Données Manquantes dans les Enquêtes

Guillaume Chauvet

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

27 avril 2015

Panorama du cours

1. Introduction et rappels
2. Traitement de la non-réponse totale
3. Traitement de la non-réponse partielle

Objectifs du cours

- 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.

Type de non-réponse

Dans le contexte des enquêtes, on distingue deux types de non-réponse :

- 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	*	*	*	*	

Introduction et rappels

Les étapes d'une enquête (Haziza, 2011)

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)

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

Rappels sur l'échantillonnage en population finie

Plan de sondage

On se place dans le cadre d'une population finie d'individus, notée U. On s'intéresse à une **variable d'intérêt** y (éventuellement vectorielle), 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 (supposée connue) sur l'ensemble des parties de U.

Plan de sondage

On suppose en particulier 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**

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

avec $d_k = 1/\pi_k$ le poids de sondage de l'unité k.

Remarque : 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 (1953)

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

avec $\Delta_{kl} = \pi_{kl} - \pi_k\pi_l$.

On peut l'estimer sans biais par

$$v_{HT}\left[\hat{t}_{y\pi}\right] = \sum_{k,l\in S} \frac{y_k}{\pi_k}\frac{y_l}{\pi_l}\frac{\Delta_{kl}}{\pi_{kl}}$$

si tous les π_{kl} sont > 0.

Le tirage poissonien

Chaque individu k est tiré dans l'échantillon avec une probabilité π_k, indépendamment des autres individus.

Le π-estimateur

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

a pour variance

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

et on l'estime sans biais par

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

Plan de taille fixe

Si le plan de sondage $p(\cdot)$ est **de taille fixe** égale à n, la variance du π-estimateur peut être alternativement obtenue par la formule de Sen-Yates-Grundy (1954)

$$V_p\left[\hat{t}_{y\pi}\right] = -\frac{1}{2}\sum_{k\neq l\in U}\left(\frac{y_k}{\pi_k}-\frac{y_l}{\pi_l}\right)^2\Delta_{kl}.$$

Un estimateur sans biais est donné par

$$v_{YG}\left[\hat{t}_{y\pi}\right] = -\frac{1}{2}\sum_{k\neq l\in S}\left(\frac{y_k}{\pi_k}-\frac{y_l}{\pi_l}\right)^2\frac{\Delta_{kl}}{\pi_{kl}}.$$

si tous les π_{kl} sont > 0.

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.

Le π-estimateur se réécrit

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

Sa variance est donnée par

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

et on l'estime sans biais par

$$v_{YG}\left[\hat{t}_{y\pi}\right] = N^2(1-f)\frac{s_y^2}{n}.$$

SRS stratifié

La population est partitionnée en H strates $U_1, \ldots, U_H$. On effectue un SRS(n_h) indépendamment dans chaque strate.

Le π-estimateur 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 on l'estime sans biais par

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

Exemple : enquêtes entreprises

Les échantillons pour les enquêtes auprès des entreprises sont souvent tirés selon des plans de sondages aléatoires simples stratifiés. La stratification est obtenue en croisant :

- un critère d'activité (nomenclature d'activités française NAF),
- un critère de taille (tranches d'effectifs salariés et/ou tranches de chiffres d'affaires).

Par exemple (voir Demoly et al., 2014), l'enquête sur les technologies de l'information et de la communication (TIC) a été tirée en stratifiant selon :

- le secteur d'activité,
- la tranche d'effectif de l'entreprise (10-19, 20-49, 50-249, 250-499, 500 et +),
- le chiffre d'affaires,

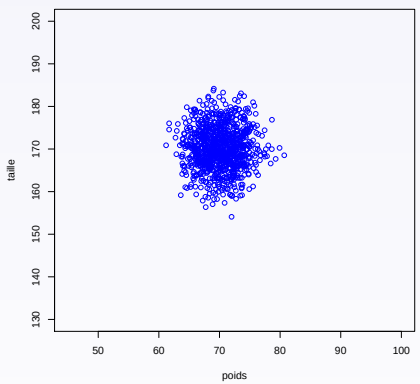
avec un seuil d'exhaustivité pour les plus grandes tranches d'effectif et les plus gros chiffres d'affaires.

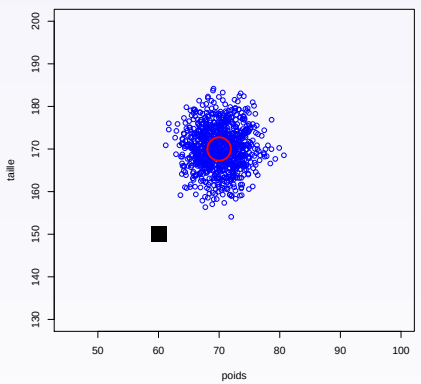
Sources d'erreur dans l'estimation

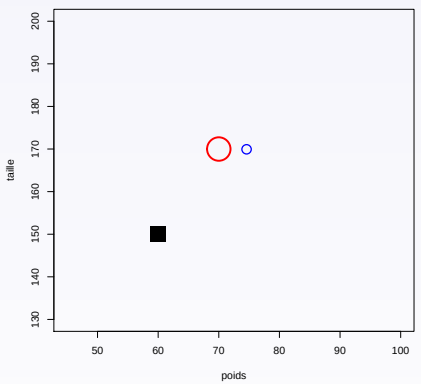
Erreur associée à l'estimateur

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

En pratique, l'erreur totale de l'estimateur, mesurée par

$$\hat{\theta} - \theta,$$

dépend des erreurs réalisées à toutes les étapes de l'enquêtes.

Ceci inclut :

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

Erreurs de couverture

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

- 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 téléphonique auprès de ménages, en utilisant une liste d'abonnés à une ligne fixe,
  - difficulté de couvrir la population-cible (enquête auprès de SDF).
- la sur-couverture (la base de sondage contient des individus qui ne sont pas dans la population-cible) :
  - échantillonnage de logements, dont le statut (RP/RS/LO/LV) n'est pas connu au moment du tirage, en vue d'une enquête en résidence principale.

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 uniquement : 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 à 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 variables d'intérêt.

Parmi les causes des erreurs de mesure :

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

Dans ce qui suit, on supposera que les erreurs de couverture et de mesure peuvent être négligées. On se focalisera sur l'erreur due à l'échantillonnage et sur l'erreur due à la non-réponse.

Les types de non-réponse

Type de non-réponse

Dans le contexte des enquêtes, on distingue deux types de non-réponse :

- 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 (partielle ou totale) 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.

z_1	z_2	...	z_q	y_1	y_2	y_3	...	...	y_p	
*	*	*	*	*	*	*	*	*	*	Réponse totale
*	*	*	*	*	*	*	*	*	*	
*	*	*	*	*	*	*	*	*	*	
*	*	*	*	*	*	*	*	*	*	
*	*	*	*	∅	∅	∅	∅	∅	∅	Non-réponse totale
*	*	*	*	∅	∅	∅	∅	∅	∅	
*	*	*	*	∅	∅	∅	∅	∅	∅	
*	*	*	*	*	*	∅	*	*	∅	Non-réponse partielle
*	*	*	*	∅	*	*	*	*	∅	
*	*	*	*	*	*	*	*	∅	∅	
*	*	*	*	∅	∅	∅	*	*	*	

Variables auxiliaires — Variables d'intérêt

Traitement de la non-réponse dans les enquêtes

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

- on supprime du fichier les non-répondants totaux,
- on augmente les poids des répondants 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 (Haziza, 2011)

- Mauvaise qualité de la base de sondage,
- Impossibilité de joindre l'individu,
- Type d'enquête (obligatoire ou volontaire),
- Fardeau de réponse,
- Méthode de collecte (interview, téléphone, courrier, ...),
- Durée de collecte,
- Suivi (et relance) des non-répondants,
- Formation des enquêteurs.

Quelques facteurs de non-réponse partielle (Haziza, 2011)

- Questionnaire mal conçu,
- Fardeau de réponse,
- Questions délicates,
- Formation des enquêteurs,
- Appel à la mémoire des enquêtés.

La prévention (ou la correction) de la non-réponse se fait à toutes les étapes de la collecte des données.

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.

On va traiter ce problème par **repondération** : on fait 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.

Identification des non-répondants

Identification des non-répondants

Un point important : la distinction entre individus **hors-champ** et individus non-répondants. Le **champ** de l'enquête désigne l'ensemble des individus statistiques auxquels on s'intéresse. Certains individus de l'échantillon sont hors-champ, et ne sont donc pas pris en compte dans l'estimation.

Les individus **non-répondants** font partie du champ de l'enquête, mais leur réponse n'est pas observée (refus de répondre, impossible à joindre, perte de questionnaire, ...) et doit être compensée.

Exemple : Enquête Logement 2006. Champ de l'enquête : logements résidences principales en 2006 (par opposition aux résidences secondaires, occasionnelles et aux logements vacants). On disposait de deux sources pour accéder à ces logements :

- Le Recensement de 1999,
- Les bases complémentaires de logements construits depuis 1999 (BSLN, issue du fichier SITADEL des permis de construire).

Schéma récapitulatif

Schéma récapitulatif

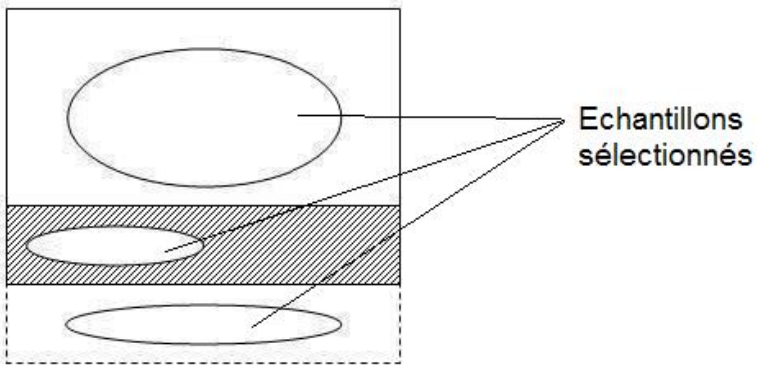


Schéma récapitulatif

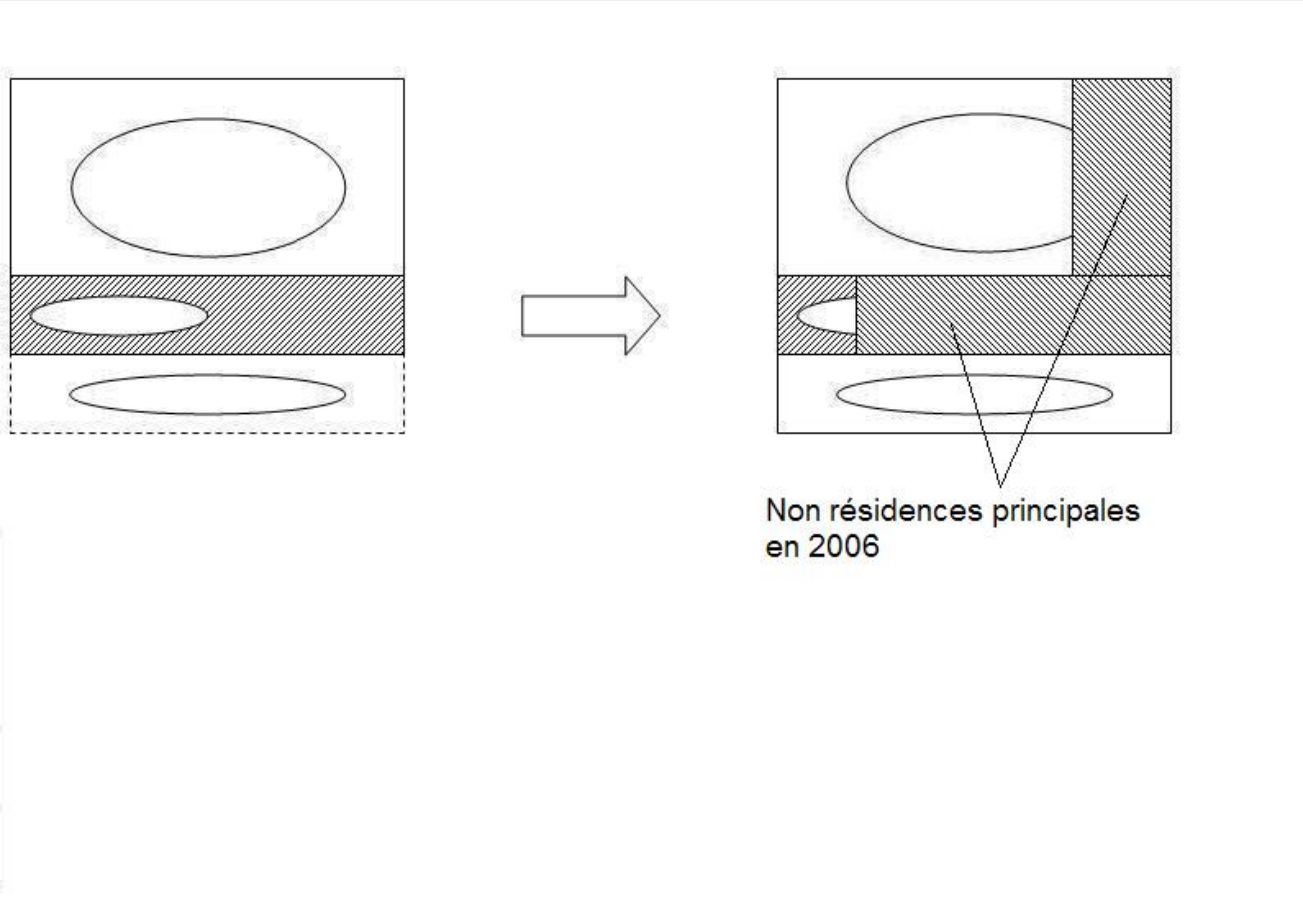


Schéma récapitulatif

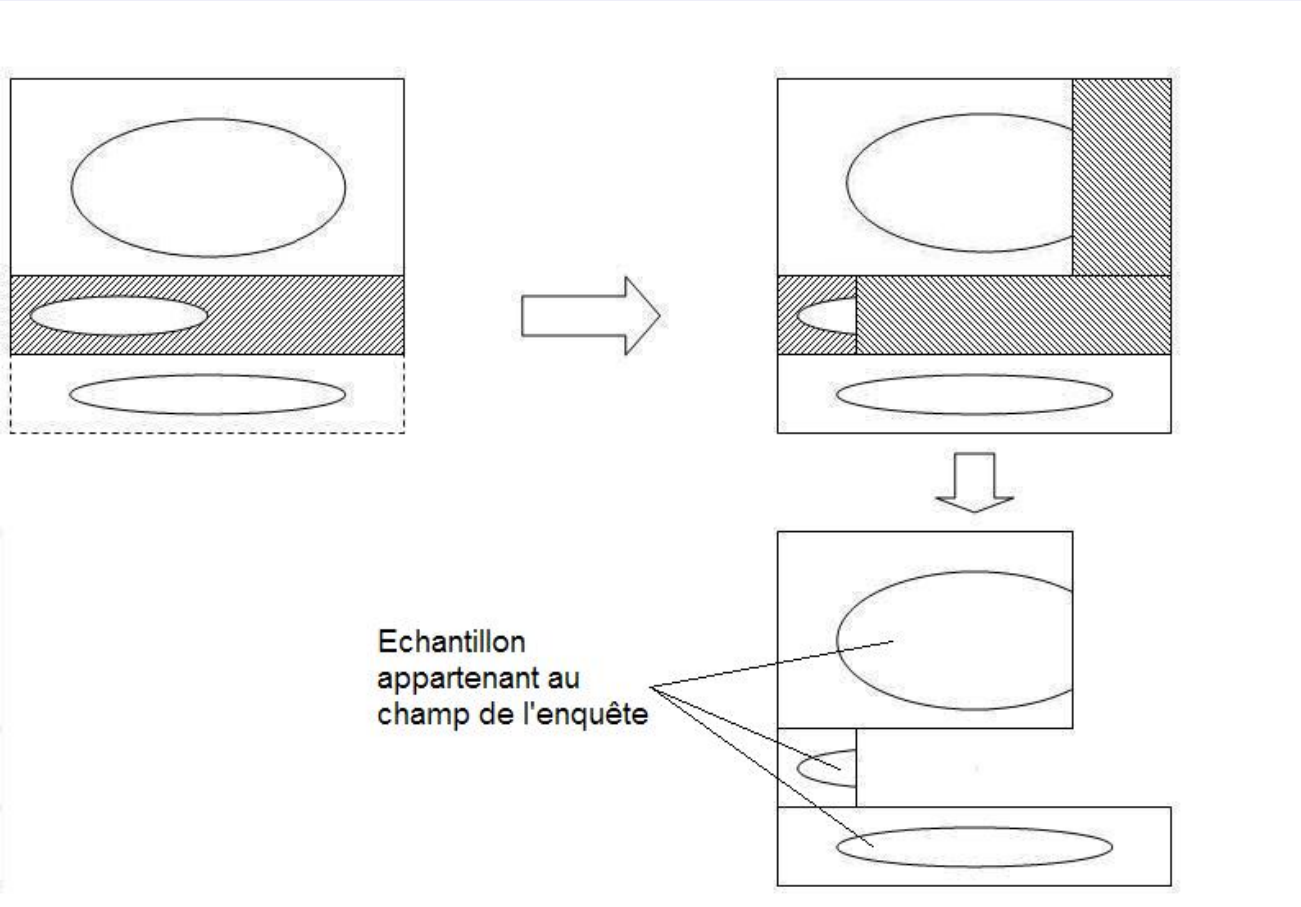
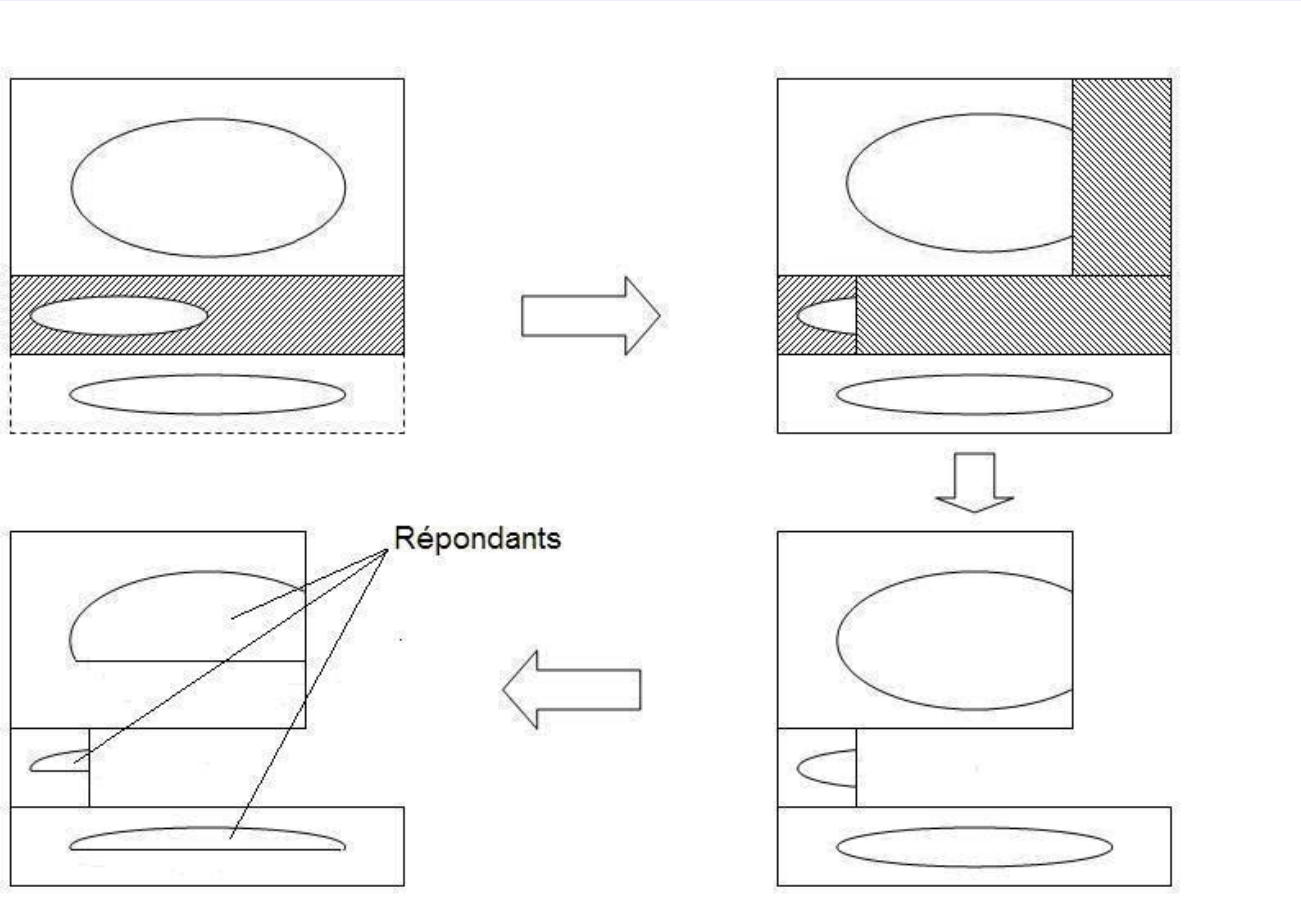


Schéma récapitulatif



Modélisation du mécanisme de non-réponse

Echantillonnage en deux phases

Dans le cadre d'une enquête, on peut être amené à sélectionner l'échantillon en deux temps :

- On sélectionne tout d'abord un gros sur-échantillon S selon un plan de sondage $p(\cdot)$.
- On tire ensuite dans S un sous-échantillon S_0 selon un plan de sondage $q(\cdot|S)$.

On parle d'échantillonnage en deux phases. Cette méthode est par exemple utilisée pour cibler une population spécifique.

Exemple : Enquête Vie Quotidienne et Santé, utilisée comme filtrage pour l'enquête Handicaps-Incapacités-Dépendances (Joinville, 2002).

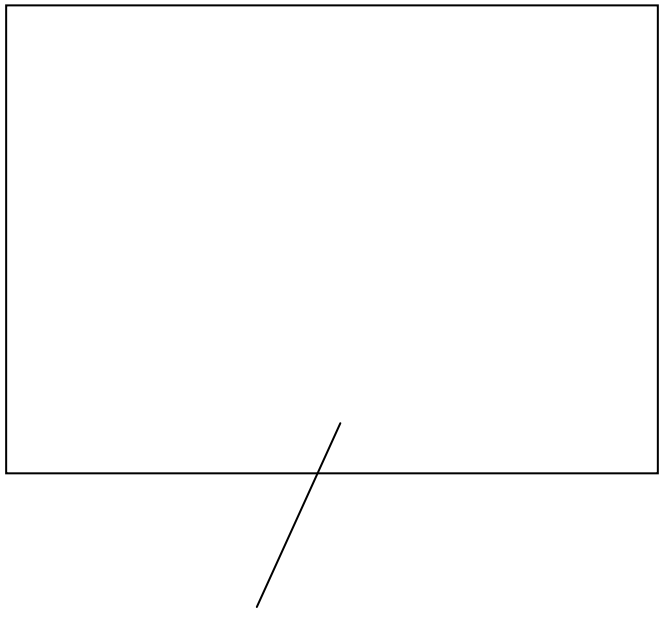
Modélisation du mécanisme de non-réponse

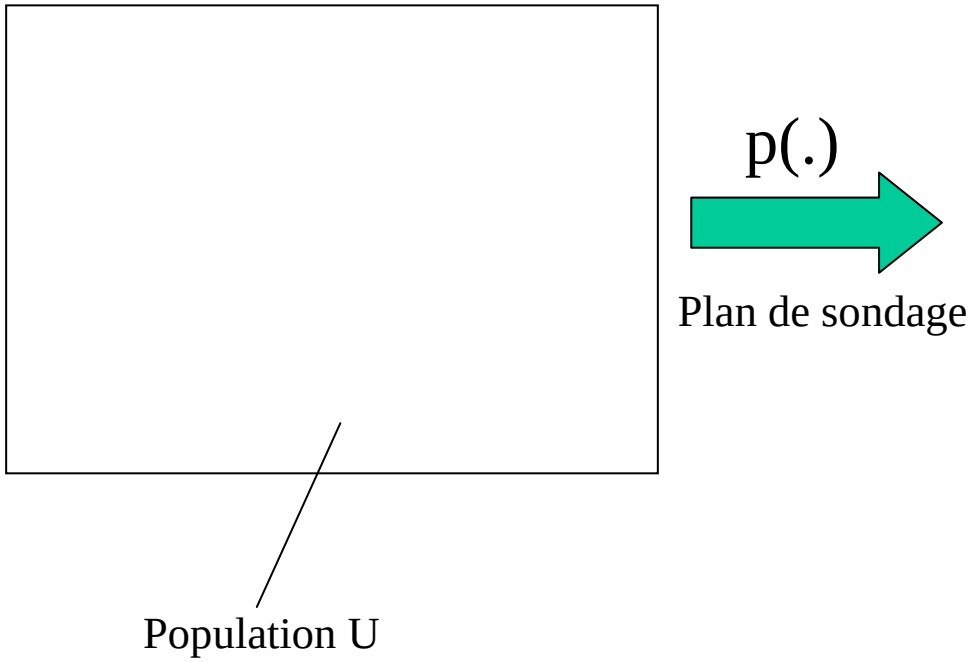
En situation de non-réponse totale :

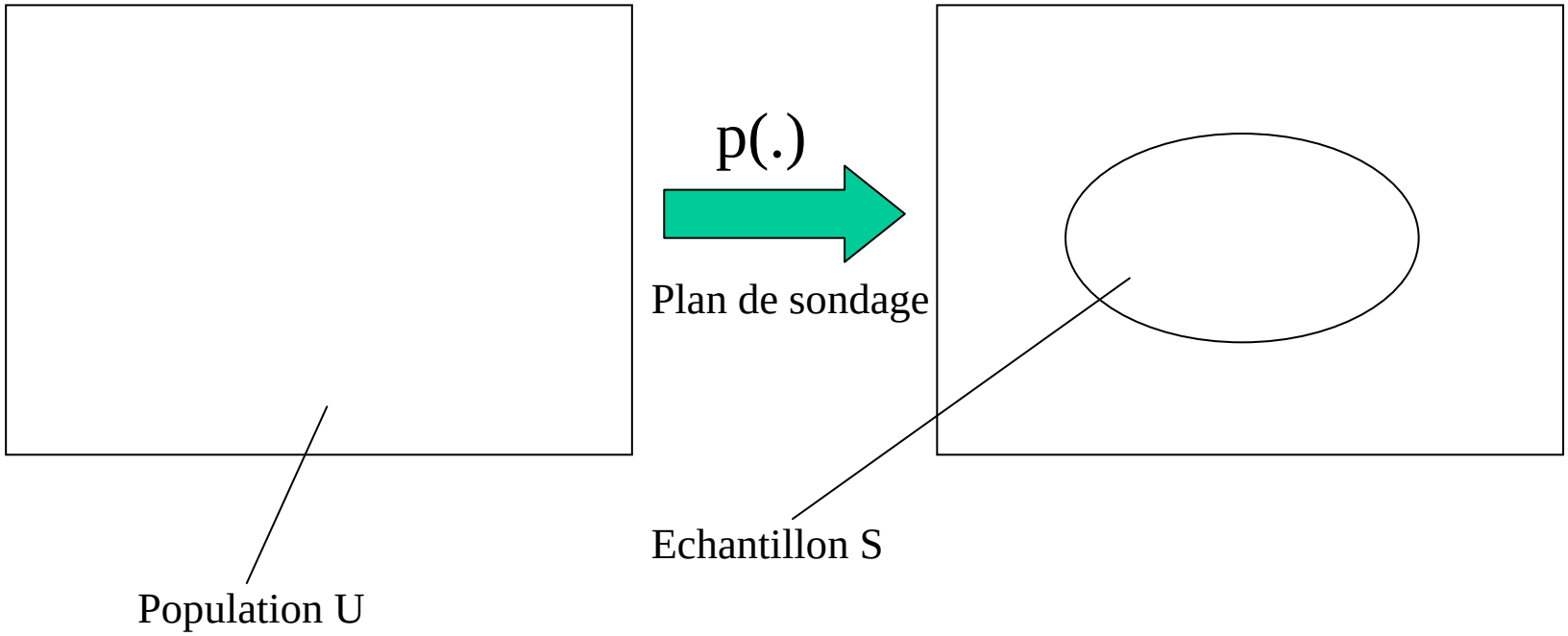
- le mécanisme de sélection de l'échantillon S est connu,
- le mécanisme de non-réponse qui conduit au sous-échantillon de répondants S_r est en revanche inconnu.

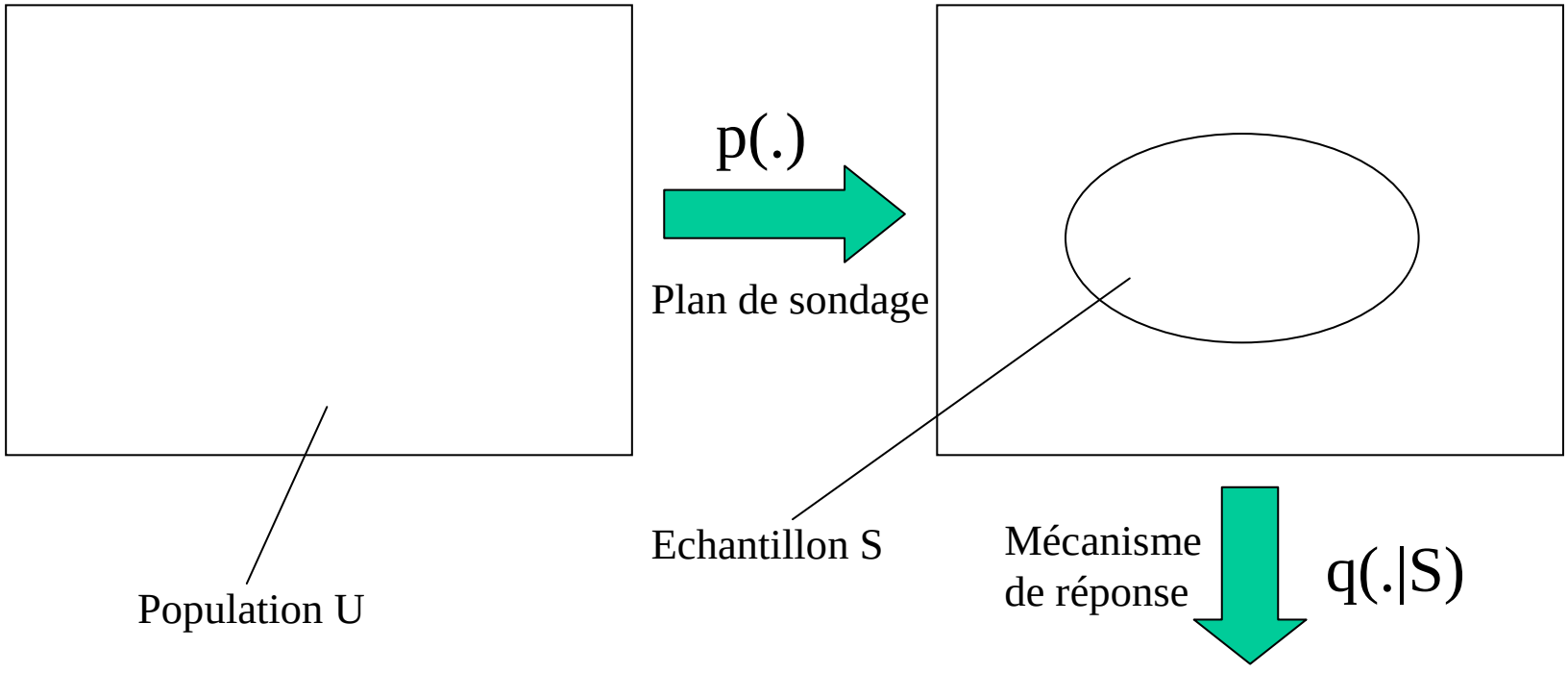
On a recours à une modélisation du mécanisme aléatoire conduisant à S_r sous la forme d'un échantillonnage en deux phases :

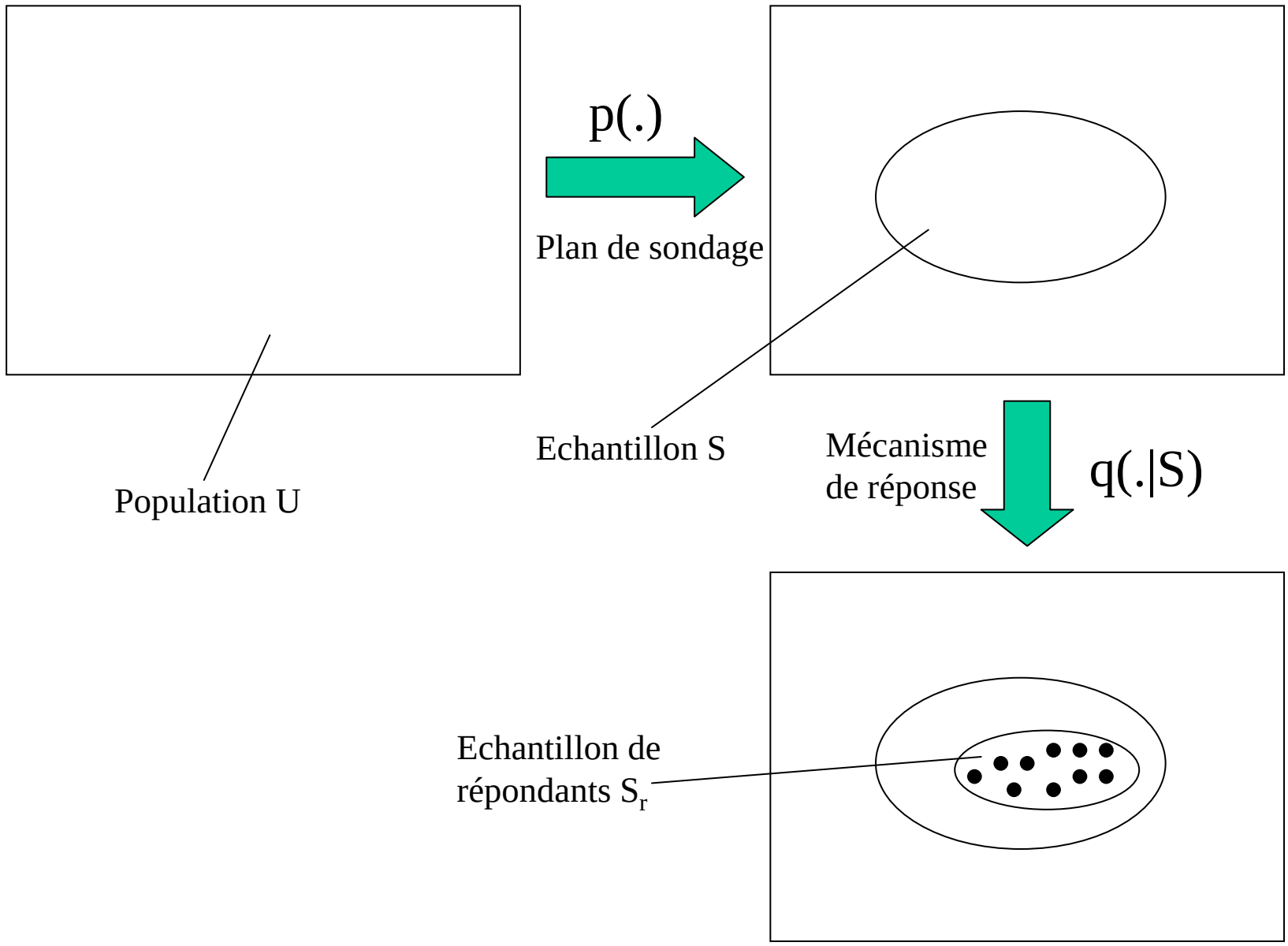
- la 1ère phase correspond à la sélection de l'échantillon S,
- la 2nde phase correspond à la "sélection" du sous-échantillon de répondants S_r
  $\Rightarrow$ mécanisme de non-réponse


Population U


p(.)
Plan de sondage
Population U


p(.)
Plan de sondage
Echantillon S
Population U


p(.)
Plan de sondage
Population U
Echantillon S
Mécanisme
de réponse
q(.|S)


p(.)
Plan de sondage
Population U
Echantillon S
Mécanisme
de réponse
q(.|S)
Echantillon de
répondants S_r

Mécanisme de non-réponse

On note r_k la variable indicatrice de réponse pour l'individu k, valant 1 si l'individu a répondu à l'enquête et 0 sinon.

On note $p_{k|S} \equiv p_k$ la probabilité de réponse pour l'unité k :

$$\begin{aligned} p_k &= \Pr(k \in S_r|S) \\ &= \Pr(r_k = 1|S). \end{aligned}$$

On fait l'hypothèse que :

- toutes les probabilités de réponse vérifient $0 < p_k \leq 1$: pas de non-répondants irréductibles,
- les individus répondent indépendamment les uns des autres :

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

Cette dernière hypothèse peut être affaiblie (Haziza et Rao, 2003 ; Skinner et D'Arrigo, 2011).

Types de mécanisme

On distingue schématiquement 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.

Types de mécanisme

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

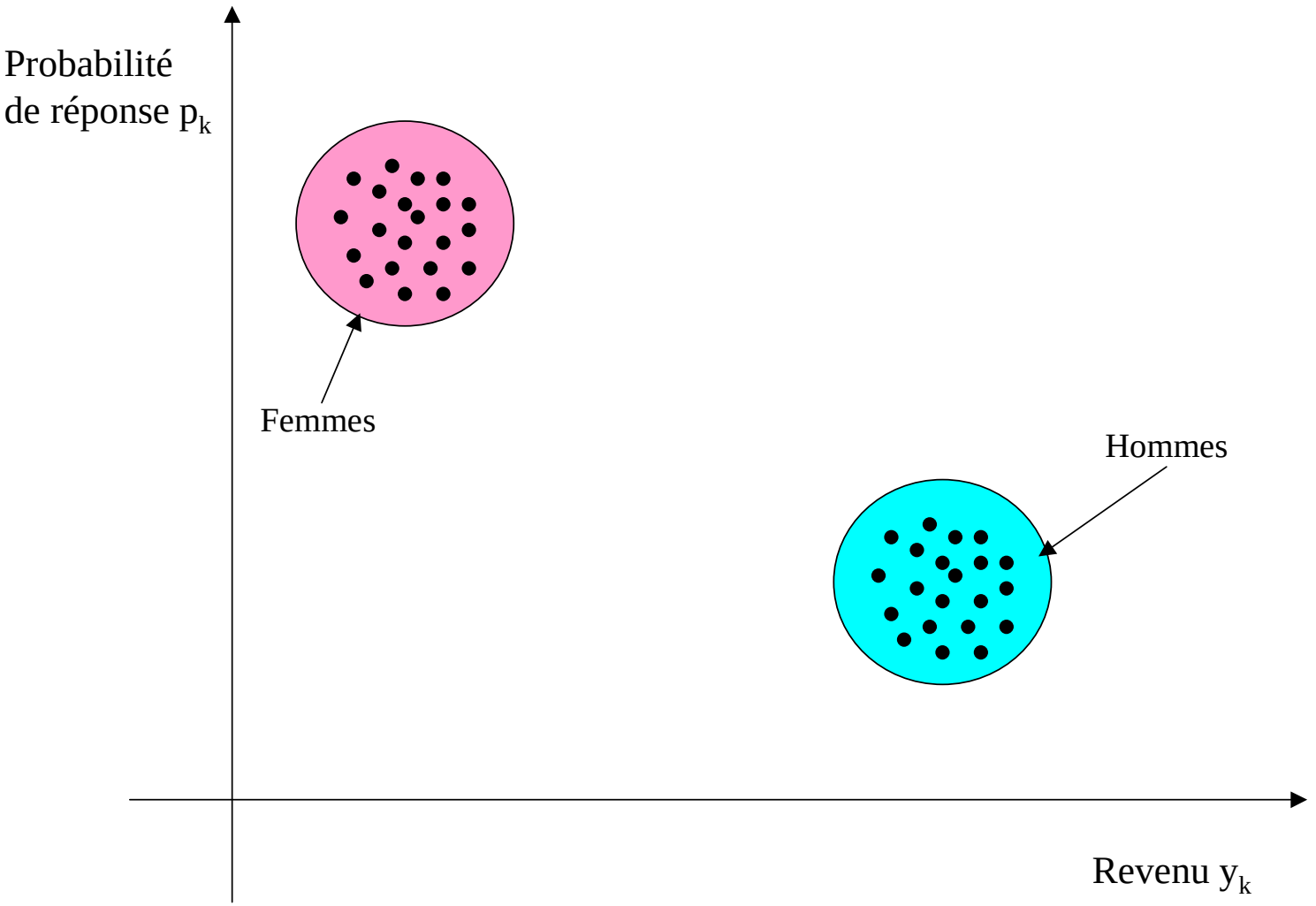
$$\Pr(r_k = 1 | y_k, \mathbf{z}_k) \quad = \quad \Pr(r_k = 1 | \mathbf{z}_k),$$

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 le sexe des individus.

Un exemple de non-réponse MAR



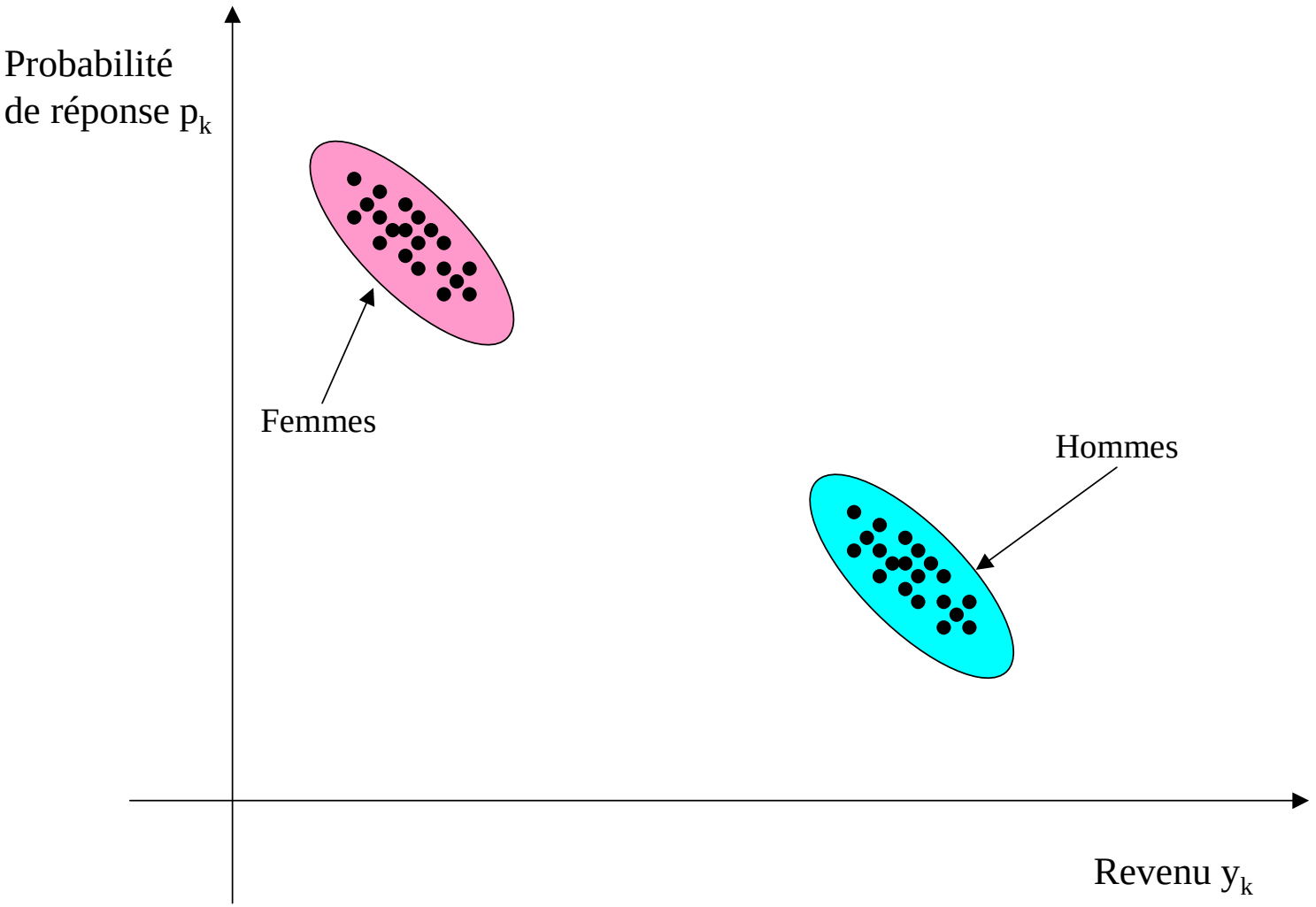
Types de mécanisme

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 très 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 sexe $\times$ revenu.

Un exemple de non-réponse NMAR



Exemple sur données simulées

On considère une population artificielle contenant 250 femmes et 250 hommes, et une variable d'intérêt y (revenu) générée selon le modèle

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

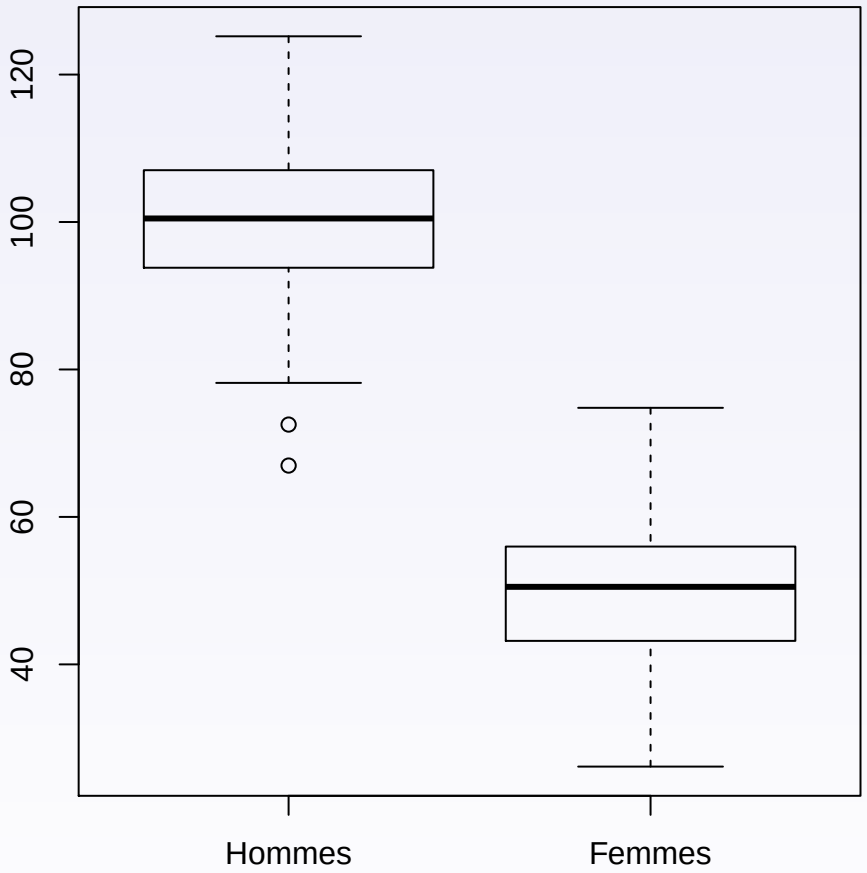
avec les ϵ_k générés selon une loi Normale$(0, 1)$.

On considère deux jeux de probabilités de réponse :

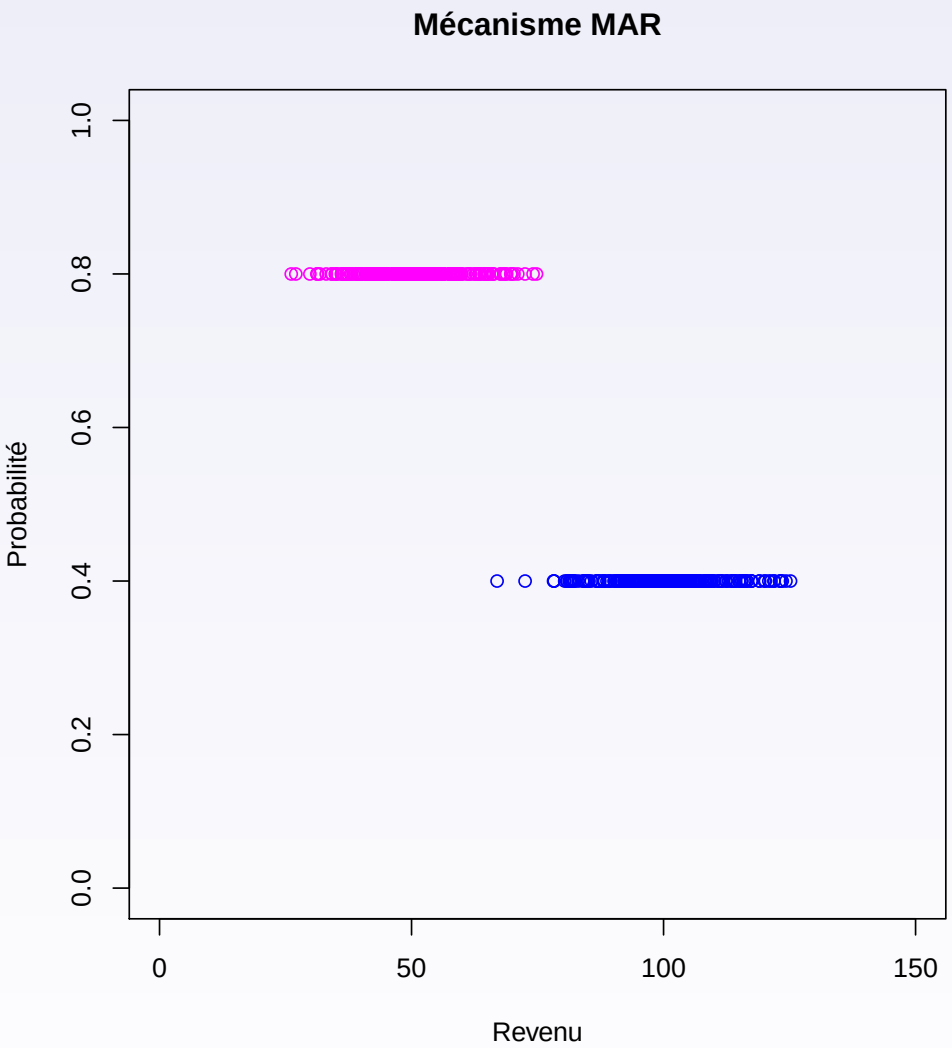
- mécanisme MAR : $p_{1k} = 0.8$ pour les femmes et $p_{1k} = 0.4$ pour les hommes,
- mécanisme NMAR : $p_{1k} = \dfrac{\exp^{8.5-0.1\times y}}{1+\exp^{8.5-0.1\times y}}$.

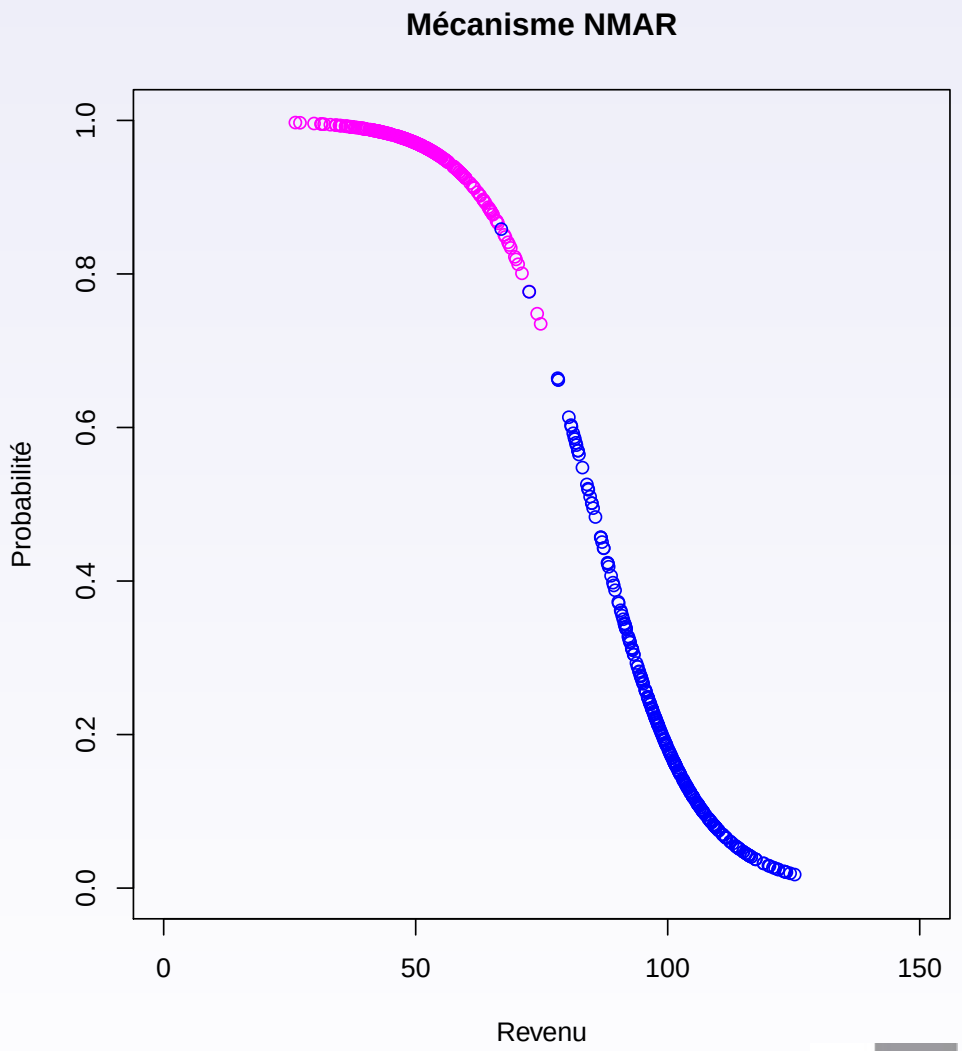
On obtient une probabilité de réponse moyenne de 0.60 environ dans chaque cas.

Distribution des revenus par sexe

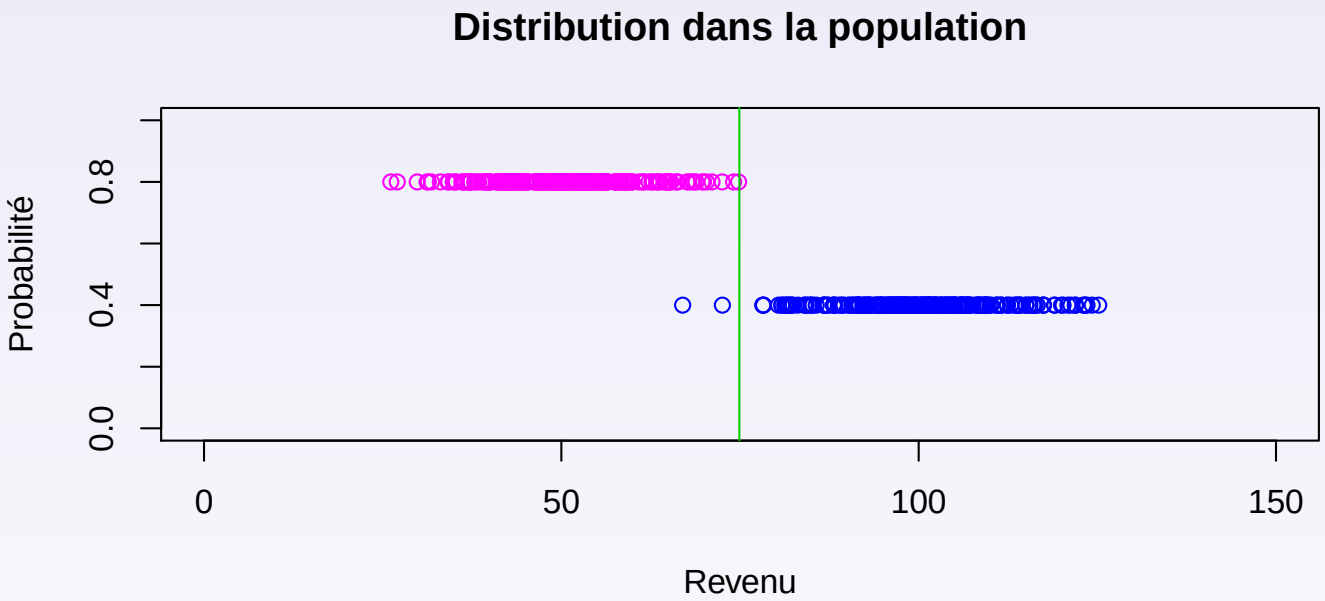


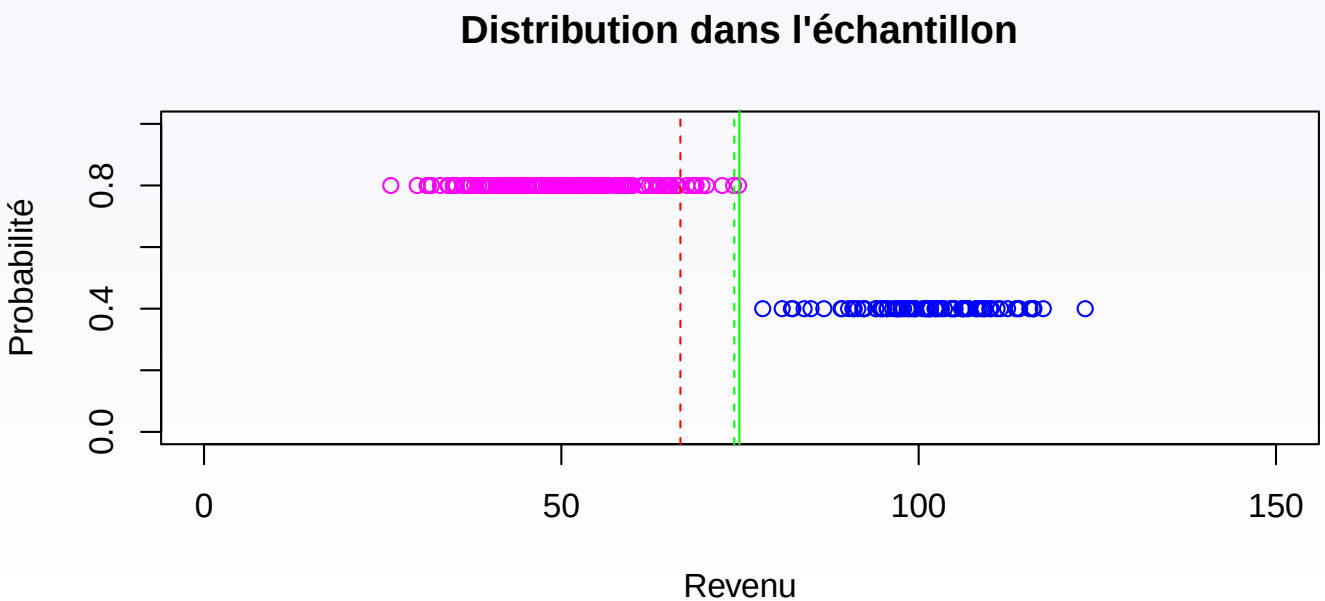
Mécanismes de réponse



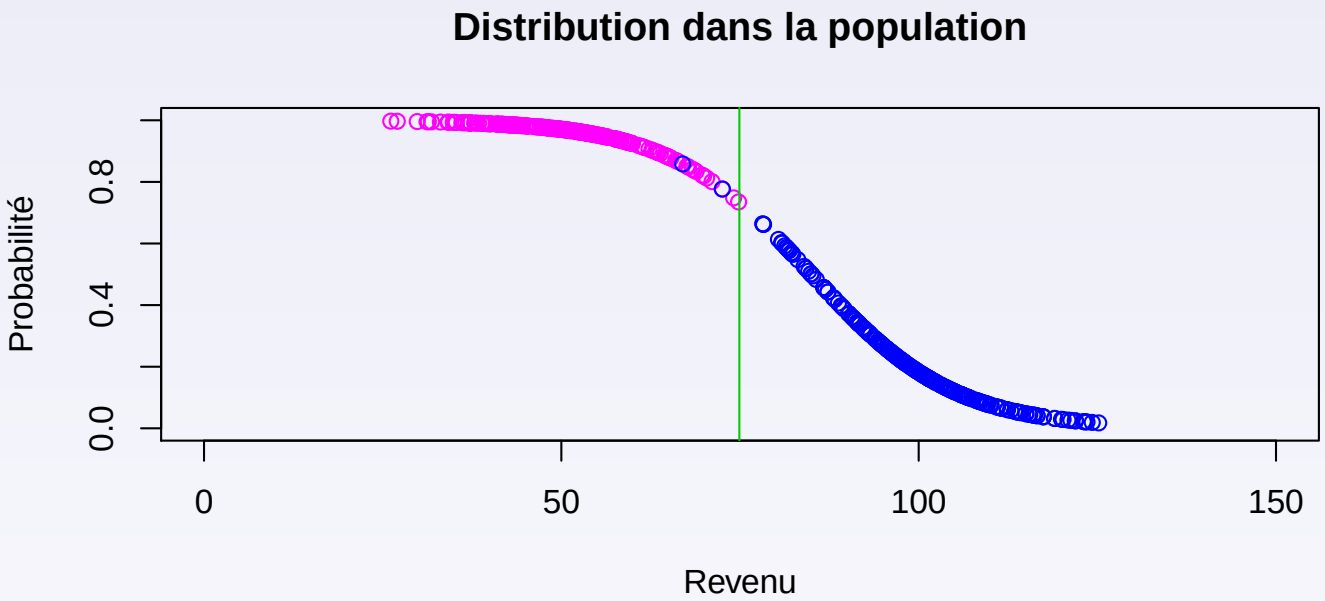


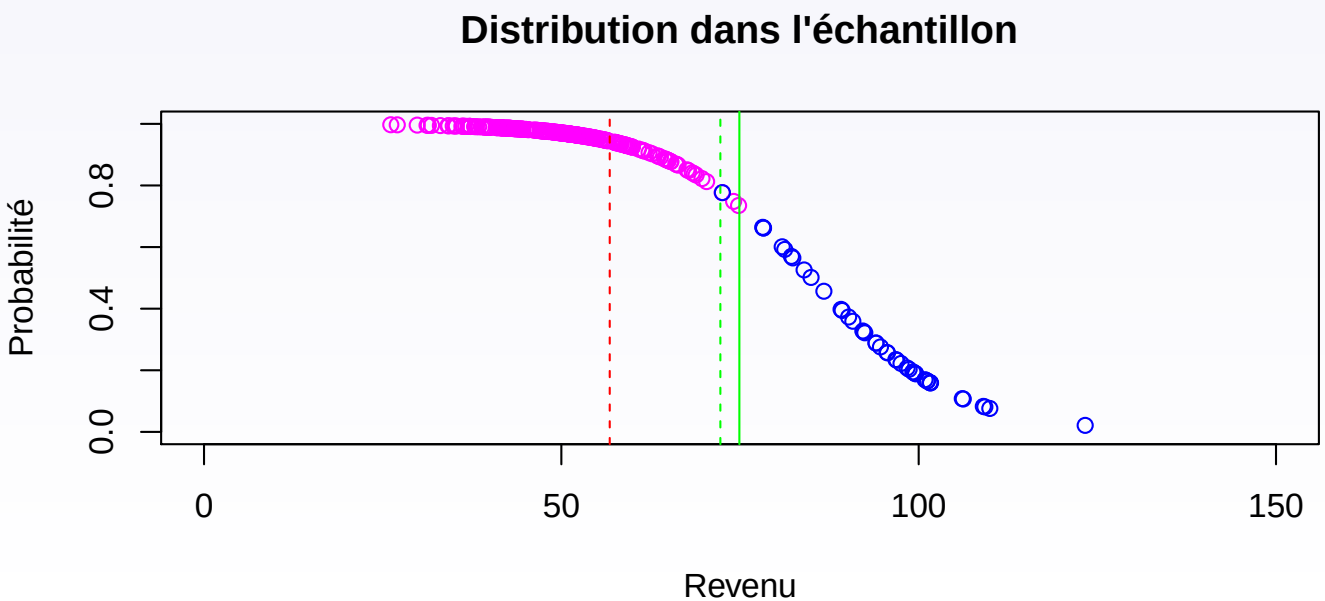
Estimation sous un mécanisme MAR





Estimation sous un mécanisme NMAR





Estimation d'un total
Cas de probabilités de réponse connues

Estimation par expansion

Si les probabilités p_k sont connues, on se trouve dans le cas d'un échantillonnage en deux phases. On peut utiliser l'**estimateur par expansion**

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

Sous la modélisation utilisée, le mécanisme de non-réponse est vu comme un plan poissonien dans l'échantillon d'origine S.

Remarque : il ne s'agit pas de l'estimateur de Horvitz-Thompson. Les probabilités d'inclusion finales

$$Pr(k \in S_r) = \sum_{s \subset U; \; k \in s} p(s) p_{k|s}$$

sont généralement impossibles à calculer.

Estimateur par expansion

L'estimateur par expansion est sans biais pour le total t_y :

$$\begin{aligned} E(\hat{t}_{ye}) &= E_p E_q(\hat{t}_{ye}|S) \\ &= E_p(\hat{t}_{y\pi}) = t_y. \end{aligned}$$

La variance de l'estimateur par expansion est donnée par :

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

Elle est donc toujours plus grande qu'en situation de réponse complète.

Estimation par expansion

Si on utilise un plan de taille fixe pour sélectionner l'échantillon S, la variance peut se réécrire :

$$\begin{aligned} V(\hat{t}_{ye}) &= -\frac{1}{2}\sum_{k\neq l\in U}\left(\frac{y_k}{\pi_k}-\frac{y_l}{\pi_l}\right)^2(\pi_{kl}-\pi_k\pi_l) \\ &+ E_p\left[\sum_{k\in S}\frac{1-p_k}{p_k}\left(\frac{y_k}{\pi_k}\right)^2\right]. \end{aligned}$$

On peut l'estimer sans biais par :

$$\begin{aligned} v(\hat{t}_{ye}) &= -\frac{1}{2}\sum_{k\neq l\in S_r}\left(\frac{y_k}{\pi_k}-\frac{y_l}{\pi_l}\right)^2\frac{\pi_{kl}-\pi_k\pi_l}{\pi_{kl}\ p_k\ p_l} \\ &+ \sum_{k\in S_r}\frac{1-p_k}{p_k^2}\left(\frac{y_k}{\pi_k}\right)^2. \end{aligned}$$

Estimation par expansion : cas d'un SRS

Dans le cas particulier d'un échantillon S tiré selon un SRS(n), on obtient :

$$V(\hat{t}_{ye}) = \frac{N^2}{n}\left[(1-f)S_y^2 + E_p\left(\frac{1}{n}\sum_{k\in S}\frac{y_k^2(1-p_k)}{p_k}\right)\right],$$

que l'on peut estimer par

$$v\left[\hat{t}_{ye}\right] = \frac{N^2}{n}\left[(1-f)s_{yr}^2 + \frac{1}{n}\sum_{k\in S_r}\frac{y_k^2(1-p_k)}{p_k^2}\right],$$

avec

$$s_{yr}^2 = \frac{1}{2n(n-1)}\sum_{k\neq l\in S_r}\frac{(y_k-y_l)^2}{p_k\ p_l}$$

un estimateur sans biais de S_y^2 calculé sur l'échantillon de répondants S_r.

Estimation d'un total
Cas de probabilités de réponse inconnues

Estimation des probabilités de réponse

En pratique, les probabilités de réponse p_k sont inconnues et doivent être estimées. On postule 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.

Le choix (couramment utilisé en pratique)

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

correspond au modèle logistique, avec $\text{logit}(p_k) = \mathbf{z}_k^\top \beta_0$.

D'autres fonctions de lien sont possibles. On peut également utiliser une modélisation non paramétrique (Da Silva et Opsomer, 2006 et 2009).

Estimation des probabilités de réponse

On peut obtenir (par exemple, à l'aide de la PROC LOGISTIC de SAS) un estimateur du paramètre β_0 en résolvant l'équation estimante :

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

On note $\hat{\beta}$ la solution de cette équation. On a

$$\hat{\beta} - \beta_0 \simeq \left[\sum_{k \in S} p_k (1 - p_k) \mathbf{z}_k \mathbf{z}_k^\top \right]^{-1} \sum_{k \in S} (r_k - p_k) \mathbf{z}_k. \qquad (2)$$

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

Estimateur du total

On obtient l'estimateur corrigé de la non-réponse totale

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

que l'on peut réécrire sous la forme :

$$\begin{aligned}
\hat{t}_{yr} &= \hat{t}_{ye} + \sum_{k \in S} \frac{r_k y_k}{\pi_k} \left(\frac{1}{\hat{p}_k} - \frac{1}{p_k} \right) \\
&\simeq \hat{t}_{ye} - \left[\sum_{k \in S} \frac{1 - p_k}{\pi_k} \mathbf{z}_k^\top y_k \right]^\top [\hat{\beta} - \beta_0] \qquad (3) \\
&\simeq \hat{\gamma}^\top \sum_{k \in S} p_k \mathbf{z}_k + \sum_{k \in S} \frac{r_k}{p_k} \left(\frac{y_k}{\pi_k} - p_k \hat{\gamma}^\top \mathbf{z}_k \right), \qquad (4)
\end{aligned}$$

avec

$$\hat{\gamma} = \left[\sum_{k \in S} p_k (1 - p_k) \mathbf{z}_k \mathbf{z}_k^\top \right]^{-1} \sum_{k \in S} \frac{1 - p_k}{\pi_k} \mathbf{z}_k y_k. \qquad (5)$$

Propriétés de l'estimateur du total

En utilisant les expressions (2) et (3), on obtient

$$E_q(\hat{t}_{yr}|S) \simeq E_q(\hat{t}_{ye}|S) = \hat{t}_{y\pi}, \tag{6}$$

et l'estimateur $\hat{t}_{yr}$ est approximativement sans biais pour t_y.

En utilisant les expressions (4) et (6), on obtient

$$\begin{aligned} V(\hat{t}_{yr}) &= V_p E_q(\hat{t}_{yr}|S) + E_p V_q(\hat{t}_{yr}|S) \\ &\simeq V_p(\hat{t}_{y\pi}) + E_p\left[\sum_{k\in S}\frac{1-p_k}{p_k}\left(\frac{y_k}{\pi_k} - p_k\hat{\gamma}^\top \mathbf{z}_k\right)^2\right]. \end{aligned}$$

Cette variance est généralement **plus faible** que celle de l'estimateur par expansion, utilisant les vraies probabilités de réponse.

Propriétés de l'estimateur du total

Si on utilise un plan de taille fixe pour sélectionner l'échantillon S, la variance peut se réécrire :

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

On peut l'estimer approximativement sans biais par :

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

avec

$$\hat{\gamma}_r = \left[\sum_{k \in S_r} (1 - \hat{p}_k) \mathbf{z}_k \mathbf{z}_k^\top \right]^{-1} \sum_{k \in S_r} \frac{1 - \hat{p}_k}{\pi_k\ \hat{p}_k} \mathbf{z}_k y_k. \tag{7}$$

Cas des groupes homogènes de réponse

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 :

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

On les appelle 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.

Détermination des GHR

En pratique, on peut constituer ces groupes 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).

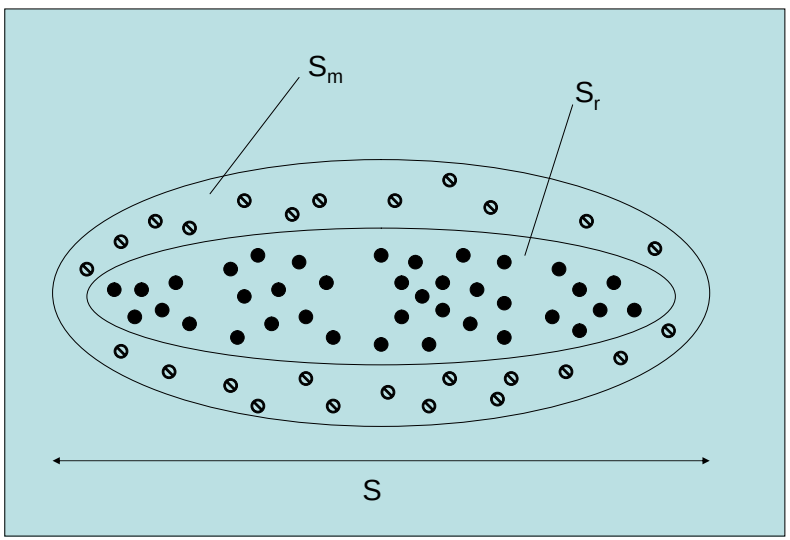
Cas des groupes homogènes de réponse

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.



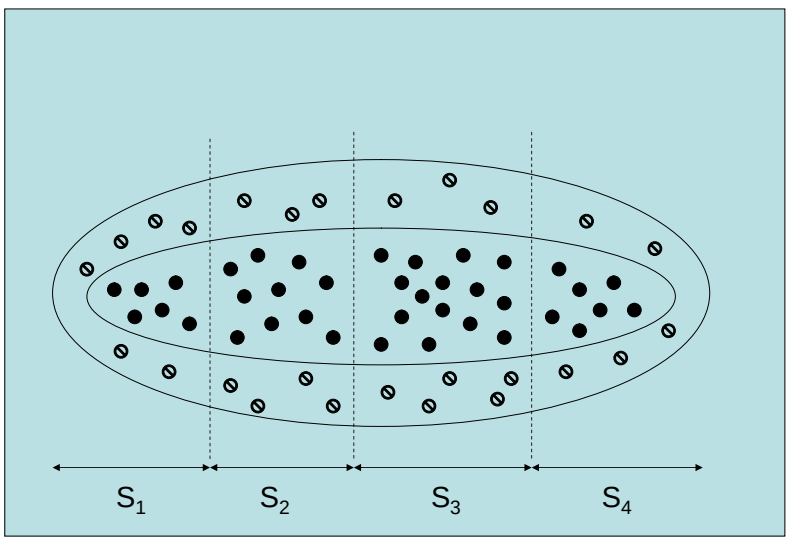
Cas des groupes homogènes de réponse

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.



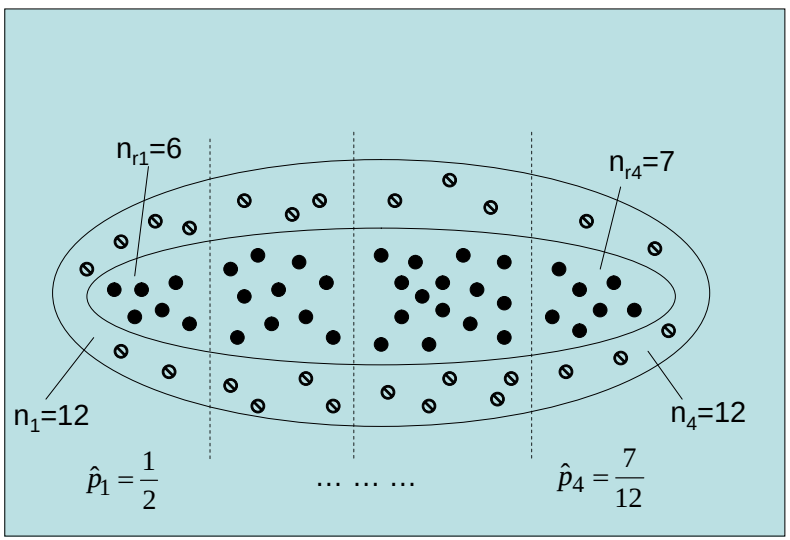
Cas des groupes homogènes de réponse

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.



Estimation

Avec le modèle correspondant aux GHR, on a :

- $\mathbf{z}_k = [1(k \in S_1), \ldots, 1(k \in S_C)]^\top$,
- $\hat{\beta} = \left[\hat{\beta}_1, \ldots, \hat{\beta}_C\right]^\top$ avec $\hat{\beta}_c = \ln\left(\frac{n_{rc}}{n_{mc}}\right)$,
- $\hat{p}_k = \hat{p}_c = \frac{n_{rc}}{n_c}$ pour $k \in S_c$,
- $\hat{\gamma} = [\hat{\gamma}_1, \ldots, \hat{\gamma}_C]^\top$ avec $\hat{\gamma}_c = \frac{1}{n_c} \sum_{k \in S_c} \frac{y_k}{\pi_k \ p_c}$,
- $\hat{\gamma}_r = [\hat{\gamma}_{r1}, \ldots, \hat{\gamma}_{rC}]^\top$ avec $\hat{\gamma}_{rc} = \frac{1}{n_{rc}} \sum_{k \in S_{rc}} \frac{y_k}{\pi_k \ \hat{p}_c}$.

On obtient $\hat{t}_{yr} = \sum_{c=1}^{C} \frac{n_c}{n_{rc}} \sum_{k \in S_{rc}} \frac{y_k}{\pi_k}$, et si on utilise un plan de taille fixe pour sélectionner S :

$$\begin{aligned} v(\hat{t}_{yr}) &= -\frac{1}{2} \sum_{k \neq l \in S_r} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2 \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl} \ \hat{p}_k \ \hat{p}_l} \\ &+ \sum_{c=1}^{C} \frac{1 - \hat{p}_c}{(\hat{p}_c)^2} \sum_{k \in S_{rc}} \left(\frac{y_k}{\pi_k} - \frac{1}{n_{rc}} \sum_{l \in S_{rc}} \frac{y_l}{\pi_l}\right)^2 . \end{aligned}$$

Estimateur redressé de la non-réponse : cas d'un SRS

Dans le cas particulier d'un échantillon S tiré selon un SRS(n), on obtient :

$$\hat{t}_{yr} = N \sum_{c=1}^{C} \frac{n_c}{n} \bar{y}_{rc} \quad \text{avec} \quad \bar{y}_{rc} = \frac{1}{n_{rc}} \sum_{k \in S_{rc}} y_k.$$

Sa variance peut être estimée par

$$v(\hat{t}_{yr}) = \frac{N^2}{n} \left[(1-f) s_{yr}^2 + \sum_{c=1}^{C} \frac{1-\hat{p}_c}{(\hat{p}_c)^2} \times \frac{n_{rc}-1}{n} s_{y,rc}^2 \right],$$

avec

$$s_{yr}^2 = \frac{1}{2n(n-1)} \sum_{k \neq l \in S_r} \frac{(y_k - y_l)^2}{\hat{p}_k \, \hat{p}_l},$$

$$s_{y,rc}^2 = \frac{1}{n_{rc}-1} \sum_{k \in S_{rc}} (y_k - \bar{y}_{rc})^2.$$

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.

Exemple sur données réelles

On considère une population de $N = 10,000$ individus extraite de l'enquête canadienne sur la santé (CCHS). On s'intéresse à l'estimation de la taille moyenne et du poids moyen des individus.

On dispose 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).

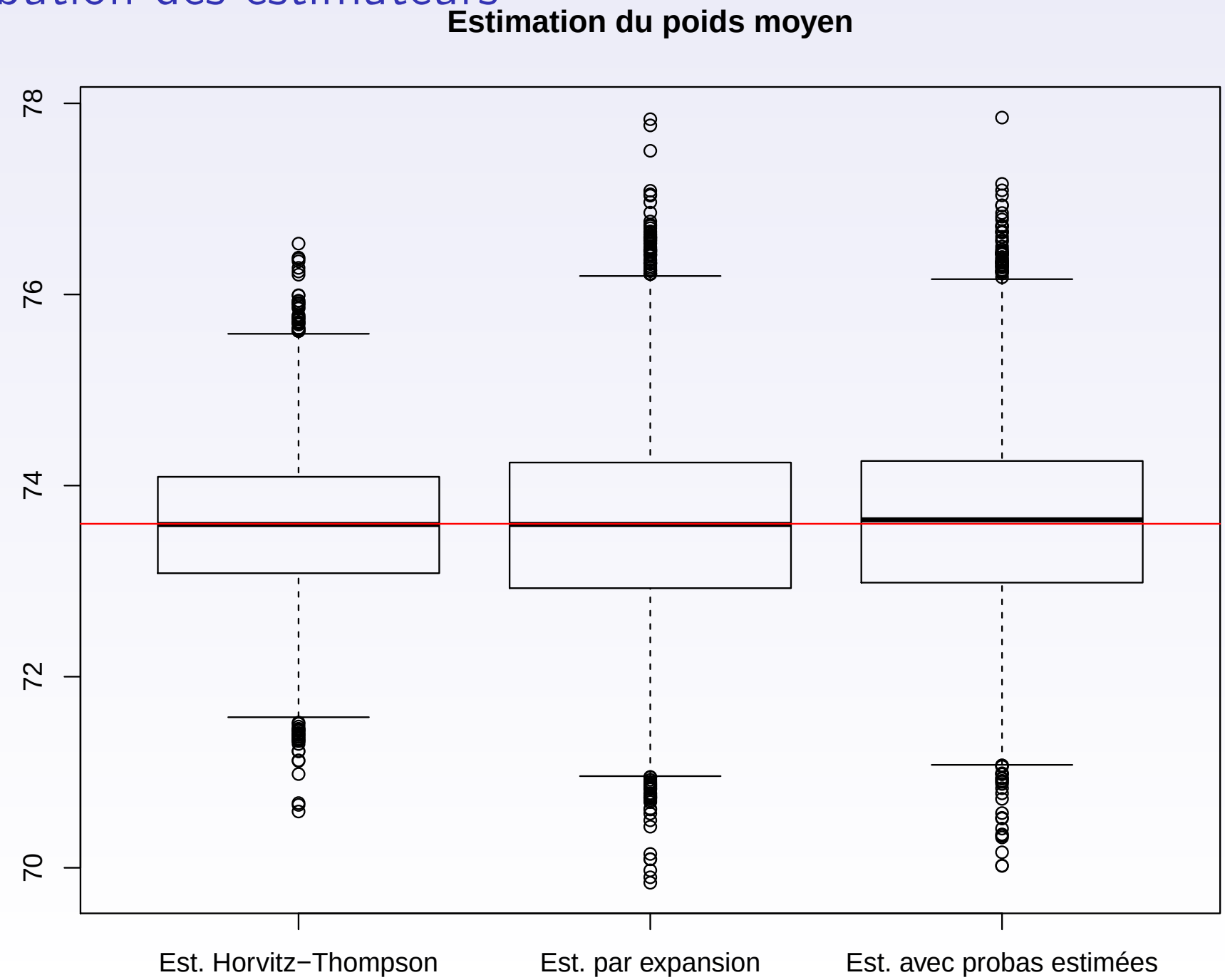
Exemple sur données réelles (2)

On sélectionne un échantillon de taille $n = 500$ selon un SRS. On considère le mécanisme de réponse (inconnu) :

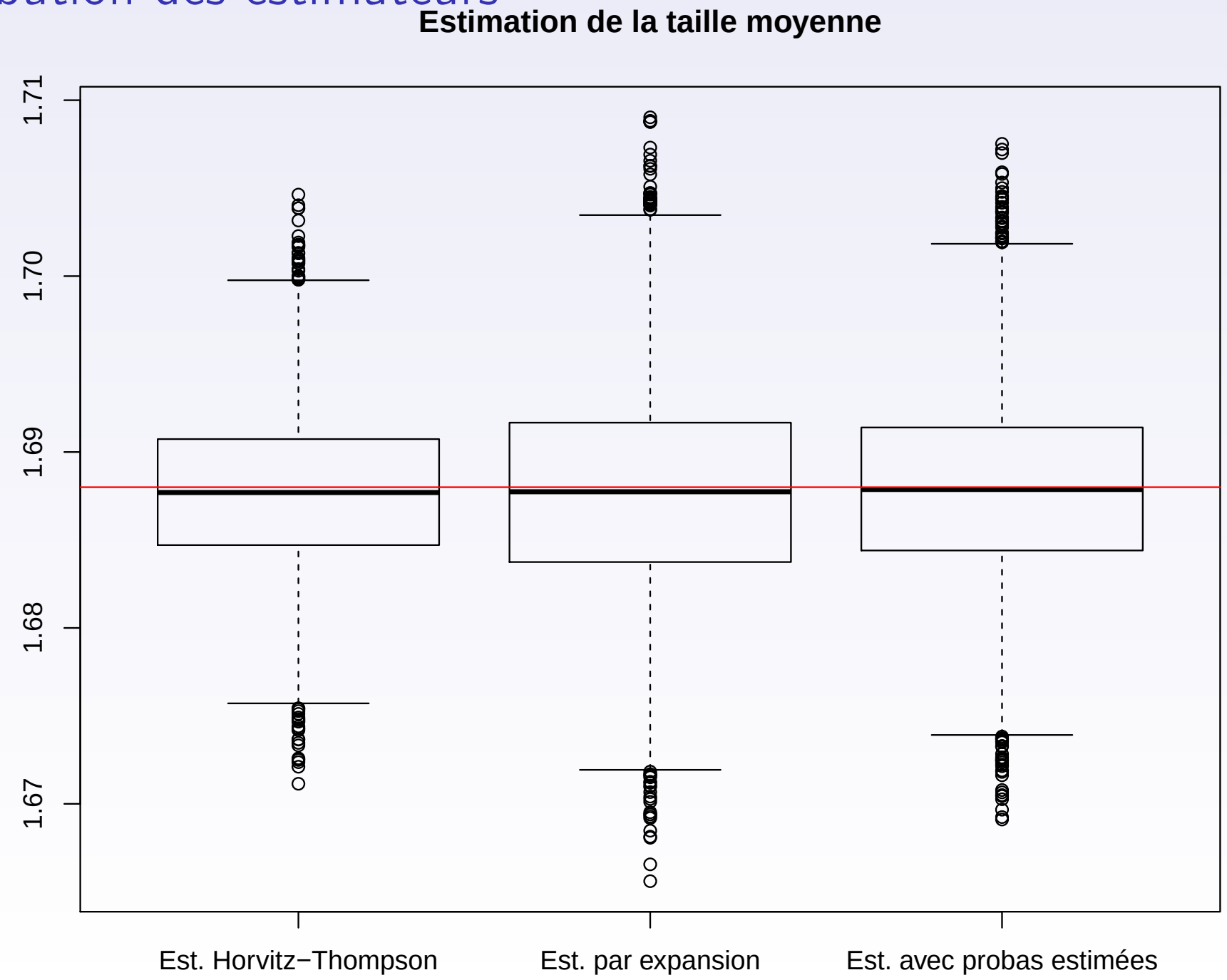
$$\begin{array}{rrrrr} & & & & +0.70(st_k = 4) \\ & & & -0.05(a_k = 3) & -0.50(st_k = 3) \\ & & -0.60(s_k = 2) & +0.15(a_k = 2) & +0.50(st_k = 2) \\ logit(p_{1k}) = & 0.80 & +0.60(s_k = 1) & -0.10(a_k = 1) & -0.70(st_k = 1) \end{array}$$

La probabilité de réponse moyenne est égale à 0.62 environ.

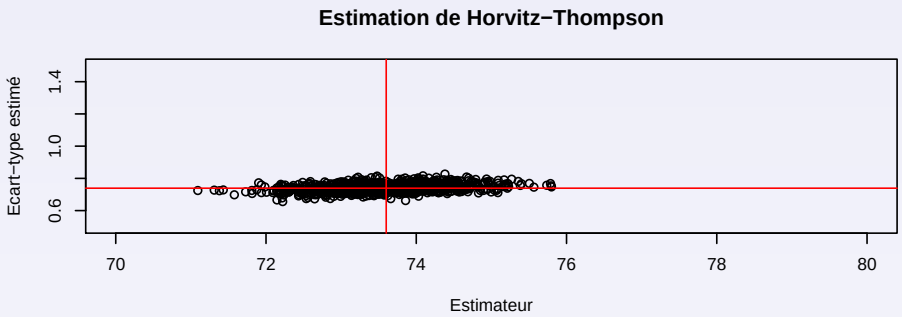
Distribution des estimateurs

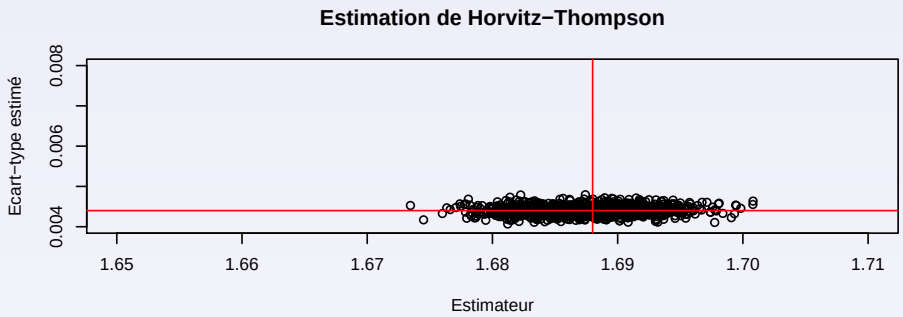


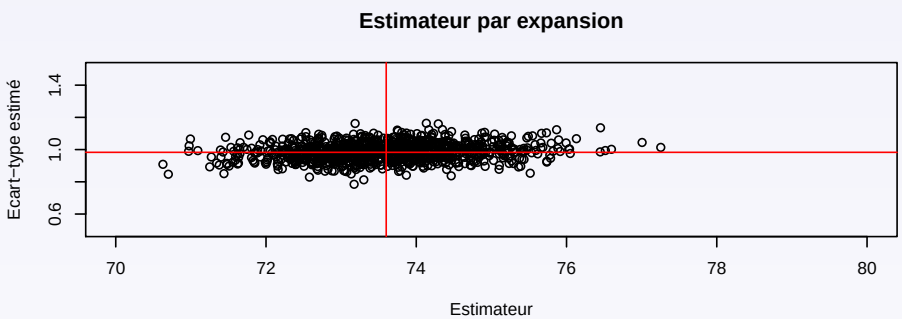
Distribution des estimateurs

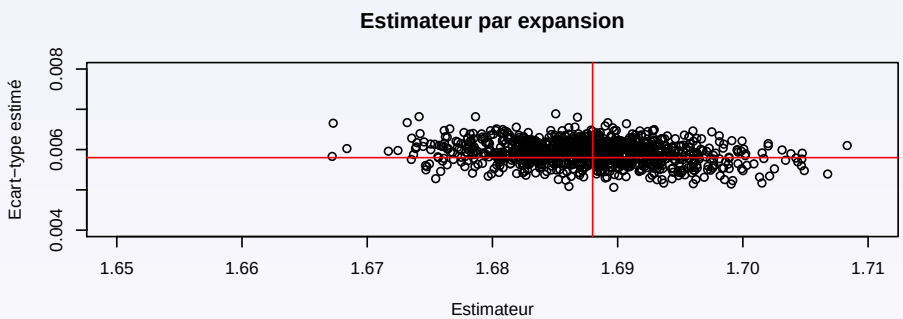


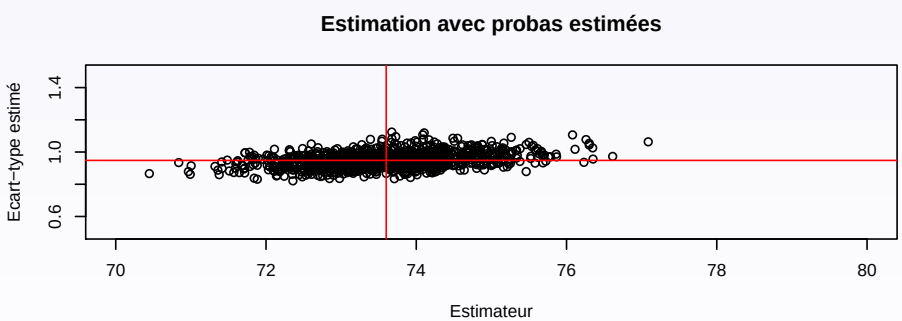
Ecart-type estimé en fonction de l'estimateur

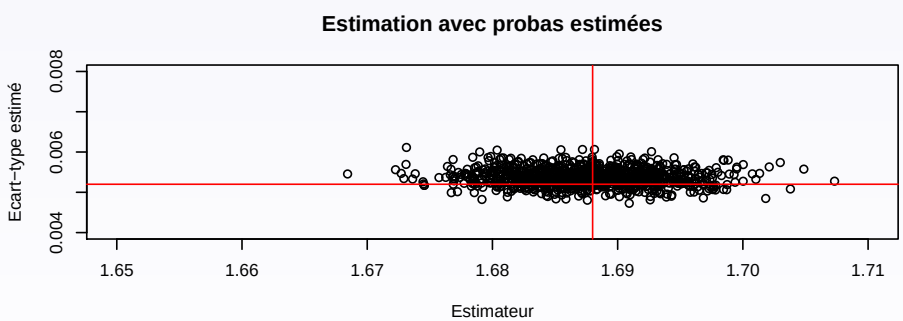












Traitement de la non-réponse partielle

Le problème

La non-réponse partielle ("item non-response") survient lorsqu'une unité répond à l'enquête, mais renseigne une partie des variables seulement.

On va traiter ce problème par imputation : une valeur manquante est remplacée par une valeur plausible. Cette imputation se justifie sous une modélisation de la variable d'intérêt appelée le modèle d'imputation.

L'imputation permet de recréer un fichier de données complet, ce qui facilite l'analyse. En revanche, elle perturbe les relations entre les variables et peut donner une impression artificielle de précision si l'imputation n'est pas prise en compte dans les calculs de variance.

Les étapes du traitement de la non-réponse 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

Deux points importants :

- 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.

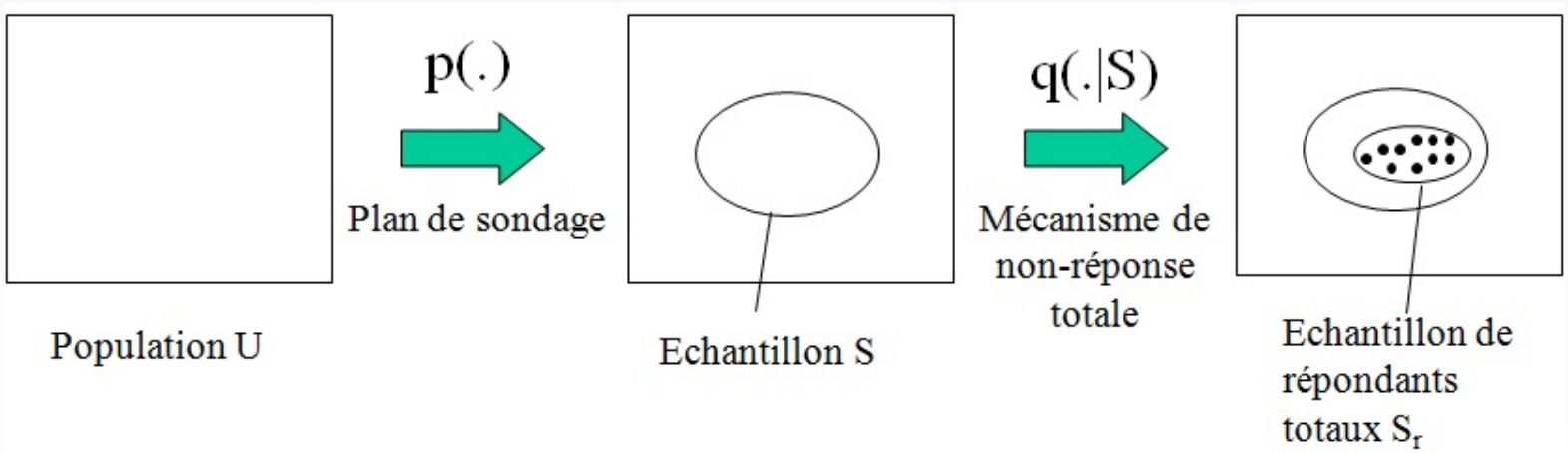
Point 2 : ne pas traiter par imputation l'absence d'une valeur y_k due à la forme du questionnaire (question filtre).

Le modèle d'imputation

Estimateur imputé

Pour simplifier, nous nous plaçons dans le cas où l'échantillon S ne présente pas de non-réponse totale ; on note d_k le poids (éventuellement calé) d'un individu k.

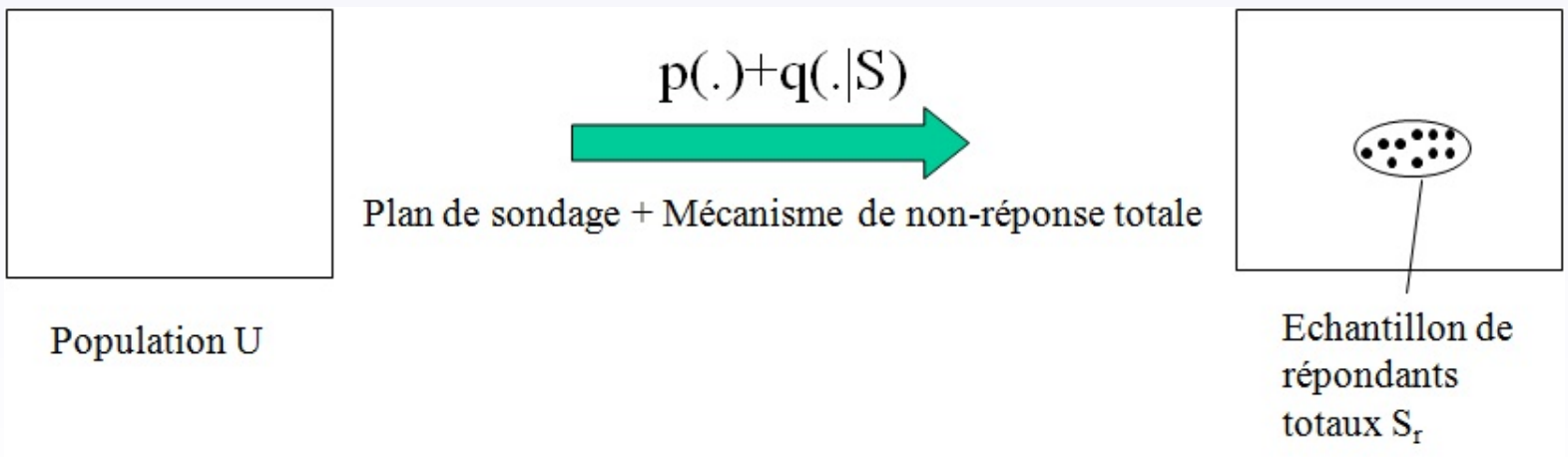
Alternativement, on peut voir S comme le résultat d'un tirage en deux phases (une correspondant au plan de sondage, l'autre au mécanisme de non-réponse totale).



Estimateur imputé

Pour simplifier, nous nous plaçons dans le cas où l'échantillon S ne présente pas de non-réponse totale ; on note d_k le poids (éventuellement calé) d'un individu k.

Alternativement, on peut voir S comme le résultat d'un tirage en deux phases (une correspondant au plan de sondage, l'autre au mécanisme de non-réponse totale) :



Estimateur imputé

On note $p(\cdot)$ le mécanisme de sélection de l'échantillon S. En l'absence de non-réponse partielle pour la variable y, le total t_y peut être estimé 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^*.

On note

- $S_{ry} \equiv S_r$ 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.

Estimateur imputé

L'estimateur imputé est donné par

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

L'erreur totale $\hat{t}_{yI} - t_y$ peut se décomposer sous la forme

$$\hat{t}_{yI} - t_y = \left(\hat{t}_y - t_y\right) + \left(\hat{t}_{yI} - \hat{t}_y\right),$$

avec

- $\hat{t}_y - t_y \Rightarrow$ erreur d'échantillonnage (+ non-réponse totale),
- $\hat{t}_{yI} - \hat{t}_y \Rightarrow$ erreur due à la non-réponse partielle et à l'imputation.

Au stade de la correction de la non-réponse partielle, le premier terme d'erreur est incompressible.

Estimateur imputé

L'objectif de l'imputation est de limiter au maximum l'erreur due à la non-réponse partielle

$$\hat{t}_{yI} - \hat{t}_y = \sum_{k \in S_m} d_k (y_k^* - y_k).$$

L'erreur d'imputation sera limitée :

- si les valeurs imputées y_k^* sont proches des valeurs réelles y_k ;
- ou si les écarts entre valeurs imputées y_k^* et valeurs réelles y_k se compensent en moyenne.

Pour créer des valeurs imputées y_k^* aussi proches que possible des valeurs réelles y_k, on va mobiliser l'information auxiliaire disponible sur S pour proposer une **modélisation raisonnable** de la variable d'intérêt.

Modèle d'imputation

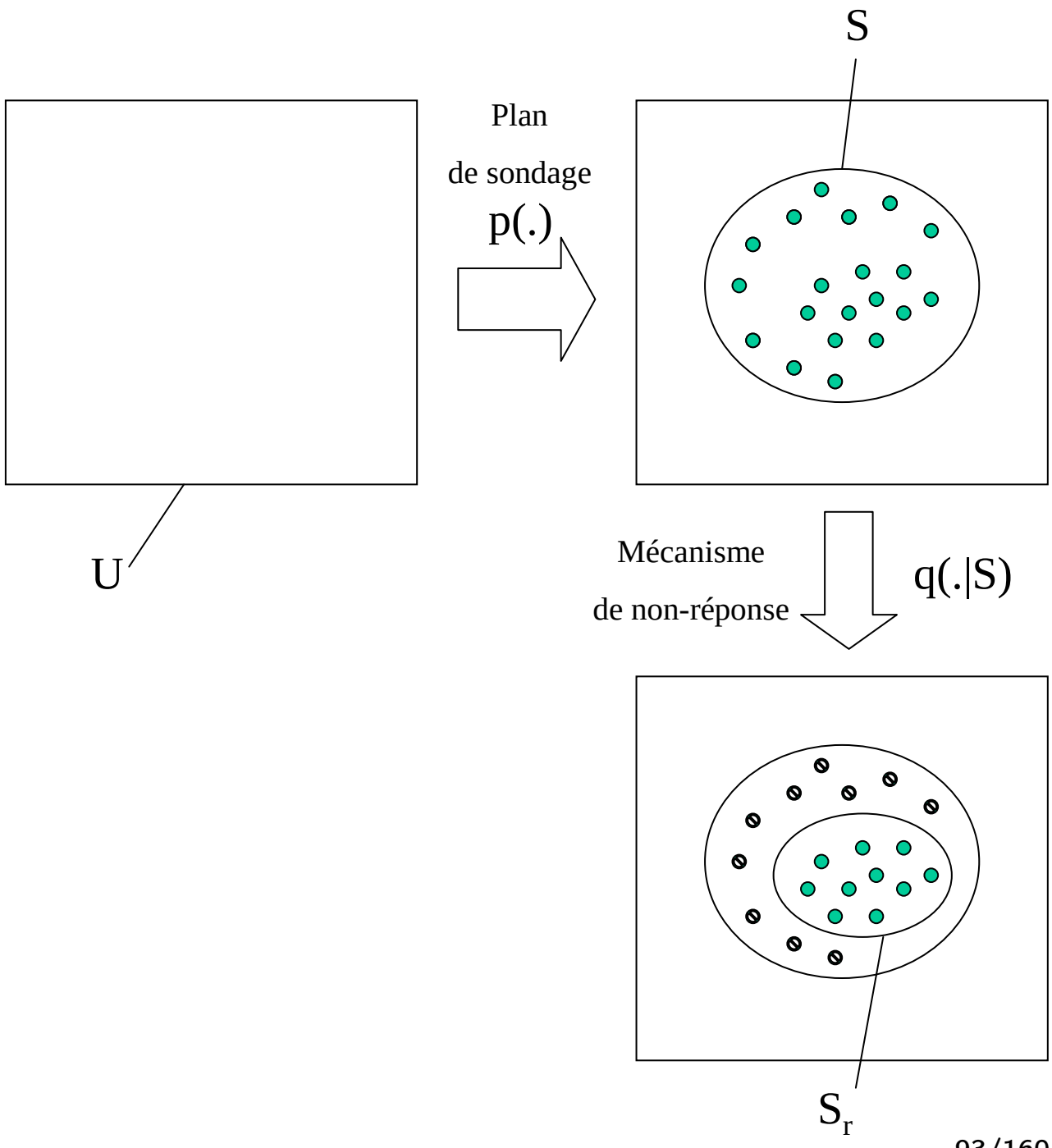
Le **mécanisme d'imputation** est généralement motivé par un **modèle d'imputation** (par exemple, un modèle de régression) qui vise à prédire la variable y_k à l'aide d'une information auxiliaire $\mathbf{z}_k$ disponible sur l'ensemble de l'échantillon.

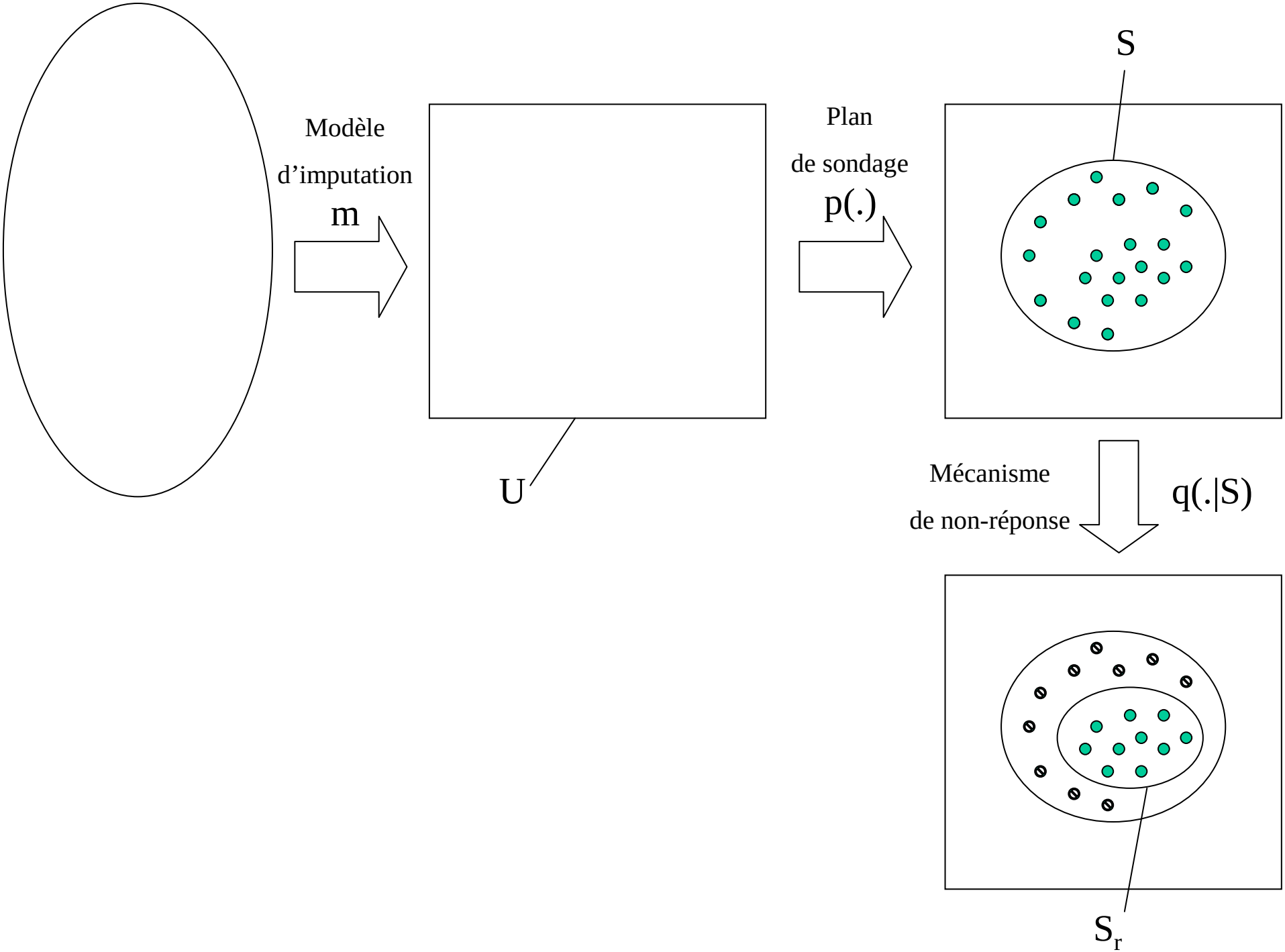
$$m \ : \ y_k \ = \ \mathbf{z}_k^\top \boldsymbol{\beta} + \sigma \sqrt{v_k} \epsilon_k \quad \text{pour} \quad k \in S. \tag{8}$$

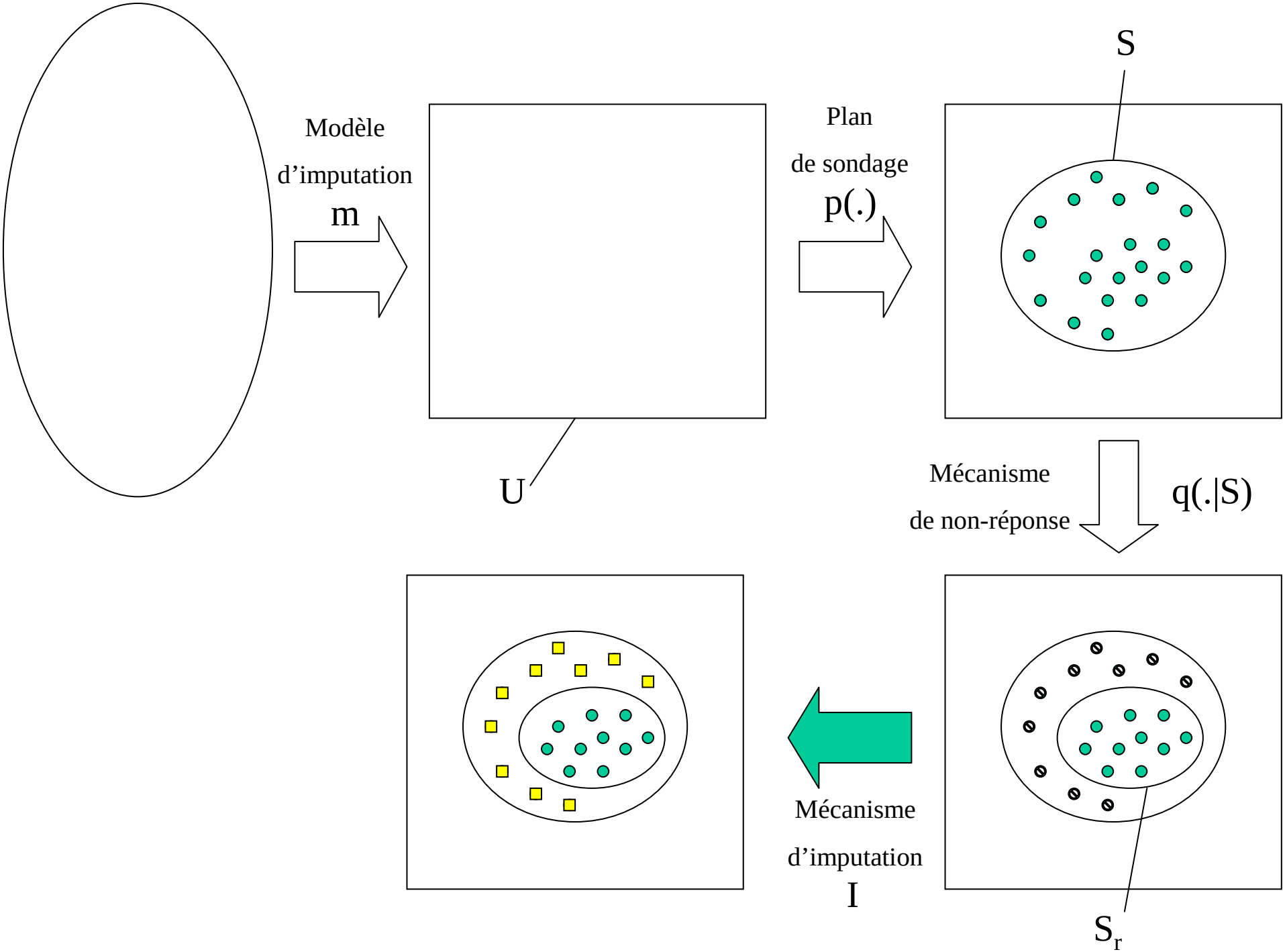
Dans ce modèle :

- $\boldsymbol{\beta}$ et σ^2 sont des paramètres inconnus,
- v_k est une constante connue,
- les résidus ϵ_k sont des variables aléatoires iid, par exemple 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
Plan
de sondage
p(.)
S
U
Mécanisme
de non-réponse
q(.|S)
Mécanisme
d'imputation
I
S_r

Propriétés de l'estimateur imputé

Sous la modélisation utilisée, le biais de l'estimateur imputé s'écrit :

$$\begin{aligned} B(\hat{t}_{yI}) &= E_{mpqI}\left(\hat{t}_{yI} - t_y\right) \\ &= E_{mp}\left(\hat{t}_y - t_y\right) + E_{mpqI}\left(\hat{t}_{yI} - \hat{t}_y\right) \\ &\simeq E_{mpqI}\left(\hat{t}_{yI} - \hat{t}_y\right). \end{aligned}$$

Le mécanisme d'imputation utilisé a pour objectif de rendre ce biais (approximativement) nul, sous des hypothèses de modélisation raisonnables.

Un objectif secondaire est de limiter la variance de l'estimateur imputé, en utilisant un mécanisme d'imputation efficace. Pour des paramètres plus complexes (tels que la médiane), il est souvent difficile de limiter à la fois le biais et la variance.

Méthodes d'imputation

Types de méthodes

On peut classer les méthodes d'imputation 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.

On peut ajouter une troisième famille de méthodes, transversale. Les **méthodes d'imputation par donneur** consistent à piocher un individu parmi les répondants, et à utiliser la valeur observée pour la variable y pour remplacer la valeur manquante.

Imputation déterministe

Mécanisme d'imputation par la régression

L'**imputation par la régression déterministe** s'appuie sur le modèle (8) :

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

avec

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

où ω_k désigne un **poids d'imputation** attaché à l'unité k (Haziza, 2009). On utilise généralement $\omega_k = 1$ (imputation non pondérée) ou $\omega_k = d_k$ (imputation pondérée par les poids de sondage).

L'estimateur imputé est égal à

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

Mécanisme d'imputation par la régression

Cet estimateur est approximativement sans biais sous la modélisation utilisée :

$$\begin{aligned} B(\hat{t}_{yI}) &\simeq E_{mpqI}\left(\hat{t}_{yI} - \hat{t}_y\right) \\ &= E_{mpq}\left[\sum_{k\in S_m} d_k(y_k^* - y_k)\right] \\ &\simeq 0. \end{aligned}$$

L'hypothèse fondamentale est que le vecteur $\mathbf{z}_k$ permette une bonne prédiction $y_k^* = \mathbf{z}_k^\top \hat{\boldsymbol{\beta}}_r$ de la valeur manquante.

La précision de l'estimateur peut être évaluée par :

- la variance totale $V_{mpq}\left(\hat{t}_{yI} - t_y\right)$,
- la **variance anticipée** $E_m V_{pq}\left(\hat{t}_{yI} - t_y\right) = E_m V_{pq}\left(\hat{t}_{yI}\right)$.

Imputation par la moyenne

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

$$m \ : \ y_k \quad = \quad \beta + \sigma \epsilon_k \quad \text{pour} \quad k \in S, \tag{9}$$

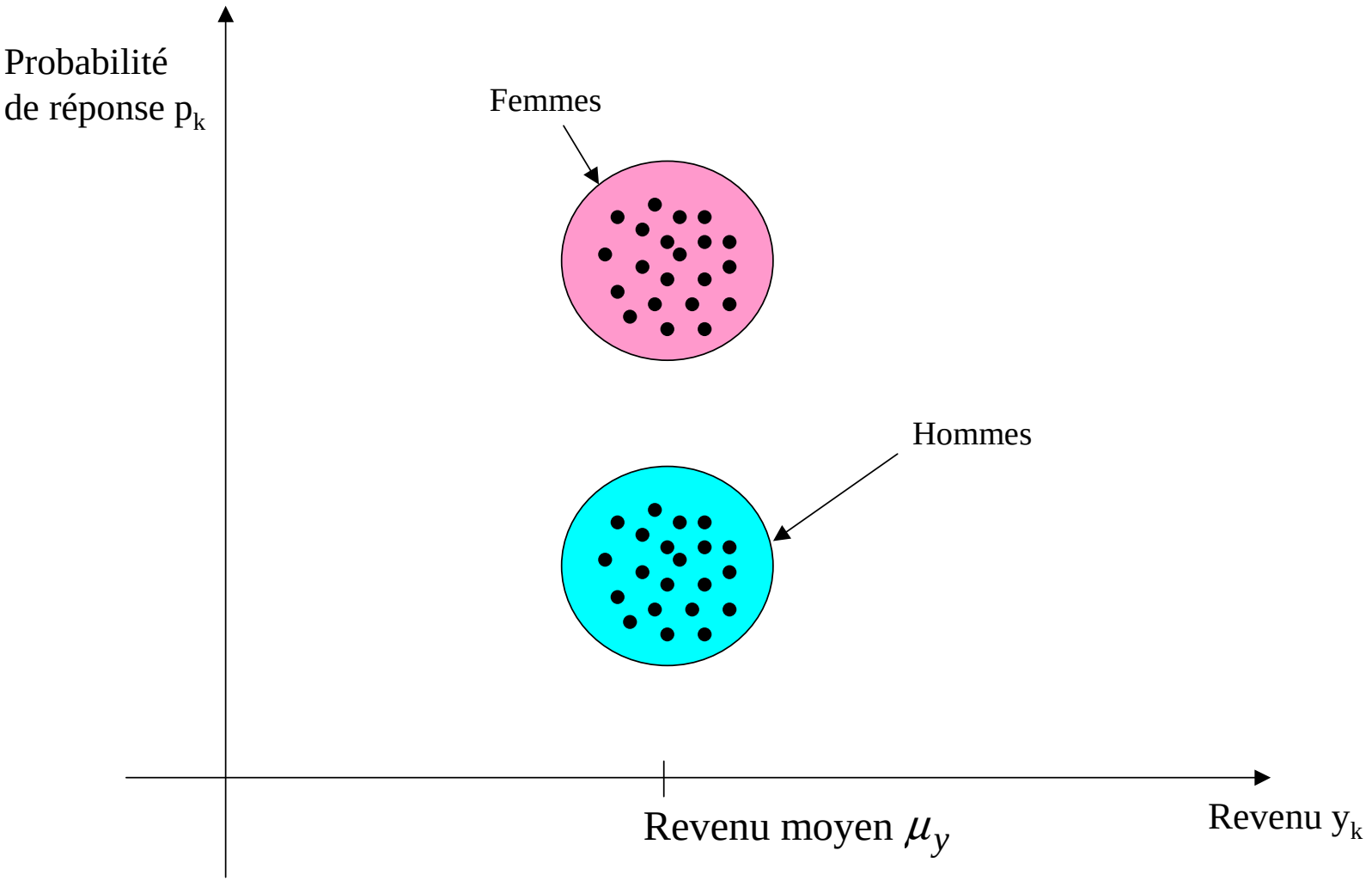
obtenu avec $\mathbf{z}_k = z_k = 1$ et $v_k = 1$. On obtient l'estimateur

$$\hat{\boldsymbol{\beta}}_r \quad = \quad \frac{\sum_{k \in S_r} \omega_k y_k}{\sum_{k \in S_r} \omega_k} \equiv \bar{y}_{\omega r}.$$

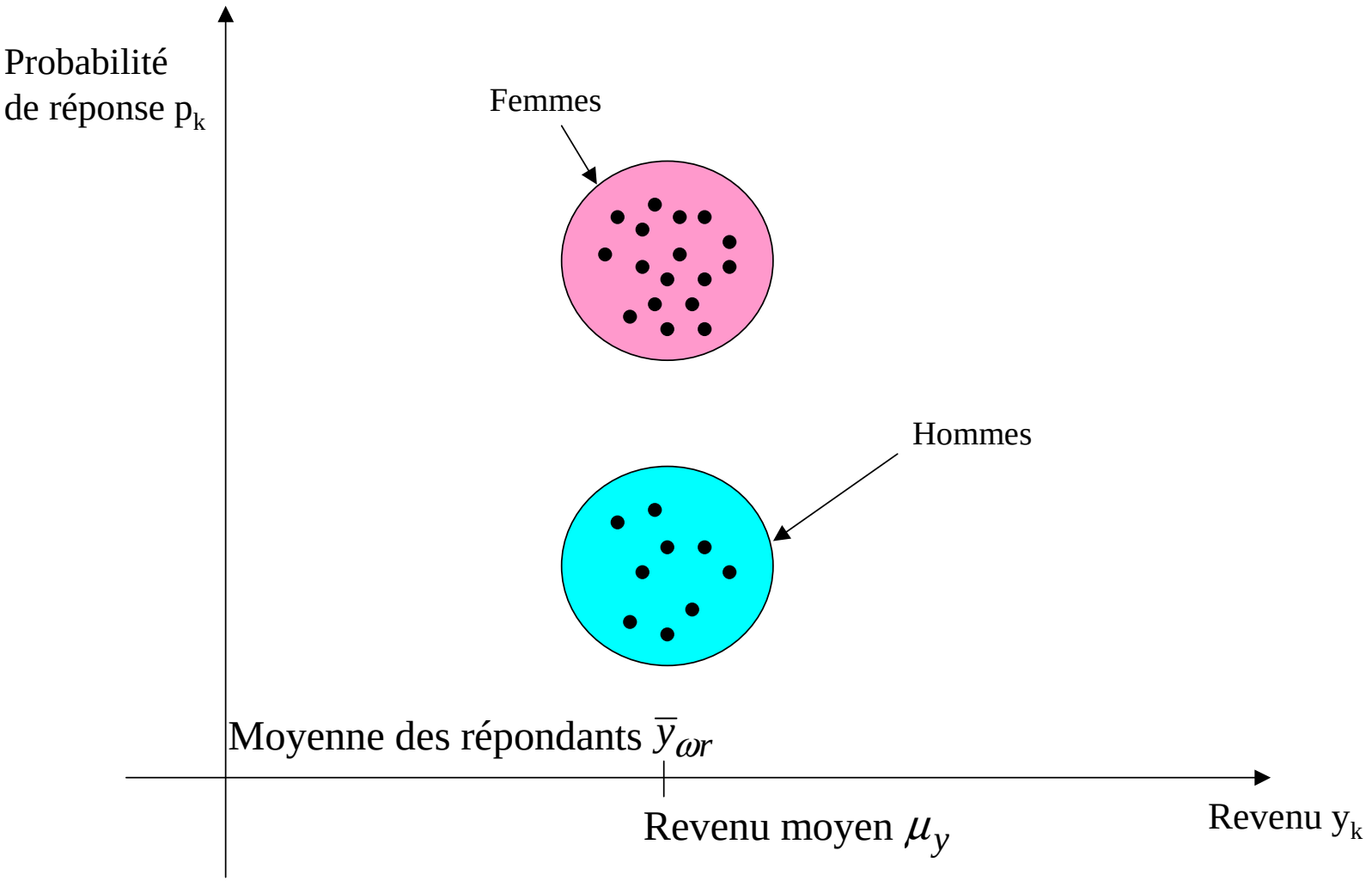
Dans le cas d'une imputation pondérée par les poids de sondage, on obtient :

$$\hat{t}_{yI} \quad = \quad \left(\frac{\sum_{k \in S} d_k}{\sum_{k \in S_r} d_k} \right) \sum_{k \in S_r} d_k y_k.$$

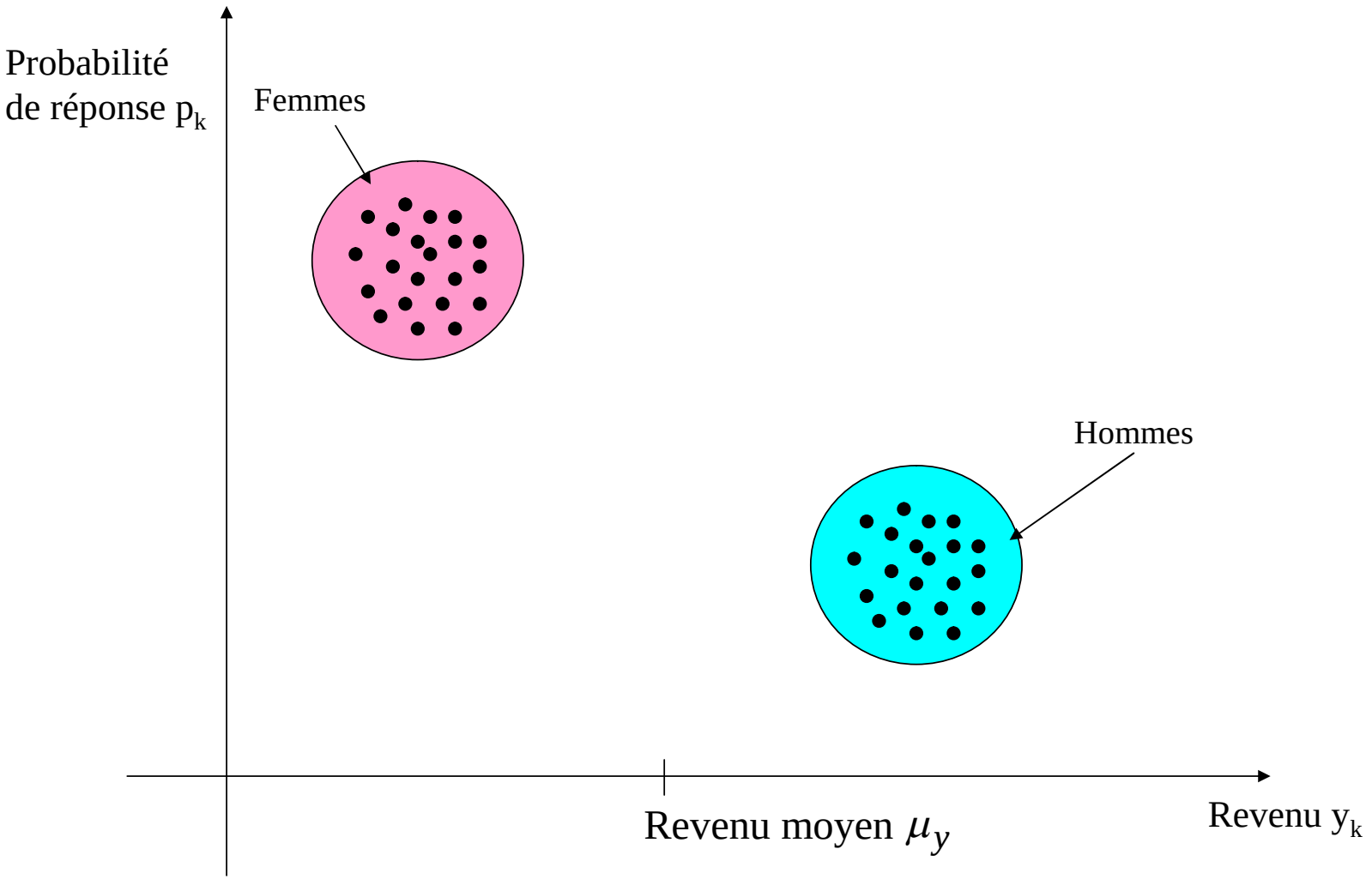
Cas favorable



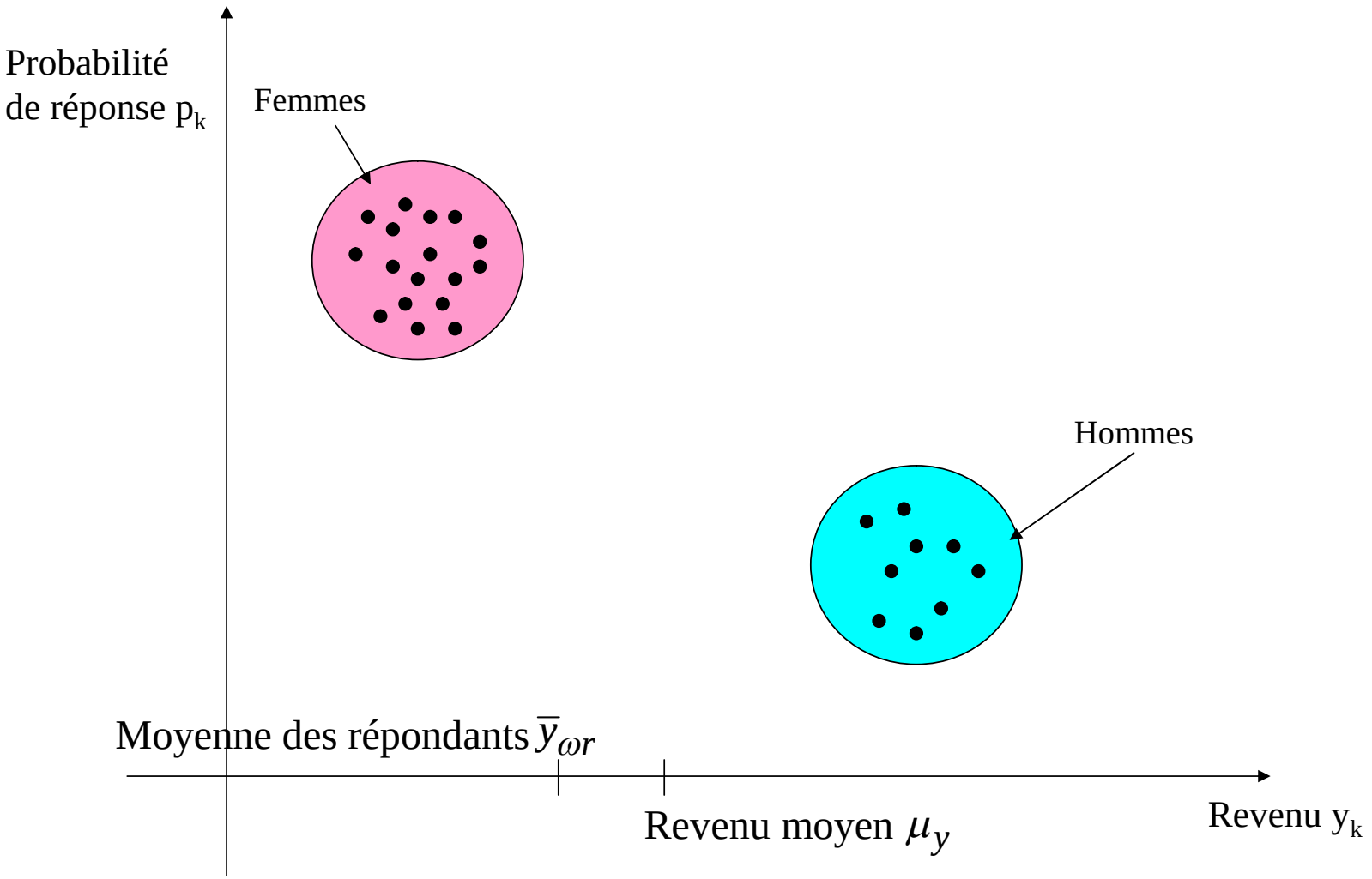
Cas favorable (suite)



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**.

En pratique, cette hypothèse est rarement vérifiée sur l'ensemble de l'échantillon. On peut en revanche essayer 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).

On impute alors par la moyenne au sein de chaque classe.

Imputation par la moyenne dans des classes

On parle d'**imputation par la moyenne dans les classes d'imputation**. Cette méthode s'appuie sur le modèle

$$m \ : \ y_k \quad = \quad \beta_h + \sigma_h \epsilon_k \quad \text{pour} \quad k \in S_h. \tag{10}$$

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, on obtient $y_k^* = \hat{\boldsymbol{\beta}}_{rh}$ avec

$$\hat{\boldsymbol{\beta}}_{rh} \quad = \quad \frac{\sum_{k \in S_{rh}} \omega_k y_k}{\sum_{k \in S_{rh}} \omega_k} \equiv \bar{y}_{\omega rh},$$

en notant $S_{rh} = S_h \cap S_r$.

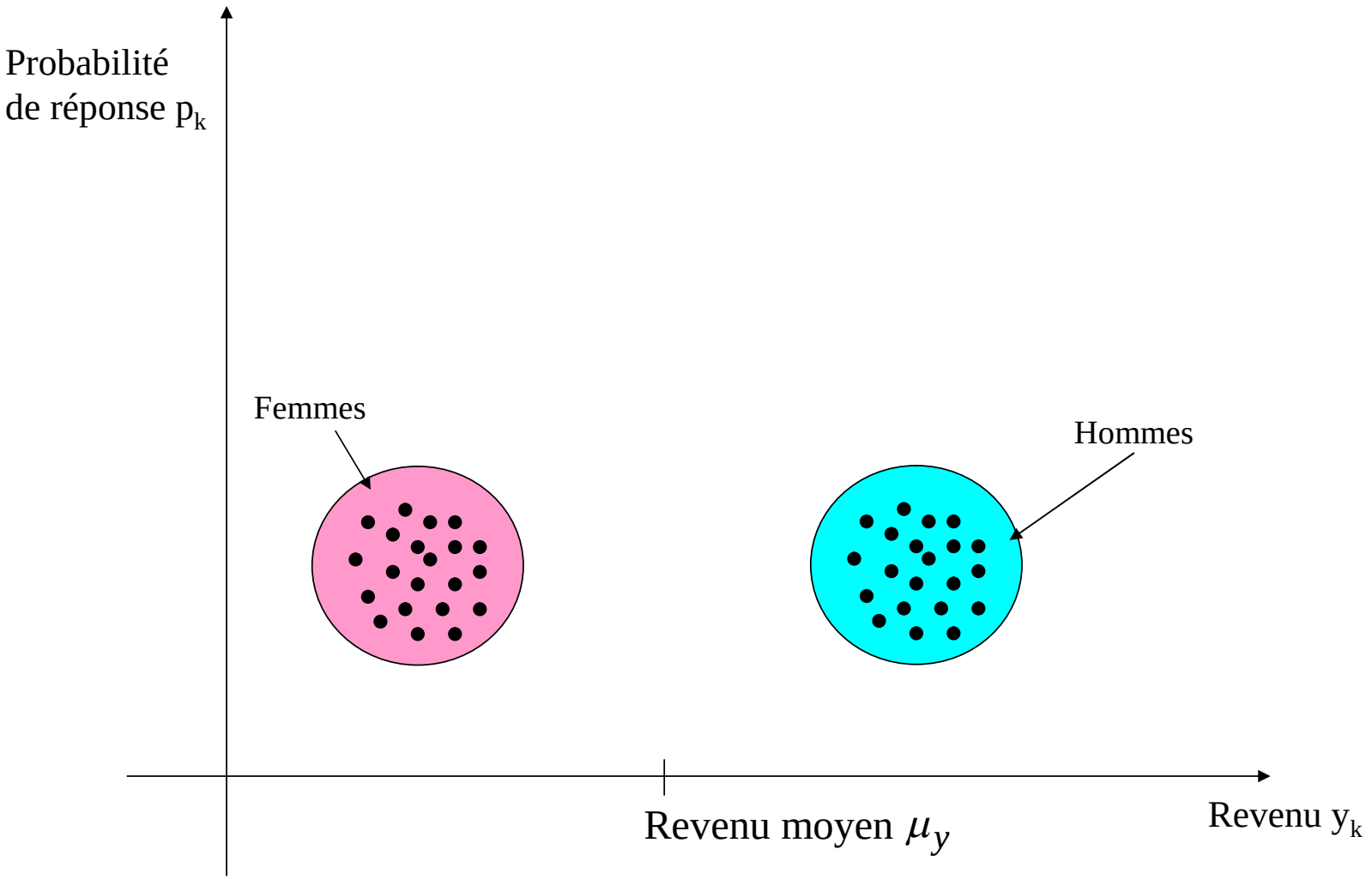
Imputation par la moyenne dans des classes

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.

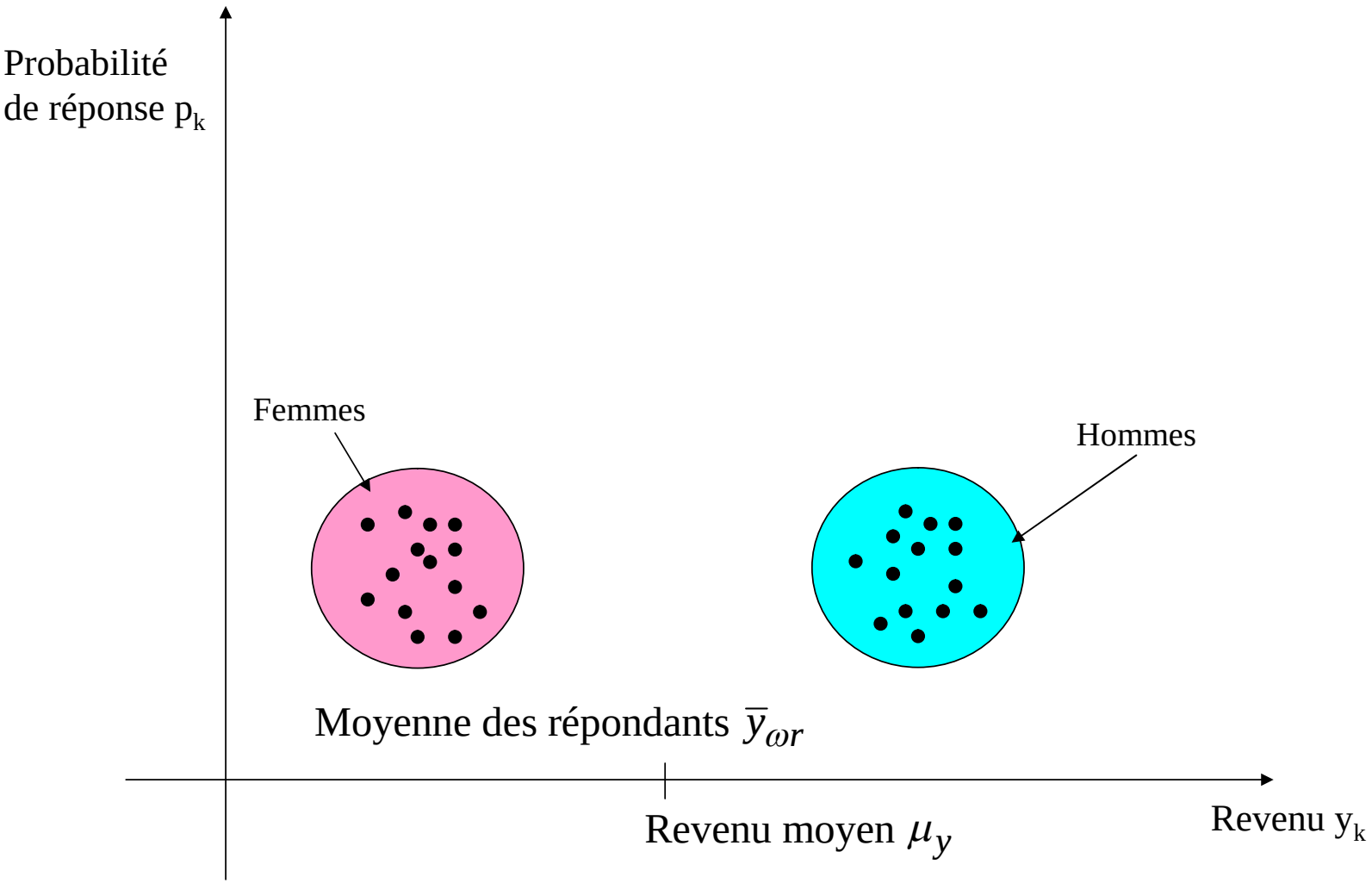
Là encore, cette hypothèse est généralement peu réaliste. Une imputation par la moyenne dans des classes conduira à une estimation approximativement non biaisée si les probabilités de réponse sont voisines au sein de chaque classe.

Il faut donc constituer les classes de façon à ce que, au sein de chaque classe, les individus soient peu différents par rapport à y et/ou les probabilités de réponse soient voisines.

Cas favorable 2



Cas favorable 2 (suite)



Construction des classes d'imputation

En pratique, on peut constituer les classes d'imputation de la façon suivante :

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

Imputation aléatoire

Mécanisme d'imputation par la régression

L'**imputation par la régression aléatoire** s'appuie sur le modèle (8) :

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

avec

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

On ajoute au terme de prédiction $\mathbf{z}_k^\top \hat{\boldsymbol{\beta}}_r$ un terme aléatoire $\hat{\sigma}\sqrt{v_k}\epsilon_k^*$, avec :

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

L'estimateur imputé est égal à

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

Mécanisme d'imputation par la régression

Cet estimateur est approximativement sans biais sous la modélisation utilisée :

$$\begin{aligned} B(\hat{t}_{yI}) &\simeq E_{mpqI}\left(\hat{t}_{yI} - \hat{t}_y\right) \\ &= E_{mpqI}\left[\sum_{k\in S_m} d_k(y_k^* - y_k)\right] \\ &\simeq 0. \end{aligned}$$

L'hypothèse fondamentale est que le vecteur $\mathbf{z}_k$ permette une bonne prédiction $y_k^* = \mathbf{z}_k^\top \hat{\boldsymbol{\beta}}_r$ de la valeur manquante.

La précision de l'estimateur peut être évaluée par :

- la variance totale $V_{mpqI}\left(\hat{t}_{yI} - t_y\right)$,
- la **variance anticipée** $E_m V_{pqI}\left(\hat{t}_{yI} - t_y\right) = E_m V_{pqI}\left(\hat{t}_{yI}\right)$.

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 simplifié

$$m \ : \ y_k \quad = \quad \beta + \sigma\epsilon_k \quad \text{pour} \quad k \in S, \tag{11}$$

obtenu avec $\mathbf{z}_k = z_k = 1$ et $v_k = 1$.

La méthode du hot-deck consiste à remplacer une valeur manquante y_k en sélectionnant au hasard et avec remise un donneur $y_j \in S_r$, avec des probabilités proportionnelles aux poids d'imputation ω_j. On obtient l'estimateur :

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

Imputation par hot-deck

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

On a

$$\begin{aligned} E_I(\hat{t}_{yI}) &= \sum_{k \in S_r} d_k y_k + \sum_{k \in S_m} d_k E_I(y_k^*) \\ &= \sum_{k \in S_r} d_k y_k + \sum_{k \in S_m} d_k \bar{y}_{\omega r}. \end{aligned}$$

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.

Le modèle d'imputation est le même que pour l'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 ;
- ou : si les probabilités de réponse sont voisines au sein d'une même classe.

Imputation par donneur

Le hot-deck est un cas particulier des méthodes d'imputation par donneur. On peut également 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)

Imputation par donneur

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).

Pour plus de détails sur les méthodes d'imputation possibles, voir Haziza (2009,2011).

Quelle méthode d'imputation utiliser ?

Dans le cas considèré ici (estimation d'un total), les méthodes d'imputation déterministes sont préférables car elles ne conduisent pas à une augmentation de la variance. Si le modèle d'imputation est correctement spécifié, l'imputation conduira à une estimation approximativement non biaisée du total.

Dans le cas général, la méthode d'imputation utilisée dépend du type de variable (quanti/quali), et de l'analyse que l'on souhaite faire : estimation d'un total, calcul d'une régression, d'une médiane, ...

Si on s'intéresse à la distribution de la variable imputée, les méthodes d'imputation déterministes ne sont généralement pas adaptées. Par exemple, l'imputation par la moyenne "écrase" de façon artificielle la variable imputée au niveau de sa valeur moyenne.

Problèmes liés à la non-réponse

L'imputation ne crée pas d'information : elle peut donner une fausse impression de précision, car elle conduit à un fichier de données complet, "comme si" on n'observait aucune non-réponse partielle.

L'imputation tend à perturber les relations entre les variables. Si l'objet de l'analyse est par exemple d'étudier une régression entre deux variables, l'imputation des données manquantes doit être réalisée de façon à préserver la relation entre ces variables.

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

On considère 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

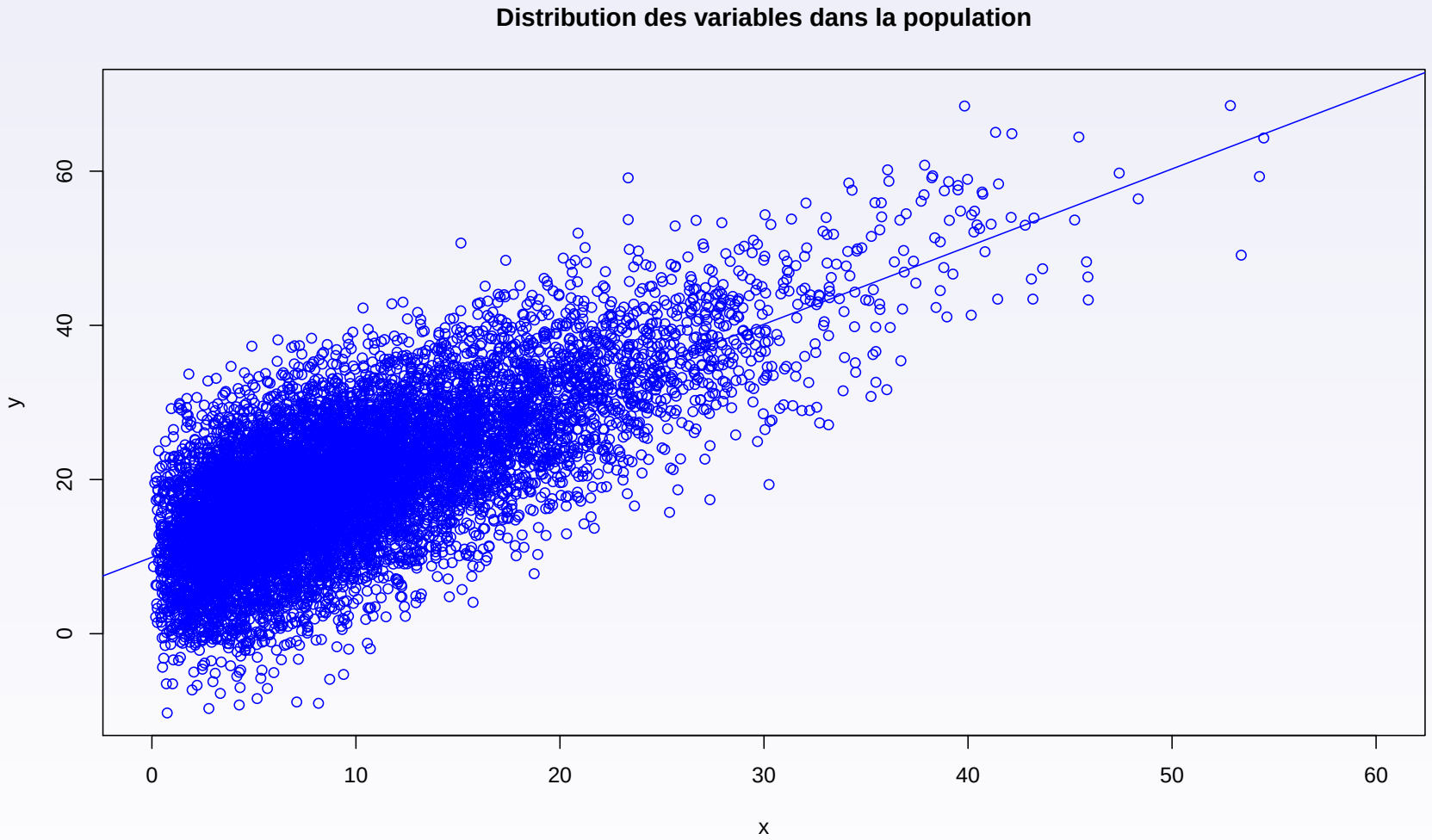
$$y_k = \beta_0 + \beta_1 \; x_k + \epsilon_k,$$

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
0
20
40
60
x
0
10
20
30
40
50
60

Estimation sur données non imputées

Estimation en situation de réponse complète

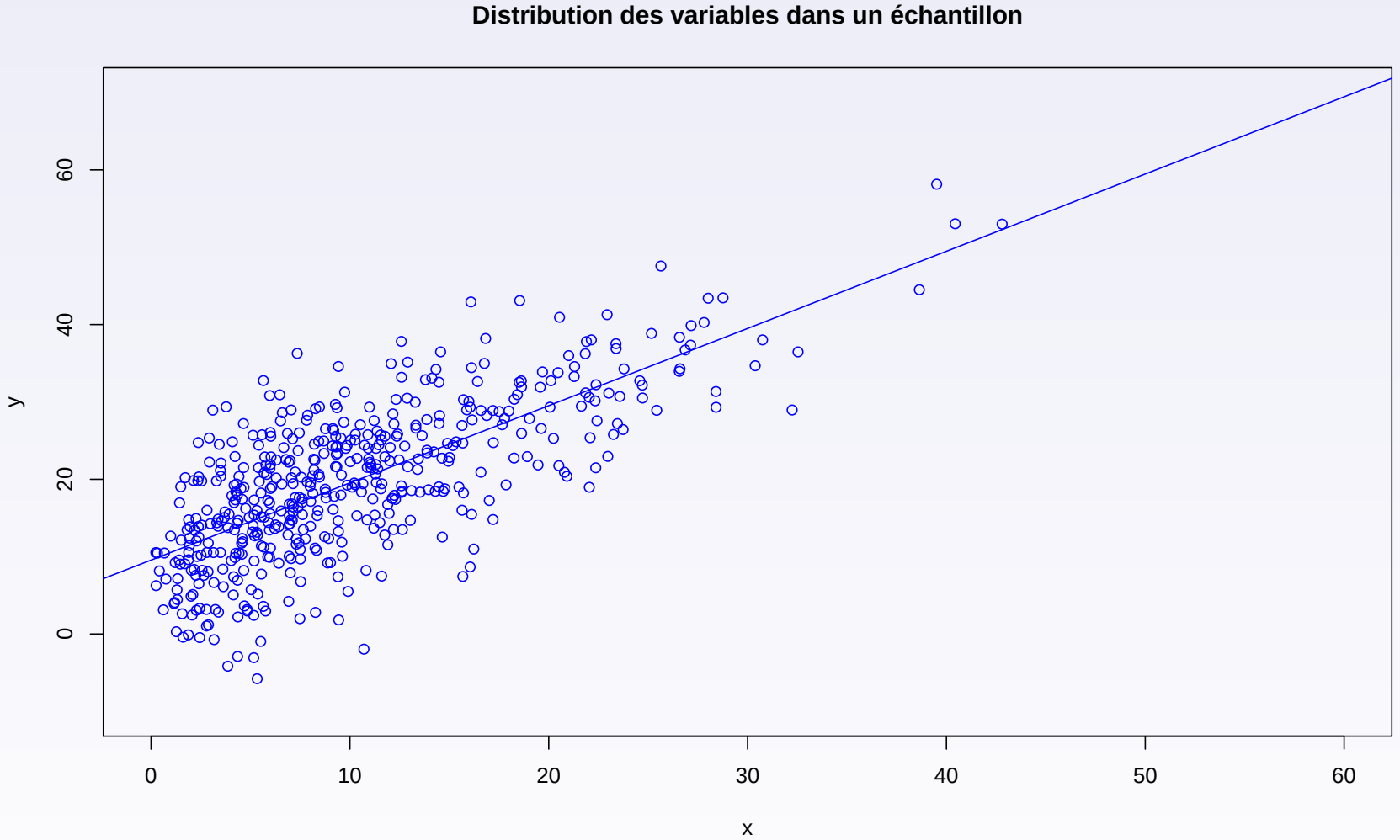
On sélectionne un échantillon S de taille $n = 500$ selon un SRS. La moyenne μ_y peut être estimée sans biais par

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

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

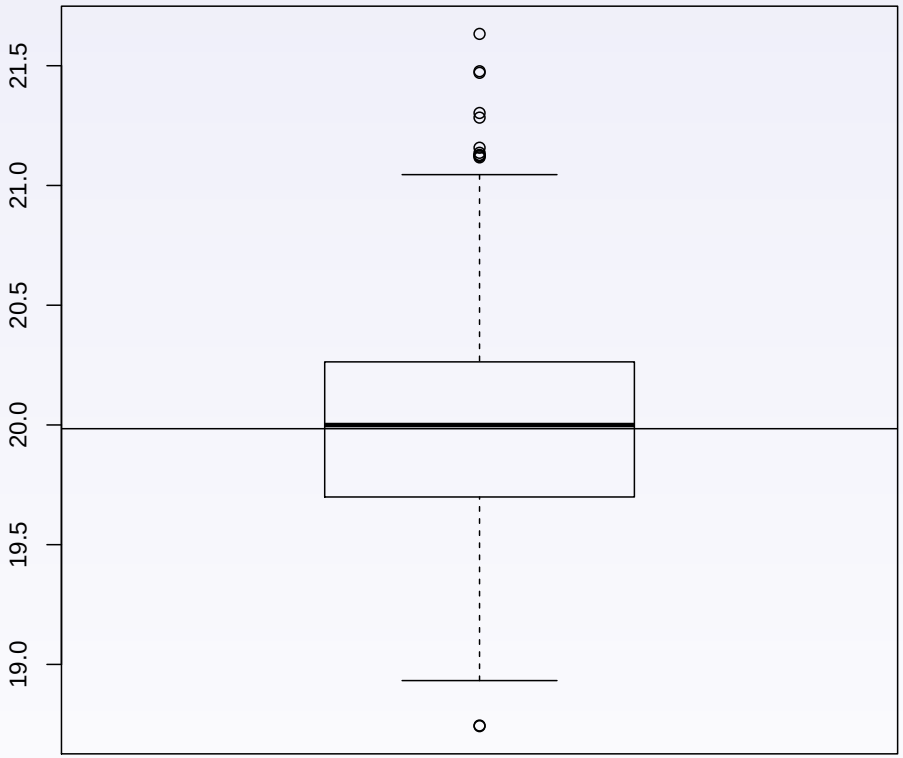
Simulations

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

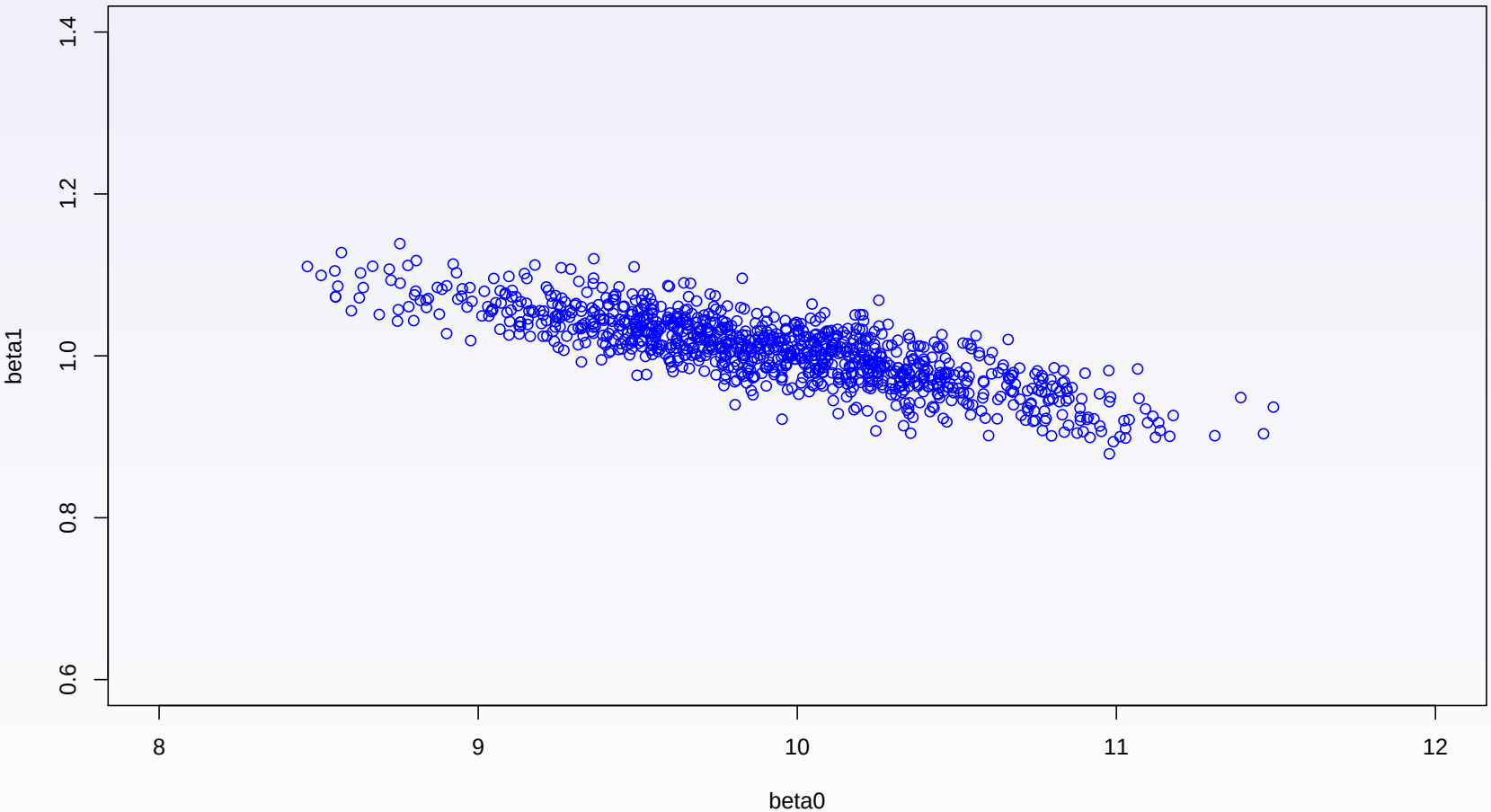
On obtient une estimation de la distribution

- 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

On suppose 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, on considère $p = 0.8$, 0.6 et 0.4.

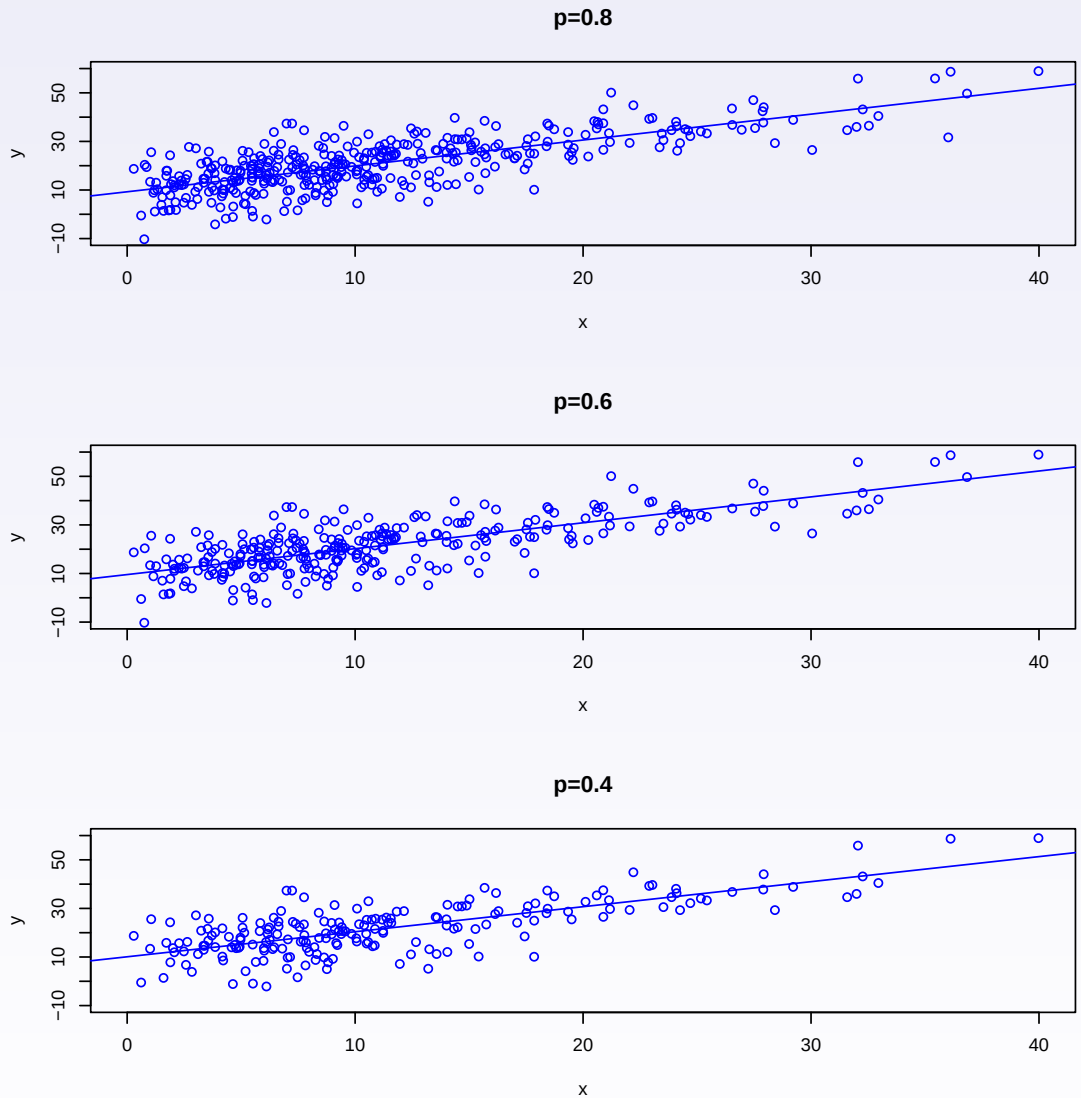
Estimation en situation de non-réponse partielle

On utilise les estimateurs basés sur les répondants

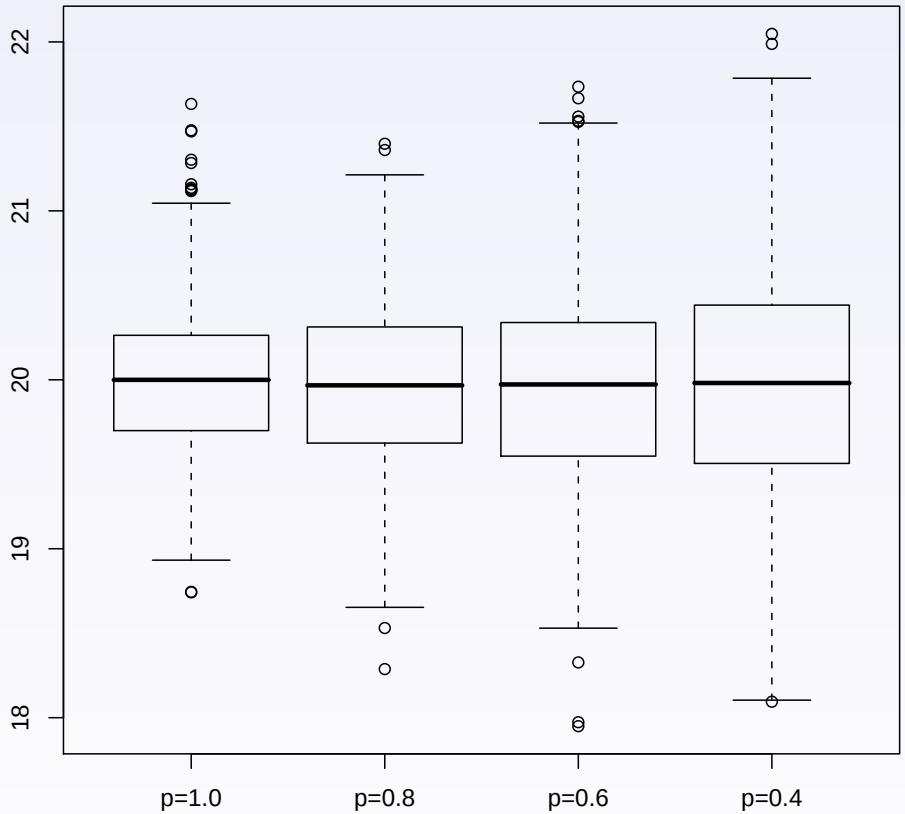
$$\begin{aligned} \bar{y}_r &= \frac{1}{n_r} \sum_{k \in S_r} y_k, \\ \hat{\beta}_r &= \left(\sum_{k \in S_r} \mathbf{x}_k \mathbf{x}_k^\top \right)^{-1} \sum_{k \in S_r} \mathbf{x}_k y_k. \end{aligned}$$

On obtient là aussi une estimation de la distribution des estimateurs $\bar{y}_r$ et $\hat{\beta}_r$ en simulant $B = 1,000$ fois le plan de sondage + le mécanisme de non-réponse.

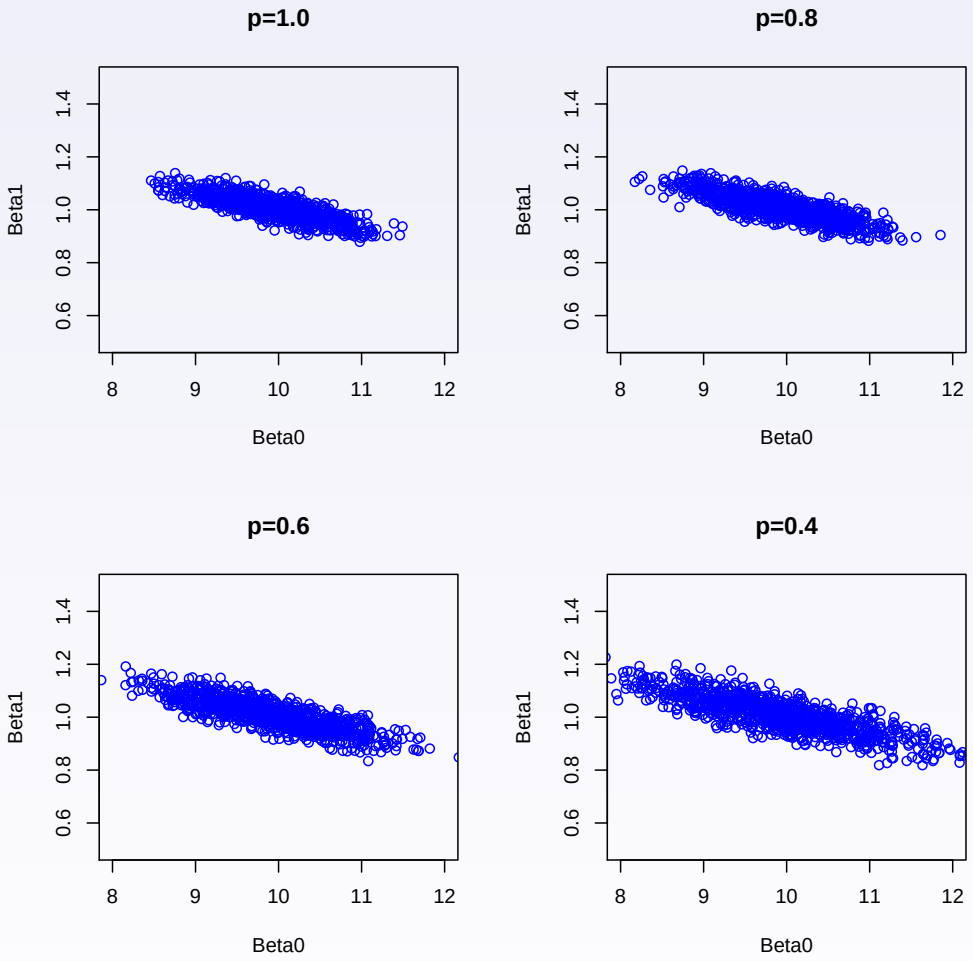
Données échantillonnées



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. On notera également

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

On obtient alors les estimateurs imputés

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

On étudie 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

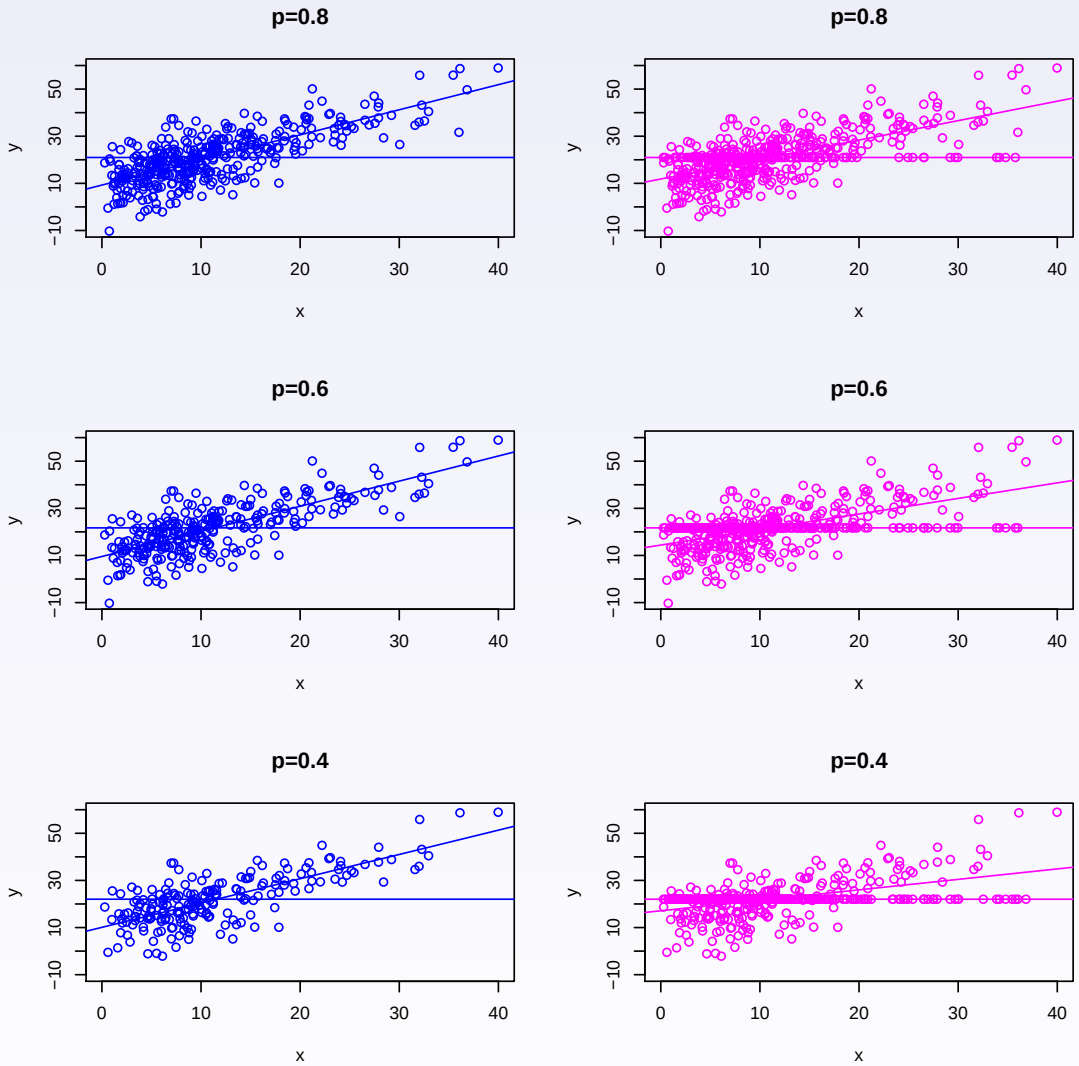
Pour un individu $k \in S_m$, on impute $y_k^* = \bar{y}_r$ avec

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

On obtient en particulier les estimateurs imputés

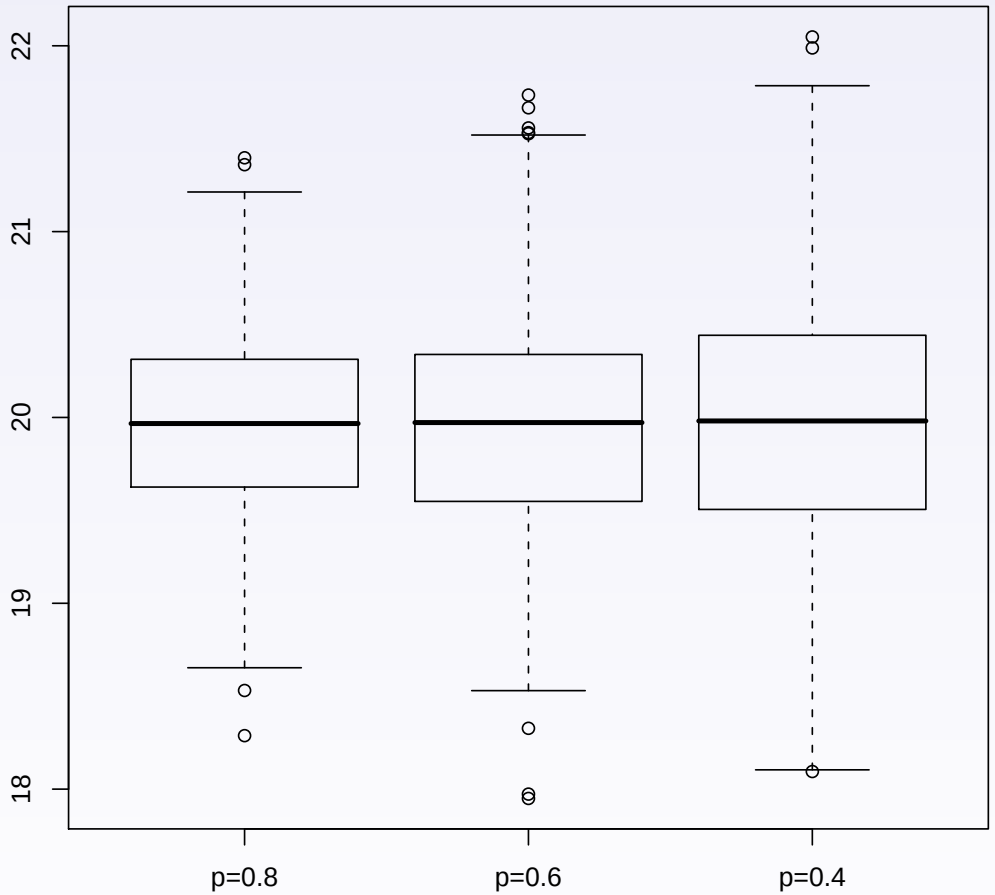
$$\begin{aligned} \bar{y}_I &= \bar{y}_r, \\ \hat{\beta}_I &= \left(\sum_{k \in S} \mathbf{x}_k \mathbf{x}_k^\top \right)^{-1} \sum_{k \in S} \mathbf{x}_k \tilde{y}_k. \end{aligned}$$

Données obtenues sur un échantillon

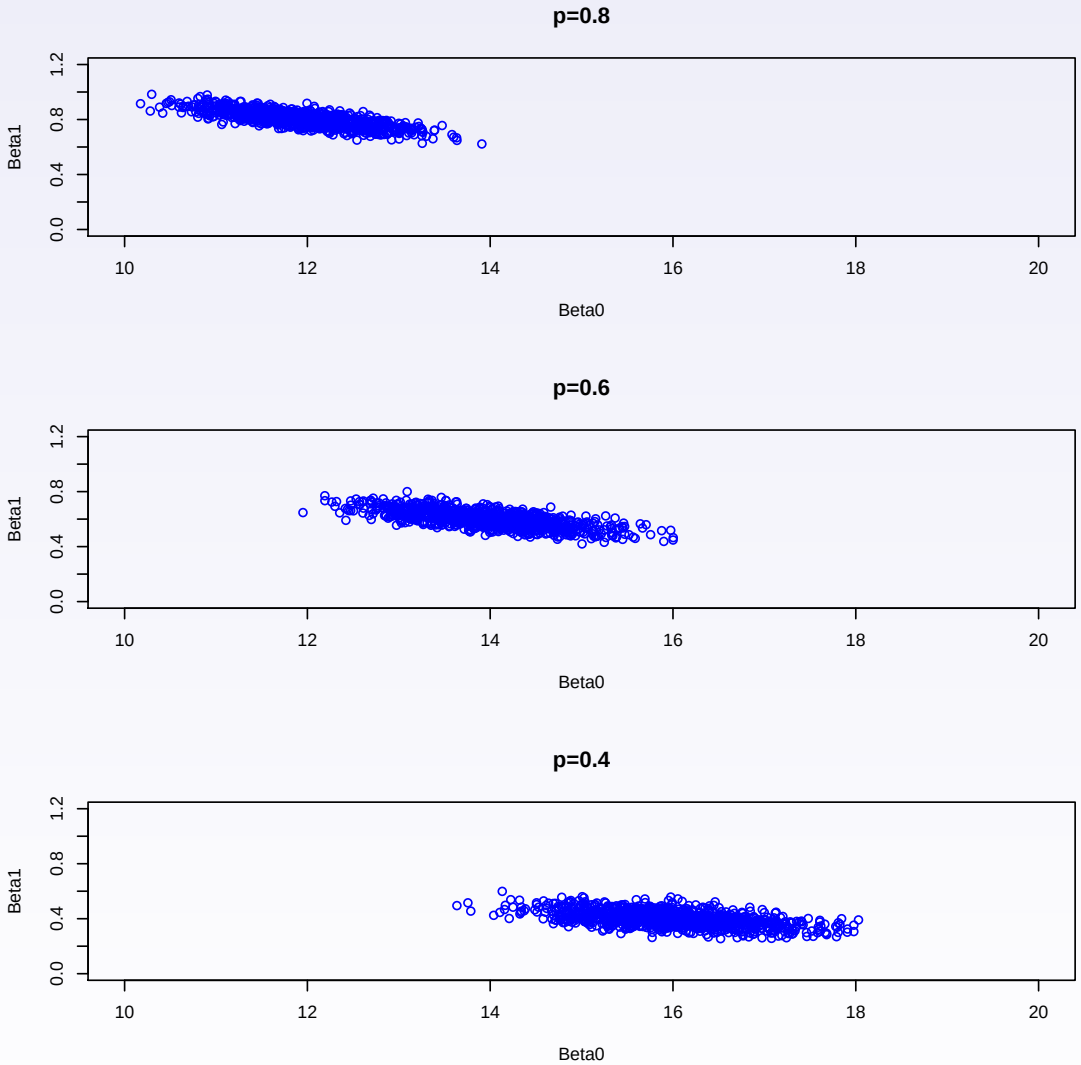


Cas complets Données imputées

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



Distribution des coefficients de régression estimés



Imputation par hot-deck

Principe

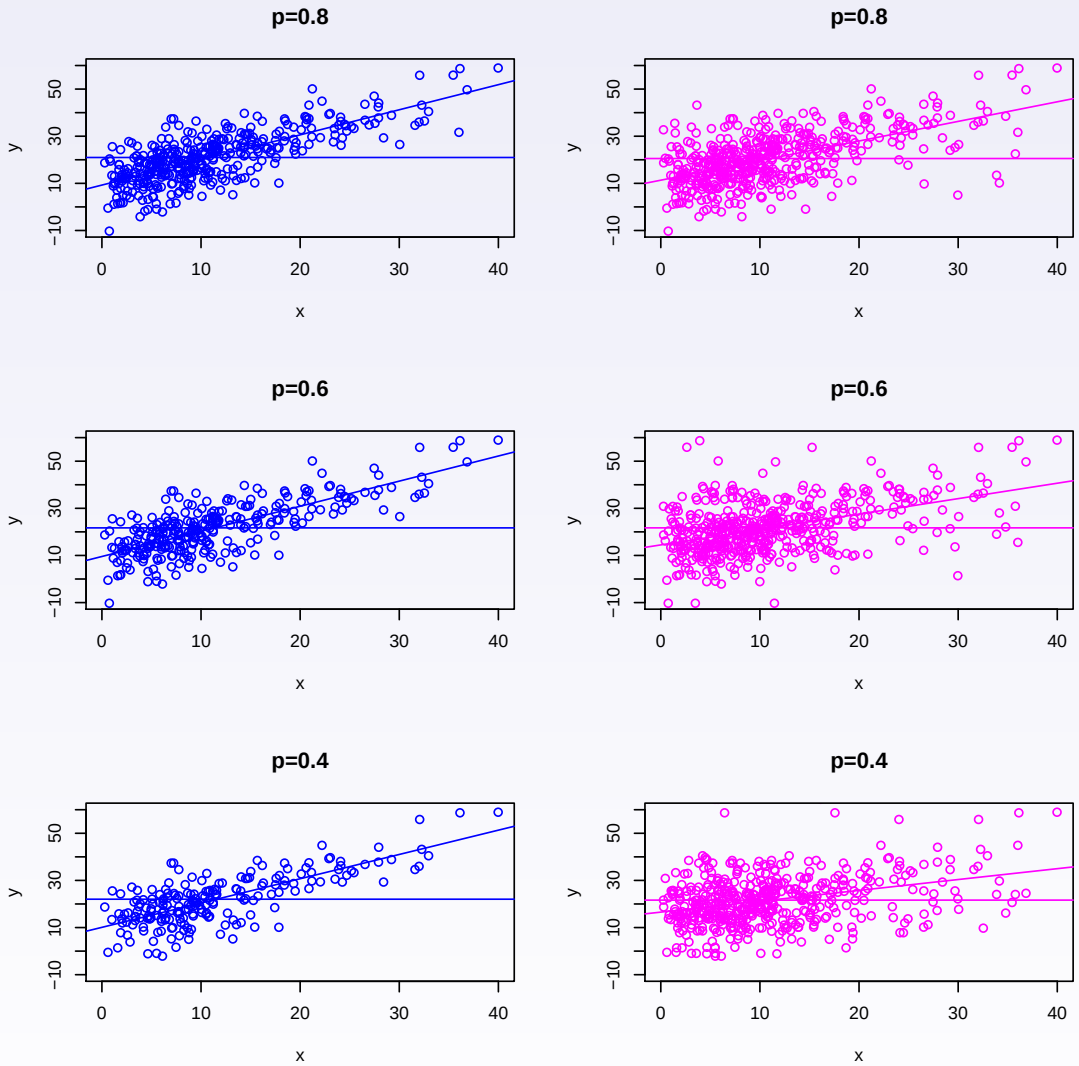
Pour un individu $k \in S_m$, la valeur y_k est remplacée en tirant au hasard et avec remise un donneur $y_{(j)} \in S_r$, avec des probabilités de tirage égales.

La valeur imputée peut encore se réécrire

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

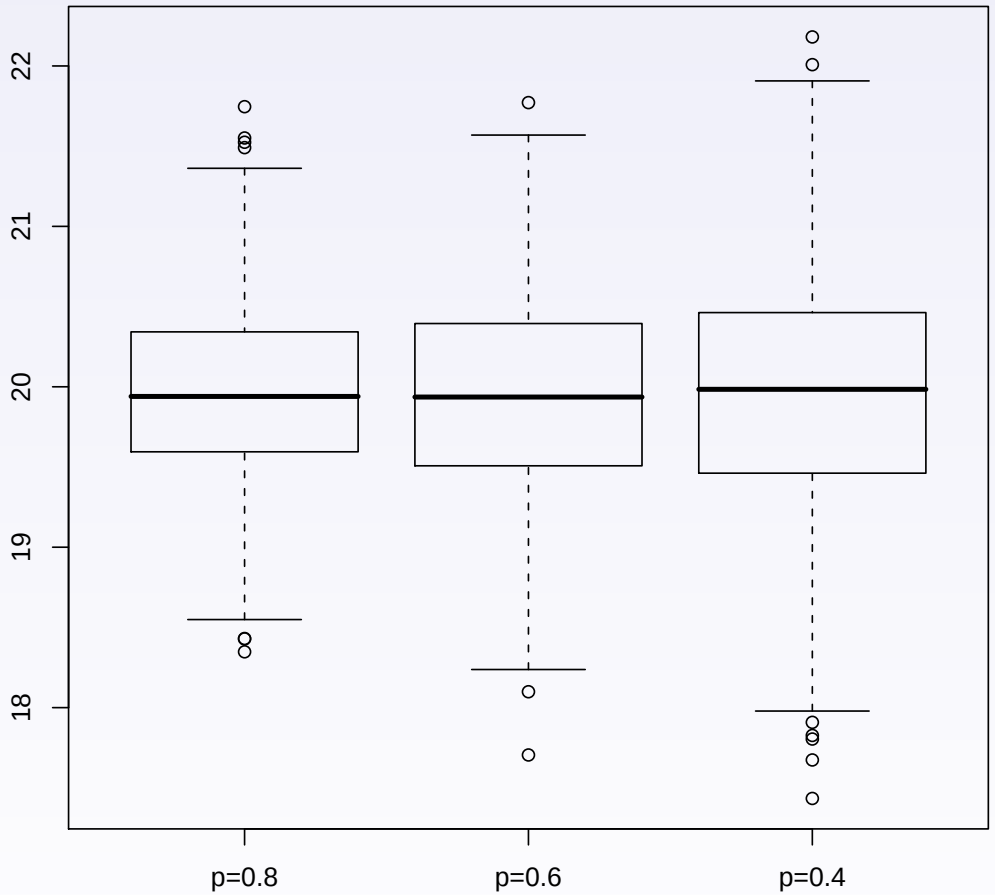
Interprétation : une valeur manquante est remplacée par la moyenne $\bar{y}_r$ des répondants, à laquelle on ajoute un résidu aléatoire (de moyenne nulle).

Données obtenues sur un échantillon

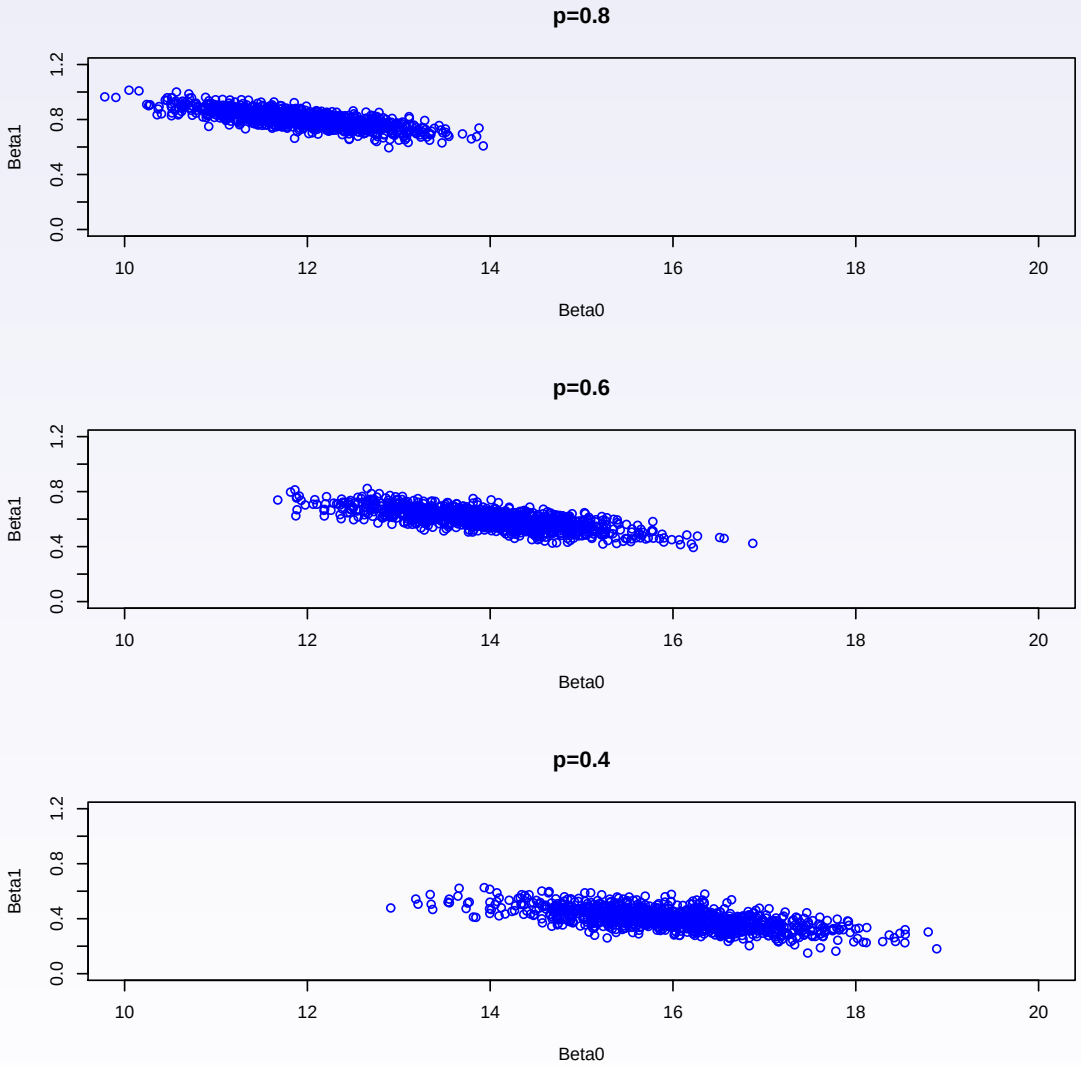


Cas complets Données imputées

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



Distribution des coefficients de régression estimés



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

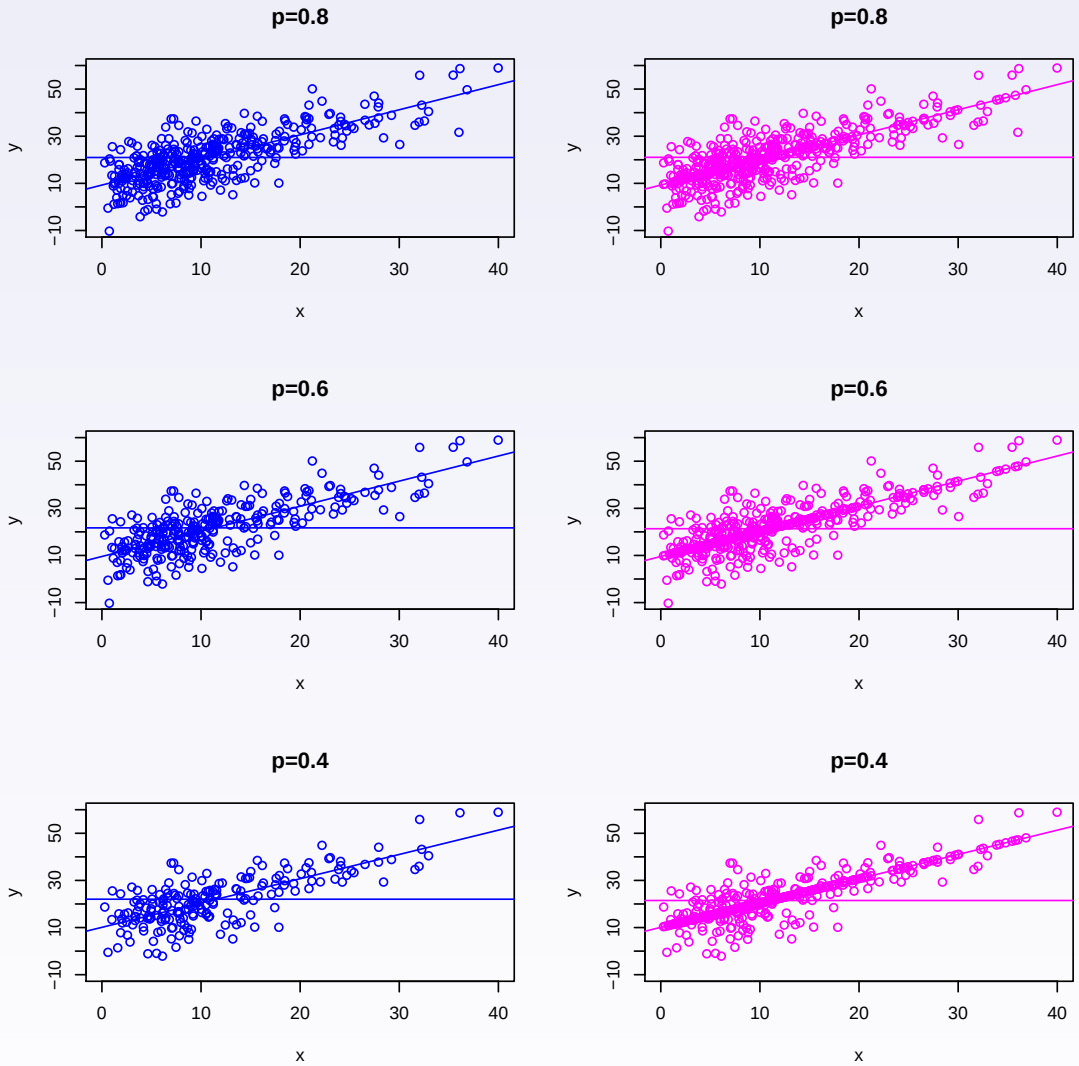
$$\hat{\beta}_r = \left(\sum_{k \in S_r} \mathbf{x}_k \mathbf{x}_k^\top \right)^{-1} \sum_{k \in S_r} \mathbf{x}_k y_k$$

le coefficient de régression estimé sur les répondants.

On obtient alors les estimateurs imputés

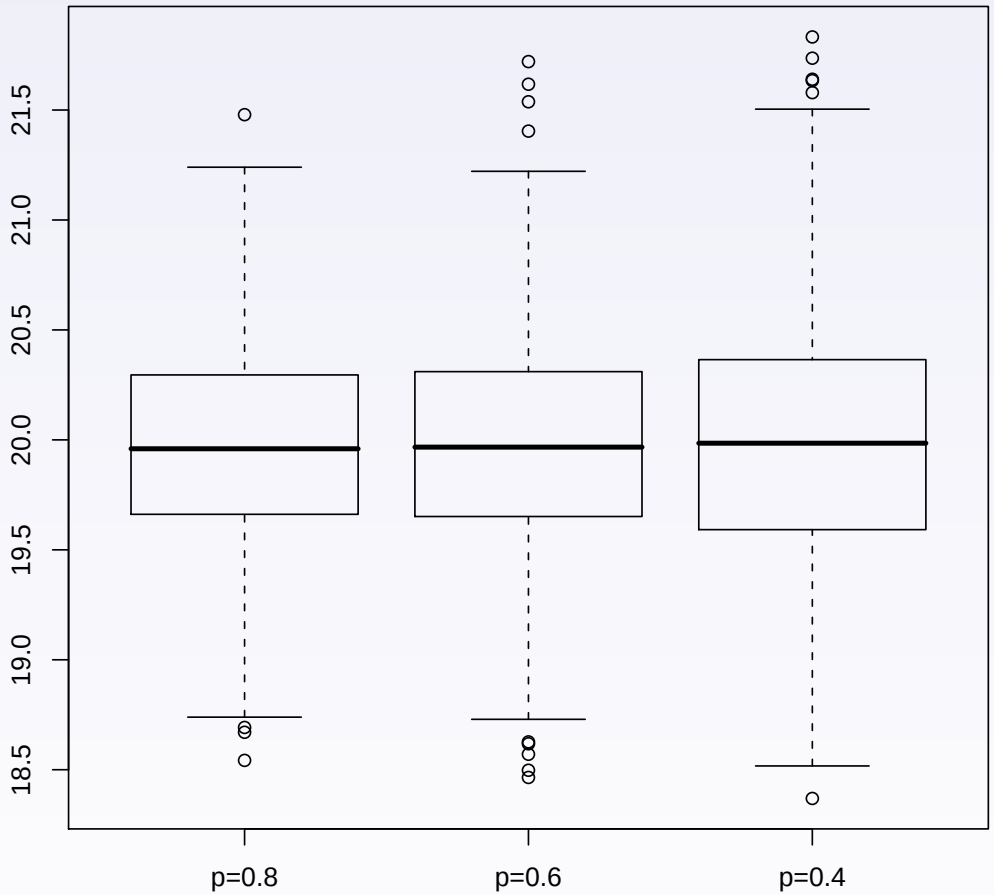
$$\begin{aligned} \bar{y}_I &= \frac{1}{n} \sum_{k \in S} \tilde{y}_k, \\ \hat{\beta}_I &= \hat{\beta}_r. \end{aligned}$$

Données obtenues sur un échantillon

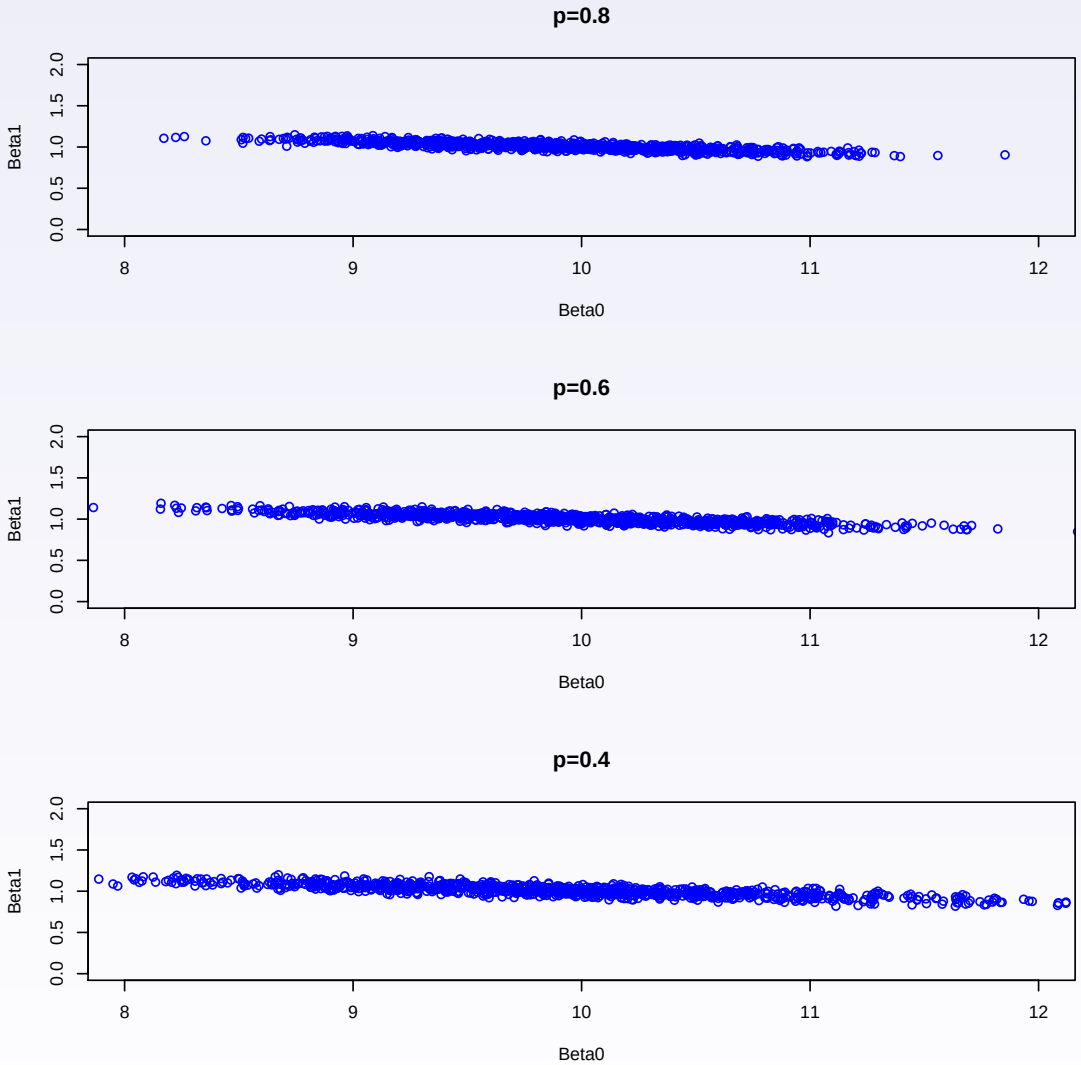


Cas complets Données imputées

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

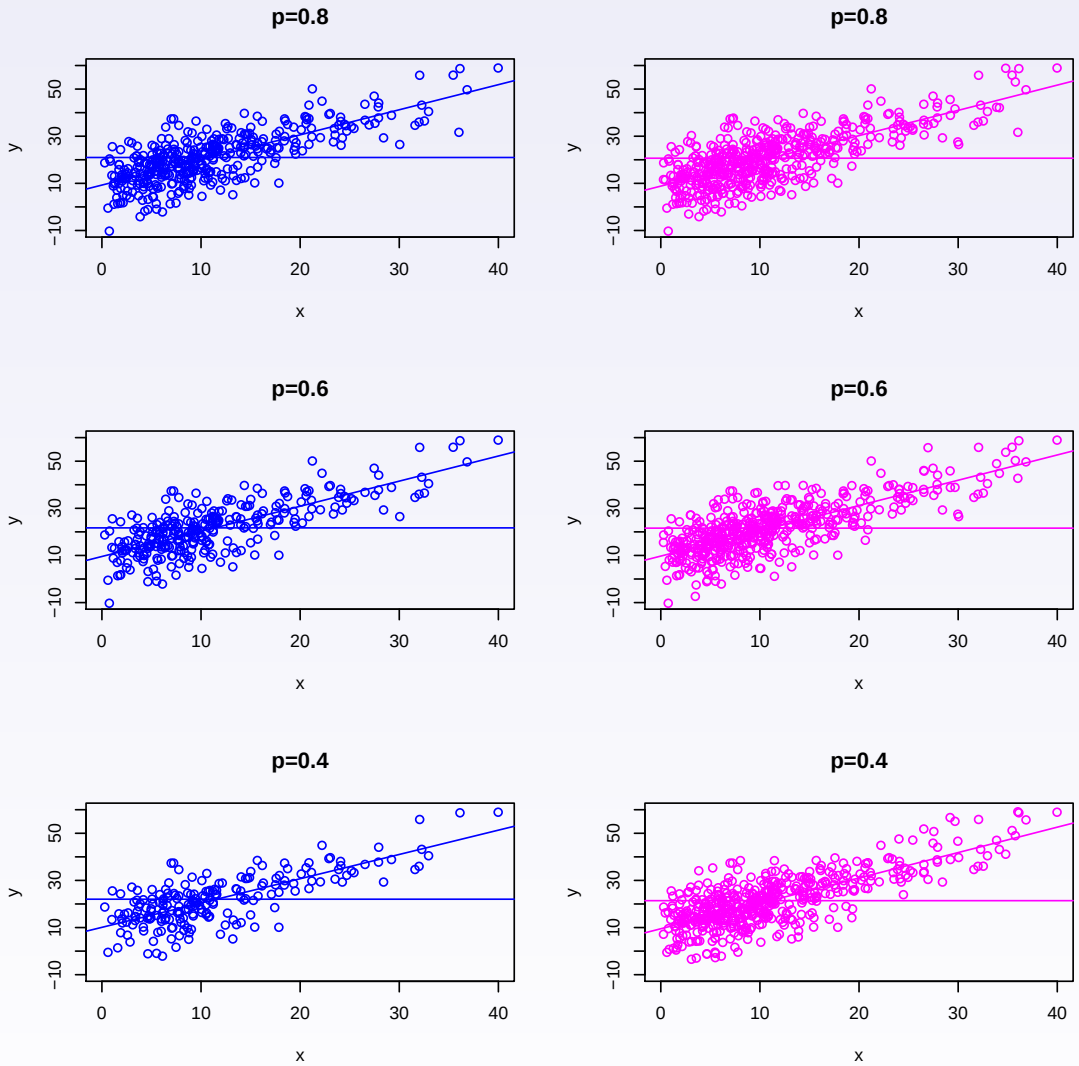
$$\eta_j = y_j - \mathbf{x}_j^\top \hat{\beta}_r \quad \text{pour} \quad j \in S_r.$$

On obtient pour $k \in S_m$ la valeur imputée

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

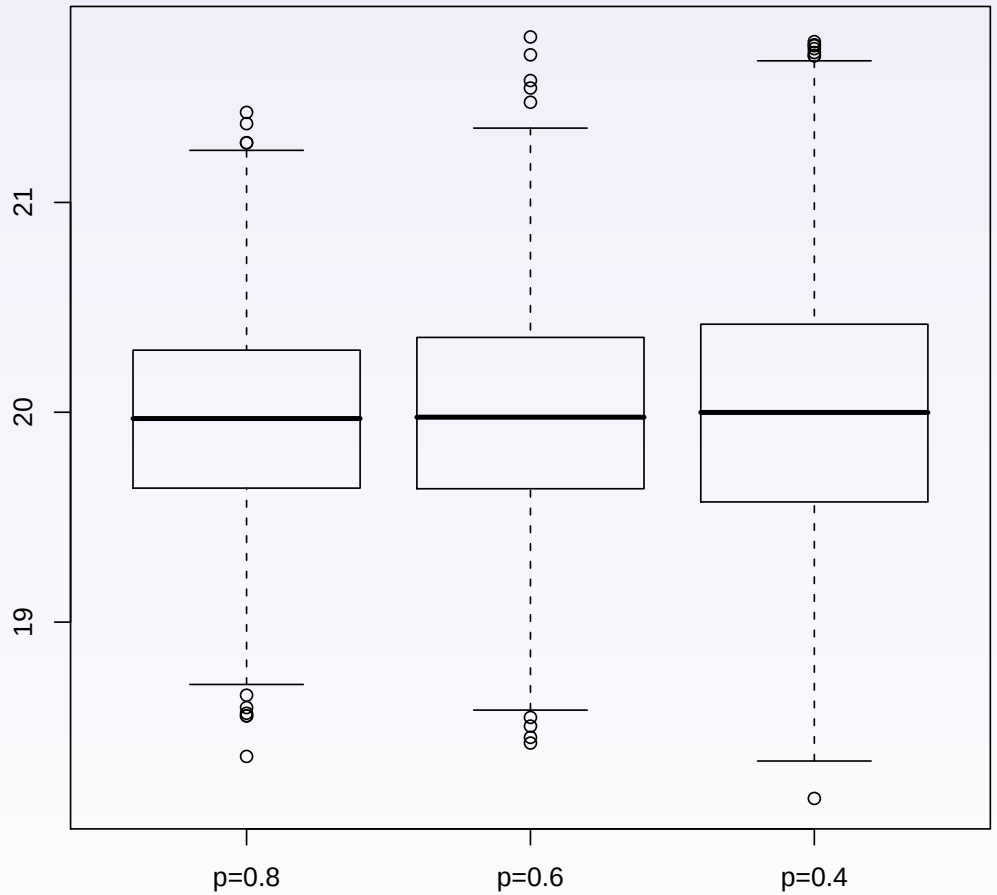
On impute donc "au plus près" du modèle (en tenant compte de ses imperfections).

Données obtenues sur un échantillon

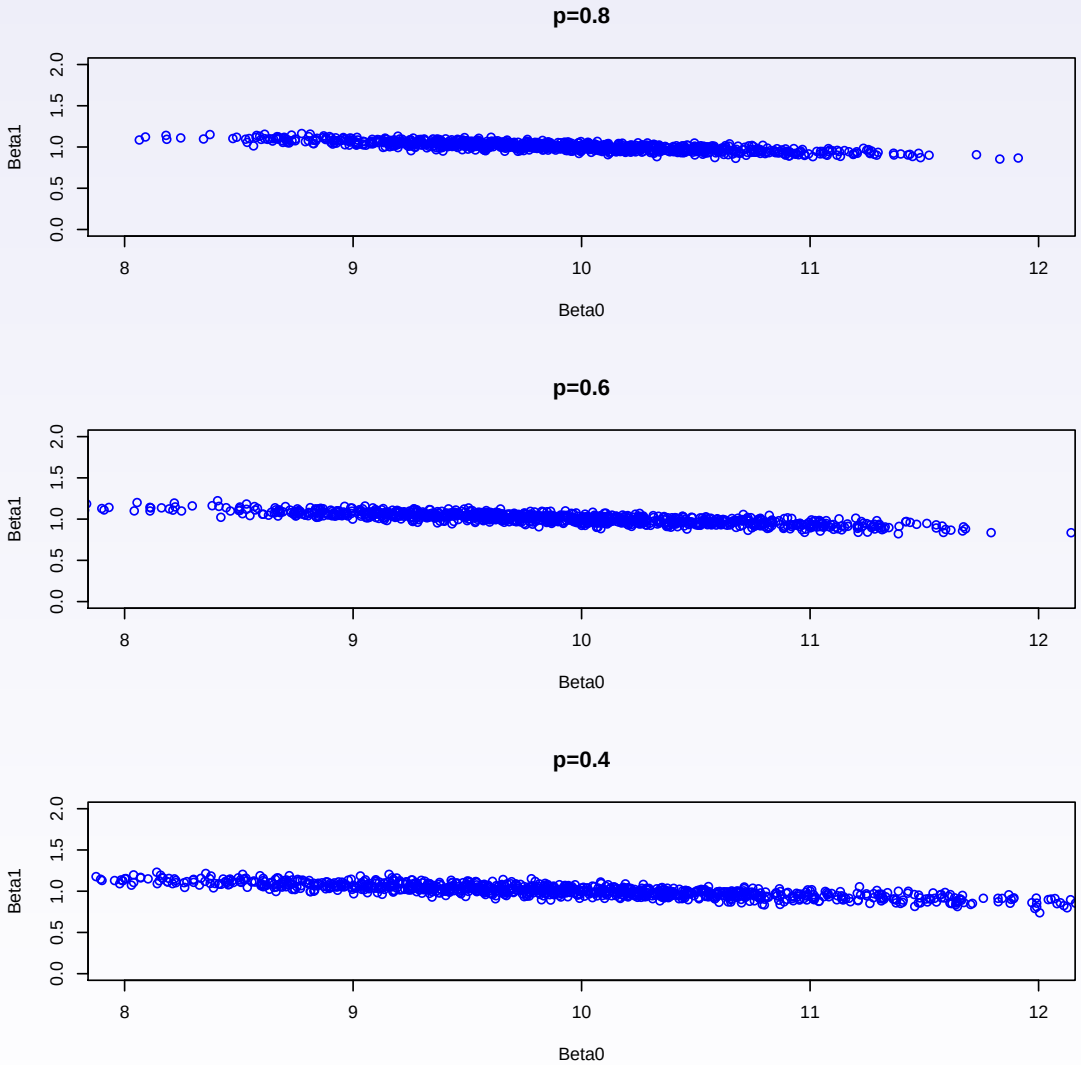


Cas complets Données imputées

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



Distribution des coefficients de régression estimés



Bibliographie

Ardilly, P. (2006), *Les Techniques de Sondage*, Technip, Paris.

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.

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*. FCDA, 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.

Skinner, C.J., et D'Arrigo, J. (2011). *Inverse probability weighting for clustered nonresponse*. Biometrika, 98, 953-966.