


INFÉRENCE EN PRÉSENCE D'IMPUTATION SIMPLE DANS LES ENQUÊTES: UN SURVOL

David Haziza
Université de Montréal

24 novembre 2006



“Pour assurer autant que possible l’exactitude du dénombrement, (...), il importe qu’une pénalité soit édictée contre les personnes qui refuseraient de les fournir, ou qui sciemment les donneraient inexacts.”

M. Legoyt.

17 Juillet 1860,

Congrès International de la Statistique



PLAN

1. INTRODUCTION
2. MÉTHODES D'IMPUTATION
3. ESTIMATION PONCTUELLE
4. CLASSES D'IMPUTATION
5. ESTIMATION DE LA VARIANCE
6. DISTORSION DES RELATIONS
7. CONCLUSIONS



NIVEAUX DE NON-RÉPONSE

Non-réponse totale:

- Absence complète d'information sur une unité.

Non-réponse partielle:

- Certaines (mais pas toutes) variables recueillies



TRAITEMENT DE LA NON-RÉPONSE

Non-réponse totale:

En général, on utilise des **méthodes de repondération** qui consiste à hausser le poids des répondants pour compenser pour les non-répondants

Non-réponse partielle:

En général, on utilise **l'imputation** qui consiste à créer une unique valeur artificielle pour boucher le trou de la valeur manquante

Toutes les méthodes de traitement ont un même but: Réduire le biais de non-réponse le plus possible!



DÉFINITIONS

- On distingue l'imputation simple de l'imputation multiple:

L'**imputation simple** est une technique qui consiste à créer une unique valeur artificielle pour “boucher le trou” de la valeur manquante

L'**imputation multiple** est une technique qui consiste à créer $M \geq 2$ valeurs artificielles pour “boucher le trou” de la valeur manquante

- L'imputation peut être effectuée par ordinateur ou manuellement



AVANTAGES DE L'IMPUTATION

- L'imputation simple mène à la création d'un **fichier de données complet**
- Les résultats issus de différentes analyses seront vraisemblablement **cohérents**
- Contrairement aux méthodes de repondération, l'imputation permet l'utilisation d'un **poids de sondage unique**



RISQUES LIÉS À L'IMPUTATION

- L'inférence n'est valide que si les hypothèses sous-jacentes sont valides
- Le fait de traiter les valeurs imputées comme si elles avaient été observées peut mener à une sous-estimation substantielle de la variance, surtout si le taux de non-réponse est grand
- L'imputation a comme effet de modifier les corrélations entre les variables



CONTEXTE

- Population finie U de taille N
- L'objectif est d'estimer le total dans la population

$$Y = \sum_{i \in U} y_i,$$

pour une variable d'intérêt y .

- On tire un échantillon aléatoire, s , de taille n , selon un plan de sondage $p(\cdot)$.



ESTIMATION: RÉPONSE COMPLÈTE

- Un estimateur de Y est l'estimateur de Horvitz-Thompson

$$\hat{Y}_{HT} = \sum_{i \in s} w_i y_i,$$

- $w_i = 1/\pi_i$ désigne le poids de sondage de l'unité i
- π_i désigne la probabilité d'inclusion de l'unité i dans l'échantillon s ; $i = 1, \dots, N$.
- Échantillonnage aléatoire simple sans remise: $w_i = N/n$

ESTIMATION: RÉPONSE COMPLÈTE

- Sous l'approche traditionnelle en sondages (**approche design-based**), le vecteur $\mathbf{y} = (y_1, \dots, y_N)'$ est traité comme fixe
- Soit δ_i la variable indicatrice de sélection dans l'échantillon

$$\delta_i = \begin{cases} 1 & \text{si } i \in s \\ 0 & \text{sinon} \end{cases}$$

- Cette variable joue un rôle crucial dans le contexte de l'inférence
- L'estimateur de HT **est sans biais sous le plan de sondage**

$$E_p(\hat{Y}_{HT}) = Y,$$

où $E_p(\cdot)$ désigne l'espérance par rapport au plan de sondage

ESTIMATEUR IMPUTÉ

- En présence de non-réponse à la variable y , on définit un **estimateur imputé** de Y

$$\hat{Y}_I = \sum_{i \in S_r} w_i y_i + \sum_{i \in S_m} w_i y_i^*,$$

où a_i est une variable indicatrice de réponse telle que

$$a_i = \begin{cases} 1 & \text{si l'unité } i \text{ a répondu à la variable } y \\ 0 & \text{sinon} \end{cases}$$

et y_i^* désigne la valeur imputée utilisée pour remplacer la valeur manquante y_i



BIAIS DE L'ESTIMATEUR IMPUTÉ

- L'estimateur imputé sera vraisemblablement biaisé si les caractéristiques des répondants sont différentes de celles des non-répondants
- Peut-on éliminer le biais de non-réponse? **Dans certains cas mais pas toujours!**
- Une chose est sûre: réduire (ou éliminer) le biais de non-réponse repose sur la disponibilité d'une **information auxiliaire** et de son **utilisation appropriée!**
- On appelle information auxiliaire un ensemble de variables qui est disponible pour toutes les unités échantillonnées (**Pas de non-réponse**)



MÉCANISME DE NON-RÉPONSE

- Soit

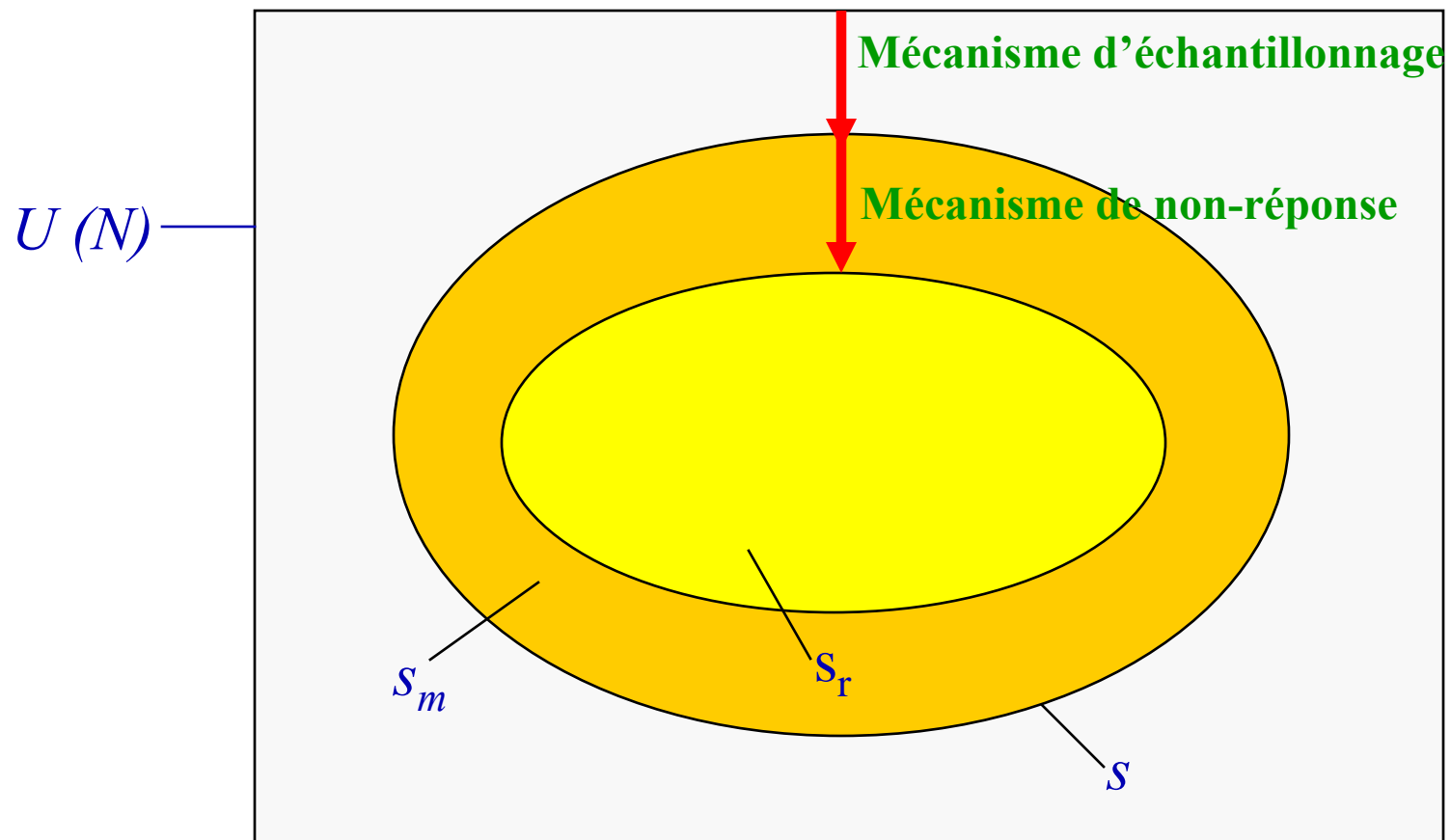
$$a_i = \begin{cases} 1 & \text{si } i \text{ a répondu à la variable } y \\ 0 & \text{sinon} \end{cases}$$

- On suppose que les unités répondent indépendamment les une des autres avec probabilité

$$p_i = P(a_i = 1)$$

- La distribution (inconnue) des a_i est appelée **mécanisme de non-réponse**

MÉCANISME DE NON-RÉPONSE





MÉCANISME DE NON-RÉPONSE

- On distingue 3 types de mécanisme de non-réponse

1. **Mécanisme uniforme:**

$$p_i = p$$

2. **Mécanisme non-confondu:** la probabilité de réponse peut dépendre d'un vecteur de variables auxiliaires \mathbf{z} mais pas de la variable d'intérêt y

$$p_i = P(a_i = 1 | y, \mathbf{z}) = P(a_i = 1 | \mathbf{z})$$

3. **Mécanisme confondu:**

$$p_i = P(a_i = 1 | y, \mathbf{z})$$



INFORMATION AUXILIAIRE

- L'imputation est avant tout **un travail de modélisation**
- Afin de réduire le biais de non-réponse, l'information auxiliaire disponible peut être utilisée à deux niveaux:
 - Peut servir à la **construction de valeurs imputées**
 - et/ou**
 - Peut servir à la **construction de classes d'imputation**



MODÈLE D'IMPUTATION VS. MODÈLE DE NON-RÉPONSE

- On distingue le modèle de non-réponse du modèle d'imputation
- **Modèle d'imputation:** C'est un modèle qui décrit la distribution de la variable d'intérêt y conditionnelle à un vecteur de variables auxiliaires \mathbf{z}_1 , $f(y | \mathbf{z}_1; \boldsymbol{\beta})$
- Par exemple, un modèle fréquemment vu en pratique est

$$m_1 : y_i = \mathbf{z}'_{1i} \boldsymbol{\beta} + \varepsilon_i$$

MODÈLE D'IMPUTATION VS. MODÈLE DE NON-RÉPONSE

- **Modèle de non-réponse:** Sert à décrire la distribution des indicateurs de réponse a_i , conditionnelle à un vecteur de variables auxiliaires \mathbf{z}_2 , $f(a | \mathbf{z}_2; \gamma)$
- Par exemple, un modèle fréquemment vu en pratique est le modèle logistique

$$m_2 : \log\left(\frac{p_i}{1 - p_i}\right) = \mathbf{z}'_{2i} \boldsymbol{\gamma},$$

$$\text{où } p_i = P(a_i = 1 | \mathbf{z}_{2i}, \boldsymbol{\gamma}).$$



RÉDUIRE LE BIAIS DE NON-RÉPONSE

- Pour réduire le biais de non-réponse, il faut que l'au moins un des deux modèles soit bien spécifié
- Est-il préférable de modéliser la variable d'intérêt y ou la probabilité de réponse à y ?
- Intuitivement, on opterait pour la modélisation de la variable d'intérêt y . Dans certains cas cependant, la modélisation peut s'avérer ardue (Haziza et Rao, 2006)



MÉTHODES D'IMPUTATION

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

- Les méthodes dites **déterministes**: Méthodes qui fournissent une valeur fixe étant donné l'échantillon
- Les méthodes dites **stochastiques ou aléatoires**: Méthodes d'imputation ayant une composante aléatoire (et donc qui ne donnent pas nécessairement la même valeur étant donné l'échantillon si la méthode est répétée)



MÉTHODES D'IMPUTATION

- La plupart des méthodes d'imputation peut être représentée par le modèle suivant:

$$m : y_i = f(\mathbf{z}_i) + \varepsilon_i,$$

$$E_m(\varepsilon_i) = 0, \quad E_m(\varepsilon_i \varepsilon_j) = 0, \quad i \neq j, \quad V_m(\varepsilon_i) = \sigma_i^2$$



MÉTHODES D'IMPUTATION

- Soit y_i^* la valeur imputée pour remplacer la valeur manquante y_i
- Imputation déterministe: $y_i^* = \hat{f}_r(\mathbf{z}_i)$
- Imputation aléatoire: $y_i^* = \hat{f}_r(\mathbf{z}_i) + e_i^*$
 - Les résidus peuvent être tirés aléatoirement parmi l'ensemble des résidus observés chez les répondants

$$e_i^* = \left[y_j - \hat{f}_r(\mathbf{z}_j) \right], \quad j \in s_r$$

MÉTHODES DÉTERMINISTES

- Imputation par régression:

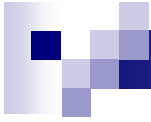
$$f(\mathbf{z}_i) = \mathbf{z}_i' \boldsymbol{\beta} \text{ et } \sigma_i^2 = \sigma^2 \boldsymbol{\lambda}' \mathbf{z}_i \Rightarrow y_i^* = \mathbf{z}_i' \hat{\boldsymbol{\beta}}_r$$

- Imputation par le ratio:

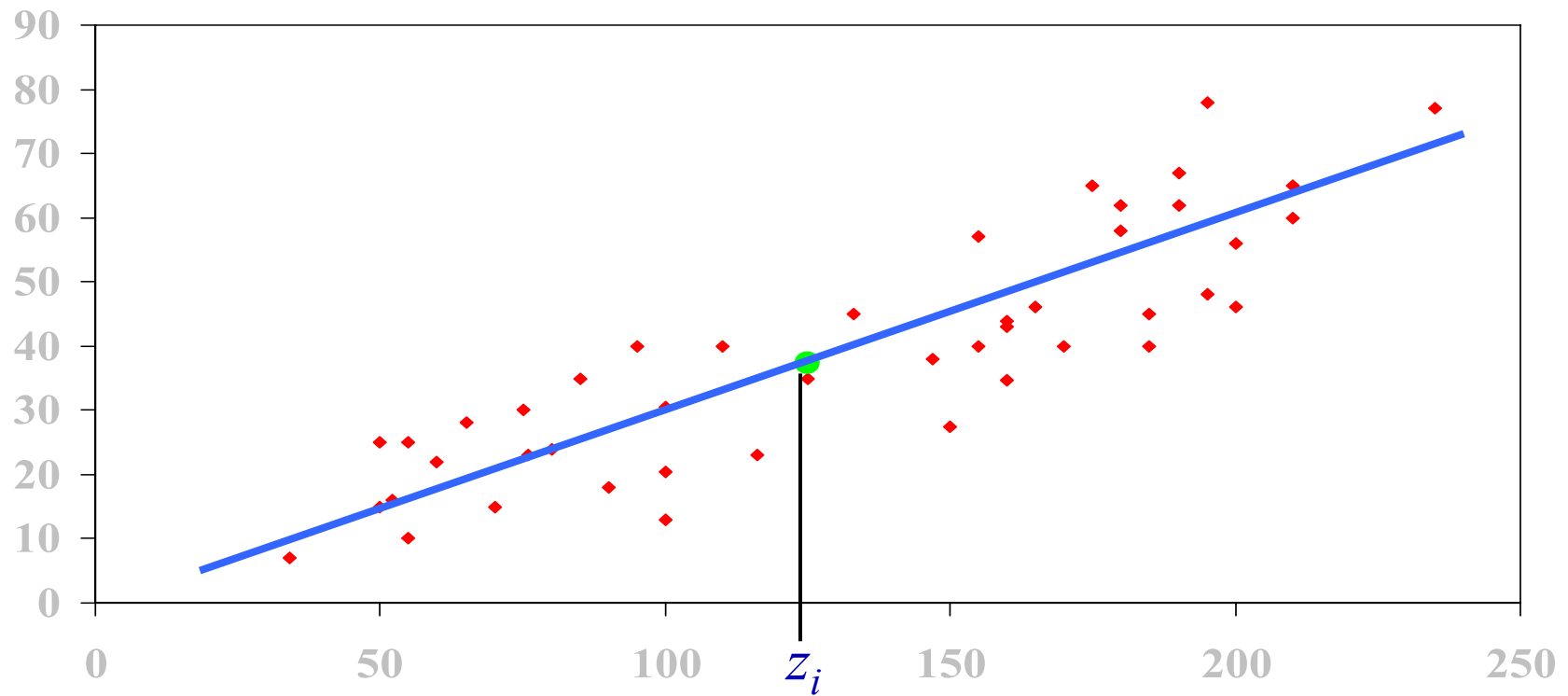
$$f(\mathbf{z}_i) = \beta z_i \text{ et } \sigma_i^2 = \sigma^2 z_i \Rightarrow y_i^* = \hat{\beta}_r z_i = \frac{\bar{y}_r}{\bar{z}_r} z_i$$

- Imputation par la moyenne:

$$z_i = 1 \forall i, \quad f(z_i) = \beta \text{ et } \sigma_i^2 = \sigma^2 \Rightarrow y_i^* = \bar{y}_r$$



IMPUTATION PAR LE RATIO



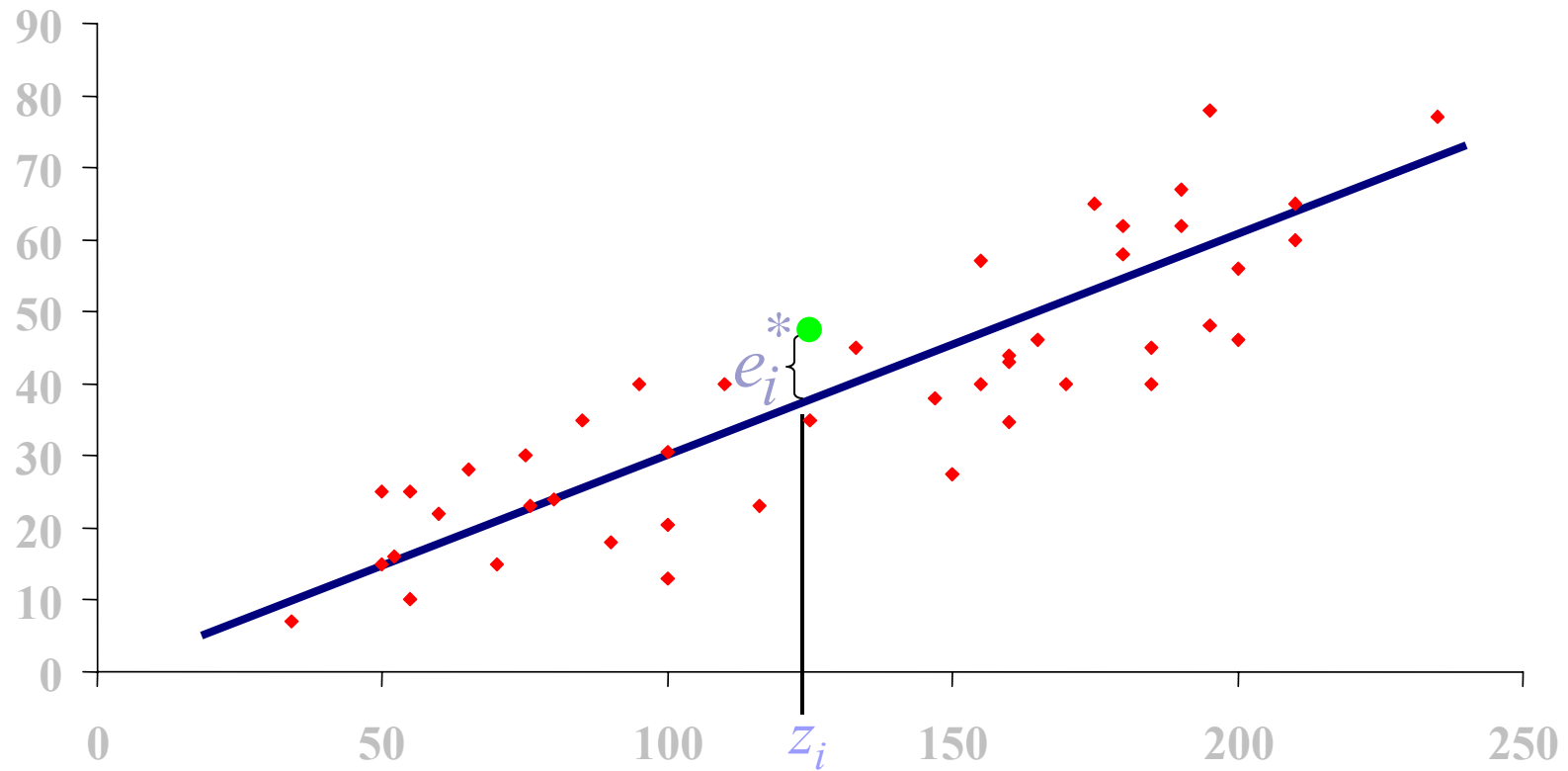
MÉTHODES ALÉATOIRES

- Imputation par hot-deck aléatoire:
 - On tire un répondant au hasard (avec remise) dans l'ensemble des répondants
 - Peut être vue comme de l'imputation par la moyenne à laquelle on a rajouté un résidu

$$y_i^* = \bar{y}_r + \underbrace{(y_j - \bar{y}_r)}_{e_i^*}, \quad j \in s_r$$

- Imputation par le ratio avec résidus: $y_i^* = \hat{\beta}_r z_i + e_i^*$

IMPUTATION PAR LE RATIO AVEC RÉSIDUS





MÉTHODES D'IMPUTATION

Méthodes déterministes

- Susceptibles de détruire la distribution des variables d'intérêt

Méthodes aléatoires

- Préservent la distribution des variables d'intérêt
- Augmente la variabilité des estimateurs

DÉCOMPOSITION DE L'ERREUR TOTALE

- L'erreur totale, $\hat{Y}_I - Y$, peut être décomposée comme suit:

$$\hat{Y}_{I23} - Y = (\hat{Y}_{23} - Y) + (\hat{Y}_{I23} - \hat{Y}_{23})$$

erreur totale erreur due à l'échantillonnage + erreur due à la non-réponse

- Dans un recensement, il n'y a pas d'erreur due à l'échantillonnage mais l'erreur due à la non-réponse est présente



BIAIS DE NON-RÉPONSE: FORMALISATION

$$\begin{aligned}\text{Biais}(\hat{Y}_I) &= E_p E_r(\hat{Y}_I - Y | s) \\ &= E_p E_r(\hat{Y} - Y | s) + E_p E_r(\hat{Y}_I - \hat{Y} | s) \\ &= E_p E_r(\hat{Y}_I - \hat{Y} | s)\end{aligned}$$

$$\text{Biais}(\hat{Y}_I) \approx 0 \Leftrightarrow E_p E_r(\hat{Y}_I - \hat{Y} | s) \approx 0$$



BIAIS QUAND LES HYPOTHÈSES NE SONT PAS VALIDES

- **Question:** Que se passe-t-il si les modèles (d'imputation ou de non-réponse) sont mal spécifiés?
- **Réponse:** L'estimateur imputé sera vraisemblablement biaisé!



ÉTUDES PAR SIMULATION

Étude 1:

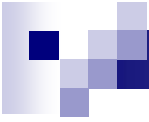
- Nous avons une population de taille $N = 1000$ comprenant deux variables y et z tel que $\rho_{yz} = 0.81$
- Nous tirons $R = 10000$ EASSR de taille $n = 100$
- Dans chaque échantillon, on génère la non-réponse de telle sorte que p_i dépend de z_i et que le taux de réponse soit 70 %



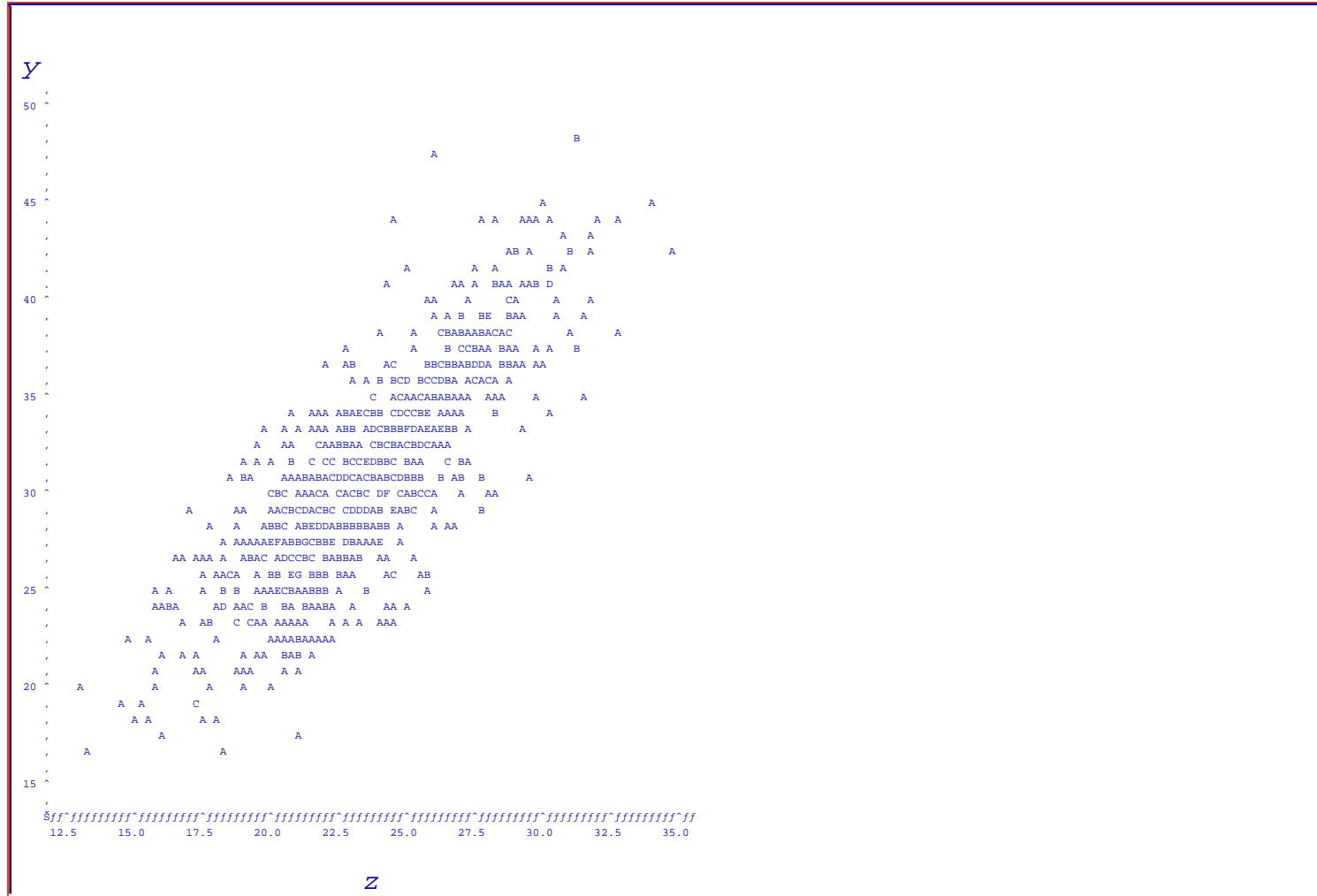
ÉTUDES DE SIMULATION

Étude 1(suite):

- Nous utilisons 3 méthodes d'imputation:
 - Imputation par la moyenne: $y_i^* = \bar{y}_r$
 - Imputation par le ratio: $y_i^* = \hat{R}_r \bar{y}_r$ et $\hat{R}_r = \bar{y}_r / \bar{z}_r$
 - Imputation par régression: $y_i^* = \hat{\beta}_{0r} + \hat{\beta}_{1r} z_i$



ÉTUDES DE SIMULATION



ÉTUDES PAR SIMULATION

Root MSE	3.151	R-Square	0.66
Dependent Mean	30.949	Adj R-Sq	0.66
Coeff Var	10.183		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	0.249	0.732	0.34	0.7333
Z	1	1.303	0.030	42.33	<.0001

ÉTUDES PAR SIMULATION

Résultats

$R = 10000$ SRS, $N = 1000$, $n = 100$ et $p = 70\%$

	Moyenne	Ratio	Régression
Biais Relatif (%)	3.99	0.038	-0.098
EQM	1.94	0.31	0.32



ÉTUDES PAR SIMULATION

Étude 2:

- Nous avons une population de taille $N = 1000$ comprenant deux variables y et z tel que $\rho_{yz} = 0.85$
- Nous tirons $R = 10000$ EASSR de taille $n = 100$
- Dans chaque échantillon, on génère la non-réponse de telle sorte que p_i dépend de z_i et que le taux de réponse soit 70 %

ÉTUDES PAR SIMULATION

Root MSE	2.893	R-Square	0.718
Dependent Mean	30.949	Adj R-Sq	0.718
Coeff Var	9.350		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	22.526	0.200	112.28	<.0001
z	1	2.241	0.046	47.88	<.0001



ÉTUDES PAR SIMULATION

Résultats

$R = 10000$ EASSR, $N = 1000$, $n = 100$ et $p = 70\%$

	Moyenne	Ratio	Régression
Biais Relatif (%)	6.58	-13.96	0.121
EQM	4.54	19.22	0.33



MORALE

- Il est important de faire un travail de modélisation minutieux afin de s'assurer que les hypothèses que l'on s'est donné au départ "tiennent la route"
- Il est important d'inclure toutes les variables auxiliaires disponibles appropriées surtout si ces variables sont corrélées avec la probabilité de réponse
- Un mauvais modèle pour le mécanisme de réponse et/ou du modèle d'imputation peut mener à des estimateurs considérablement biaisés



CLASSES D'IMPUTATION

- En pratique, on forme des classes avant d'imputer
 - car c'est plus pratique lorsque il y a plusieurs variables à imputer
 - Ça amène une certaine robustesse par rapport à l'imputation par régression si le modèle d'imputation est mal spécifié
- L'objectif des classes est de réduire (du mieux qu'on peut) le biais dû à la non-réponse



JUSTIFICATION THÉORIQUE

- Soit U une population de taille N ;
- On veut estimer la moyenne dans la population

$$\bar{Y} = \frac{1}{N} \sum_{i \in U} y_i$$

- On tire un échantillon aléatoire s selon un plan de sondage $p(\cdot)$
- On suppose que $a_i \sim B(1, p_i), i = 1, \dots, N$.



JUSTIFICATION THÉORIQUE

- Un estimateur imputé de \bar{y} est défini par

$$\bar{y}_{I,1} = \frac{1}{N} \left[\sum_{i \in s_r} w_i y_i + \sum_{i \in s_m} w_i y_i^* \right]$$

- L'indice 1 dans $\bar{y}_{I,1}$ signifie que l'estimateur est basé sur 1 classe d'imputation (c'est-à-dire, l'échantillon s)

JUSTIFICATION THÉORIQUE

- On peut montrer que, dans le cas d'imputation par la moyenne, $\bar{y}_{I,1}$ est biaisé. Le biais est donné par

$$\text{Biais}(\bar{y}_{I,1}) = \frac{1}{N\bar{P}} \sum_{i \in U} (p_i - \bar{P})(y_i - \bar{Y})$$

où $\bar{P} = \frac{1}{N} \sum_P p_i$.

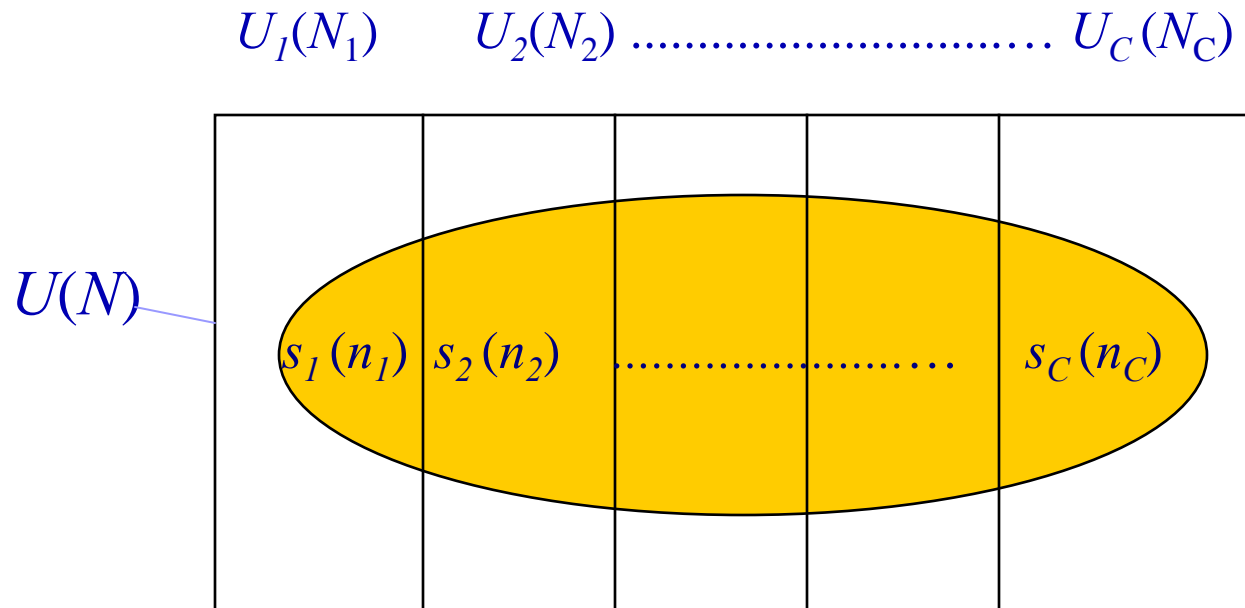
- Le biais est égal à 0 quand la covariance entre la probabilité de réponse et la variable d'intérêt est 0 dans la population



JUSTIFICATION THÉORIQUE

- Cette condition est irréaliste en pratique; c'est pourquoi, on appellera l'estimateur imputé $\bar{y}_{I,1}$ un estimateur "non-ajusté".
- L'objectif sera donc de créer des classes en partitionnant la population.

JUSTIFICATION THÉORIQUE



JUSTIFICATION THÉORIQUE

- À l'intérieur de chacune des classes, on utilise l'imputation par la moyenne, ce qui mène à l'estimateur "ajusté" suivant:

$$\bar{y}_{I,C} = \sum_{v=1}^C w'_v \bar{y}_v$$

$$\text{où } w'_v = \sum_{s_v} w_i / \sum_s w_i,$$

$$\text{et } \bar{y}_v = \frac{1}{\sum_{s_v} w_i} \left[\sum_{s_{rv}} w_i y_i + \sum_{s_{mv}} w_i y_i^* \right].$$

JUSTIFICATION THÉORIQUE

- L'estimateur ajusté $\bar{y}_{I,C}$ est lui aussi biaisé mais dans ce cas, le biais est donné par

$$\text{Biais}(\bar{y}_{I,C}) = \frac{1}{N} \sum_{v=1}^C \frac{1}{\bar{P}_v} \sum_{P_v} (p_i - \bar{P}_v)(y_i - \bar{Y}_v)$$

$$\text{où } \bar{P}_v = \frac{1}{N_v} \sum_{P_v} p_i \text{ et } \bar{Y}_v = \frac{1}{N_v} \sum_{P_v} y_i$$

- Le biais est égal à 0 si la covariance entre la probabilité de réponse et la variable d'intérêt y est 0 dans chacune des classes.



JUSTIFICATION THÉORIQUE

- L'objectif est donc de créer des classes telles que, à l'intérieur de chaque classe, les unités aient approximativement la même probabilité de réponse ET/OU les unités aient approximativement la même valeur pour la variable d'intérêt
- Les classes sont alors **HOMOGÈNES** par rapport aux probabilités de réponse ET/OU à la variable d'intérêt



CONSTRUCTION DES CLASSES

- En pratique, plusieurs méthodes sont utilisées pour former les classes dont
 - classe = strate
 - croisement de variables auxiliaires catégoriques
 - classes formées au moyen des valeurs prédites \hat{y}_i et \hat{p}_i

Little (1986) et Haziza et Beaumont (2007)



ESTIMATION DE LA VARIANCE

Pourquoi estimer la variance?

- Permet de mesurer la qualité (précision) des estimations
- Aide à tirer les bonnes conclusions
- Permet d'informer correctement les utilisateurs
- En présence de valeurs imputées, permet de fournir l'heure juste et de connaître l'impact de l'imputation
- Afin de mieux répartir les ressources entre l'échantillon et les procédures d'imputation/de suivi

ESTIMATION DE LA VARIANCE

Cas de 100% réponse (EASSR):

$$V_p(\bar{y}) = \left(\frac{1}{n} - \frac{1}{N} \right) S_y^2 \quad \text{avec} \quad S_y^2 = \frac{1}{N-1} \sum_{i \in U} (y_i - \bar{Y})^2$$



estimation

$$v(\bar{y}) = \left(\frac{1}{n} - \frac{1}{N} \right) s_y^2 \quad \text{avec} \quad s_y^2 = \frac{1}{n-1} \sum_{i \in S} (y_i - \bar{y})^2$$



ESTIMATION DE LA VARIANCE

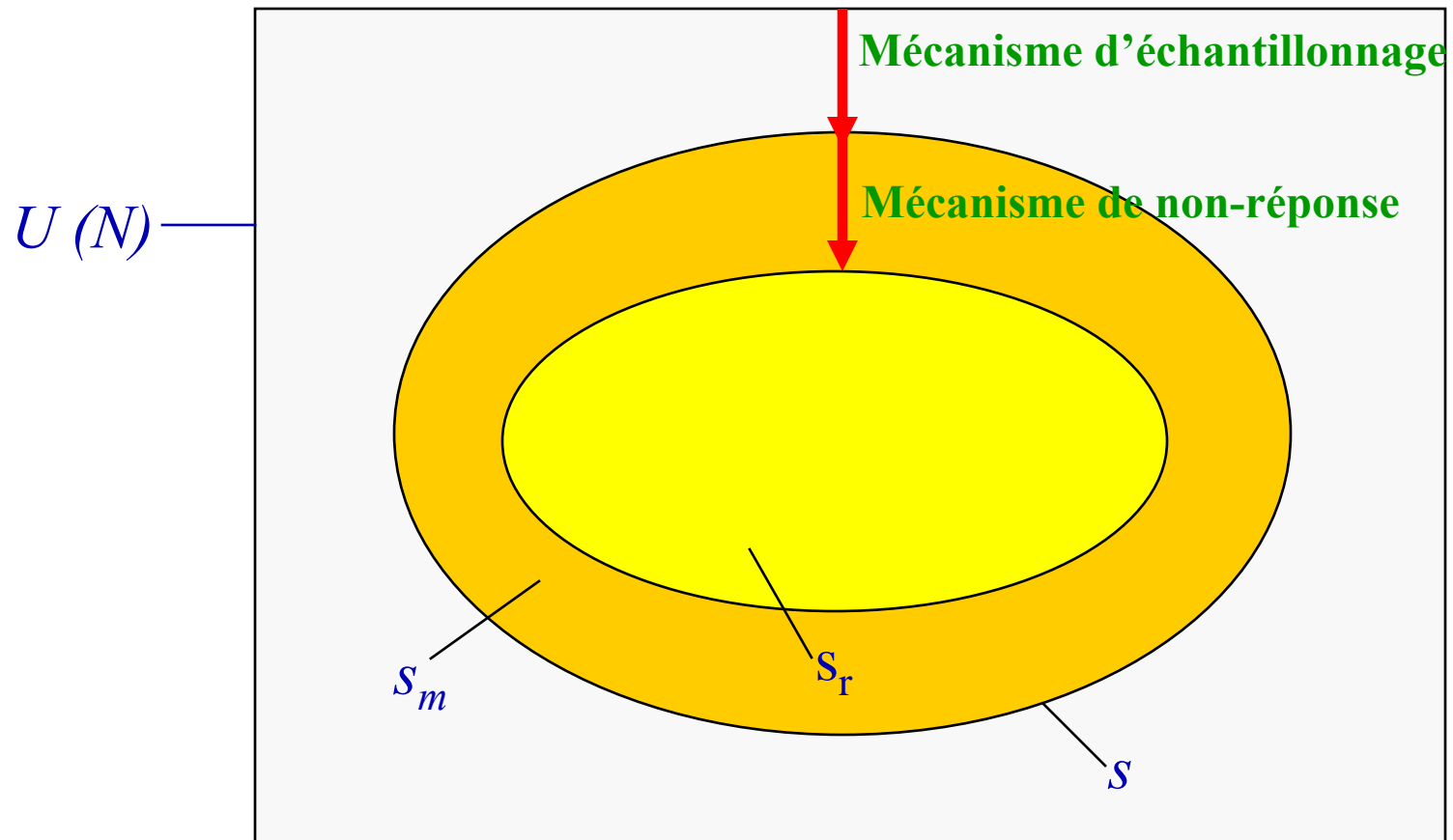
- L'approche deux-phases

$$U \longrightarrow s \longrightarrow (s_r, s_m)$$

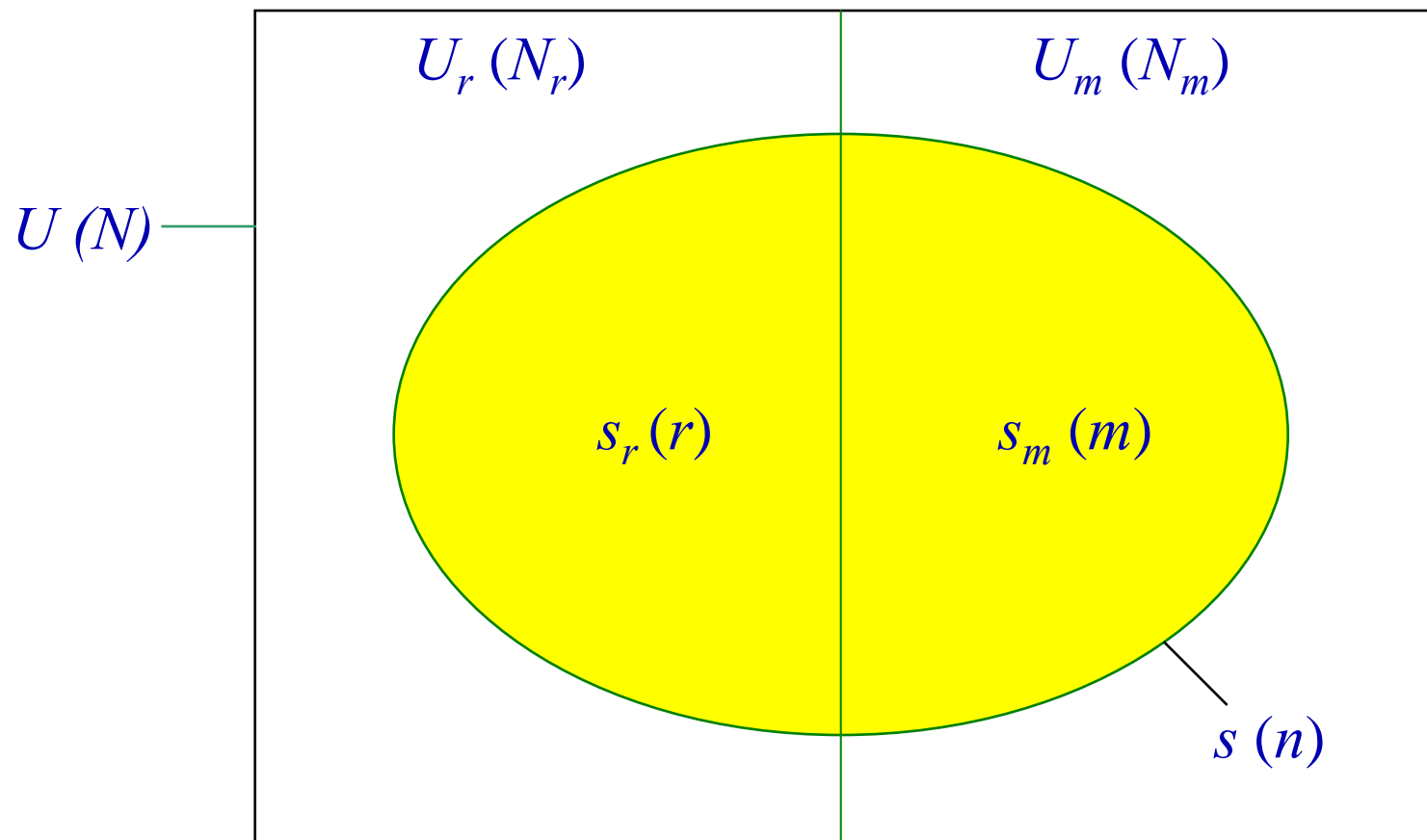
- L'approche renversée (Fay, 1991)

$$U \rightarrow (U_r, U_m) \rightarrow (s_r, s_m)$$

APPROCHE DEUX-PHASES



APPROCHE RENVERSÉE



DEUX-PHASES: MODÈLE DE NON-RÉPONSE

- Sous un mécanisme de non-réponse uniforme, la variance de l'estimateur imputé est donnée par

$$V(\hat{Y}_I) = V_p E_r(\hat{Y}_I | s, r) + E_p V_r(\hat{Y}_I | s, r)$$

$$\approx N^2 \left(\frac{1}{E(r)} - \frac{1}{N} \right) S_y^2$$

- Un estimateur correct de la variance est donné par

$$v_{cor}(\hat{Y}_I) = N^2 \left(\frac{1}{r} - \frac{1}{N} \right) S_{yr}^2$$

DEUX-PHASES: MODÈLE DE NON-RÉPONSE

- Si l'on traite les valeurs imputées comme si elles avaient été observées, on obtient l'estimateur incorrect de la variance

$$v_{inc}(\hat{Y}_I) = N^2 \left(\frac{1}{n} - \frac{1}{N} \right) \frac{r-1}{n-1} s_{yr}^2$$

- On a

$$\frac{v_{cor}(\hat{Y}_I)}{v_{inc}(\hat{Y}_I)} \approx \left(\frac{n}{r} \right)^2$$

- Si le taux de réponse est 50%, alors $\frac{v_{cor}(\hat{Y}_I)}{v_{inc}(\hat{Y}_I)} \approx 4$.

DEUX-PHASES: MODÈLE D'IMPUTATION

- Cette approche est due à Särndal (1990)
- Basée sur la décomposition:

$$\underbrace{\hat{Y}_I - Y}_{\text{Erreur totale}} = \underbrace{\left(\hat{Y}_2 - Y \right)}_{\substack{\text{Erreur due} \\ \text{à l'échantillonnage}}} + \underbrace{\left(\hat{Y}_I - \hat{Y}_2 \right)}_{\substack{\text{Erreur due} \\ \text{à la non-réponse}}}$$



DEUX-PHASES: MODÈLE D'IMPUTATION

La variance de l'estimateur imputé est donnée par

$$\begin{aligned} V(\hat{Y}_I - Y) &= E(\hat{Y}_I - Y)^2 \\ &= V_{éch} + V_{nr} + V_{mix} \end{aligned}$$

Pour des méthodes d'imputation simples, $V_{mix} = 0$



L'APPROCHE RENVERSÉE

Approche deux-phases:

$$V(\hat{Y}_I) = V_p E_r(\hat{Y}_I | s) + E_p V_r(\hat{Y}_I | s)$$

Approche renversée:

$$V(\hat{Y}_I) = E_r V_p(\hat{Y}_I | a_i) + V_r E_p(\hat{Y}_I | a_i)$$

L'APPROCHE RENVERSÉE

On estime chacune des composantes séparément




- Estimation de $V_1 = E_r V_p (\bar{y}_I - \bar{Y} | a_i)$
- Estimation de $V_2 = V_r E_p (\bar{y}_I - \bar{Y} | a_i)$
- Lorsque la fraction de sondage n/N est négligeable, la composante V_2 est négligeable par rapport à V_1
- L'estimateur v_1 de V_1 ne dépend pas du mécanisme de réponse et/ou du modèle d'imputation → robuste

L'APPROCHE RENVERSÉE

Estimation de V_I

Exemple: Imputation par la moyenne, EASSR

$$\hat{Y}_I = N\bar{y}_r = N \frac{\sum_{i \in s} a_i y_i}{\sum_{i \in s} a_i}$$

- Estimation de V_I  Linéarisation de Taylor
-  Jackknife
-  Bootstrap



LE JACKKNIFE

- Soit $\hat{\theta}$ un estimateur d'un paramètre "lisse" θ
- L'approche jackknife fonctionne comme suit:
 - (i) Enlever l'unité j
 - (ii) Ajuster les poids de sondage
 - (iii) Calculer $\hat{\theta}$ avec les poids ajustés $\longrightarrow \hat{\theta}_{(j)}$
 - (iv) Replacer l'unité enlevée à l'étape (i), enlever la prochaine unité et recalculer $\hat{\theta}$
 - (v) Répéter (i)-(iv) jusqu'à ce que toutes les unités aient été enlevées

LE JACKKNIFE

- La variance jackknife de $\hat{\theta}$ est alors obtenue en estimant la variabilité des $\hat{\theta}_{(j)}$, c'est-à-dire,

$$v_J(\hat{\theta}) = \frac{n-1}{n} \sum_{j=1}^n (\hat{\theta}_{(j)} - \hat{\theta})^2$$

$$\text{où } w_{i(j)} = \begin{cases} \frac{n}{n-1} w_i & \text{si } i \neq j \\ 0 & \text{si } i = j \end{cases}$$

LE JACKKNIFE

- En présence de non-réponse, si l'on traite les valeurs imputées comme si elles avaient été observées, le jackknife "traditionnel" mène généralement à une sous-estimation de la variance de l'estimateur imputé
- **Exemple:** $\theta = Y$ et imputation par la moyenne

$$v_J(\hat{Y}_I) = N^2 \frac{r-1}{n-1} \frac{s_r^2}{n} = v_{inc}(\hat{Y}_I)$$



LE JACKKNIFE

Le Jackknife ajusté: (Rao-Shao, 1992) Le Jackknife ajusté est calculé de la même manière que le jackknife traditionnel sauf que

- lorsqu'une unité répondante, $j \in s_r$, est éliminée, chacune des valeurs imputées y_i^* est ajustée
- lorsque qu'une unité non-répondante, $j \in s_m$, les valeurs imputées sont laissées telles quelles

LE JACKKNIFE

Imputation par la moyenne:

$$\begin{aligned} v_{JRS}(\hat{Y}_I) &= \frac{n-1}{n} \sum_{j \in S} (\hat{Y}_{I(j)}^a - \hat{Y}_I)^2 \\ &= N^2 \frac{s_r^2}{r} = v_{cor}(\hat{Y}_I) \end{aligned}$$

- Le jackknife de Rao-Shao est un estimateur de V_I dans l'approche renversée



LE JACKKNIFE

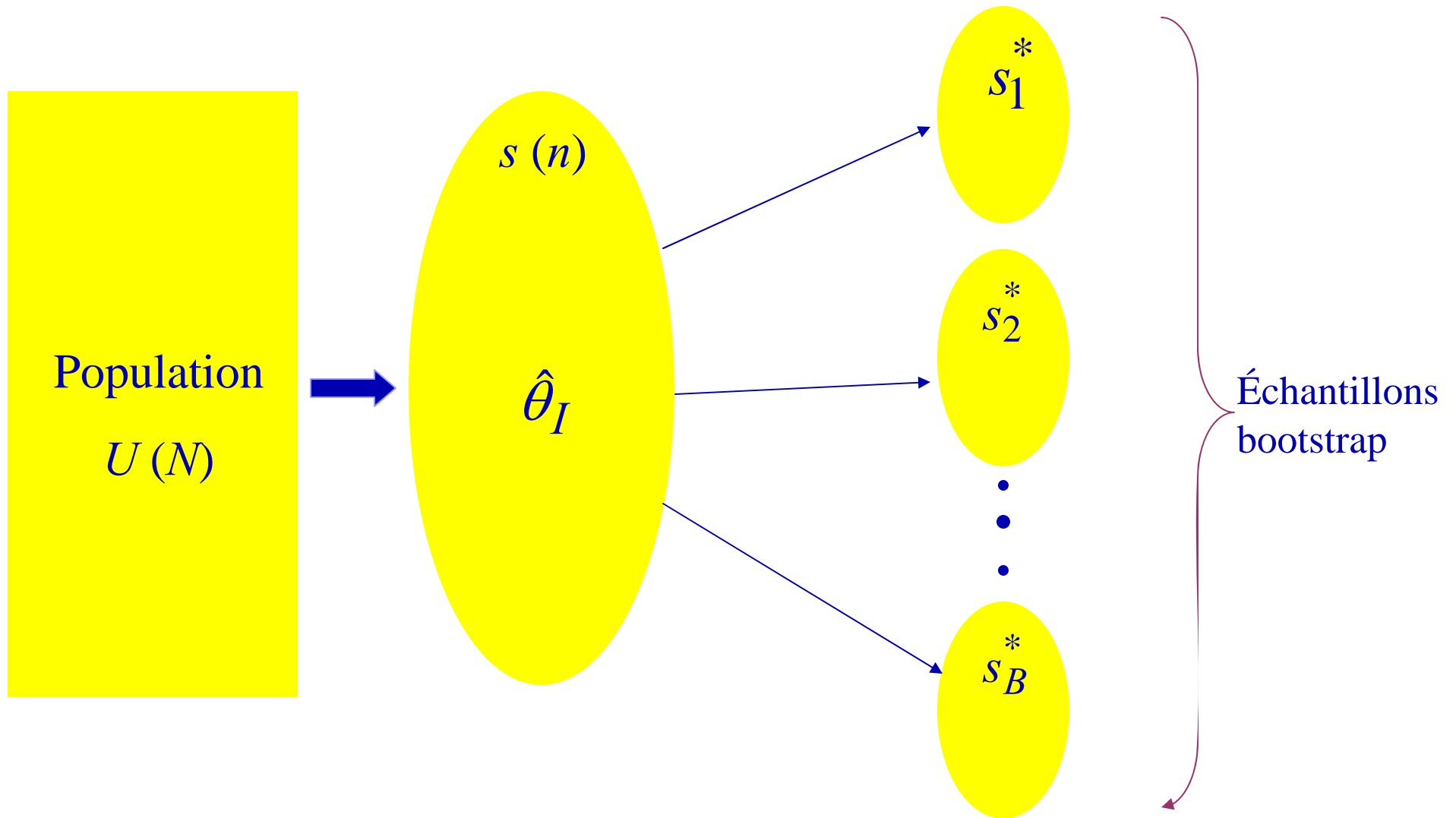
- La méthode peut être appliquée à plusieurs méthodes d'imputation (hot-deck aléatoire, moyenne, ratio, régression,...)
- La méthode peut être utilisée pour des plans complexes (stratifiés à degrés multiples)
- La méthode ne peut être utilisée pour des fonctions non-lisses de totaux telles médiane, quantiles,.....
- **Suppose que les unités ont été tirées avec remise ou que la fraction de sondage est négligeable**



LE BOOTSTRAP

- L'adaptation du bootstrap en présence d'imputation a été proposée par Shao et Sitter (1996)
- L'application du bootstrap "traditionnel" mène généralement à une sous-estimation de la variance de l'estimateur imputé
- Shao et Sitter ont proposé de réimputer dans chaque échantillon bootstrap en utilisant la même méthode/procédure qui a été utilisée pour imputer les valeurs dans le fichier de données original

LE BOOTSTRAP



LE BOOTSTRAP

- La variance bootstrap de $\hat{\theta}_I$ est donnée par

$$v_B(\hat{Y}_I) = \frac{1}{B-1} \left[\sum_{b=1}^B \left(\hat{Y}_{I(b)}^* - \bar{\hat{Y}}_I^* \right)^2 \right]$$

$$\text{où } \bar{\hat{Y}}_I^* = \frac{1}{B} \sum_{b=1}^B \hat{Y}_{I(b)}^*$$

- $v_B(\hat{Y}_I)$ est un estimateur de V_I dans le cas de l'approche renversée



LE BOOTSTRAP

- La méthode peut être appliquée à plusieurs méthodes d'imputation (hot-deck aléatoire, moyenne, ratio, régression,...)
- La méthode peut être utilisée pour des plans complexes (stratifiés à degrés multiples)
- La méthode peut être utilisée pour des fonctions non-lisses de totaux telles médiane, quantiles,....
- Suppose que les unités ont été tirées avec remise ou que la fraction de sondage est négligeable

PARAMÈTRES COMPLEXES

- Moyenne d'un domaine: $\bar{Y}_d = \sum_{i \in U} x_i y_i / \sum_{i \in U} x_i$
- Coefficient de régression: $\mathbf{B}_N = \left(\sum_{i \in U} \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \sum_{i \in U} \mathbf{x}_i y_i$
- Coefficient de corrélation: $\rho_{xy} = \frac{1}{N-1} \frac{\sum_{i \in U} x_i y_i - N \bar{X} \bar{Y}}{S_x S_y}$

DOMAINES

- Des estimations au niveau des domaines sont presque toujours requises

- Moyenne d'un domaine:
$$\bar{Y}_d = \frac{\sum_{i \in U} x_i y_i}{\sum_{i \in U} x_i}$$

où $x_i = \begin{cases} 1 & \text{si l'unité } i \text{ appartient au domaine } d \\ 0 & \text{sinon} \end{cases}$

DOMAINES

- Un estimateur imputé de \bar{Y}_d est donné par

$$\bar{y}_{dI} = \frac{1}{\sum_{i \in S} w_i x_i} \left[\sum_{i \in S_r} w_i x_i y_i + \sum_{i \in S_m} w_i x_i y_i^* \right]$$

- Les domaines ne sont pas toujours connus au stade de l'imputation
- Lorsque l'on impute, on peut tenir compte ou pas des domaines pour construire les valeurs imputées (ou construire les classes d'imputation)

DOMAINES

Étude de simulation: Données provenant de l'EPA

- Population de taille $N = 11270$
- Revenu hebdomadaire moyen dans la population: $\bar{Y} = \$555$

Age	15-19	20-24	25-29	30-34	35-39
Revenu hebdomadaire	139.7	343.6	513.9	587.2	625.6

Age	40-44	45-49	50-59	60-64	65+
Revenu hebdomadaire	661.5	704.5	692.4	629.6	549.2



DOMAINES

- Nous avons tiré $R = 5000$ EASSR de taille $n = 500$ de la population
- Dans chaque échantillon, la non-réponse a été générée selon un mécanisme uniforme avec probabilité 0.7
- Pour imputer nous utilisons deux méthodes:
 - moyenne des répondants \bar{y}_r (sans tenir compte des domaines)
 - moyenne des répondants à l'intérieur du domaine \bar{y}_{dr}

DOMAINES

- Résultats:

Biais relatif (%) de l'estimateur imputé \bar{y}_{dl}

	$y_i^* = \bar{y}_{dr}$	$y_i^* = \bar{y}_r$
Domaine 1 15-19	0.5	88
Domaine 4 30-34	0.4	-2.5



COEFFICIENT DE CORRÉLATION

$$\rho_{xy} = \frac{1}{N-1} \frac{\sum_{i \in P} x_i y_i - N\bar{X}\bar{Y}}{S_x S_y}$$

- Les deux variables x et y sont susceptibles d’être manquantes
- Shao et Wang (2002) ont proposé une méthode d’imputation qui mène à un estimateur imputé approximativement sans biais
- Skinner et Rao (2002) et Haziza et Rao (2002b) ont proposé un estimateur ajusté après imputation “traditionnelle”



CONCLUSIONS

- L'imputation est un travail de modélisation
- Beaucoup a été fait! Beaucoup reste à faire!
 - Paramètres complexes (quantiles, etc)
 - Plans complexes
 - Préserver la structure multivariée des données