in S_r} \frac{y_k}{\hat{p}_k}}{\sum_{k \in S_r} \frac{1}{\hat{p}_k}} = \frac{\sum_{h=1}^{2} \frac{N_h}{n_{rh}} \sum_{k \in S_{rh}} y_k}{\sum_{h=1}^{2} \frac{N_h}{n_{rh}} \sum_{k \in S_{rh}} 1}$$

$$= \sum_{h=1}^{2} \frac{N_h}{N} \bar{y}_{rh}.$$

Estimation redressé de la non-réponse : cas MAR (2)

Nous verrons un peu plus loin comment étudier les propriétés d'un estimateur utilisant des probabilités de réponse estimées. Dans le cas traité ici, nous utilisons les résultats suivants :

- Dans chaque strate U_h, l'échantillon S_{rh} est obtenu par tirage de Poisson à probabilités égales.
- (ADMIS) Conditionnellement au nombre de répondants n_{rh}, le sous-échantillon S_{rh} est tiré selon un SRS(n_{rh}).

En utilisant les propriétés du STSRS, nous obtenons donc

$$E_{q_1}\left(\bar{y}_{red} - \mu_y \middle| n_{r1}, n_{r2}\right) = E_{q_1}\left(\sum_{h=1}^{2} \frac{N_h}{N}\bar{y}_{rh} - \mu_y \middle| n_{r1}, n_{r2}\right) = 0$$

$$\Rightarrow E_{q_1}\left(\bar{y}_{red} - \mu_y\right) = E_{q_1} E_{q_1}\left(\bar{y}_{red} - \mu_y \middle| n_{r1}, n_{r2}\right) = 0.$$

L'estimateur redressé est donc sans biais pour le revenu moyen μ_y.

Estimation redressé de la non-réponse : cas NMAR

Le mécanisme de réponse NMAR déséquilibre l'échantillon directement par rapport au revenu (les revenus plus faibles sont sur-représentés parmi les répondants).

Comme le revenu y_k n'est pas observé pour tous les individus de la population, il n'est pas possible de redresser directement par rapport à cette variable. Nous pouvons cependant utiliser l'estimateur $\bar{y}_{red}$ redressé sur l'âge.

Estimation redressé de la non-réponse : cas NMAR (2)

Sous le mécanisme de réponse NMAR :

$$
\begin{aligned}
E_{q_2}\left(\bar{y}_{red}-\mu_y\right) &= E_{q_2}\left\{\sum_{h=1}^{H}\frac{N_h}{N}\left(\bar{y}_{rh}-\mu_{yh}\right)\right\} \\
&\simeq \sum_{h=1}^{H}\frac{N_h}{N}\left(\frac{\sum_{k\in U_h} p_k y_k}{\sum_{k\in U_h} p_k}-\mu_{yh}\right) \\
&= \sum_{h=1}^{H}\frac{N_h}{N}\frac{\frac{1}{N_h}\sum_{k\in U_h}(y_k-\mu_{yh})(p_k-\mu_{ph})}{\frac{1}{N_h}\sum_{k\in U_h} p_k} \equiv \sum_{h=1}^{H}\frac{N_h}{N}\frac{S_{py,h}}{\mu_{ph}}.
\end{aligned}
$$

Rappelons qu'avec l'estimateur naïf :

$$
E_{q_2}\left(\bar{y}_r-\mu_y\right) \simeq \frac{S_{py}}{\mu_p}.
$$

Le biais de l'estimateur redressé est donc plus faible que celui de l'estimateur naïf si la post-stratification permet d'expliquer (au moins) une partie de la corrélation entre la variable d'intérêt y_k et la probabilité de réponse p_k.

Estimation sous un mécanisme MAR

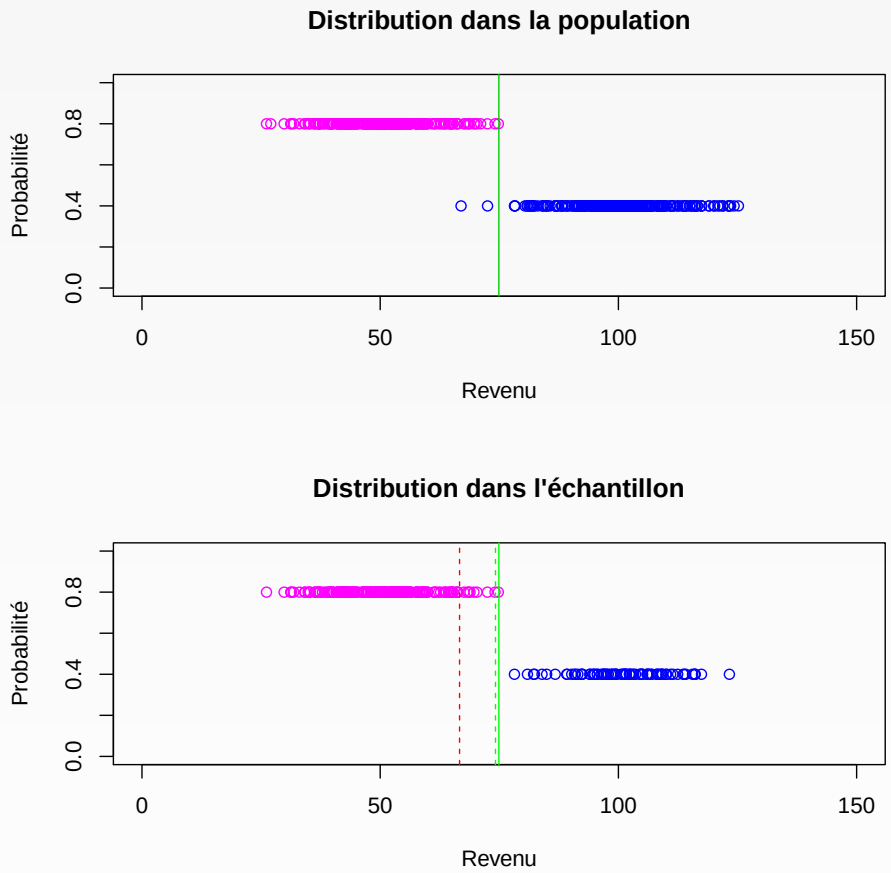


Figure – Distribution du revenu dans la population (en haut) et dans l'échantillon (en bas). Le trait vert plein représente la moyenne μ_y, le trait rouge en pointillés la moyenne des répondants $\bar{y}_r$, et le trait vert en pointillés l'estimateur redressé de la non-réponse $\bar{y}_{red}$.

Estimation sous un mécanisme NMAR

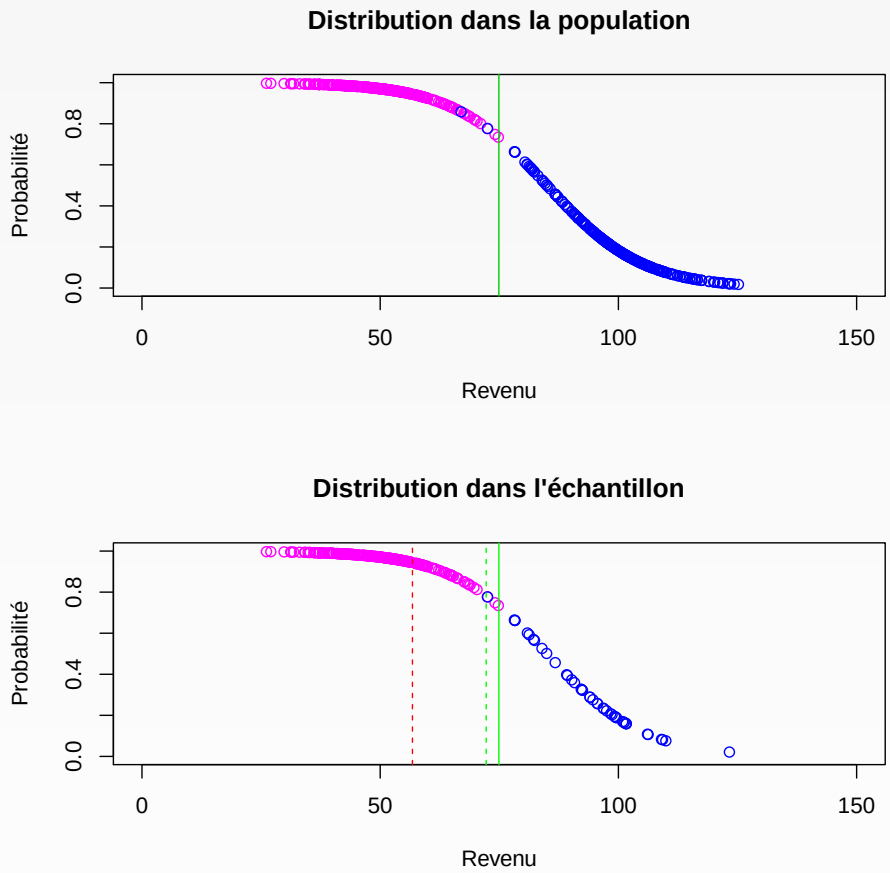


Figure – Distribution du revenu dans la population (en haut) et dans l'échantillon (en bas). Le trait vert plein représente la moyenne μ_y, le trait rouge en pointillés la moyenne des répondants $\bar{y}_r$, et le trait vert en pointillés l'estimateur redressé de la non-réponse $\bar{y}_{red}$.

Etapes 3 et 4
Estimation des probabilités de réponse
Calcul des poids redressés de la non-réponse totale

Estimation des probabilités de réponse

Les probabilités de réponse sont inconnues et doivent être estimées. Nous postulons alors un **modèle de réponse** de la forme

$$p_k \quad = \quad f(\mathbf{z}_k, \beta_0), \text{ avec}$$

- $\mathbf{z}_k$ un vecteur de variables auxiliaires connu sur S,
- $f(\cdot, \cdot)$ une fonction connue,
- β_0 un paramètre inconnu.

Nous considérerons la fonction de lien

$$f(\mathbf{z}_k, \beta) \quad = \quad \frac{\exp\left(\mathbf{z}_k^\top \beta\right)}{1 + \exp\left(\mathbf{z}_k^\top \beta\right)} = \text{expit}(\mathbf{z}_k^\top \beta),$$

qui correspond au modèle logistique, avec $\text{logit}(p_k) = \mathbf{z}_k^\top \beta_0$. D'autres fonctions de lien sont possibles (Da Silva et Opsomer, 2006 et 2009).

Estimation des probabilités de réponse (2)

Sous un mécanisme de réponse de Poisson, nous obtenons la log-vraisemblance

$$\begin{aligned}\mathcal{L}(r_1,\ldots,r_N) &= \sum_{k\in U}\{r_k\ln(p_k)+(1-r_k)\ln(1-p_k)\}\\ &= \sum_{k\in U} r_k \ln\left(\frac{p_k}{1-p_k}\right)+\sum_{k\in U}\ln(1-p_k)\\ &= \sum_{k\in U} r_k(\mathbf{z}_k^\top\beta)-\sum_{k\in U}\ln\left\{1+\exp\left(\mathbf{z}_k^\top\beta\right)\right\} \text{ sous le mod. log.}\end{aligned}$$

En différenciant par rapport à β, nous obtenons une équation estimante pour l'estimateur $\hat{\beta}$

$$U(\beta)\equiv\sum_{k\in S}\{r_k-f(\mathbf{z}_k,\beta)\}\mathbf{z}_k \quad = \quad 0.$$

Nous pouvons remplacer les probabilités inconnues $p_k=f(\mathbf{z}_k,\beta_0)$ par leurs estimations $\hat{p}_k=f(\mathbf{z}_k,\hat{\beta})$.

Estimation des probabilités de réponse (3)

Par linéarisation, nous obtenons

$$\begin{aligned}
\underbrace{U(\hat{\beta})}_{=0} - U(\beta_0) &\simeq U'(\beta_0) \times \left(\hat{\beta} - \beta_0\right) \\
\Rightarrow \hat{\beta} - \beta_0 &\simeq -\left\{U'(\beta_0)\right\}^{-1} U(\beta_0) \\
&= \underbrace{\left\{\sum_{k \in S} p_k(1-p_k)\mathbf{z}_k\mathbf{z}_k^\top\right\}^{-1}}_{A_s^{-1}} \sum_{k \in S}(r_k - p_k)\mathbf{z}_k. \quad (11)
\end{aligned}$$

Sous l'approximation (11), nous obtenons

$$\begin{aligned}
E\left(\hat{\beta} - \beta_0 \middle| S\right) \simeq 0 &\Rightarrow E\left(\hat{\beta} - \beta_0\right) \simeq 0, \\
V\left(\hat{\beta} - \beta_0 \middle| S\right) \simeq A_s^{-1} &\Rightarrow V\left(\hat{\beta} - \beta_0\right) \simeq E(A_s^{-1}) = O(n^{-1}).
\end{aligned}$$

Estimation en situation de non-réponse

Si les probabilités p_k étaient connues, nous serions dans le cas d'un échant. en deux phases. Nous pourrions utiliser l'**estimateur par expansion**

$$\hat{t}_{ye} = \sum_{k \in S_r} \frac{y_k}{\pi_k \; p_k} = \sum_{k \in U} \frac{I_k}{\pi_k} \frac{r_k}{p_k} y_k.$$

En remplaçant les probabilités de réponse par leurs estimations, nous obtenons l'estimateur corrigé de la non-réponse totale

$$\hat{t}_{yr} = \sum_{k \in S_r} \frac{y_k}{\pi_k \; \hat{p}_k} = \sum_{k \in U} \frac{I_k}{\pi_k} \frac{r_k}{\hat{p}_k} y_k.$$

Les poids corrigés de la non-réponse totale sont donc donnés par :

$$d_{rk} = \frac{1}{\pi_k \; \hat{p}_k} = \frac{d_k}{\hat{p}_k} \text{ pour } k \in S_r.$$

Cas des groupes homogènes de réponse

Un modèle de non-réponse couramment utilisé en pratique consiste à supposer que la probabilité de réponse p_k est constante au sein de groupes $S_1, \ldots, S_C$ partitionnant l'échantillon S :

$$\forall k \in S_c \quad p_k = p_c.$$

Ils sont appelés les **groupes homogènes de réponse** (GHR). Cette modélisation a l'avantage :

- d'être simple à mettre en oeuvre,
- d'offrir une certaine robustesse contre une mauvaise spécification du modèle de non-réponse.

Exemple : enquête sur le revenu + GHR définis en croisant sexe et tranche d'âge.

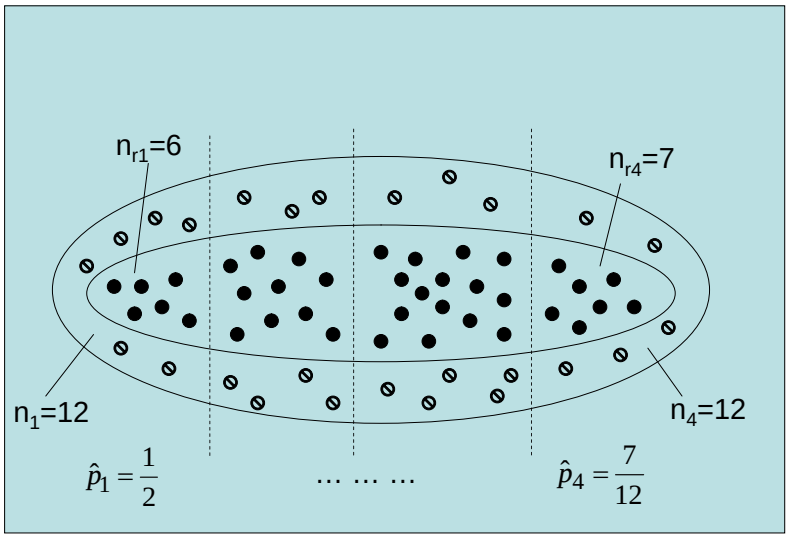
Cas des groupes homogènes de réponse

Au sein de chaque GHR S_c, la probabilité p_c est estimée par

$$\hat{p}_c = \frac{n_{rc}}{n_c},$$

en notant

- n_c le nombre d'individus dans S_c,
- n_{rc} le nombre de répondants dans S_c.



Détermination des GHR

En pratique, ces groupes peuvent être constitués de la façon suivante :

1. On effectue une régression logistique afin d'expliquer les probabilités de réponse en fonction de l'information auxiliaire disponible.
2. On peut ensuite :
   - soit ordonner les individus k selon les $\hat{p}_k$ (**méthode des scores**), puis diviser l'échantillon en groupes de tailles approximativement égales (méthode des quantiles égaux) ;
   - soit utiliser les variables qui resortent de façon significative dans la régression logistique, et les croiser pour définir les groupes (**méthode par croisement**).

Méthode des scores : exemple

k	r_k	$\hat{p}_k$
1	1	0.54
2	1	0.78
3	1	0.28
4	0	0.48
5	1	0.47
6	0	0.32
7	1	0.82
8	0	0.75
9	0	0.52
10	1	0.83
11	0	0.29
12	1	0.81

$\Rightarrow$

k	r_k	$\hat{p}_k$	$\hat{p}_c$
3	1	0.28	
11	0	0.29	0.33
6	0	0.32	
5	1	0.47	
4	0	0.48	
9	0	0.52	0.50
1	1	0.54	
8	0	0.75	
2	1	0.78	
12	1	0.81	0.80
7	1	0.82	
10	1	0.83	

Détermination des GHR (2)

La procédure précédent de détermination des GHR consiste à :

1. ajuster un modèle paramétrique de régression logistique,
2. constituer des GHR en fonction des résultats de ce modèle.

Une alternative strictement non paramétrique consiste à utiliser une méthode de segmentation par arbre (e.g., Deroyon, 2017 ; Gelein et al., 2019).

A chaque étape, la méthode détermine la variable et ses modalités qui séparent le mieux l'ensemble des données par rapport à l'indicatrice de réponse r_k, et par rapport à un critère de sélection (e.g., test du chi-deux avec CHAID ; Opsomer et Riddles, 2023).

La procédure permet de construire un arbre dont les feuilles constituent les GHR.

Etape 5
Estimation ponctuelle et estimation de précision

Estimation d'un total - probabilités de réponse connues

Nous nous trouvons alors dans le cas d'un échantillonnage en deux phases. Nous utilisons l'**estimateur par expansion**

$$\hat{t}_{ye} = \sum_{k \in S_r} \frac{y_k}{\pi_k \, p_k} = \sum_{k \in U} \frac{I_k}{\pi_k} \frac{r_k}{p_k} y_k.$$

Il est sans biais pour le total t_y, et sa variance est toujours plus grande qu'en situation de réponse complète :

$$\begin{aligned} E(\hat{t}_{ye}) &= E_p E_q(\hat{t}_{ye}|S) = E_p(\hat{t}_{y\pi}) = t_y, \\ V(\hat{t}_{ye}) &= V_p E_q(\hat{t}_{ye}|S) + E_p V_q(\hat{t}_{ye}|S) \\ &= \underbrace{V_p(\hat{t}_{y\pi})}_{\text{Variance Echantillonnage}} + \underbrace{E_p\left[\sum_{k \in S}\left(\frac{y_k}{\pi_k}\right)^2 \frac{1-p_k}{p_k}\right]}_{\text{Variance Non Réponse}}. \end{aligned}$$

Estimation d'un total - probabilités de réponse inconnues

Les probabilités de réponse sont estimées :

$$p_k = f(\mathbf{z}_k, \beta_0) \quad \Rightarrow \quad \hat{p}_k = f(\mathbf{z}_k, \hat{\beta}).$$

Nous obtenons l'estimateur corrigé de la non-réponse totale

$$\hat{t}_{yr} = \sum_{k \in S_r} \frac{y_k}{\pi_k \, \hat{p}_k} = \sum_{k \in U} \frac{I_k}{\pi_k} \frac{r_k}{\hat{p}_k} y_k.$$

En utilisant le développement de Taylor (11) obtenu pour $\hat{\beta} - \beta_0$, nous obtenons après calcul l'approximation

$$\hat{t}_{yr} \simeq \hat{t}_{ye} - \sum_{k \in S} (r_k - p_k) \mathbf{z}_k^\top \lambda_s, \tag{12}$$

$$\text{avec } \lambda_s = \left\{ \sum_{k \in S} p_k (1 - p_k) \mathbf{z}_k \mathbf{z}_k^\top \right\}^{-1} \left\{ \sum_{k \in S} \frac{1 - p_k}{\pi_k} \mathbf{z}_k y_k \right\}.$$

Estimation d'un total - probabilités de réponse inconnues

Nous obtenons :

$$
\begin{aligned}
E(\hat{t}_{yr}) &\simeq E_p E_q(\hat{t}_{ye}|S) = t_y, \\
V(\hat{t}_{yr}) &= V_p E_q(\hat{t}_{yr}|S) + E_p V_q(\hat{t}_{yr}|S) \\
&\simeq \underbrace{V_p(\hat{t}_{y\pi})}_{\text{Variance Echantillonnage}} + \underbrace{E_p\left[\sum_{k\in S}\frac{1-p_k}{p_k}\left(\frac{y_k}{\pi_k} - p_k\lambda_s^\top \mathbf{z}_k\right)^2\right]}_{\text{Variance Non Réponse}}.
\end{aligned}
$$

L'estimateur $\hat{t}_{yr}$ est approximativement sans biais pour t_y, et sa variance est plus faible que celle de l'estimateur par expansion.

Estimation ponctuelle et estimation de variance

La variance peut se réécrire :

$$V(\hat{t}_{yr}) \simeq \underbrace{\sum_{k,l \in U} \Delta_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}}_{V_p(\hat{t}_{yr})} + \underbrace{E_p \left[\sum_{k \in S} \frac{1-p_k}{p_k} \left(\frac{y_k}{\pi_k} - p_k \lambda_s^\top \mathbf{z}_k \right)^2 \right]}_{V_{nr}(\hat{t}_{yr})}.$$

Nous pouvons l'estimer approximativement sans biais par :

$$v(\hat{t}_{yr}) = \sum_{k,l \in S_r} \frac{\Delta_{kl}}{\pi_{kl} \hat{p}_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} + \sum_{k \in S_r} \frac{1-\hat{p}_k}{\hat{p}_k^2} \left(\frac{y_k}{\pi_k} - \hat{p}_k \hat{\lambda}_r^\top \mathbf{z}_k \right)^2 \tag{13}$$

avec

$$\hat{\lambda}_r = \left[\sum_{k \in S_r} (1-\hat{p}_k) \mathbf{z}_k \mathbf{z}_k^\top \right]^{-1} \sum_{k \in S_r} \frac{1-\hat{p}_k}{\pi_k \ \hat{p}_k} \mathbf{z}_k y_k. \tag{14}$$

L'estimateur $\hat{\lambda}_r$ s'obtient à partir de λ_s en remplaçant chaque $\sum_{k \in S} \diamond$ par $\sum_{k \in S_r} \frac{\diamond}{p_k}$ (estimation par substitution), puis en remplaçant la probabilité de réponse p_k par son estimateur $\hat{p}_k$.

Cas des groupes homogènes de réponse

Estimateur redressé de la non-réponse totale

Dans le cas des GHR, la probabilité de réponse p_k est supposée constante au sein de groupes $S_1, \ldots, S_C$ partitionnant l'échantillon S :

$$\forall k \in S_c \quad p_k = p_c.$$

Nous nous trouvons dans le cas d'un modèle de régression logistique, avec $\mathbf{z}_k = \{1(k \in S_1), \ldots, 1(k \in S_C)\}^\top$. Nous obtenons

$$\begin{aligned} \hat{t}_{yr} &= \sum_{k \in S_r} \frac{y_k}{\pi_k \hat{p}_k} = \sum_{c=1}^{C} \sum_{k \in S_{rc}} \frac{y_k}{\pi_k \underbrace{\hat{p}_k}_{\frac{n_{rc}}{n_c}}} \\ &= \sum_{c=1}^{C} \frac{n_c}{n_{rc}} \sum_{k \in S_{rc}} \frac{y_k}{\pi_k}. \end{aligned}$$

Variance d'échantillonnage

La variance d'échantillonnage

$$V_p(\hat{t}_{yr}) = \sum_{k,l \in U} \Delta_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}$$

est estimée approximativement sans biais par

$$v_p(\hat{t}_{yr}) = \sum_{k,l \in S_r} \frac{\Delta_{kl}}{\pi_{kl} \hat{p}_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}.$$

Estimateur de la variance de non-réponse

Avec $\mathbf{z}_k = \{1(k \in S_1), \ldots, 1(k \in S_C)\}^\top$, nous obtenons d'abord

$$\begin{aligned}
\hat{\lambda}_r &= \left[\sum_{k\in S_r}(1-\hat{p}_k)\mathbf{z}_k\mathbf{z}_k^\top\right]^{-1}\sum_{k\in S_r}\frac{1-\hat{p}_k}{\pi_k\ \hat{p}_k}\mathbf{z}_k y_k \\
&= \begin{pmatrix}(1-\hat{p}_1)n_{r1} & & \\ & \ddots & \\ & & (1-\hat{p}_c)n_{rc}\end{pmatrix}^{-1}\begin{pmatrix}\frac{1-\hat{p}_1}{\hat{p}_1}\sum_{k\in S_{r1}}\frac{y_k}{\pi_k} \\ \vdots \\ \frac{1-\hat{p}_C}{\hat{p}_C}\sum_{k\in S_{rC}}\frac{y_k}{\pi_k}\end{pmatrix} \\
&= \left(\frac{1}{n_{r1}\hat{p}_1}\sum_{k\in S_{r1}}\frac{y_k}{\pi_k}, \cdots, \frac{1}{n_{rC}\hat{p}_C}\sum_{k\in S_{rC}}\frac{y_k}{\pi_k}\right)^\top.
\end{aligned}$$

Estimateur de la variance de non-réponse (2)

En injectant cette expression dans l'estimateur de la variance due à la non-réponse

$$v_{nr}(\hat{t}_{yr}) = \sum_{k \in S_r} \frac{1-\hat{p}_k}{\hat{p}_k^2} \left(\frac{y_k}{\pi_k} - \hat{p}_k \hat{\lambda}_r^\top \mathbf{z}_k \right)^2,$$

nous obtenons dans le cas des GHR

$$v_{nr}(\hat{t}_{yr}) = \sum_{c=1}^{C} \frac{1-\hat{p}_c}{(\hat{p}_c)^2} \sum_{k \in S_{rc}} \left(\frac{y_k}{\pi_k} - \frac{1}{n_{rc}} \sum_{l \in S_{rc}} \frac{y_l}{\pi_l} \right)^2. \quad (15)$$

Application au SRS stratifié

Nous supposons que la population U est partitionnée en H strates de tailles $N_1, \ldots, N_H$, et que nous sélectionnons indépendamment dans chaque strate U_h un échantillon S_h selon un SRS de taille n_h.

Nous supposons que le mécanisme de non-réponse peut être modélisé par des GHR $S_1, \ldots, S_C$ partitionnant l'échantillon S.

L'estimateur corrigé de la non-réponse totale s'écrit

$$\begin{aligned}\hat{t}_{yr} &= \sum_{k \in S_r} \frac{1}{\pi_k} \frac{1}{\hat{p}_k} y_k \\ &= \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{c=1}^{C} \frac{n_c}{n_{rc}} \sum_{k \in S_{rch}} y_k \text{ avec } S_{rch} = S_{rc} \cap U_h.\end{aligned}$$

Application au SRS stratifié

Variance d'échantillonnage

La variance d'échantillonnage est donnée par

$$V_p(\hat{t}_{yr}) \simeq \sum_{h=1}^{H} N_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) S_{yh}^2 \quad \text{avec} \quad S_{yh}^2 = \frac{1}{N_h - 1} \sum_{k \in U_h} (y_k - \mu_{yh})^2.$$

En l'absence de non-réponse totale, nous pouvons l'estimer sans biais par

$$\tilde{v}_p(\hat{t}_{yr}) = \sum_{h=1}^{H} N_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) s_{yh}^2 \quad \text{avec} \quad s_{yh}^2 \simeq \frac{1}{n_h} \sum_{k \in S_h} (y_k - \bar{y}_h)^2.$$

En situation de non-réponse totale, nous pouvons estimer s_{yh}^2 par

$$s_{yhr}^2 = \frac{\sum_{k \in S_{hr}} \frac{1}{\hat{p}_k} (y_k - \bar{y}_{hr})^2}{\sum_{k \in S_{hr}} \frac{1}{\hat{p}_k}} \quad \text{avec} \quad \bar{y}_{hr} = \frac{\sum_{k \in S_{hr}} \frac{y_k}{\hat{p}_k}}{\sum_{k \in S_{hr}} \frac{1}{\hat{p}_k}}.$$

En résumé

1. Identification des non-répondants
   $\Rightarrow$ séparation des individus hors-champ et des non-répondants
2. Recherche des facteurs explicatifs de la non-réponse
   $\Rightarrow$ e.g., régression logistique pour identifier les $\mathbf{z}_k$ explicatifs
3. Estimation des probabilités de réponse
   $\Rightarrow$ e.g., méthode des scores ou méthode par croisement pour définir les GHR
4. Calcul des poids corrigés de la non-réponse totale
5. (**Calage** des estimateurs)
6. Estimation ponctuelle et estimation de variance

Application

Enquête PPV

Contexte de l'enquête

L'enquête Panel Politique de la Ville (PPV) a été mise en place pour étudier les conditions de vie des habitants des quartiers de la politique de la ville. Les quatre vagues d'enquête (entre 2011 et 2014) visent à appréhender :

- la mobilité résidentielle dans les quartiers,
- la perception des politiques publiques,
- l'impact des politiques publiques sur les bénéficiaires.

L'échantillon initial est tiré selon un plan à 3 degrés (Couvert el al., 2016) :

- tirage d'un échantillon de quartiers de la politique de la ville, stratifié selon le degré d'avancement du programme de rénovation urbaine,
- dans les quartiers, tirage d'un échantillon de logements à l'aide d'une base de sondage constituée à partir des **EAR**[1],
- dans les logements, tirage d'une **unité de vie** ($\simeq$ un ménage) et de tous les individus de cette unité de vie.

1. Enquêtes annuelles de recensement

Pondération

L'enquête comporte un questionnaire au niveau ménage et un questionnaire au niveau individuel. L'échantillon va donc comporter un double jeu de pondération.

Chaque ménage $k \in$ logement $j \in$ quartier i possède un poids initial de sondage

$$d_k = \underbrace{d_i}_{\text{pds quartier}} \times \underbrace{d_{j|i}}_{\text{pds log. ds quartier}} \times \underbrace{d_{k|i,j}}_{\text{pds mén. ds logement}} . \quad (16)$$

Chaque individu $l \in$ ménage k possède un poids initial de sondage

$$d_l = \underbrace{d_k}_{\text{pds ménage}} \times \underbrace{1}_{\text{pds ind. ds ménage}} = d_k. \quad (17)$$

Non-réponse au niveau ménage

La non-réponse totale sur l'échantillon S^{men} de ménages conduit au sous-échantillon de répondants S_r^{men}.

La correction de la non-réponse totale des ménages s'est faite par la méthode des GHR :

- variables explicatives identifiées par régression logistique : nb de pièces, HLM (oui/non), type d'habitation, année de construction.
- constitution des GHR par la méthode des scores (8 groupes).

Chaque ménage $k \in S_r^{men}$ possède le poids redressé de la non-réponse

$$d_{rk} = \underbrace{d_k}_{\text{pds de sondage}} \times \underbrace{\frac{1}{\hat{p}_k}}_{\text{pds de NR ménage}} . \quad (18)$$

Non-réponse au niveau individuel

La non-réponse totale sur l'échantillon S^{ind} d'individus conduit au sous-échantillon d'individus répondants S_r^{ind}.

La correction de la non-réponse totale des individus s'est faite par la méthode des GHR :

- variables explicatives identifiées par régression logistique : âge, lieu de naissance, statut matrimonial.
- constitution des GHR par la méthode des scores (5 groupes).

Chaque individu $l \in S_r^{ind}$ possède le poids redressé de la non-réponse

$$d_{rl} = \underbrace{d_{rk}}_{\text{pds ménage}} \times \underbrace{\frac{1}{\hat{p}_l}}_{\text{pds de NR individuel}} . \quad (19)$$

Simulations sur données réelles

Cadre

Nous considérons une population de $N = 10,000$ individus extraite de l'enquête canadienne sur la santé (CCHS). Nous nous intéressons à l'estimation de la taille moyenne et du poids moyen des individus.

Nous disposons des variables auxiliaires :

- âge : 3 modalités (12-17, 18-64, 65 et +),
- sexe : 2 modalités,
- statut matrimonial : 4 modalités (married, common law, widow/sep/div, single/never married),
- province : 11 modalités,
- consommation d'alcool : 4 modalités (regular, occasional, former, never drank).

Exemple sur données réelles (2)

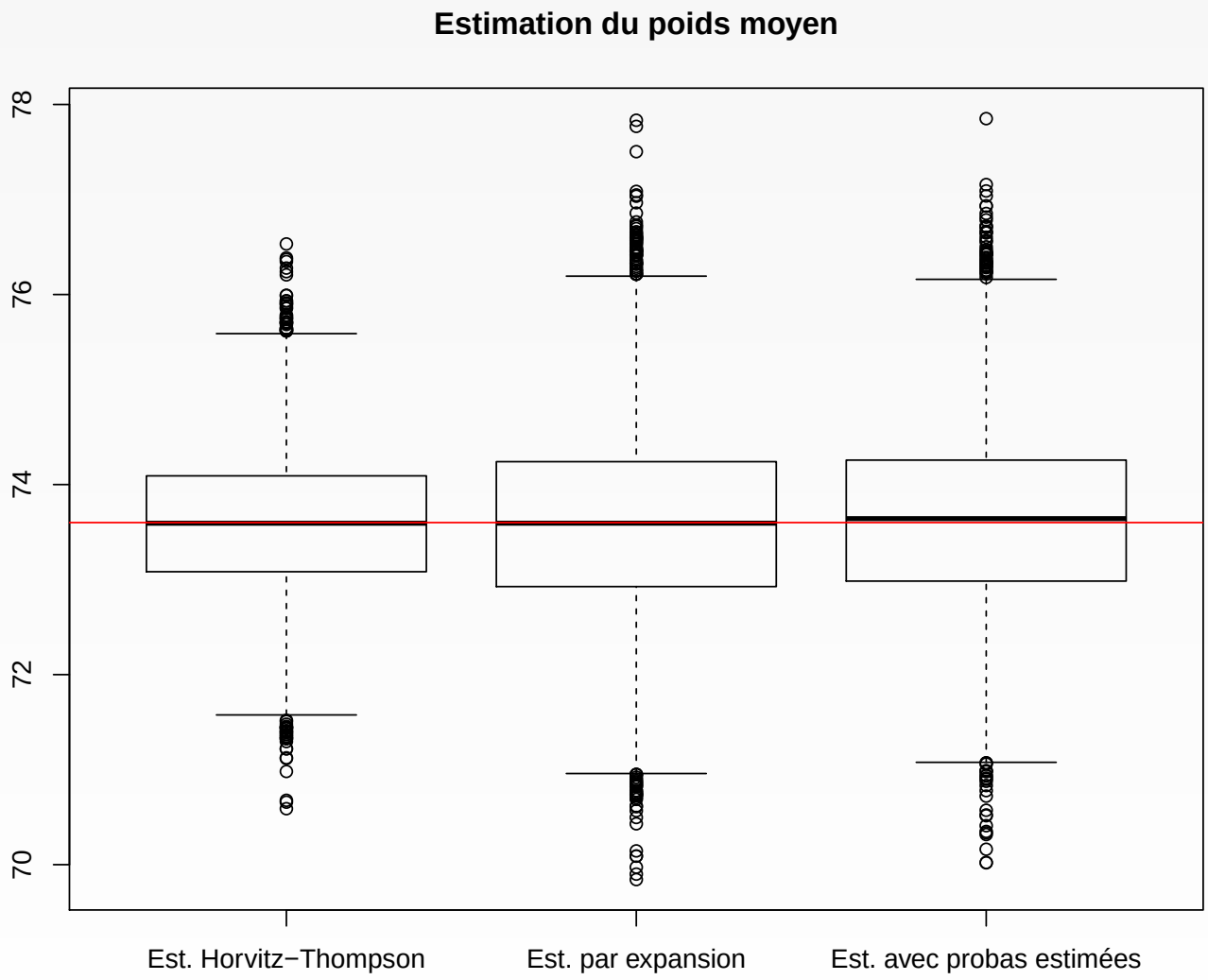
Nous sélectionnons un échantillon de taille $n = 500$ selon un SRS. Nous considérons le mécanisme de réponse (inconnu) :

$$\begin{aligned} & & & & & +0.70(st_k = 4) \\ & & & & -0.05(a_k = 3) & -0.50(st_k = 3) \\ & & & -0.60(s_k = 2) & +0.15(a_k = 2) & +0.50(st_k = 2) \\ logit(p_{1k}) = & & 0.80 & +0.60(s_k = 1) & -0.10(a_k = 1) & -0.70(st_k = 1) \end{aligned}$$

La probabilité de réponse moyenne est égale à 0.62 environ.

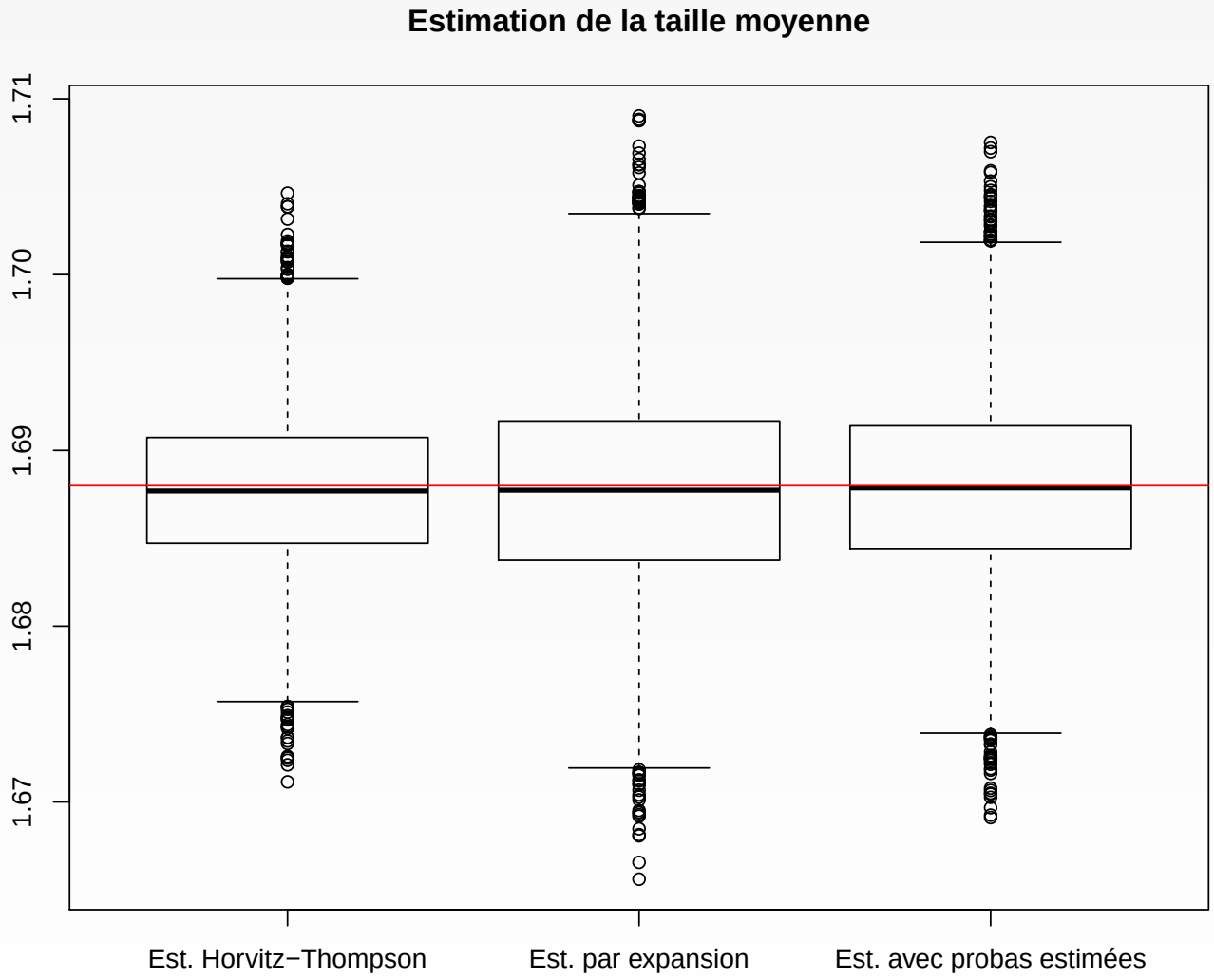
Distribution des estimateurs

Les estimateurs de totaux sont non biaisés



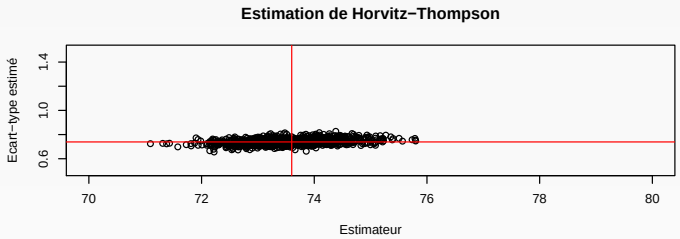
Distribution des estimateurs

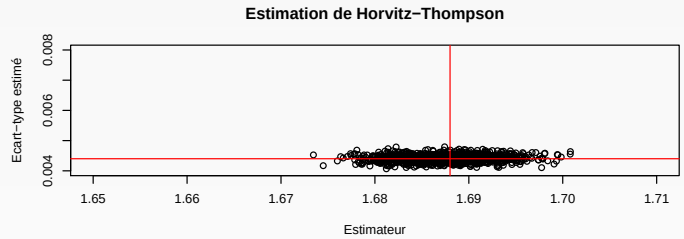
Les estimateurs de totaux sont non biaisés

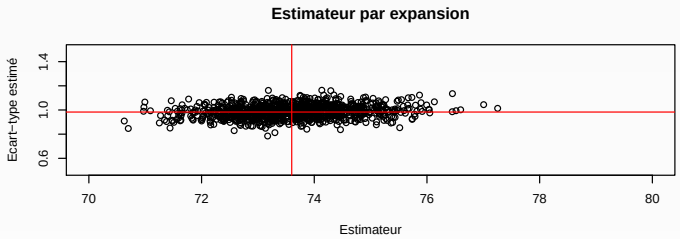


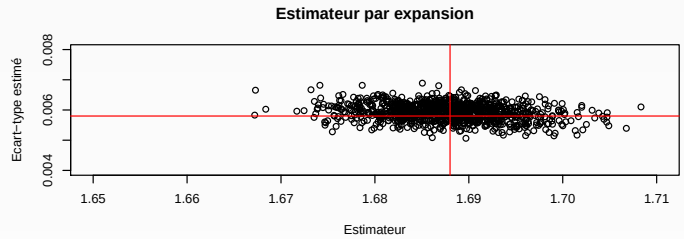
Ecart-type estimé en fonction de l'estimateur

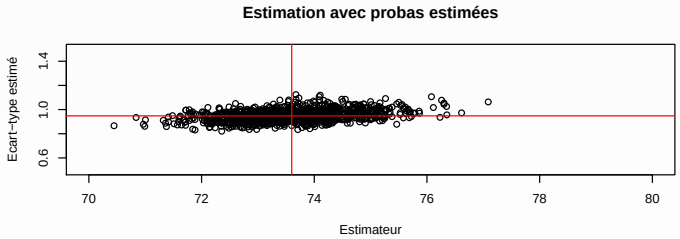
La non-réponse augmente la variance

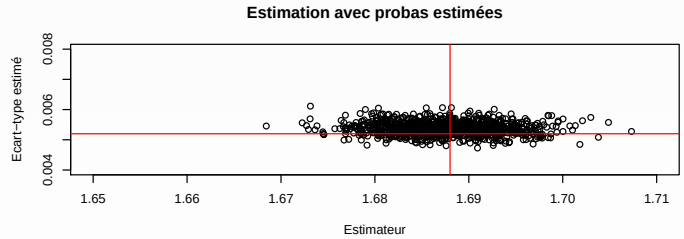












Traitement de la non-réponse partielle

Contexte

Type de non-réponse

Dans le contexte des enquêtes, nous distinguons :

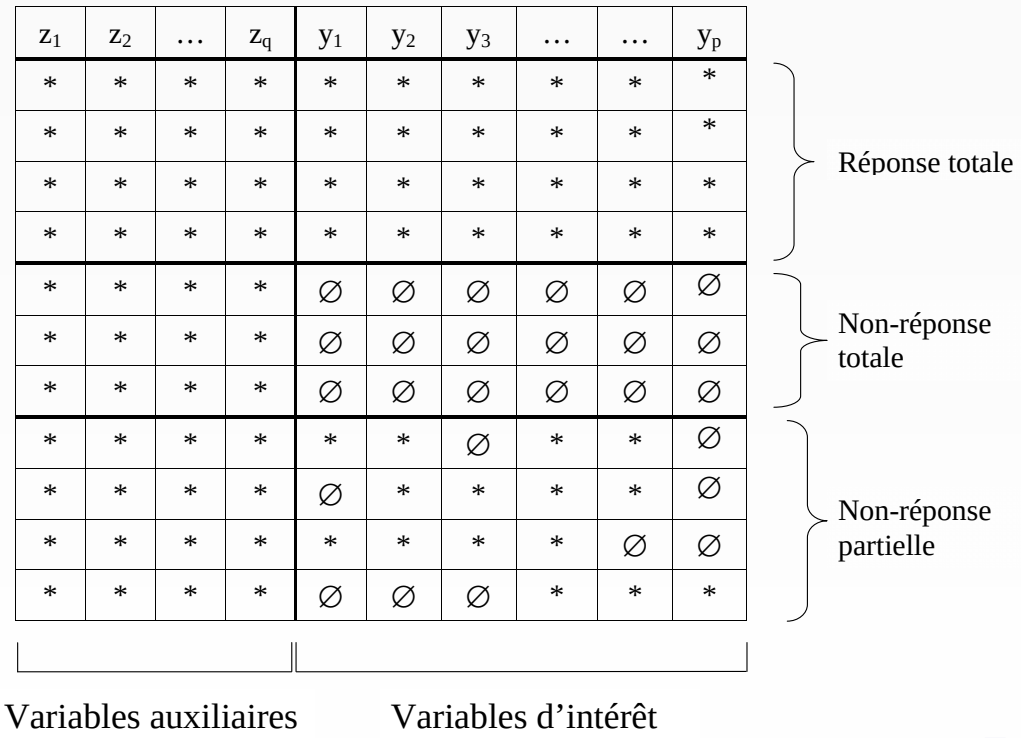
- la non-réponse totale ("unit non-response") : aucune information n'est relevée pour une unité,
- la non-réponse partielle ("item non-response") : une partie seulement de l'information est relevée pour une unité.

y_1	y_2	y_3	y_4	…	…	…	…	…	y_p	
*	*	*	*	*	*	*	*	*	*	Réponse totale
*	*	*	*	*	*	*	*	*	*	
*	*	*	*	*	*	*	*	*	*	
*	*	*	*	*	*	*	*	*	*	
ϕ	ϕ	ϕ	ϕ	ϕ	ϕ	ϕ	ϕ	ϕ	ϕ	Non-réponse totale
ϕ	ϕ	ϕ	ϕ	ϕ	ϕ	ϕ	ϕ	ϕ	ϕ	
ϕ	ϕ	ϕ	ϕ	ϕ	ϕ	ϕ	ϕ	ϕ	ϕ	
*	*	ϕ	*	ϕ	*	ϕ	*	*	ϕ	Non-réponse partielle
ϕ	*	*	*	ϕ	*	ϕ	*	*	ϕ	
*	*	*	*	*	*	*	*	ϕ	ϕ	
ϕ	ϕ	ϕ	*	*	ϕ	*	*	*	*	

Type de non-réponse

La correction de la non-réponse passe par la connaissance d'**information auxiliaire** connue sur l'ensemble de l'échantillon S, et qui soit

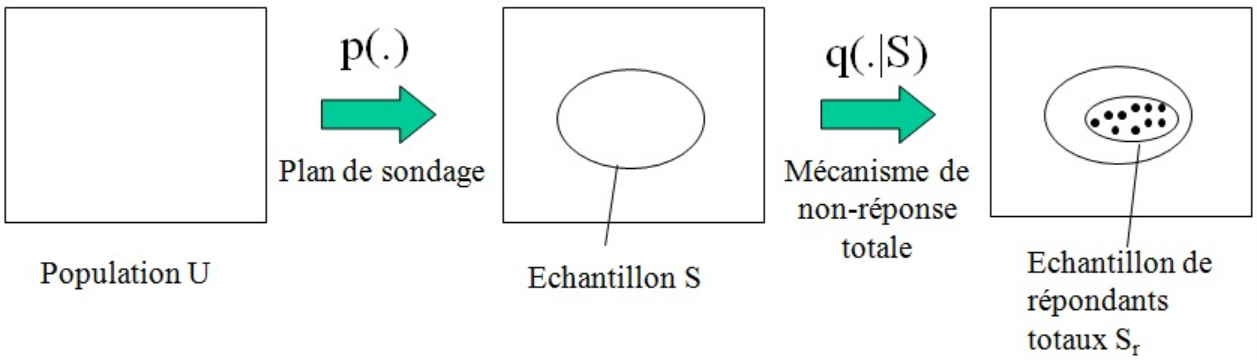
- explicative de la probabilité de répondre (traitement NR totale),
- et/ou explicative de la variable d'intérêt (traitement NR partielle).



Le problème

Nous traitons la NR partielle par imputation : une valeur manquante est remplacée par une valeur plausible. Cette imputation se justifie sous une modélisation de la variable d'intérêt : le **modèle d'imputation**.

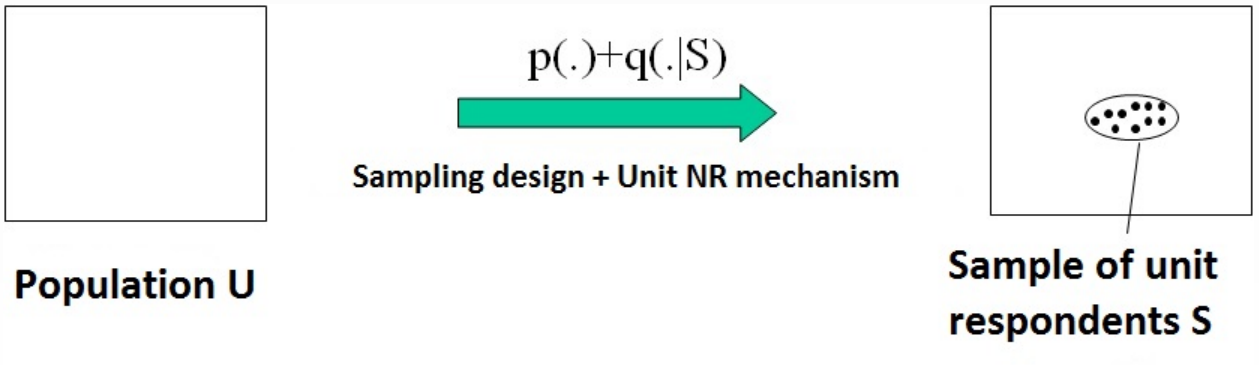
La correction de la NR partielle intervient généralement après la correction de la NR totale. Pour simplifier, nous notons S l'ensemble des répondants totaux, et d_k les poids redressés de la NR totale. L'échantillon S est donc le résultat d'un tirage en deux phases (plan de sondage+NR totale).



Le problème

Nous traitons la NR partielle par imputation : une valeur manquante est remplacée par une valeur plausible. Cette imputation se justifie sous une modélisation de la variable d'intérêt : le **modèle d'imputation**.

La correction de la NR partielle intervient généralement après la correction de la NR totale. Pour simplifier, nous notons S l'ensemble des répondants totaux, et d_k les poids redressés de la NR totale. L'échantillon S est donc le résultat d'un tirage en deux phases (plan de sondage+NR totale).



Objectifs

En raison de la NR partielle, nous obtenons un fichier à trous. Objectifs de l'imputation :

- pouvoir utiliser l'ensemble de l'information collectée : ne travailler que sur les **cas complets** peut conduire à travailler avec une taille d'échantillon fortement diminuée,
- obtenir un **fichier de données rectangulaire** : difficulté des logiciels statistiques à traiter les fichiers à trous,
- **limiter le biais de non-réponse** (différence de profil entre répondants et non-répondants),
- **éviter de perturber l'analyse du fichier de données** : généralement difficile, et suppose de connaître au moment de l'imputation l'analyse qui va être faite en aval.

Remarques

L'imputation **ne va pas amener plus d'information que celle disponible sur les répondants** : elle ne crée pas d'information. L'imputation vise à obtenir un jeu de données exploitable en évitant de perturber les relations entre variables.

L'imputation recrée un jeu de données artificiellement complet. Elle peut donc donner une fausse impression de précision, si l'alea associé à la non-réponse et à l'imputation n'est pas pris en compte dans les calculs d'intervalles de confiance.

Les étapes du traitement de la NR partielle

1. Identification des valeurs manquantes,
2. Choix d'un modèle d'imputation,
3. Recherche des facteurs explicatifs de la variable d'intérêt,
4. Choix du mécanisme d'imputation,
5. Imputation des valeurs manquantes.

Identification des valeurs manquantes

Il faut en particulier :

- distinguer les non-répondants partiels des non-répondants totaux,
- distinguer la non-réponse partielle des valeurs manquantes dues à la forme du questionnaire.

Point 1 : l'imputation ne concerne que les individus qui ont répondu globalement à l'enquête (répondants totaux), mais pas spécifiquement à la variable d'intérêt y (non-répondant partiel). Les deux mécanismes de non-réponse sont généralement différents.

Point 2 : ne pas traiter par imputation l'absence d'une valeur y_k due à la forme du questionnaire (question filtre).

Choix d'un modèle d'imputation

Notations

Nous notons $p(\cdot)$ le mécanisme de sélection de l'échantillon S. En l'absence de non-réponse partielle pour la variable y, le total t_y est estimé approximativement sans biais par

$$\hat{t}_y = \sum_{k \in S} d_k y_k.$$

En situation de non-réponse partielle, deux mécanismes supplémentaires interviennent :

- le **mécanisme de réponse** à la variable y, noté $q(\cdot)$, avec p_k la probabilité que y_k soit renseigné ;
- le **mécanisme d'imputation**, noté I, qui remplace une valeur manquante y_k par une valeur artificielle y_k^*.

Nous notons

- $S_{ry} \equiv S_r$ le sous-échantillon d'individus ayant renseigné la variable y,
- $S_{my} \equiv S_m$ le sous-échantillon d'individus n'ayant pas renseigné la variable y.

Modèle d'imputation

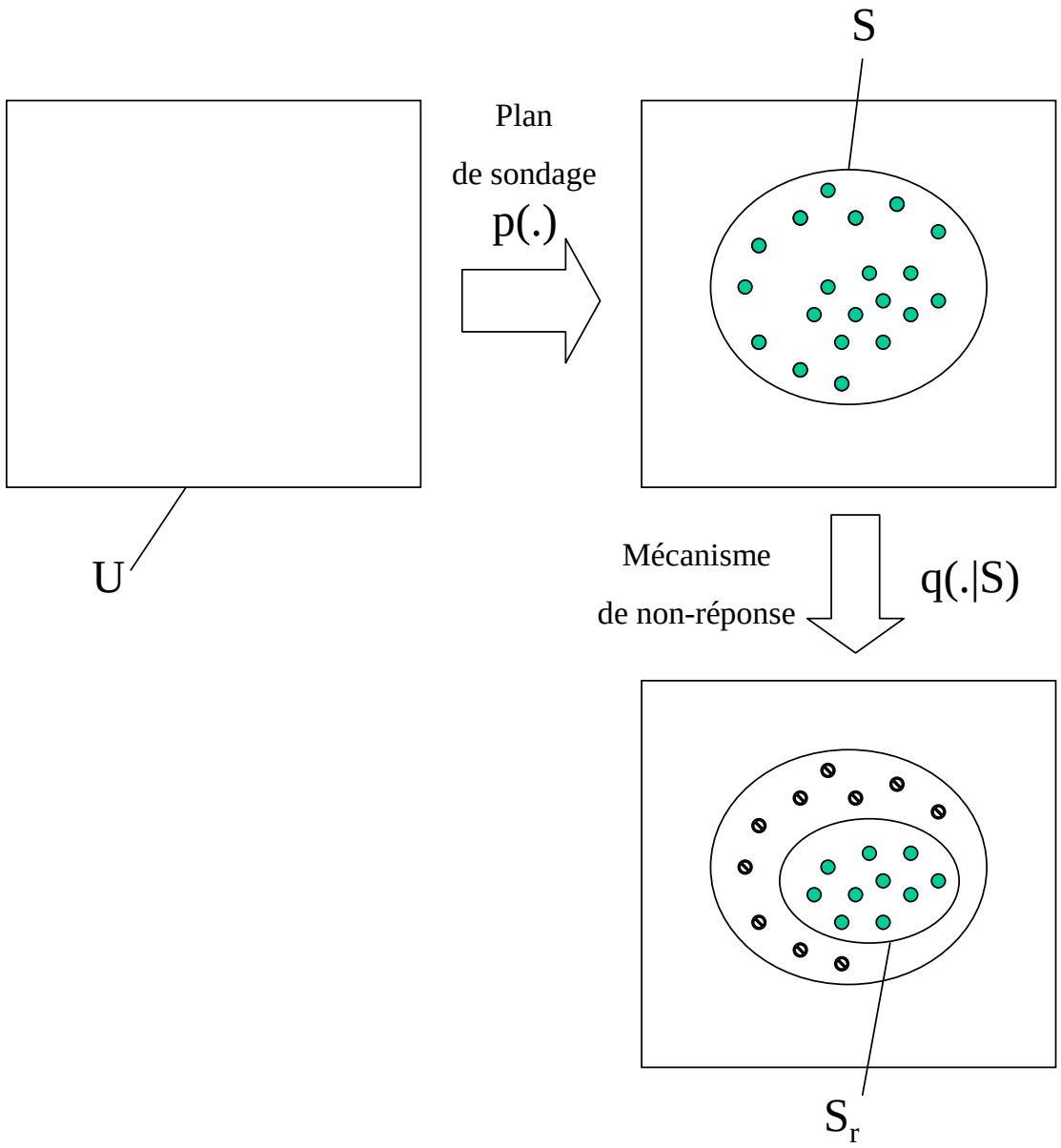
Le **mécanisme d'imputation** est motivé par un **modèle d'imputation** (par exemple, un modèle de régression) qui vise à prédire la variable y_k à l'aide d'une information auxiliaire $\mathbf{z}_k$ disponible sur l'ensemble de l'échantillon.

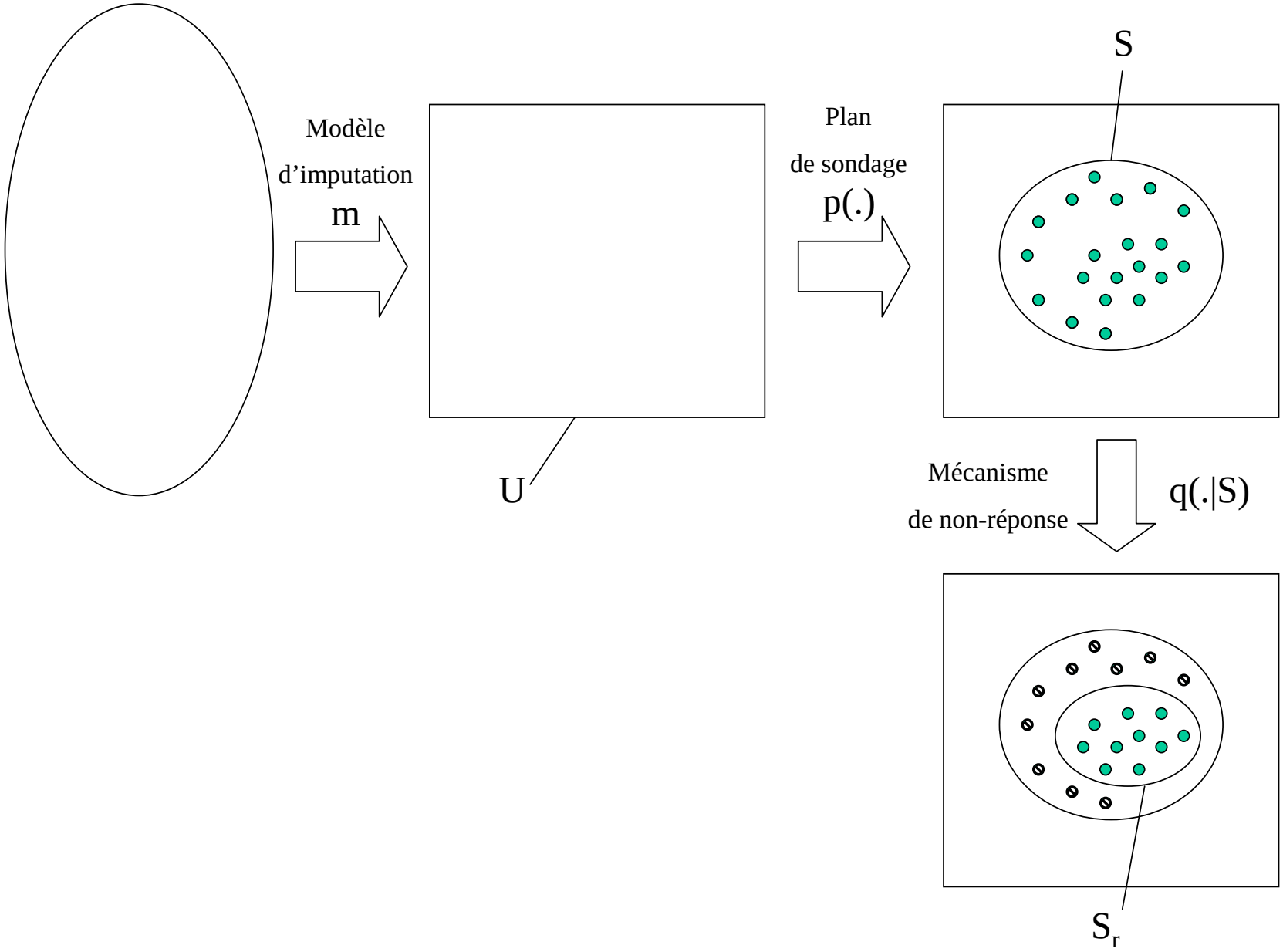
$$m \ : \ y_k \ = \ \mathbf{z}_k^\top \beta + \sigma \sqrt{v_k} \epsilon_k \quad \text{pour} \quad k \in S. \tag{20}$$

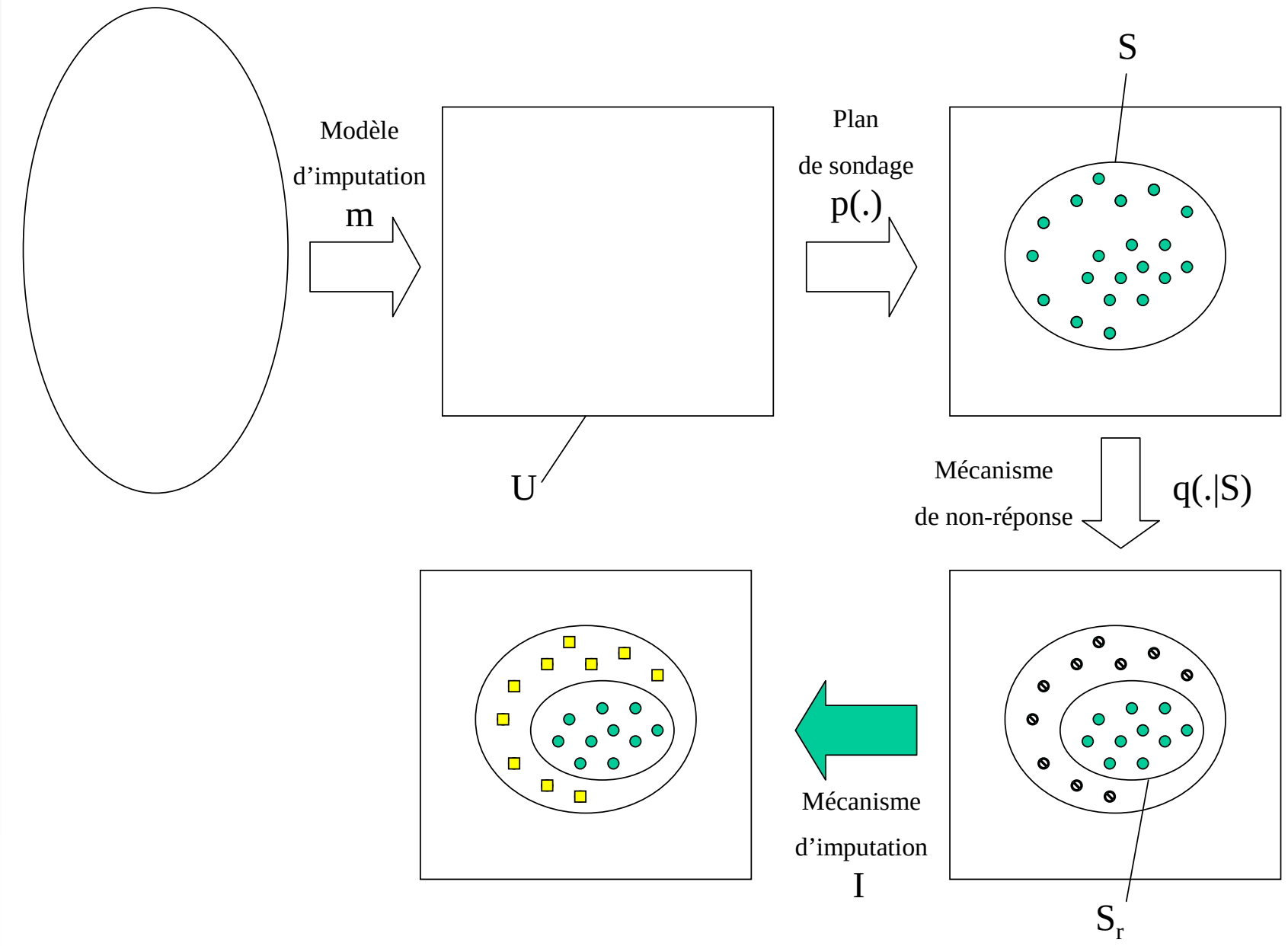
Dans ce modèle :

- β et σ^2 sont des paramètres inconnus,
- v_k est une constante connue,
- les résidus ϵ_k sont des variables aléatoires iid, centrées réduites.

Le modèle d'imputation utilisé doit être adapté au type de variable traité. Le mécanisme d'imputation doit être adapté à l'analyse que l'on souhaite réaliser sur l'échantillon.


S
Plan
de sondage
p(.)
U
Mécanisme
de non-réponse
q(.|S)
S_r


Modèle
d'imputation
m
U
Plan
de sondage
p(.)
S
Mécanisme
de non-réponse
q(.|S)
S_r


Modèle
d'imputation
m
U
Plan
de sondage
p(.)
S
Mécanisme
de non-réponse
q(.|S)
Mécanisme
d'imputation
I
S_r

Exemples de modèles d'imputation

Exemple 1 : avec $\mathbf{z_k} = z_k = 1$ et $v_k = 1$, modèle constant

$$m \ : \ y_k \ = \ \beta + \sigma\epsilon_k \quad \text{pour} \quad k \in S. \tag{21}$$

Exemple 2 : avec $\mathbf{z_k} = [1(k \in S_1), \ldots, 1(k \in S_H)]^\top$ et $v_k = v_h$ pour $k \in S_h$, modèle constant par classes

$$m \ : \ y_k \ = \ \beta_h + \sigma_h\epsilon_k \quad \text{pour} \quad k \in S_h. \tag{22}$$

Exemple 3 : avec $\mathbf{z_k} = z_k$ et $v_k = z_k$, modèle ratio

$$m \ : \ y_k \ = \ \beta z_k + \sigma\sqrt{z_k}\epsilon_k \quad \text{pour} \quad k \in S. \tag{23}$$

Les deux derniers modèles sont couramment utilisés en pratique.

Choix d'un mécanisme d'imputation

Types de méthodes

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

- les **méthodes déterministes** : elles conduisent à la même valeur imputée si le mécanisme d'imputation est répété,
- les **méthodes aléatoires** : la valeur imputée inclut une composante aléatoire, et peut donc changer si le mécanisme d'imputation est répété.

Il existe une troisième famille de méthodes, transversale : celle des **méthodes d'imputation par donneur**, qui consistent à piocher un individu parmi les répondants, et à utiliser la valeur observée pour la variable y pour remplacer la valeur manquante.

Mécanismes d'imputation déterministes

Imputation par la régression déterministe

Ce mécanisme s'appuie sur le modèle (20) :

$$\begin{aligned} m &: y_k = \mathbf{z}_k^\top \beta + \sigma\sqrt{v_k}\epsilon_k \\ \Rightarrow I &: y_k^* = \mathbf{z}_k^\top \hat{\beta}_r \quad \text{pour} \quad k \in S_m, \end{aligned}$$

avec

$$\hat{\beta}_r = \left(\sum_{k \in S_r} \omega_k v_k^{-1} \mathbf{z}_k \mathbf{z}_k^\top \right)^{-1} \sum_{k \in S_r} \omega_k v_k^{-1} \mathbf{z}_k y_k,$$

où ω_k désigne un **poids d'imputation** attaché à l'unité k (Haziza, 2009). Nous nous limiterons au cas $\omega_k = 1$, qui conduit à l'estimateur des MCG sous le modèle m.

Dans le cas d'un total, l'estimateur imputé est égal à

$$\hat{t}_{yI} = \sum_{k \in S_r} d_k y_k + \sum_{k \in S_m} d_k \left[\mathbf{z}_k^\top \hat{\beta}_r \right].$$

Imputation par la régression déterministe (2)

Vérifions que l'estimateur imputé $\hat{t}_{yI}$ est sans biais sous le modèle. Nous supposons que le plan de sondage et le mécanisme de réponse sont non informatifs, i.e. ne dépendent pas directement de la variable d'intérêt y (Deville et Särndal, 1994). Alors :

$$\begin{aligned}
E(\hat{t}_{yI} - t_y) &= E_m E_p E_q \left(\hat{t}_{yI} - \hat{t}_{y\pi}\right) + E_m E_p E_q \left(\hat{t}_{y\pi} - t_y\right) \\
&= E_m E_p E_q \left(\hat{t}_{yI} - \hat{t}_{y\pi}\right) + E_m \underbrace{E_p \left(\hat{t}_{y\pi} - t_y\right)}_{=0} \\
&= E_m E_p E_q \{ \sum_{k \in S_m} d_k (\mathbf{z_k}^\top \hat{\beta}_r - y_k) \} \\
&= E_p E_q E_m \{ \sum_{k \in S_m} d_k (\mathbf{z_k}^\top \hat{\beta}_r - y_k) \} \text{ (non info.)}.
\end{aligned}$$

Sous le modèle m, nous avons $E_m(\mathbf{z_k}^\top \hat{\beta}_r) = E_m(y_k) = \mathbf{z_k}^\top \beta$, ce qui donne le résultat.

Imputation par la moyenne

L'**imputation par la moyenne** est un cas particulier d'imputation par la régression. Elle s'appuie sur le modèle (21) :

$$m \; : \; y_k \quad = \quad \beta + \sigma \epsilon_k \quad \text{pour} \quad k \in S.$$

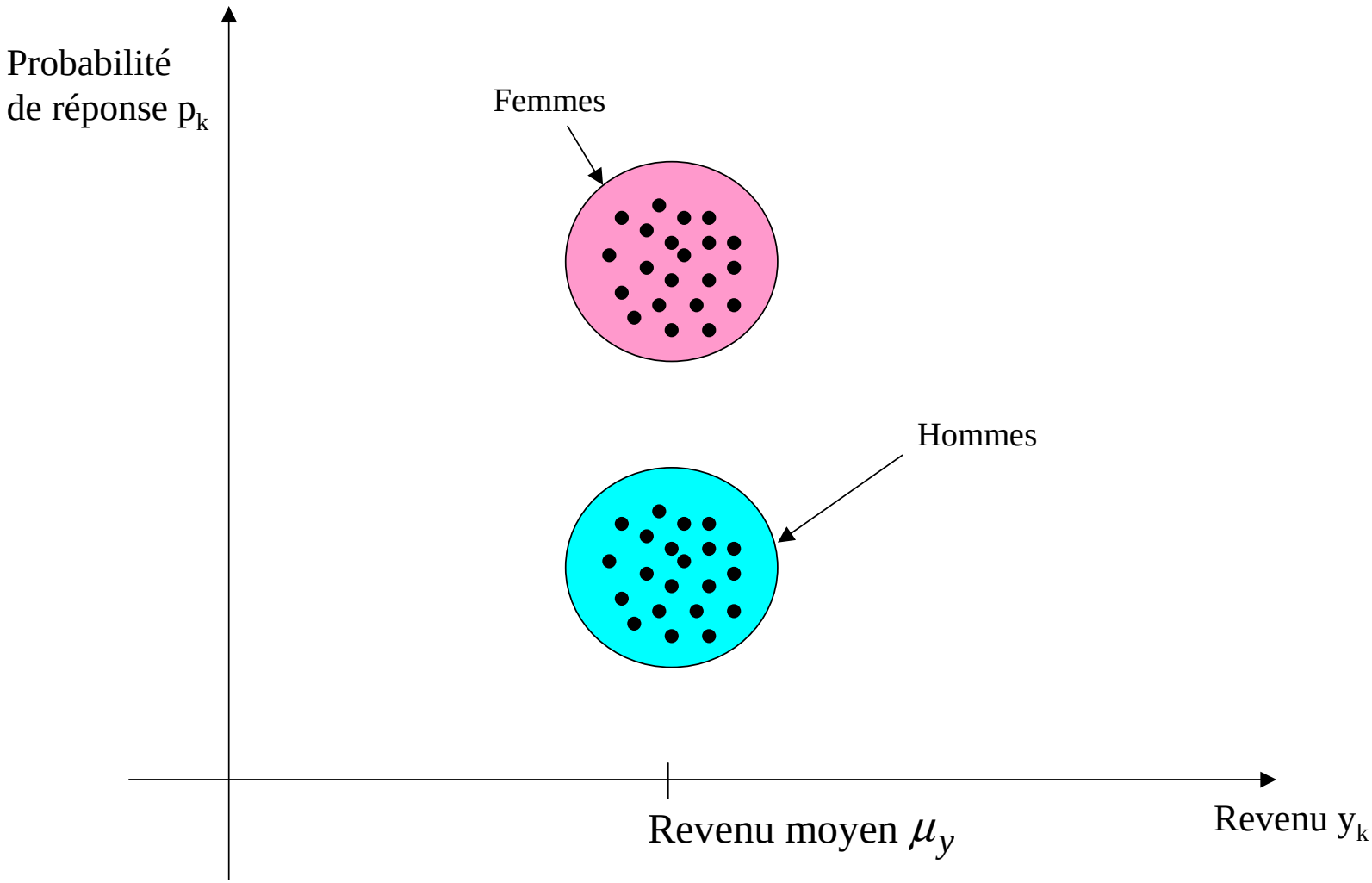
obtenu avec $\mathbf{z}_k = z_k = 1$ et $v_k = 1$. Nous obtenons l'estimateur

$$\hat{\boldsymbol{\beta}}_r \quad = \quad \frac{1}{n_r} \sum_{k \in S_r} y_k \equiv \bar{y}_r.$$

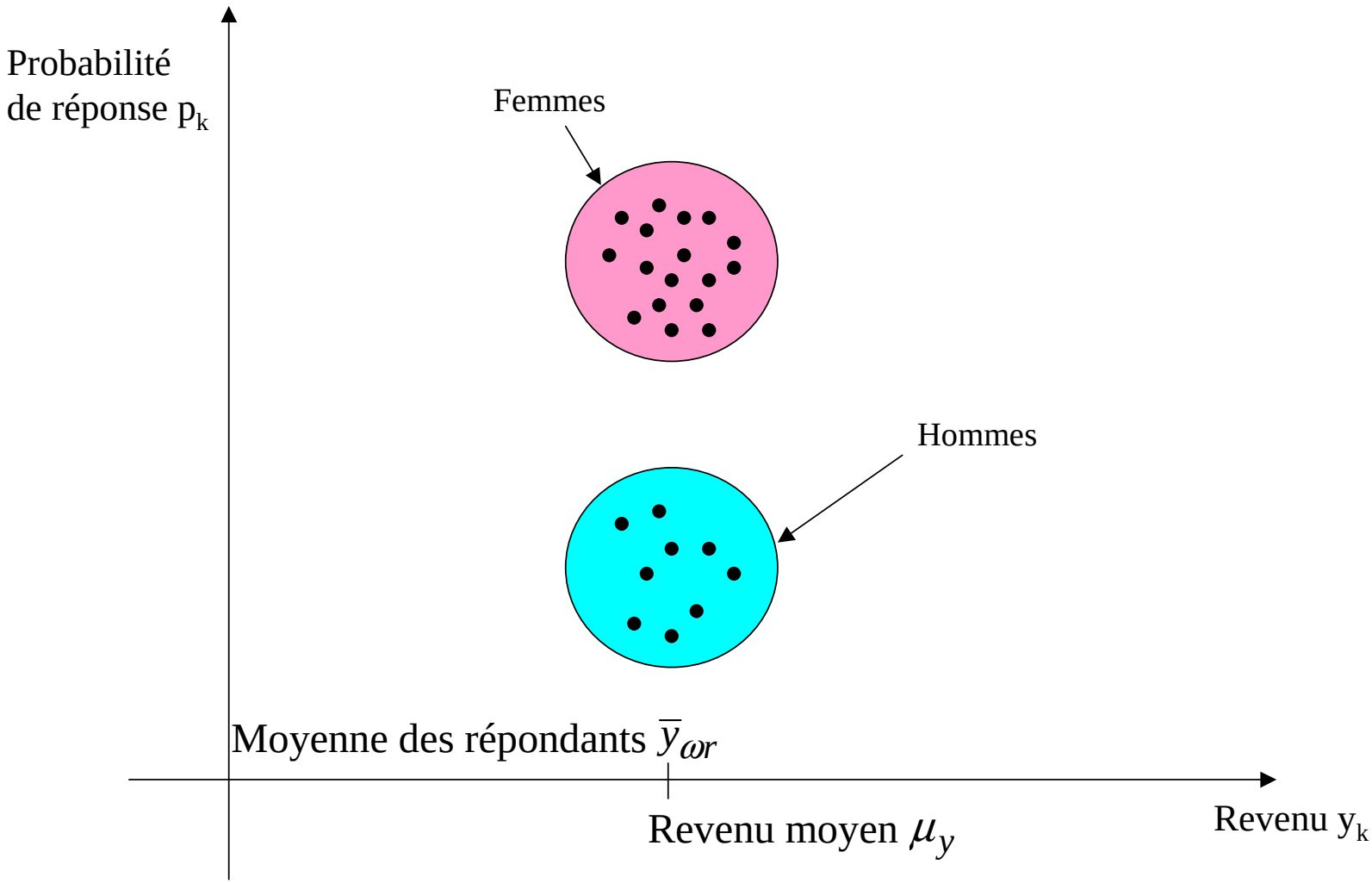
Dans le cas d'un total, l'estimateur imputé est égal à

$$\hat{t}_{yI} \quad = \quad \sum_{k \in S_r} d_k y_k + \sum_{k \in S_m} d_k \bar{y}_r.$$

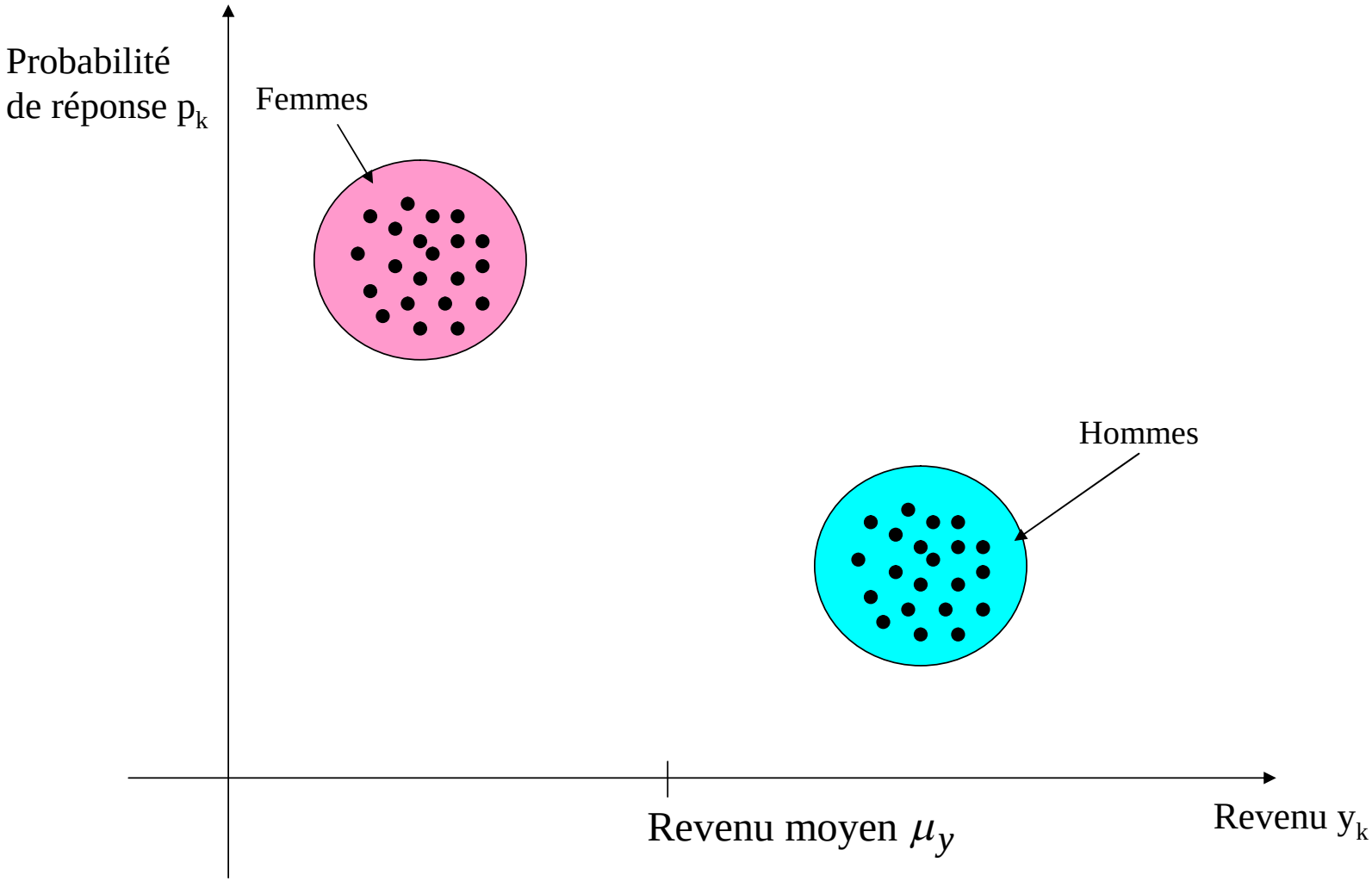
Cas favorable



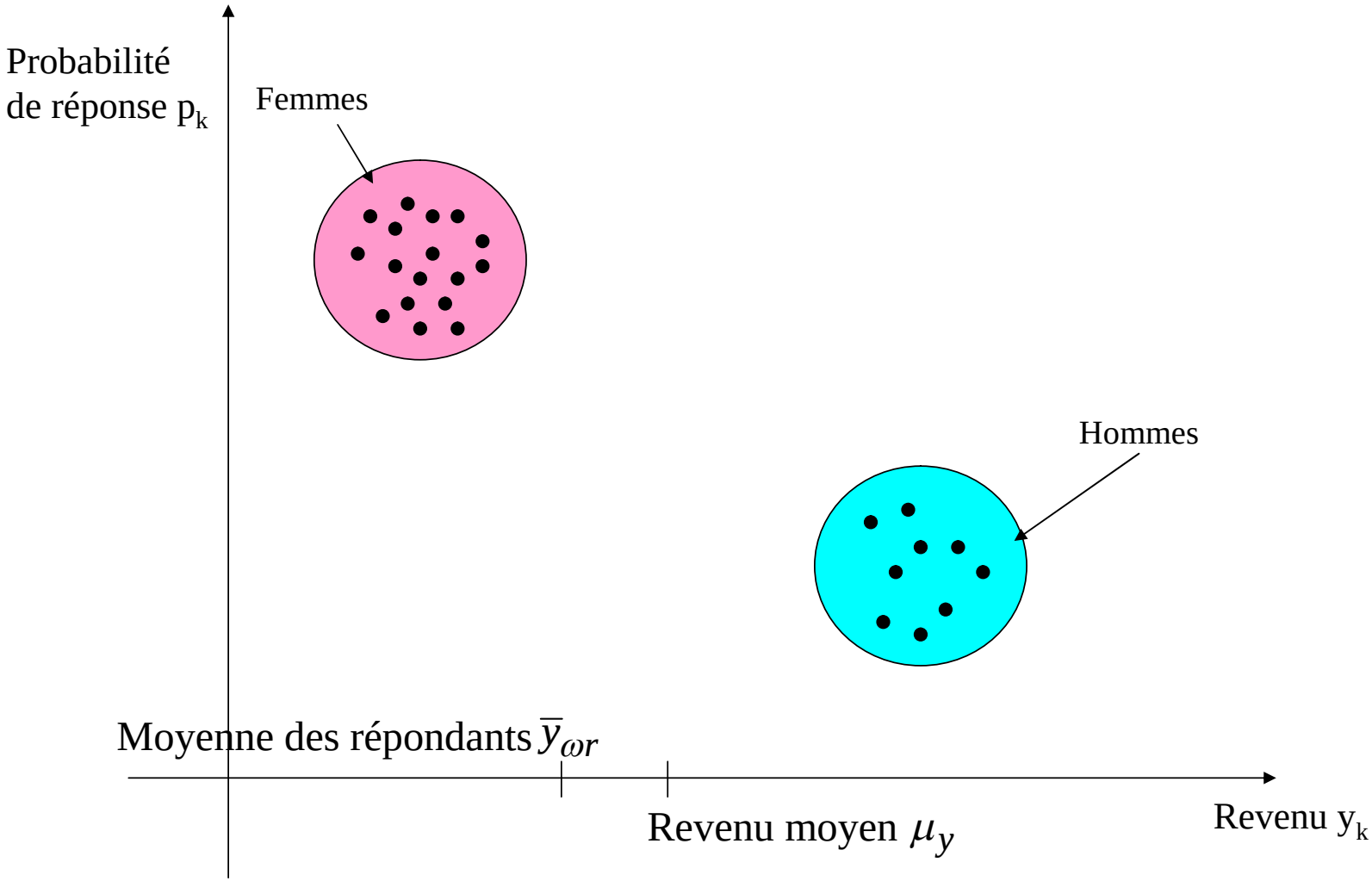
Cas favorable (suite)



Cas défavorable



Cas défavorable (suite)

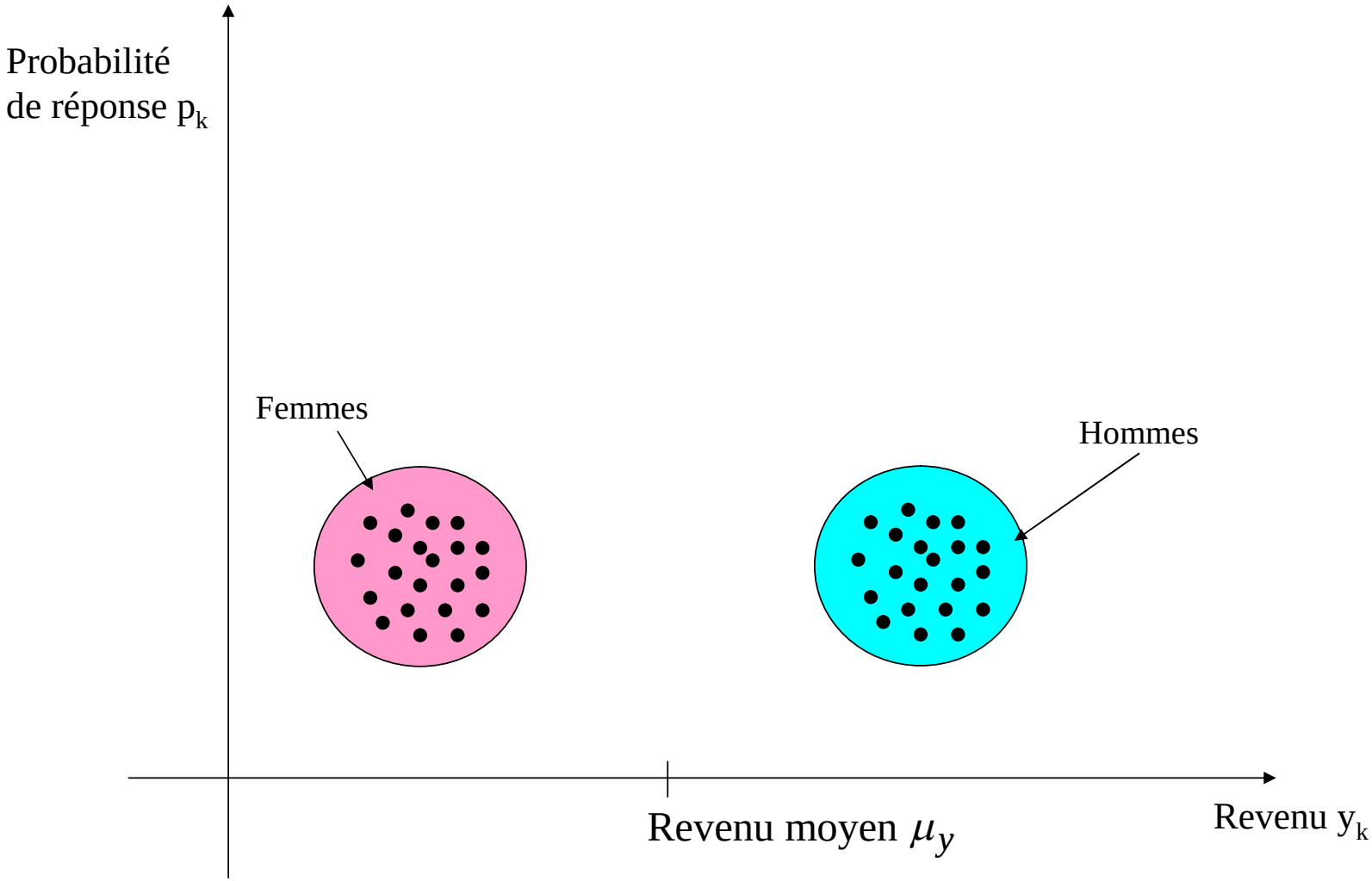


Imputation par la moyenne

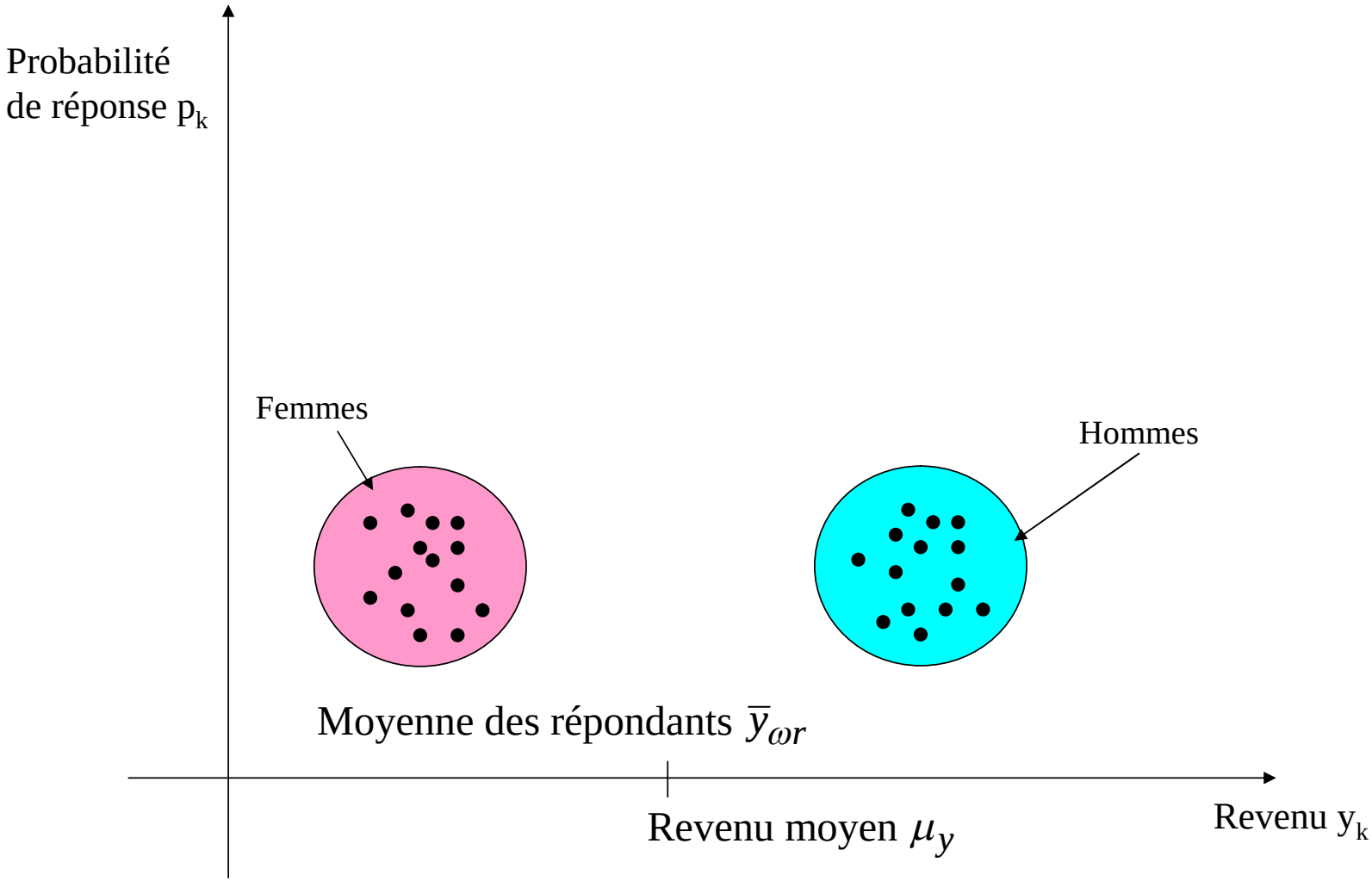
Compte-tenu du modèle d'imputation utilisé, l'imputation par la moyenne conduit à une estimation approximativement non biaisée du total **si tous les individus de l'échantillon sont peu différents par rapport à la variable d'intérêt**.

L'imputation par la moyenne conduira également à une estimation (approximativement) non biaisée si le comportement moyen des individus de S_r ne diffère pas du comportement moyen des individus de S, par rapport à la variable y. Ce sera le cas en particulier si les probabilités de réponse sont voisines.

Autre cas favorable



Autre cas favorable (suite)



Imputation par la moyenne dans des classes

Ces deux hypothèses (variable y_k approximativement constante ou probabilité de réponse p_k approximativement constante) sont rarement vérifiées sur l'ensemble de l'échantillon.

En revanche, il est possible de partitionner l'échantillon en classes $S_1, \ldots, S_H$ de façon à ce que au sein de chaque classe les individus soient :

- peu différents par rapport à y (même logique que pour la stratification),
- et/ou peu différents par rapport aux probabilités de réponse (même logique que pour les GHR).

Nous pouvons alors imputer par la moyenne au sein de chaque classe.

Imputation par la moyenne dans des classes

Nous obtenons une **imputation par la moyenne dans les classes d'imputation**. Cette méthode s'appuie sur le modèle (22)

$$m \ : \ y_k \ = \ \beta_h + \sigma_h \epsilon_k \quad \text{pour} \quad k \in S_h.$$

Exemple : imputation de la variable revenu par la moyenne, dans des classes définies selon le sexe.

Pour un individu k non-répondant de la classe S_h, nous obtenons $y_k^* = \hat{\beta}_{rh}$ avec

$$\hat{\beta}_{rh} \ = \ \frac{1}{n_{rh}} \sum_{k \in S_{rh}} y_k \equiv \bar{y}_{rh},$$

en notant $S_{rh} = S_h \cap S_r$, de taille n_{rh}.

Construction des classes d'imputation

Il est possible de constituer les classes de la façon suivante :

1. soit en modélisant la variable y :
   - Régression linéaire afin d'obtenir une prédiction $\hat{y}_k$ de y_k, en fonction de l'information auxiliaire disponible.
   - Les classes d'imputation sont obtenues en ordonnant les individus selon les $\hat{y}_k$, ou en croisant les variables qui resortent de façon significative.
2. soit en modélisant la probabilité de réponse à la variable y :
   - Régression logistique afin d'obtenir une prédiction des probabilités de réponse $\hat{p}_{yk}$.
   - Les classes d'imputation sont obtenues en ordonnant les individus selon les $\hat{p}_{yk}$, ou en croisant les variables qui resortent de façon significative.

Il est également possible de définir les classes d'imputation par croisement des variables explicatives de y_k et de la probabilité de réponse.

Mécanismes d'imputation aléatoires

Mécanisme d'imputation par la régression aléatoire

L'**imputation par la régression aléatoire** s'appuie sur le modèle (20) :

$$\begin{aligned} m &: y_k = \mathbf{z}_k^\top \beta + \sigma\sqrt{v_k}\epsilon_k \\ \Rightarrow I &: y_k^* = \mathbf{z}_k^\top \hat{\beta}_r + \hat{\sigma}\sqrt{v_k}\epsilon_k^* \quad \text{pour} \quad k \in S_m, \end{aligned}$$

avec

$$\hat{\beta}_r = \left(\sum_{k \in S_r} v_k^{-1} \mathbf{z}_k \mathbf{z}_k^\top \right)^{-1} \sum_{k \in S_r} v_k^{-1} \mathbf{z}_k y_k.$$

Ajout au terme de prédiction $\mathbf{z}_k^\top \hat{\boldsymbol{\beta}}_r$ d'un terme aléatoire $\hat{\sigma}\sqrt{v_k}\epsilon_k^*$ t.q. :

- $\hat{\sigma}$ est un estimateur de σ,
- ϵ_k^* est un **résidu aléatoire** centré réduit.

L'estimateur imputé du total est égal à

$$\hat{t}_{yI} = \sum_{k \in S_r} d_k y_k + \sum_{k \in S_m} d_k \left\{ \mathbf{z}_k^\top \hat{\beta}_r + \hat{\sigma}\sqrt{v_k}\epsilon_k^* \right\}.$$

Mécanisme d'imputation par la régression aléatoire (2)

Par rapport à l'imputation déterministe, l'ajout d'un résidu aléatoire permet de mieux respecter la distribution de y_k, en évitant un trop bon ajustement avec les $\mathbf{z}_k$ (R^2 surévalué).
En contrepartie, la variance augmente en raison de l'alea d'imputation.

Plusieurs méthodes sont possibles pour générer les résidus ϵ_k^* :

1. Postuler une distribution centrée réduite $\mathcal{L}(0,1)$ pour les résidus du modèle ϵ_k, et générer les résidus ϵ_k^* selon la même distribution lors de l'imputation. Un travail de modélisation est nécessaire, une approximation gaussienne n'étant pas forcément appropriée (e.g., variable de revenu).
2. Tirer le résidu aléatoire ϵ_k^* au hasard (et généralement avec remise) parmi les résidus estimés

$$e_k = \frac{y_k - \mathbf{z}_k^\top \hat{\beta}_r}{\hat{\sigma}\sqrt{v_k}} \text{ pour } k \in S_r.$$

Imputation par la régression aléatoire (3)

Vérifions que l'estimateur imputé $\hat{t}_{yI}$ est sans biais sous le modèle. Nous avons

$$\begin{aligned} E\left(\hat{t}_{yI}^{ra} - t_y\right) &= E_m E_p E_q E_I \left(\hat{t}_{yI}^{rd} - t_y\right) + E_m E_p E_q E_I \left(\hat{\sigma} \sum_{k \in S_m} d_k \sqrt{v_k} \epsilon_k^*\right) \\ &= \underbrace{E_m E_p E_q \left(\hat{t}_{yI}^{rd} - t_y\right)}_{=0} + E_m E_p E_q \left\{\hat{\sigma} \sum_{k \in S_m} d_k \sqrt{v_k} \underbrace{E_I(\epsilon_k^*)}_{=0}\right\} \\ &= 0. \end{aligned}$$

Nous avons également

$$V(\hat{t}_{yI}^{ra} - t_y) = V(\hat{t}_{yI}^{rd} - t_y) + E_m E_p E_q V_I \left(\hat{\sigma} \sum_{k \in S_m} d_k \sqrt{v_k} \epsilon_k^*\right).$$

La variance est donc logiquement plus grande que sous une imputation par la régression déterministe.

Imputation par hot-deck

L'**imputation par hot-deck** est un cas particulier d'imputation par la régression aléatoire. Elle s'appuie sur le modèle (21) :

$$m \ : \ y_k \ = \ \beta + \sigma\epsilon_k \quad \text{pour} \quad k \in S.$$

La méthode du hot-deck consiste à remplacer une valeur manquante y_k en sélectionnant au hasard et avec remise un donneur $y_j \in S_r$, à probabilités égales. Nous obtenons l'estimateur :

$$\hat{t}_{yI} \ = \ \sum_{k \in S_r} d_k y_k + \sum_{k \in S_m} d_k y_k^*.$$

Imputation par hot-deck

C'est la version aléatoire de l'imputation par la moyenne. Elle s'appuie sur le même modèle d'imputation : nous supposons que les individus de la population ont en moyenne le même comportement par rapport à la variable y.

Le hot-deck a l'avantage d'aller chercher une valeur effectivement observée : en particulier, la méthode est applicable pour une variable catégorielle.

En revanche, il s'agit d'une méthode d'**imputation aléatoire** : elle conduit donc à une augmentation de la variance.

Imputation par hot-deck dans des classes

Comme l'imputation par la moyenne, l'imputation par hot-deck est généralement réalisée au sein de classes d'imputation : une valeur manquante y_k est remplacée en sélectionnant au hasard un donneur parmi les répondants de la même classe.

Ce mécanisme d'imputation s'appuie sur le modèle (22) :

$$m \ : \ y_k \ = \ \beta_h + \sigma_h \epsilon_k \quad \text{pour} \quad k \in S.$$

Comme dans le cas d'une imputation par la moyenne dans des classes, l'estimateur imputé sera approximativement non biaisé :

- si les individus d'une même classe sont peu différents par rapport à y ;
- et/ou si les probabilités de réponse sont voisines au sein d'une même classe.

Méthodes d'imputation par donneur

Imputation par donneur

Le hot-deck est un cas particulier des méthodes d'imputation par donneur. Il est également possible d'utiliser :

- l'**imputation par la valeur précédente** : une valeur manquante $y_{k,t}$ est remplacée par la valeur observée à une date précédente $y_{k,t-1}$,
  $\Rightarrow$ efficace si la variable mesurée évolue peu dans le temps,
- l'**imputation par le plus proche voisin** : une valeur manquante y_k est remplacée en choisissant le donneur le plus proche du non-répondant k, au sens d'une fonction de distance à définir (en fonction des variables auxiliaires disponibles)

Les méthodes par donneurs ont l'avantage

- d'imputer des valeurs effectivement observées,
- de pouvoir être utilisées pour les variables catégorielles,
- de permettre d'imputer plusieurs variables à la fois (aide à préserver le lien entre les variables).

Quelle méthode d'imputation utiliser ?

Pour des paramètres descriptifs (totaux, ratios), les méthodes d'imputation déterministes sont préférables car elles ne conduisent pas à une augmentation de la variance.
Pour une analyse du fichier de données d'enquête (modèle linéaire, modèle linéaire généralisé), les méthodes d'imputation aléatoires sont préférables car elles préservent mieux les distributions.

Comme une seule méthode peut être utilisée par variable, une bonne pratique consiste à inclure dans le fichier des données d'enquête des "imputation flags", indiquant a minima les valeurs réelles et les valeurs imputées.
Eurostat recommande d'accompagner les données d'enquête de métadonnées explicitant par exemple les méthodes d'imputation utilisées ou le volume de données manquantes.

Estimation de paramètres après imputation

Objectifs

Etudier dans le cadre de données simulées les conséquences de la non-réponse sur l'estimation d'un paramètre univarié (moyenne) ou multivarié (ajustement d'une régression).

Etudier les conséquences de l'imputation sur :

- le biais des estimateurs,
- la variance des estimateurs,
- la préservation des relations entre les variables.

Le cadre

Nous considérons une population artificielle de taille $N = 10,000$ contenant deux variables x et y. La variable x a été générée selon une loi Gamma$(2, 5)$. La variable y est générée selon le modèle

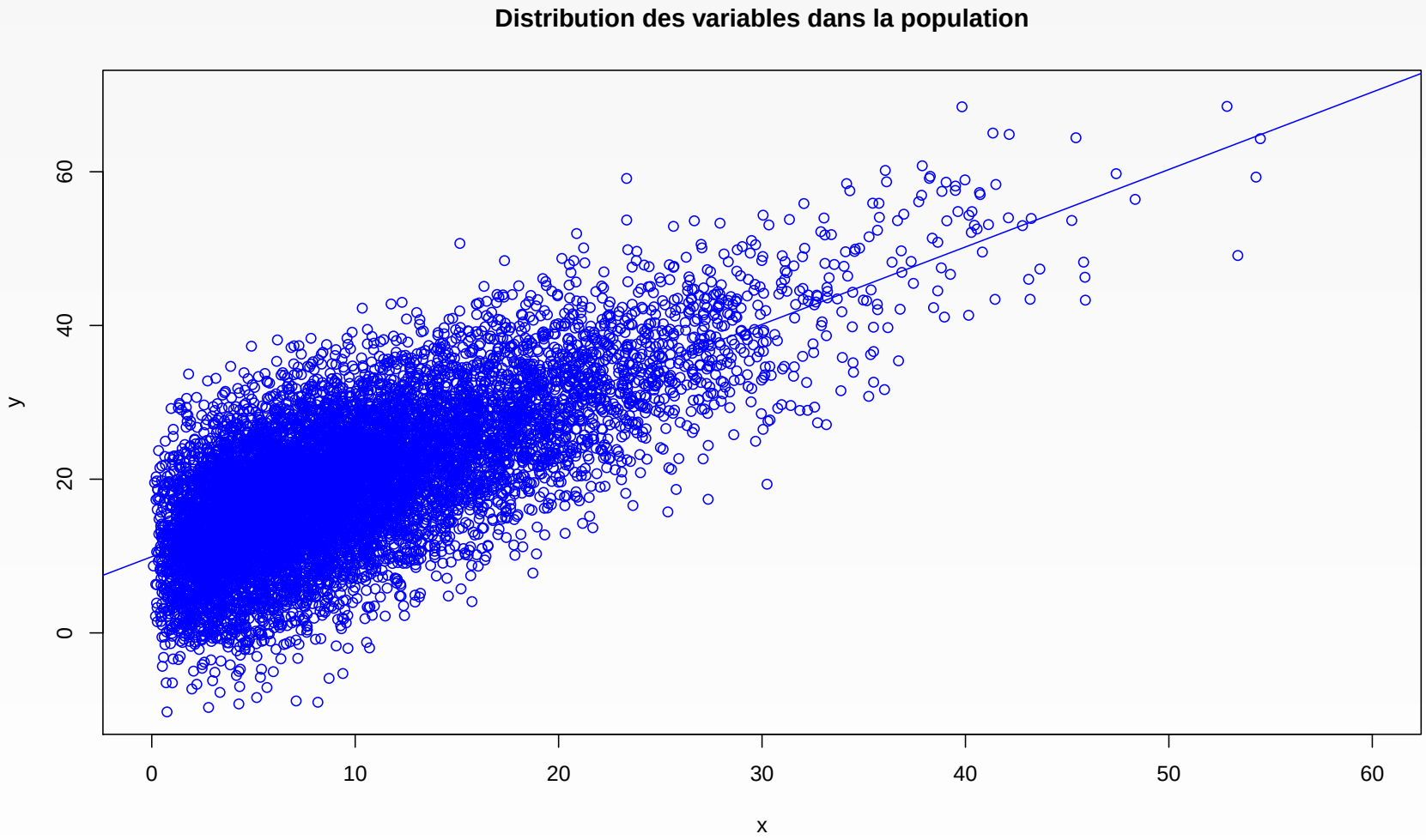
$$y_k = \beta_0 + \beta_1 \, x_k + \epsilon_k,$$

avec les ϵ_k générés selon une loi Normale$(0, \sigma^2)$.

Le R^2 du modèle est égal à 0.5. Les paramètres d'intérêt sont :

- le vecteur des coefficients de régression $\beta = (\beta_0, \beta_1) = (10, 1)$,
- la moyenne $\mu_y = 19.98$.

Les données


Distribution des variables dans la population
y
x
0
20
40
60
0
10
20
30
40
50
60

Estimation sur données non imputées

Estimation en situation de réponse complète

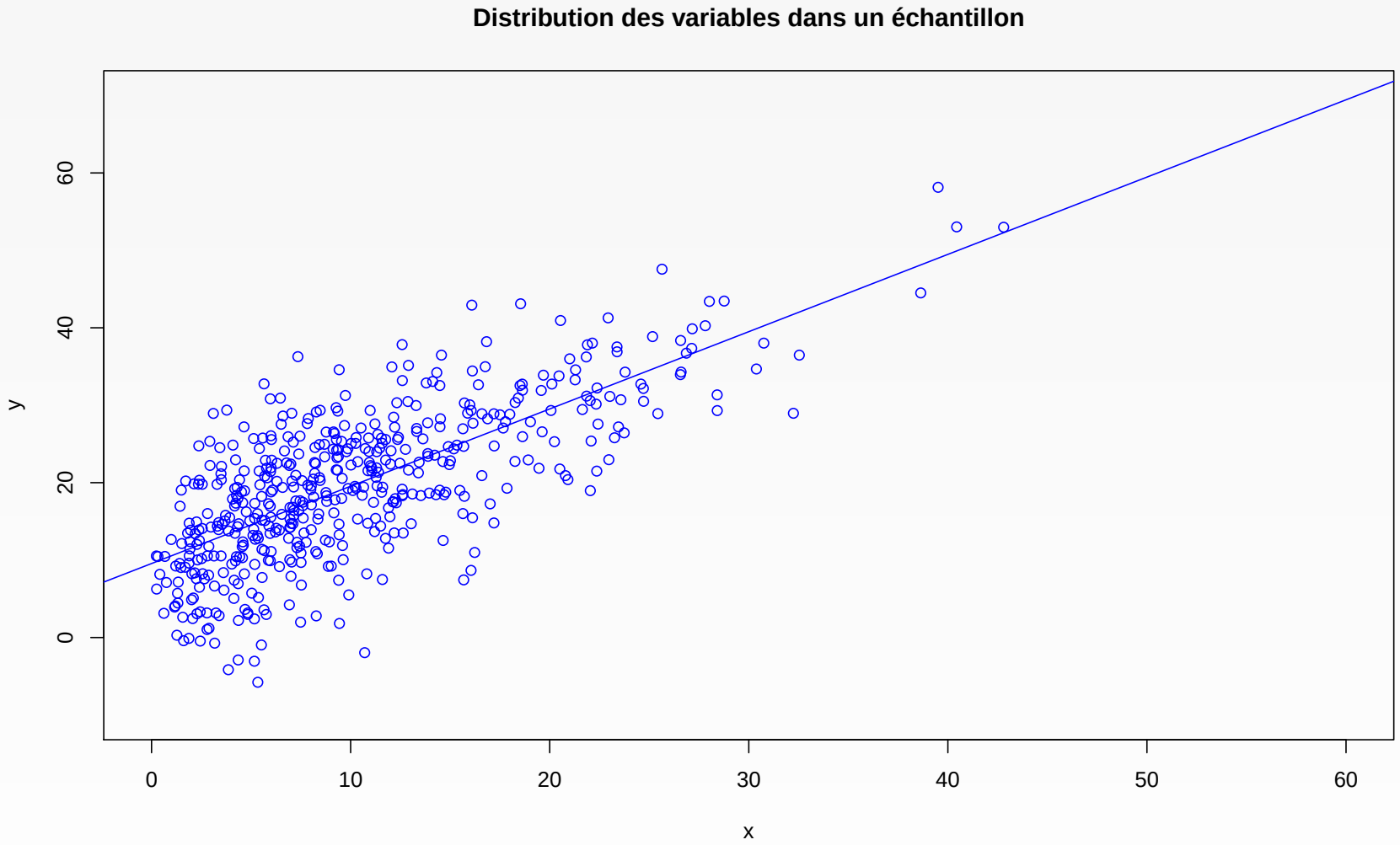
Nous sélectionnons un échantillon S de taille $n = 500$ selon un SRS. La moyenne μ_y peut être estimée sans biais par

$$\bar{y} = \frac{1}{n}\sum_{k\in S} y_k.$$

Le vecteur β est estimé approximativement sans biais par

$$\begin{aligned}\hat{\beta}_\pi &= \left(\sum_{k\in S} d_k \mathbf{x}_k \mathbf{x}_k^\top\right)^{-1} \sum_{k\in S} d_k \mathbf{x}_k y_k \\ &= \left(\sum_{k\in S} \mathbf{x}_k \mathbf{x}_k^\top\right)^{-1} \sum_{k\in S} \mathbf{x}_k y_k\end{aligned}$$

avec $\mathbf{x}_k = (1, x_k)^\top$ et $d_k = N/n$ le poids de sondage.


Distribution des variables dans un échantillon
y
x
0
20
40
60
0
10
20
30
40
50
60

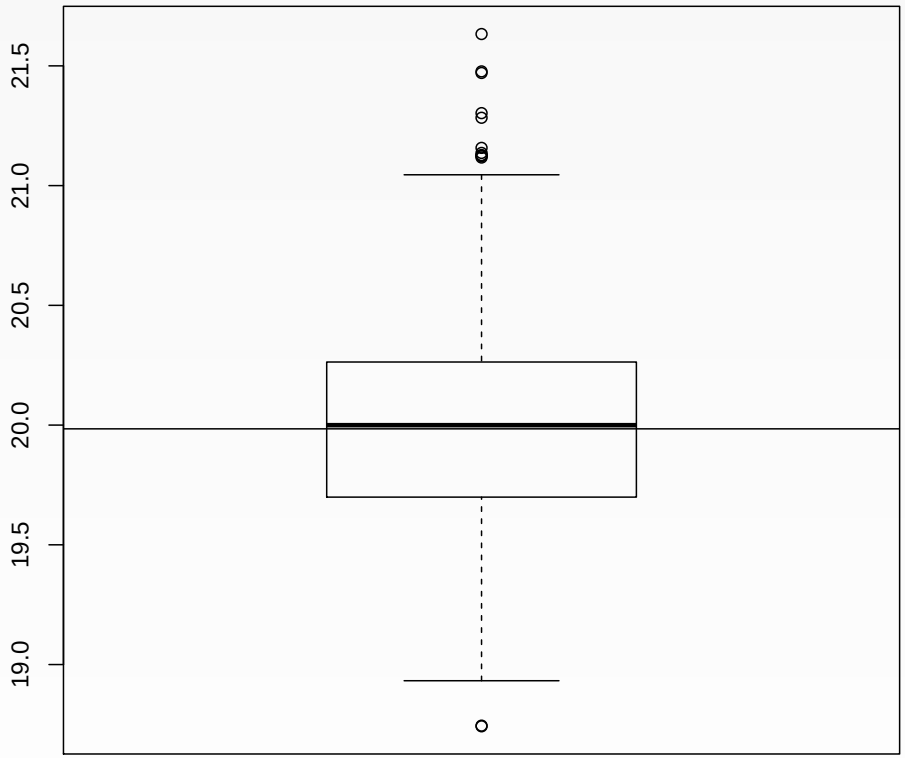
Simulations

Nous répétons $B = 1,000$ fois la procédure de sélection et d'estimation des paramètres.

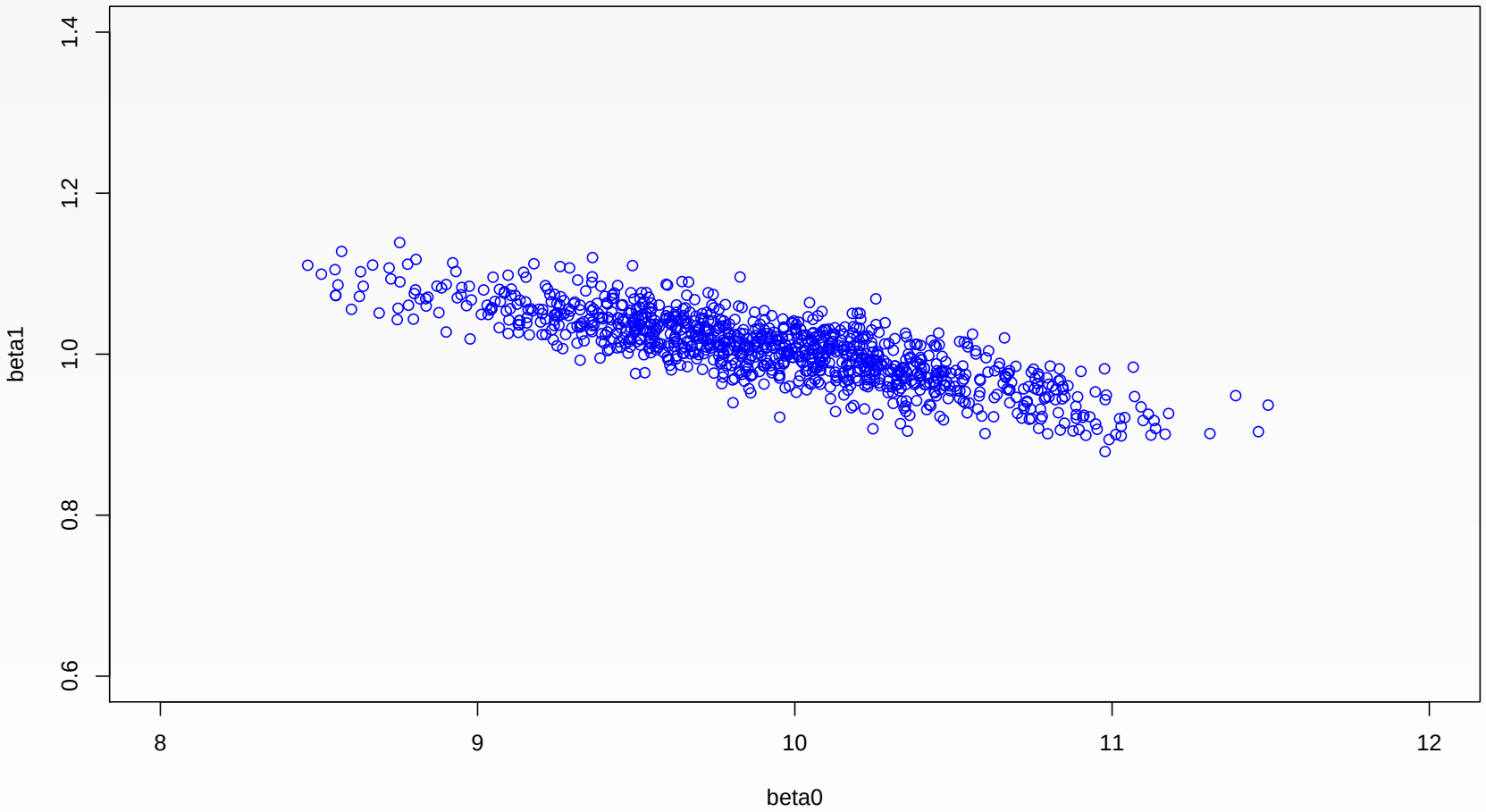
Nous obtenons une estimation de la distribution

- de l'estimateur de moyenne $\bar{y}$ (boxplot),
- de l'estimateur $\hat{\beta}_\pi$ des coefficients de régression (nuage de points).

Distribution de l'estimateur $\bar{y}$



Distribution des coefficients de régression estimés $\hat{\beta}_\pi$



Estimation en situation de non-réponse partielle

Nous supposons maintenant :

- que la variable x est renseignée pour chaque individu $k \in S$,
- que la variable y est affectée par de la non-réponse partielle, et n'est observée que sur un sous-échantillon de répondants S_r.

Ici, chaque individu de l'échantillon renseigne la variable y avec une probabilité p. Il s'agit donc d'un mécanisme MCAR. Dans ce qui suit, nous considérons $p = 0.8$, 0.6 et 0.4.

Estimation en situation de non-réponse partielle

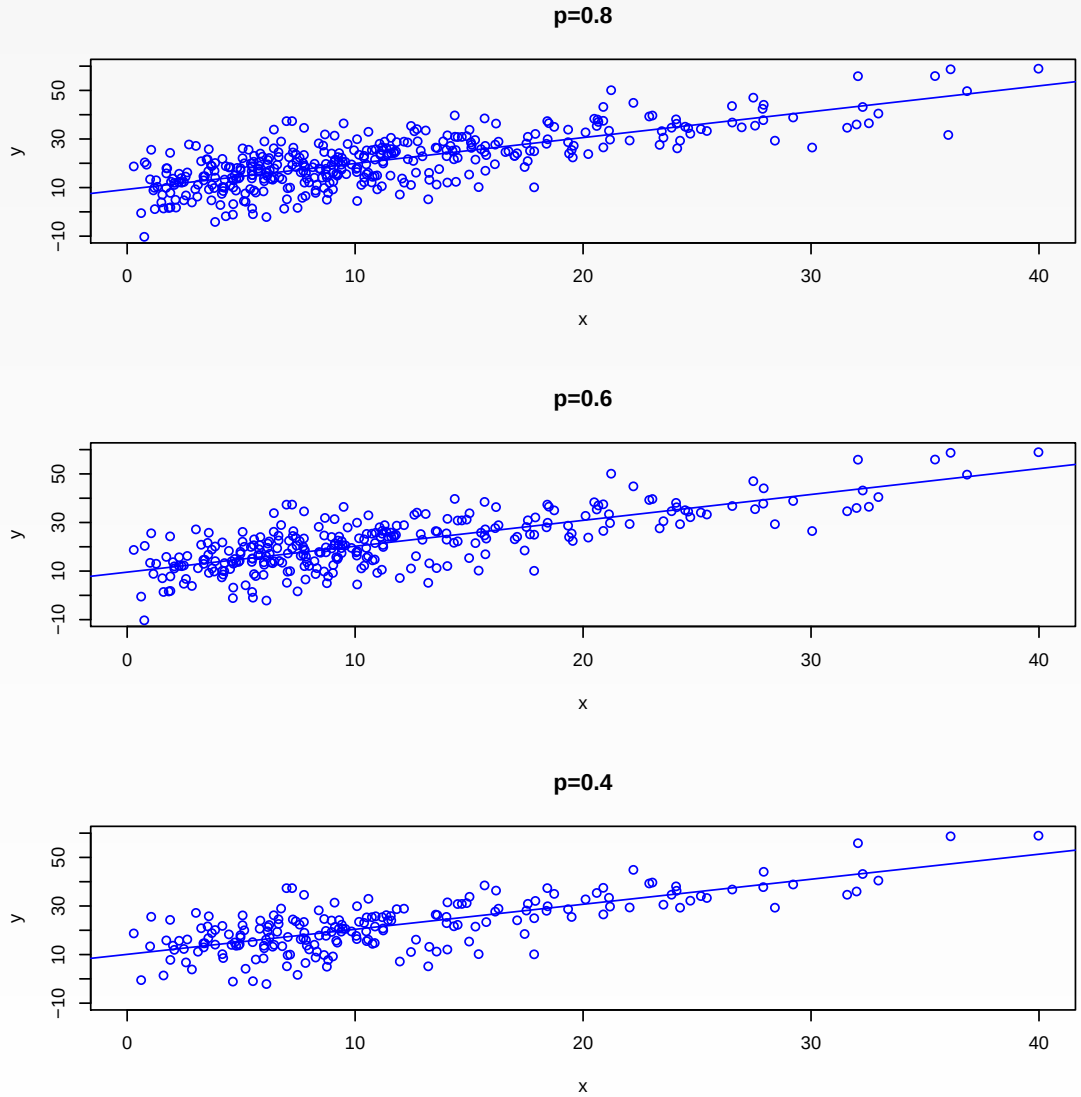
Nous utilisons les estimateurs basés sur les répondants

$$\begin{aligned} \bar{y}_r &= \frac{1}{n_r} \sum_{k \in S_r} y_k, \\ \hat{\beta}_r &= \left(\sum_{k \in S_r} \mathbf{x}_k \mathbf{x}_k^\top \right)^{-1} \sum_{k \in S_r} \mathbf{x}_k y_k. \end{aligned}$$

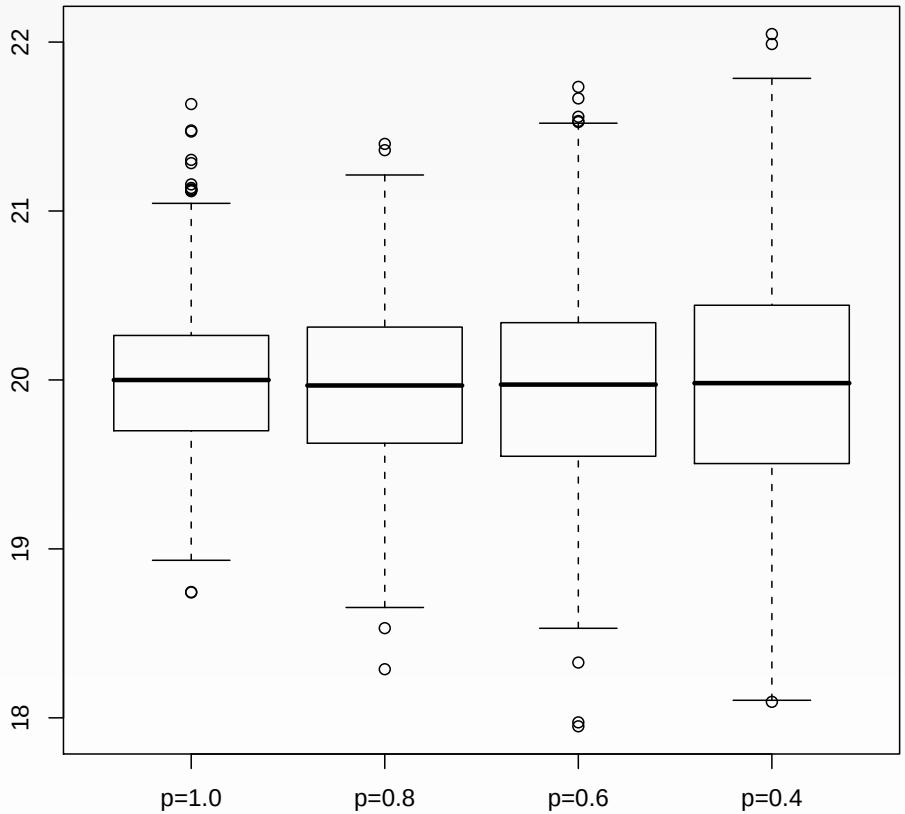
Sous un SRS suivi d'un mécanisme de réponse MCAR, l'échantillon de répondants S_r peut être vu comme ici d'un sondage aléatoire simple de taille n_r, conditionnellement à n_r.

Les estimateurs ci-dessus sont donc non biaisés, mais avec une variance plus grande. Nous obtenons là aussi une estimation de leur distribution en simulant $B = 1,000$ fois le plan de sondage + le mécanisme de non-réponse.

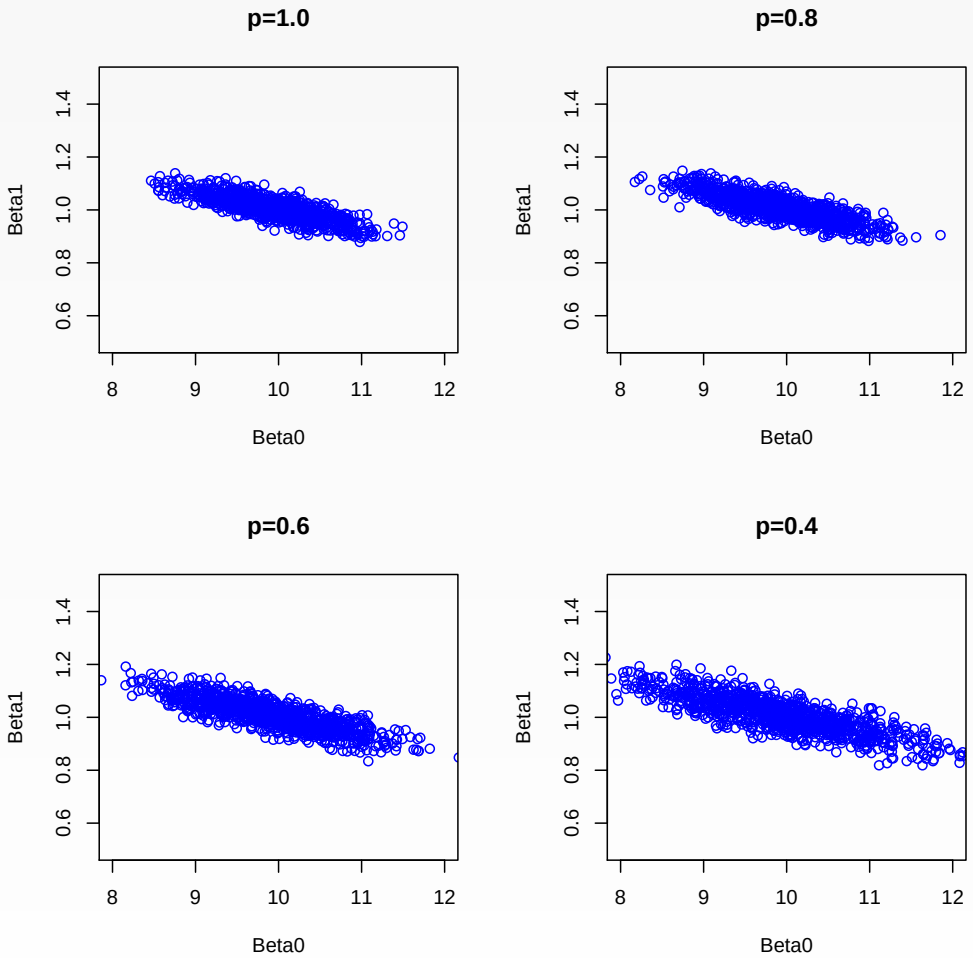
Données échantillonnées



Distribution de l'estimateur $\bar{y}_r$



Distribution des coefficients de régression estimés $\hat{\beta}_r$



Estimation sur données imputées

Estimateurs imputés

Pour un individu $k \in S_m$, soit y_k^* la valeur imputée pour remplacer y_k. Nous noterons également

$$\tilde{y}_k = \begin{cases} y_k & \text{si } k \in S_r, \\ y_k^* & \text{si } k \in S_m. \end{cases}$$

Nous obtenons les estimateurs imputés

$$\begin{aligned} \bar{y}_I &= \frac{1}{n} \sum_{k \in S} \tilde{y}_k, \\ \hat{\beta}_I &= \left(\sum_{k \in S} \mathbf{x}_k \mathbf{x}_k^\top \right)^{-1} \sum_{k \in S} \mathbf{x}_k \tilde{y}_k. \end{aligned}$$

Nous étudions le comportement de ces estimateurs en simulant $B = 1,000$ fois : plan de sondage + mécanisme de non-réponse + mécanisme d'imputation.

Imputation par la moyenne

Principe

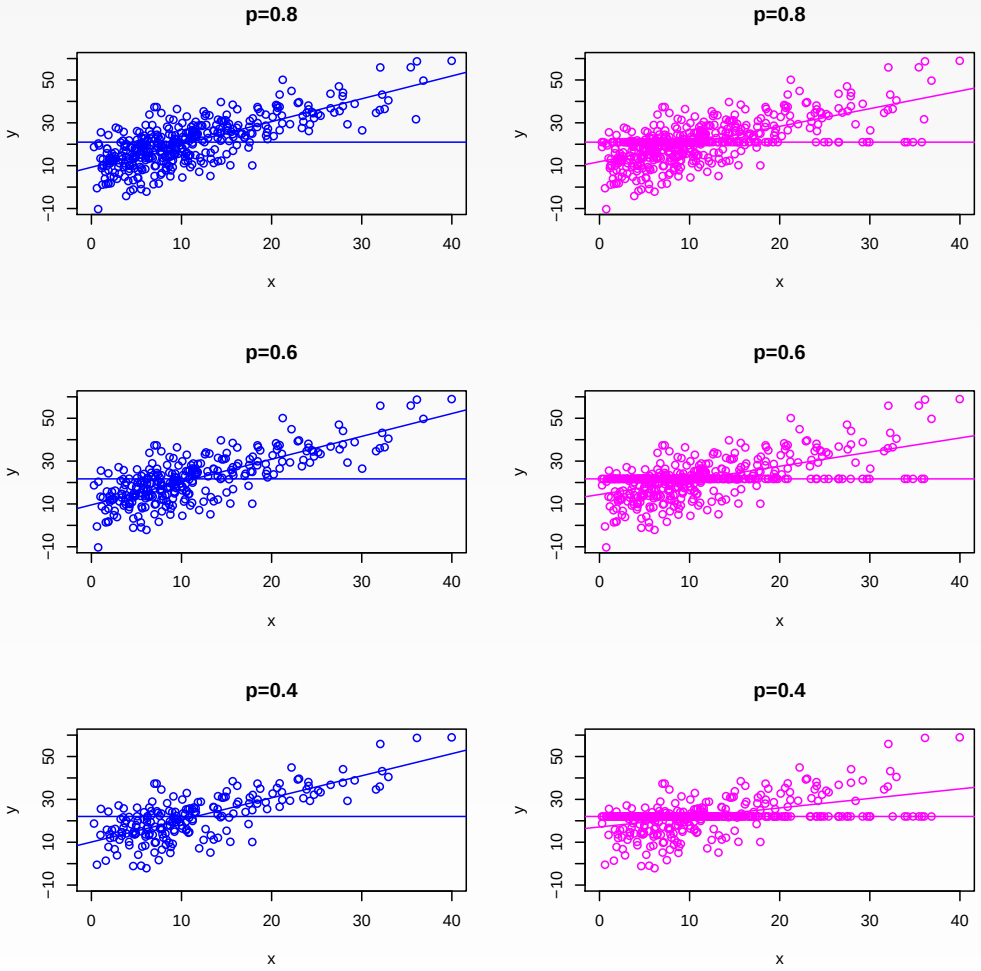
Pour un individu $k \in S_m$, nous utilisons $y_k^* = \bar{y}_r$ avec

$$\bar{y}_r = \frac{1}{n_r} \sum_{k \in S_r} y_k.$$

Nous obtenons les estimateurs imputés

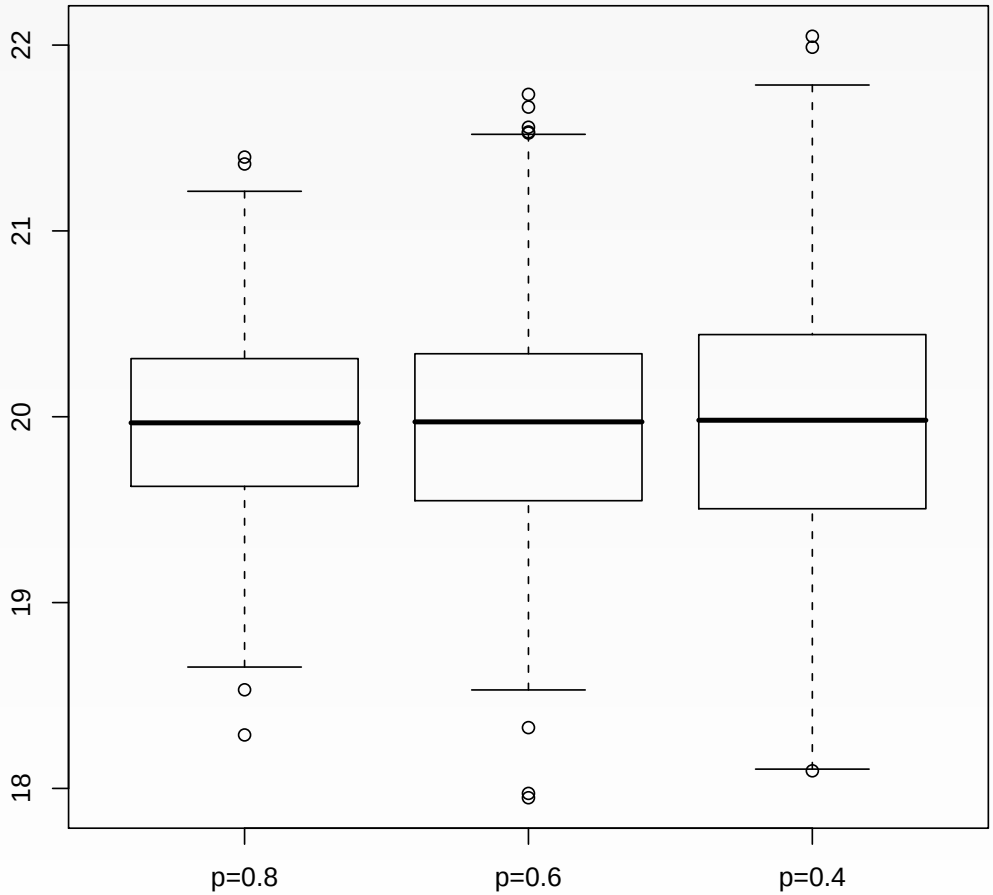
$$\begin{aligned} \bar{y}_I &= \bar{y}_r, \\ \hat{\beta}_I &= \left(\sum_{k \in S} \mathbf{x}_k \mathbf{x}_k^\top \right)^{-1} \sum_{k \in S} \mathbf{x}_k \tilde{y}_k. \end{aligned}$$

Données obtenues sur un échantillon

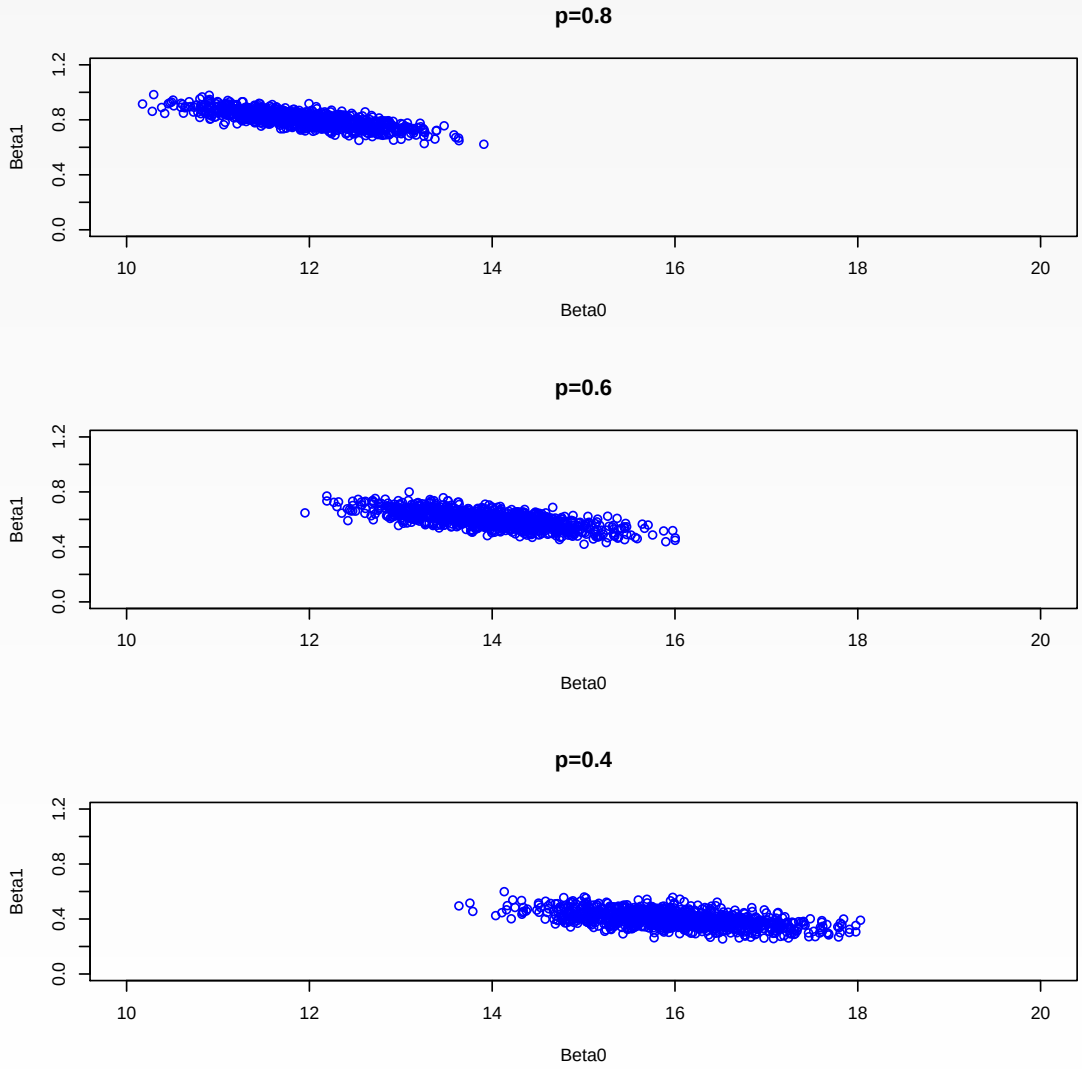


Cas complets Données imputées

Distribution de l'estimateur $\bar{y}_I$



Distribution des coefficients de régression estimés



Etude de l'estimateur imputé $\hat{\beta}_I$

En utilisant l'expression

$$\hat{\beta}_I = \left(\sum_{k \in S} \mathbf{x}_k \mathbf{x}_k^\top \right)^{-1} \left[\sum_{k \in S_r} \mathbf{x}_k y_k + \sum_{k \in S_m} \mathbf{x}_k \mathbf{x}_k^\top \begin{pmatrix} \bar{y}_r \\ 0 \end{pmatrix} \right],$$

il est possible de montrer que

$$E_p E_q(\hat{\beta}_I) \simeq p \begin{pmatrix} B_{0,U} \\ B_{1,U} \end{pmatrix} + (1-p) \begin{pmatrix} \mu_y \\ 0 \end{pmatrix},$$

$$\text{avec } \begin{pmatrix} B_{0,U} \\ B_{1,U} \end{pmatrix} = \left(\sum_{k \in U} \mathbf{x}_k \mathbf{x}_k^\top \right)^{-1} \sum_{k \in U} \mathbf{x}_k y_k.$$

En espérance, l'estimateur $\hat{\beta}_I$ est donc égal à un mélange entre l'estimateur des MCO du coefficient de régression du modèle, calculé sur la population entière, et le vecteur $\begin{pmatrix} \mu_y \\ 0 \end{pmatrix}$ associé à la droite horizontale d'ordonnée μ_y.

Imputation par hot-deck

Principe

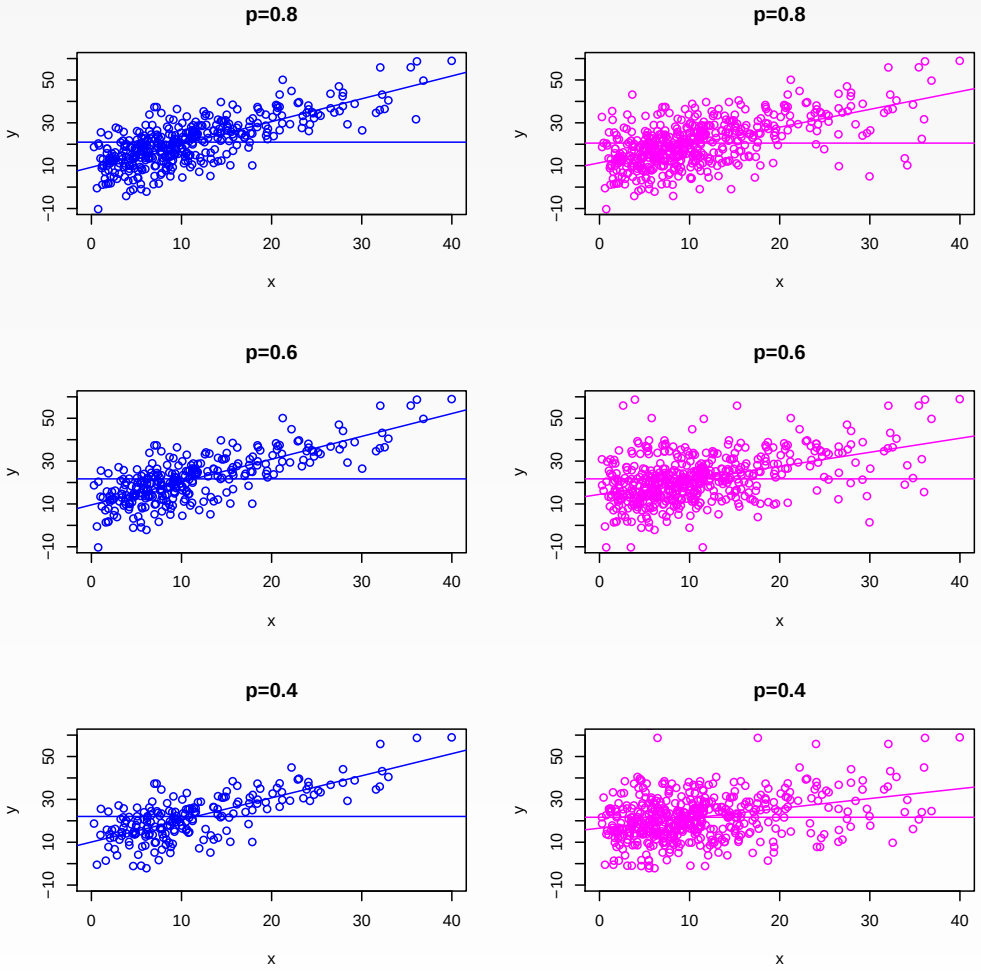
Pour un individu $k \in S_m$, la valeur y_k est remplacée en tirant au hasard et avec remise un donneur $y_{(j)} \in S_r$, avec des probabilités de tirage égales.

La valeur imputée peut encore se réécrire

$$y_k^* \quad = \quad \bar{y}_r + \left[y_{(j)} - \bar{y}_r \right].$$

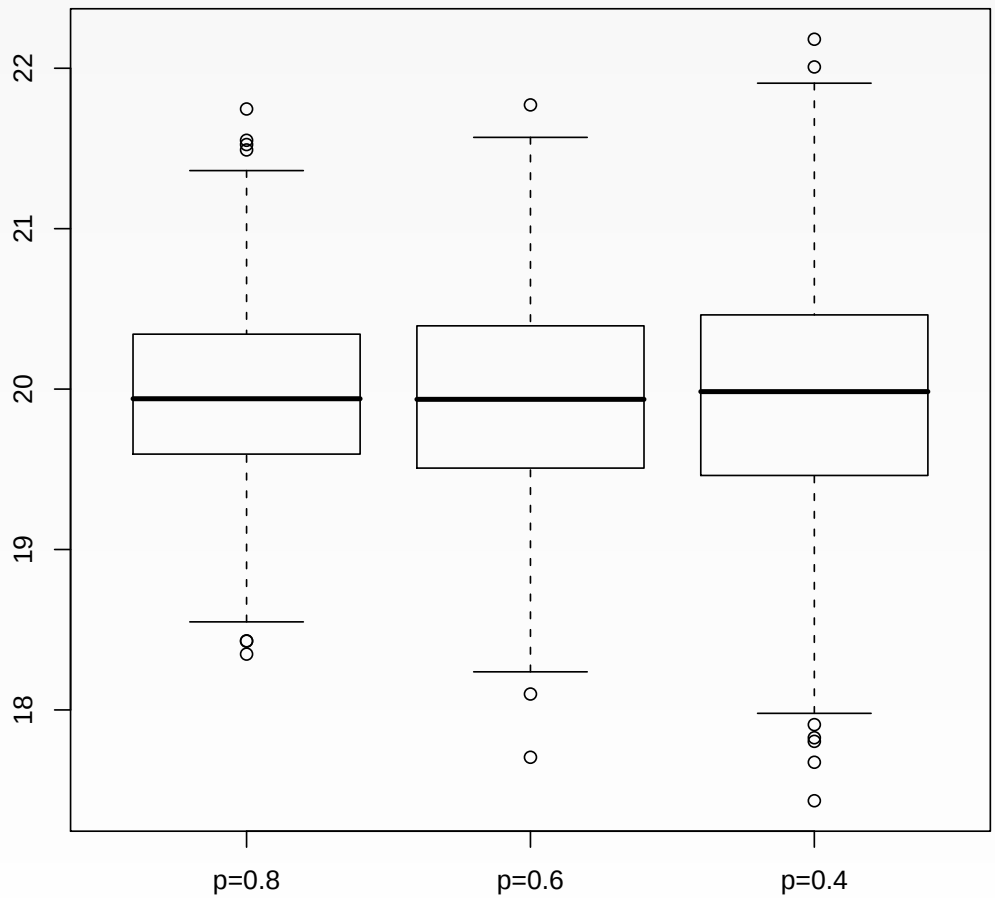
Interprétation : une valeur manquante est remplacée par la moyenne $\bar{y}_r$ des répondants, à laquelle on ajoute un résidu aléatoire (de moyenne nulle).

Données obtenues sur un échantillon

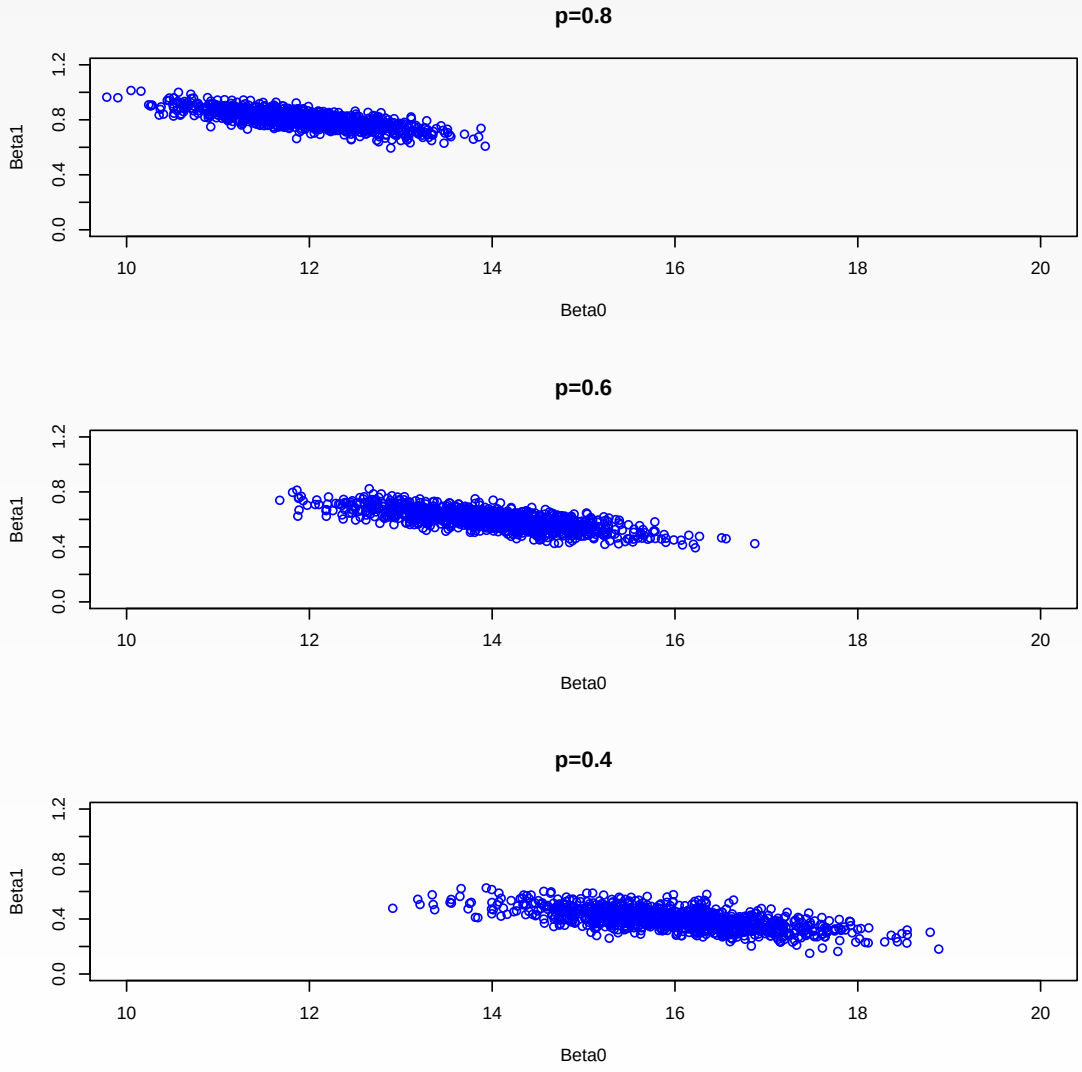


Cas complets Données imputées

Distribution de l'estimateur $\bar{y}_I$



Distribution des coefficients de régression estimés



Etude de l'estimateur imputé $\hat{\beta}_I$

Les estimateurs imputés possèdent en espérance le même comportement qu'avec une imputation par la moyenne.

De plus, la variance augmente en raison de l'alea d'imputation.

Imputation par la régression déterministe

Principe

Pour un individu $k \in S_m$, la valeur y_k est remplacée par la prédiction $y_k^* = \mathbf{x}_k^\top \hat{\beta}_r$, avec

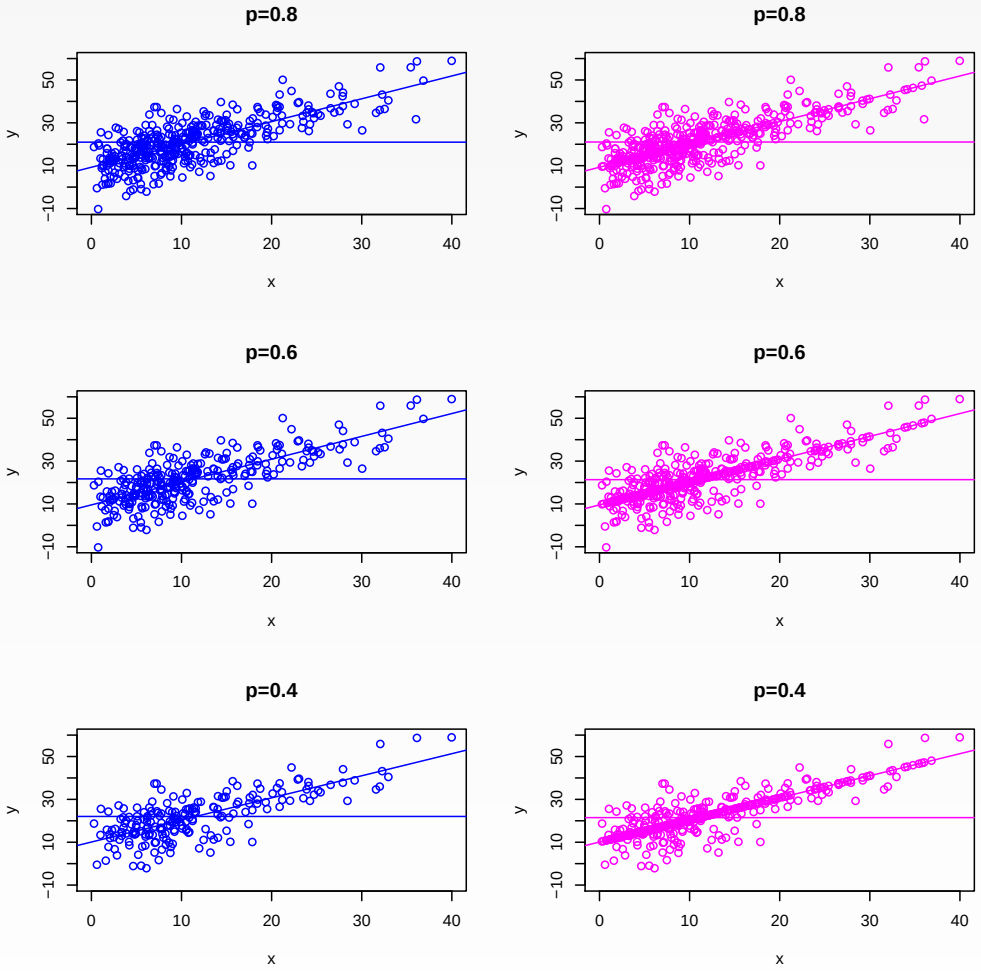
$$\hat{\beta}_r = \left(\sum_{k \in S_r} \mathbf{x}_k \mathbf{x}_k^\top \right)^{-1} \sum_{k \in S_r} \mathbf{x}_k y_k$$

le coefficient de régression estimé sur les répondants.

Nous obtenons les estimateurs imputés

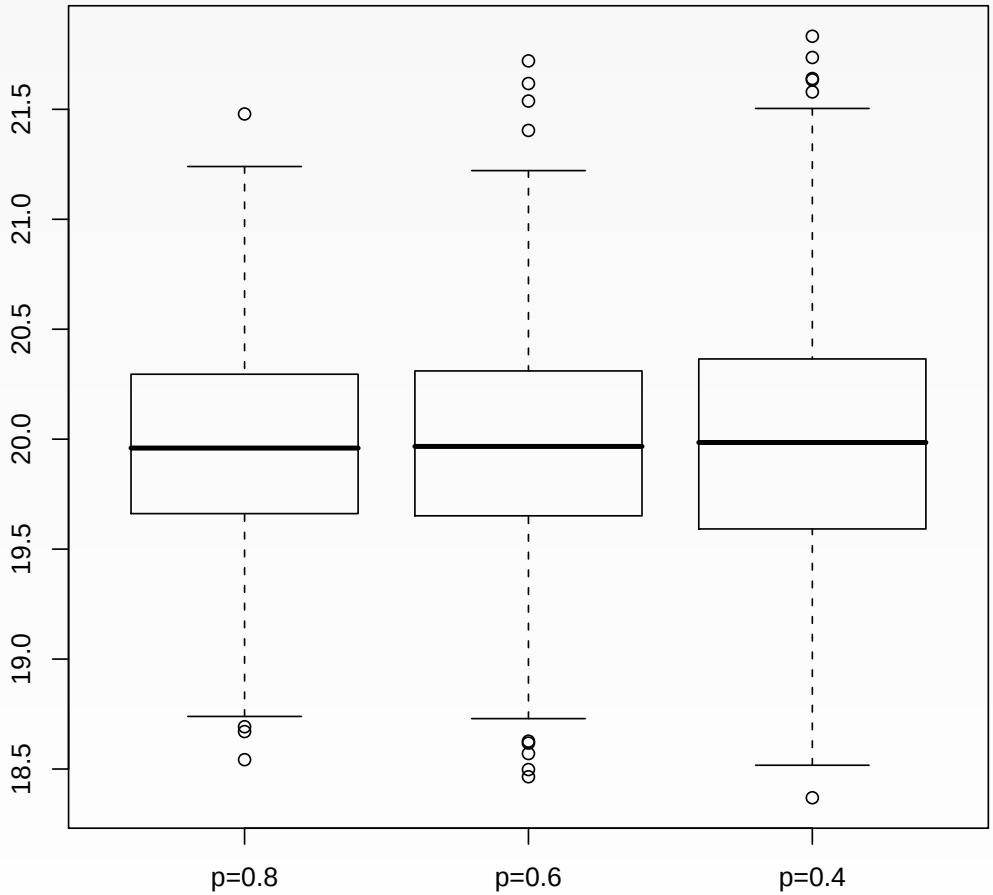
$$\begin{aligned} \bar{y}_I &= \frac{1}{n} \sum_{k \in S} \tilde{y}_k, \\ \hat{\beta}_I &= \hat{\beta}_r. \end{aligned}$$

Données obtenues sur un échantillon

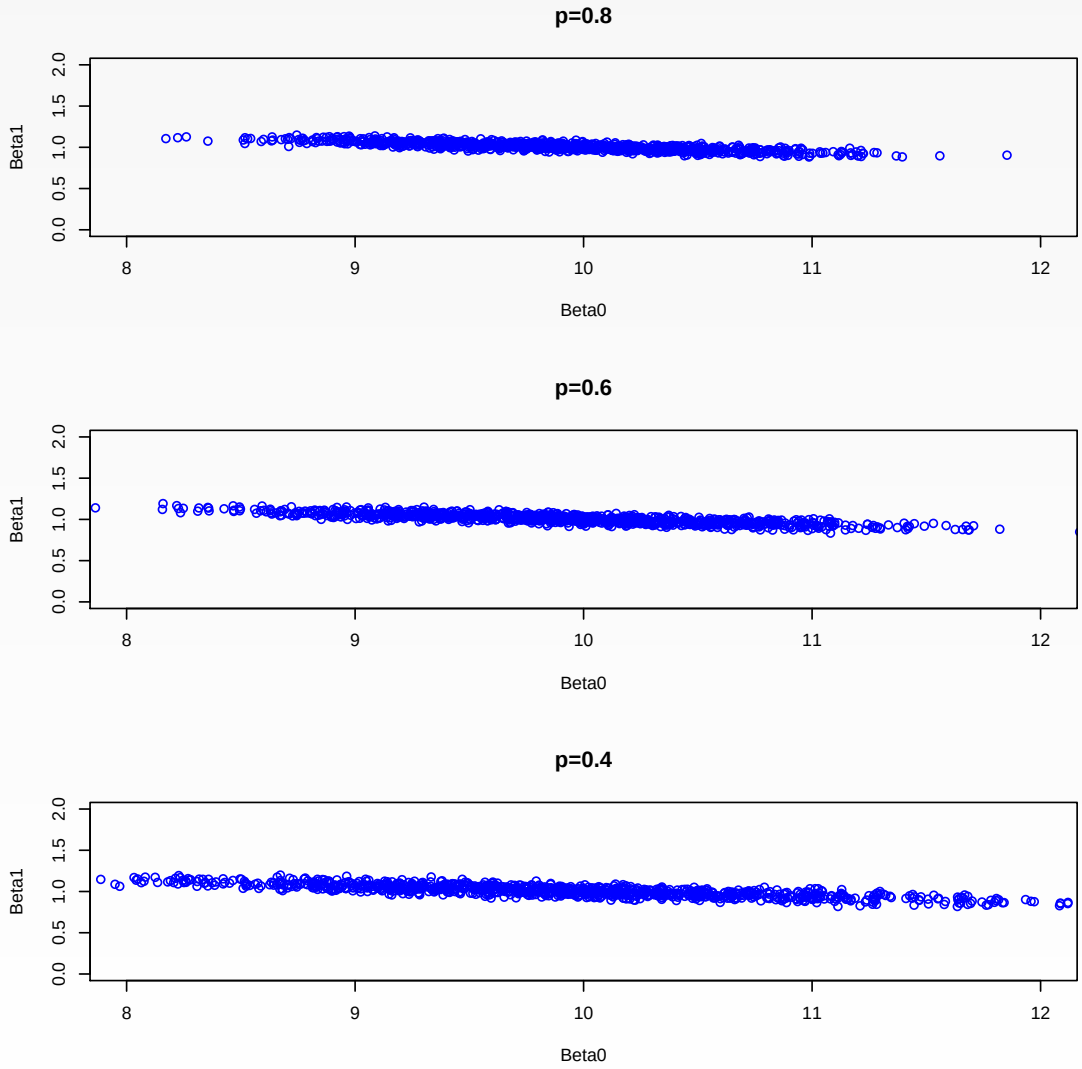


Cas complets Données imputées

Distribution de l'estimateur $\bar{y}_I$



Distribution des coefficients de régression estimés



Imputation par la régression aléatoire

Principe

Pour un individu $k \in S_m$, la valeur y_k est remplacée par la prédiction $\mathbf{x}_k^\top \hat{\beta}_r$, à laquelle on ajoute un résidu aléatoire $\eta_{(j)}$.

Ce résidu aléatoire est tiré, avec remise et à probabilités égales, parmi les résidus effectivement observés

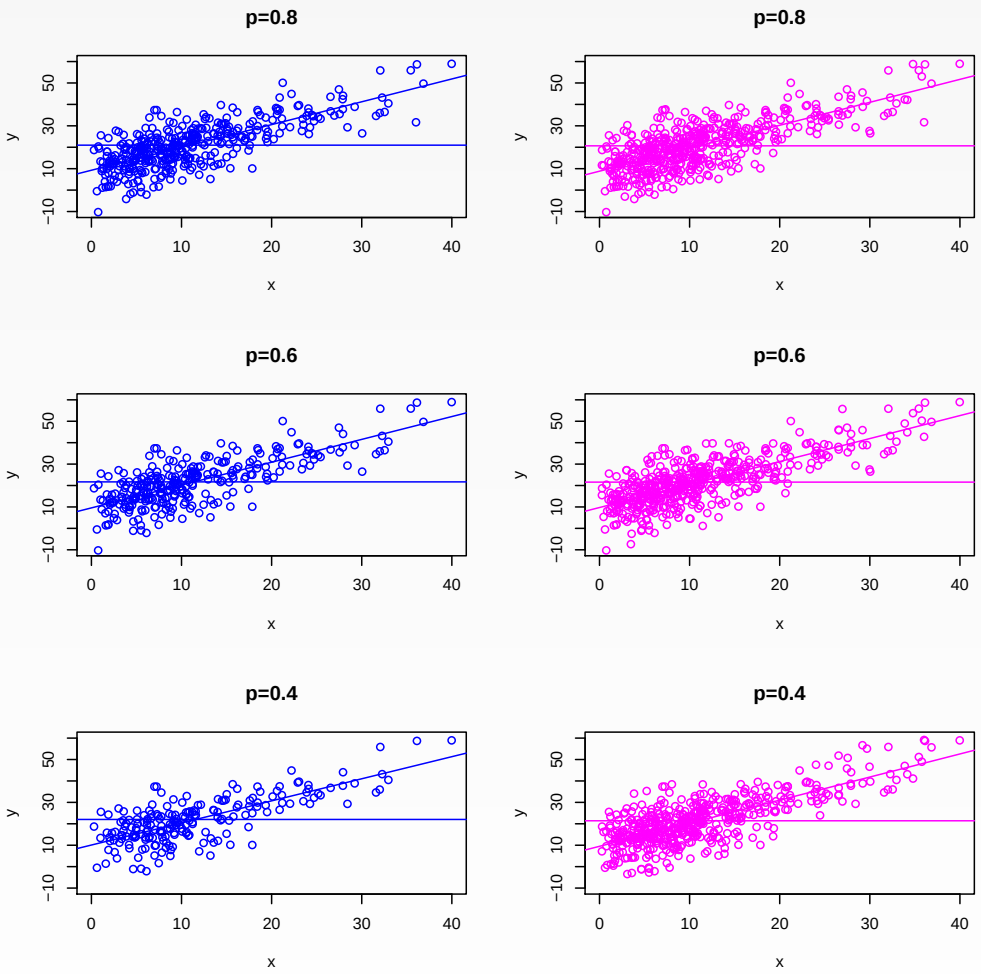
$$\eta_j = y_j - \mathbf{x}_j^\top \hat{\beta}_r \quad \text{pour} \quad j \in S_r.$$

Nous obtenons pour $k \in S_m$ la valeur imputée

$$y_k^* = \mathbf{x}_k^\top \hat{\beta}_r + \eta_{(j)}.$$

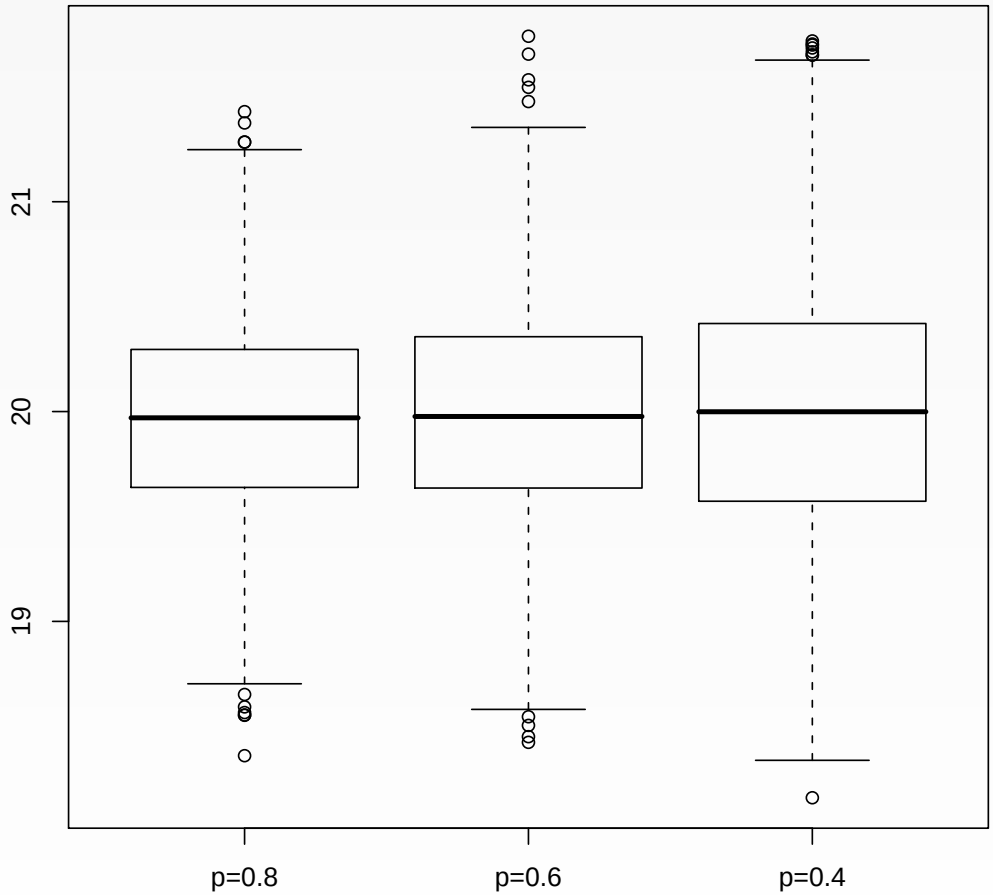
Nous imputons donc "au plus près" du modèle (en tenant compte de ses imperfections).

Données obtenues sur un échantillon

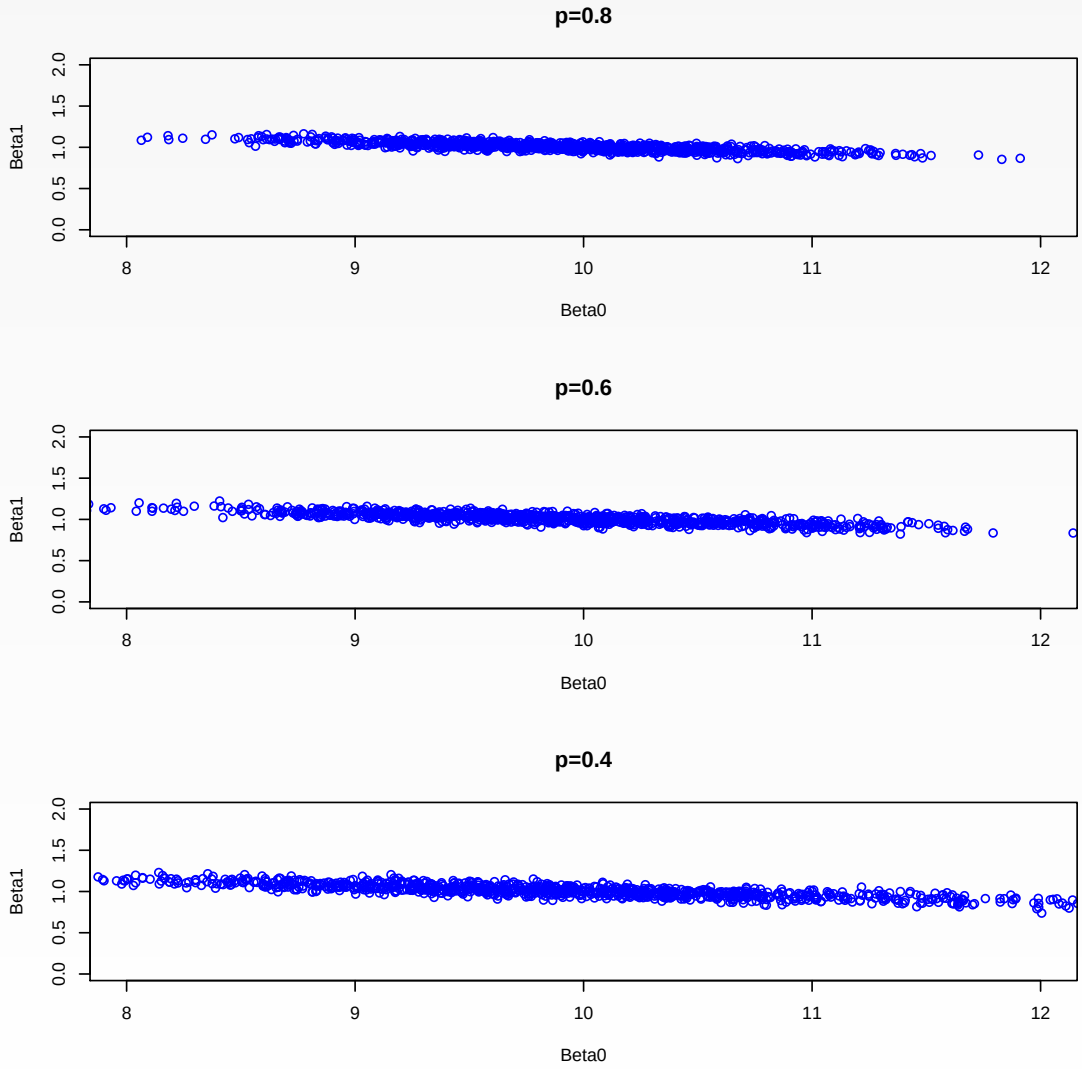


Cas complets Données imputées

Distribution de l'estimateur $\bar{y}_I$



Distribution des coefficients de régression estimés



Bibliographie

- Ardilly, P. (2006), *Les Techniques de Sondage*, Technip, Paris.
- Couvert, N., Dieusaert, P., et Henry, M. (2016), *Enquête Panel Politique de la Ville 4ème vague*, Dossier Comité du Label du Cnis.
- Da Silva, D.N., et Opsomer, J.D. (2006). *A kernel smoothing method to adjust for unit nonresponse in sample surveys*. Canadian Journal of Statistics, 34, 563-579.
- Da Silva, D.N., et Opsomer, J.D. (2009). *Nonparametric propensity weighting for survey nonresponse through local polynomial regression*. Survey Methodology, 35, 165-176.
- Demoly, E., Fizzala, A., et Gros, E. (2014). *Méthodes et pratiques des enquêtes entreprises à l'Insee*. Journal de la SFdS, 155, 134-159.
- Deroyon, T. (2017). *non-response correction through reweighting*. Document de travail, Département des méthodes statistiques, Insee.
- Gelein, B., Haziza, D., Causeur, D. (2018), *Propensity weighting for survey nonresponse through machine learning*. Journées de méthodologie statistique, Insee.
- Haziza, D. (2009). *Imputation and inference in the presence of missing data*, Handbook of Statistics, vol. 29, chap. 10.
- Haziza, D. (2011). *Traitement de la non-réponse totale et partielle dans les enquêtes*. Master de Statistique Publique, Ensai.
- Haziza, D., et Rao, J.N.K. (2003). *Inference for population means under unweighted imputation for missing survey data*. Survey Methodology, 29, 81-90.
- Joinville, O. (2002). *Mise en oeuvre du logiciel POULPE pour estimer la précision de l'enquête HID*, Rapport de Stage de 2nde année, ISUP.
- Opsomer, J., Riddles, M. (2023). *Fitting Classification Trees to Complex Survey Data*, Webinar IASS.
- Skinner, C.J., et D'Arrigo, J. (2011). *Inverse probability weighting for clustered nonresponse*. Biometrika, 98, 953-966.