INFÉRENCE EN PRÉSENCE D'IMPUTATION: UN SURVOL

David Haziza

Statistique Canada

JMS

16 Décembre 2002

"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. QUELQUES MÉTHODES D’IMPUTATION
3. ESTIMATION PONCTUELLE
4. CLASSES D’IMPUTATION
5. ESTIMATION DE LA VARIANCE
6. DISTORSION DES RELATIONS

NIVEAUX DE NON-RÉPONSE

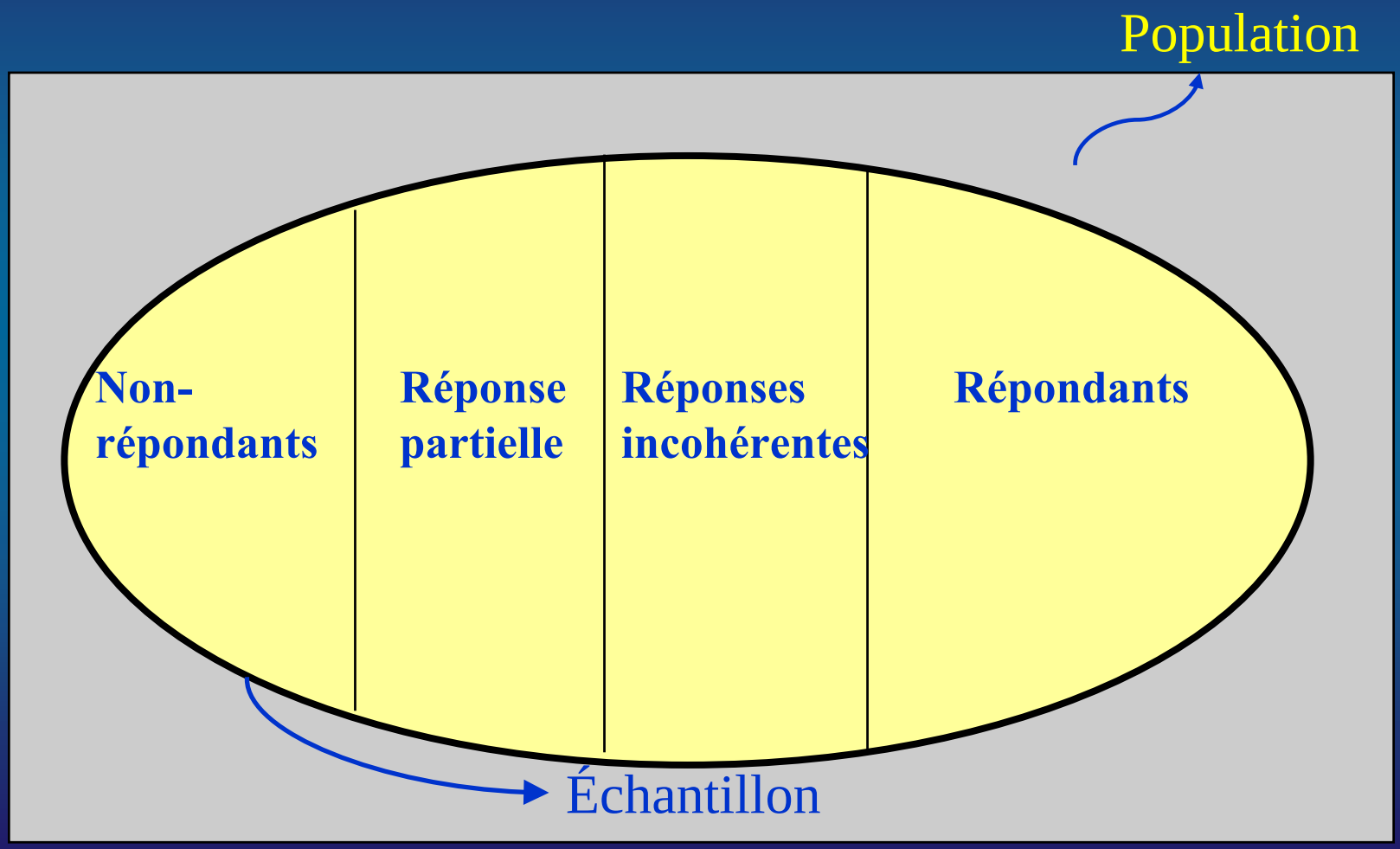
1. Non-réponse totale (par unité):

   - Absence complète d'information sur une unité

2. Non-réponse partielle (par item):

   - Certaines (mais pas toutes) variables recueillies

NIVEAUX DE NON-RÉPONSE



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 2 valeurs artificielles pour "boucher le trou" de la valeur manquante (Rubin,1978)
- L'imputation peut être effectuée par ordinateur ou manuellement

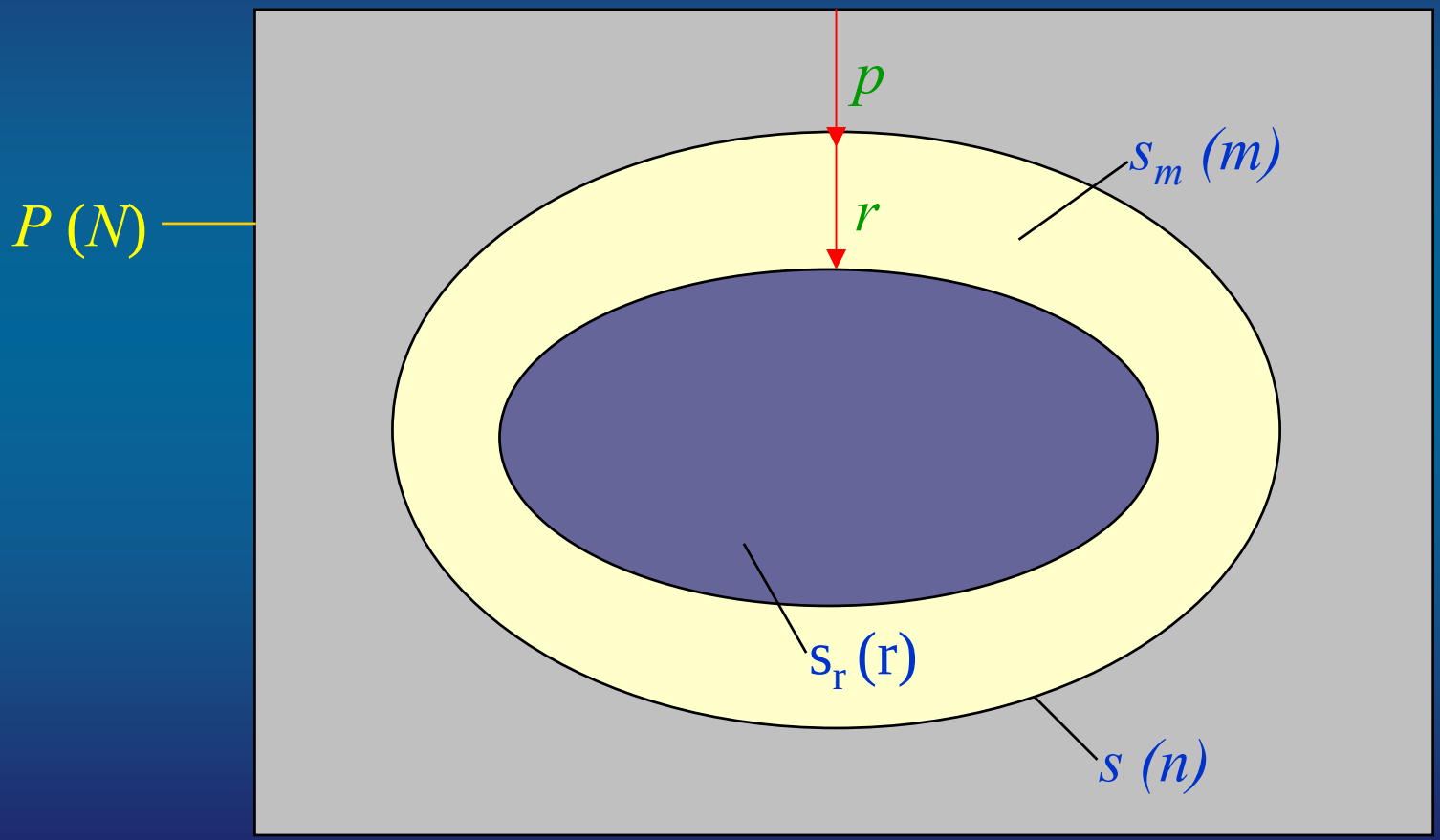
POURQUOI IMPUTE-T-ON ?

- 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 re-pondération, 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

NON-RÉPONSE VS DEUX-PHASES



MÉCANISMES DE NON-RÉPONSE

- Supposons que les unités répondent indépendamment les une des autres avec probabilité p_i, c'est-à-dire que le mécanisme de réponse est décrit par

$$a_i \sim B(1, p_i), i = 1, \ldots, N$$

- On distingue 3 types de mécanismes de non-réponse:

  (i) uniforme (MCAR)

  (ii) ignorable (MAR, non-confondu)

  (iii) non-ignorable (NMAR, confondu)

MÉCANISMES DE NON-RÉPONSE

1. Un mécanisme de non-réponse est dit uniforme quand pour toutes les unités dans la population $p_i = p$
2. Ce mécanisme, pris tel quel, n'est pas réaliste en pratique. Cependant, on peut supposer un mécanisme uniforme à l'intérieur de classes d'imputation
3. Si $P(a_i = 1 | y, \mathbf{z_i}) = P(a_i = 1 | \mathbf{z_i})$ alors le mécanisme de réponse est ignorable (Rubin, 1976)
4. Un mécanisme de réponse est dit non-ignorable s'il n'est pas ignorable

MÉCANISMES DE NON-RÉPONSE

1. Le problème lorsque l'on est en présence de non-réponse non-ignorable est le biais
2. Dans ce cas, l'élimination du biais requière des techniques plus sophistiquées (Qin, Leung et Shao, 2002)
3. Pour réduire le biais, il est important d'inclure toute l'information auxiliaire disponible (si appropriée)

INFORMATION AUXILIAIRE

- L’imputation est avant tout un travail de modélisation
- La qualité des estimations repose donc sur la disponibilité d’information auxiliaire de qualité
- 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

MÉTHODES D'IMPUTATION

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

- Plusieurs méthodes d'imputation peuvent être représenter par le modèle suivant (Kalton et Kasprzyk, 1986):

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

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

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

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

CAS PARTICULIERS

- Imputation par régression:

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

- 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:

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

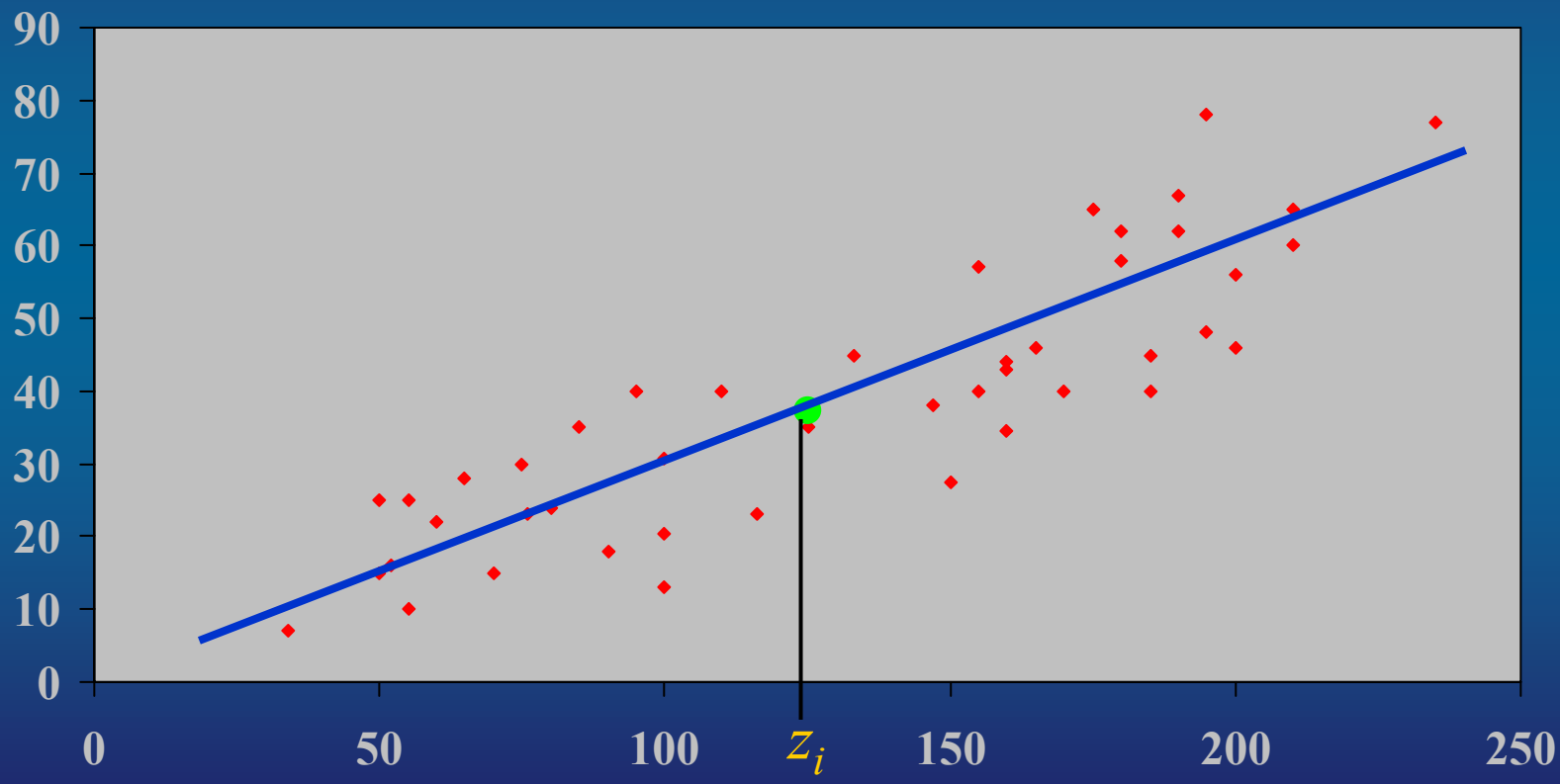
CAS PARTICULIERS

- 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

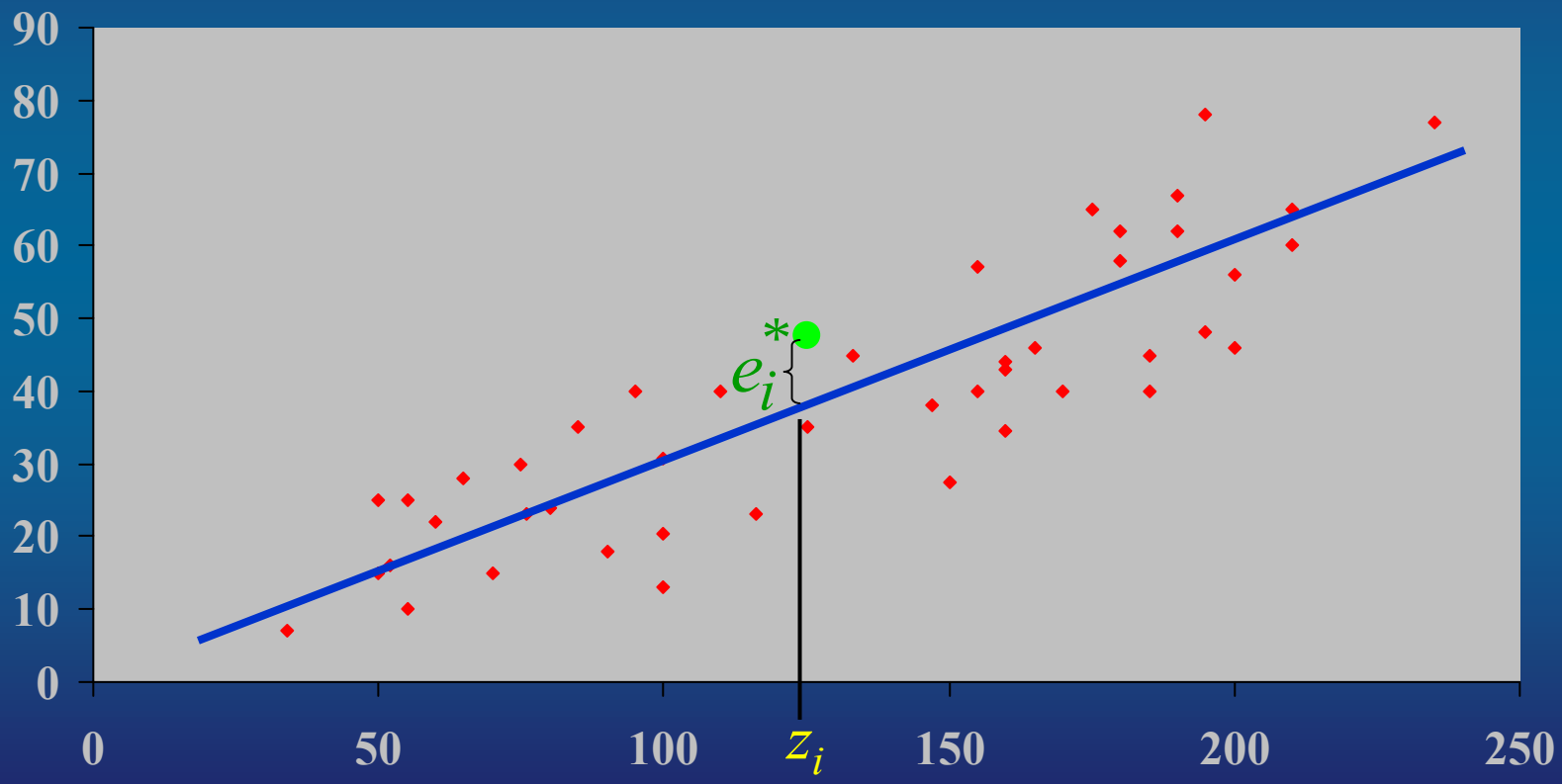
$$y_i^* = \bar{y}_r + \underbrace{\left(y_j - \bar{y}_r\right)}_{e_i^*}, \qquad j \in s_r$$

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

Imputation par le ratio



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

• Une bonne description des méthodes d'imputation est donnée dans Kovar et Whitridge (1995)

ESTIMATION PONCTUELLE

APPROCHES POUR L'INFERENCE

- Pour faire de l'inférence en présence d'imputation simple, deux approches ont été proposées:

  (1) L'approche basée sur le plan de sondage (BP) (Rao, 1990)

  (2) L'approche basée sur un modèle (BM) (Särndal, 1990,1992)

APPROCHES POUR L'INFERENCE

1. Approche BP: On suppose qu'à l'intérieur de chaque classe, le mécanisme de réponse est uniforme.
2. Approche BM: On suppose, qu'à l'intérieur de chaque classe, le mécanisme de réponse est ignorable. On fait alors appel à un modèle d'imputation généralement de la forme

$$m: y_i = \mathbf{z}_i'\boldsymbol{\beta} + \varepsilon_i,$$

$$E_m(\varepsilon_i) = 0; \quad E_m(\varepsilon_i \varepsilon_j) = 0, \ i \neq j; \quad V_m(\varepsilon_i) = \sigma^2 \boldsymbol{\lambda}' \mathbf{z}_i;$$

ESTIMATION PONCTUELLE

- En l'absence de non-réponse, un estimateur approximativement sans biais de la moyenne de la population $\bar{Y} = \frac{1}{N}\sum_{i \in P} y_i$ est donné par

$$\bar{y} = \frac{1}{\sum_{i \in s} w_i} \sum_{i \in s} w_i y_i$$

où $w_i = 1/\pi_i$ est le poids de sondage de l'unité i et $\pi_i = P(i \in s)$.

- Échantillon aléatoire simple sans remise: $w_i = N/n$

ESTIMATION PONCTUELLE

- Soit s_r l'ensemble des répondants à l'item y, de taille r et s_m l'ensemble des non-répondants à l'item y, de taille m

- En présence de non-réponse à la variable y, on définit un estimateur imputé de la moyenne $\bar{Y} = \frac{1}{N}\sum_{i \in P} y_i$ par

$$\bar{y}_I = \frac{1}{\sum_{i \in s} w_i}\left[\sum_{i \in s_r} w_i y_i + \sum_{i \in s_m} w_i y_i^*\right]$$

BIAIS DE L'ESTIMATEUR IMPUTÉ

Question: L'estimateur imputé $\bar{y}_I$ est-il sans biais pour la moyenne de la population $\bar{Y}$?

Réponse: Cela dépend d'au moins un facteur:

- la validité des hypothèses à propos du mécanisme de non-réponse et/ou du modèle d'imputation

BIAIS DE L'ESTIMATEUR IMPUTÉ

Exemple: Imputation par régression: $y_i^* = \mathbf{z}_i'\hat{\boldsymbol{\beta}}_r$

(i) Approche BP

$$E(\bar{y}_I - \bar{Y}) = E_p E_r(\bar{y}_I - \bar{Y} \mid s) \approx 0$$

(ii) Approche BM

$$E(\bar{y}_I - \bar{Y}) = E_p E_m(\bar{y}_I - \bar{Y} \mid s) = 0$$

⟹ Si les hypothèses sont satisfaites, l'estimateur imputé sera approximativement sans biais

BIAIS QUAND LES HYPOTHÈSES NE SONT PAS VALIDES

- Question: Qu'en est-il si les hypothèses à propos du mécanisme de non-réponse et/ou du modèle d'imputation ne sont pas valides?

- Réponse: L'estimateur imputé sera vraisemblablement biaisé!

ÉTUDES DE SIMULATION

Étude 1:

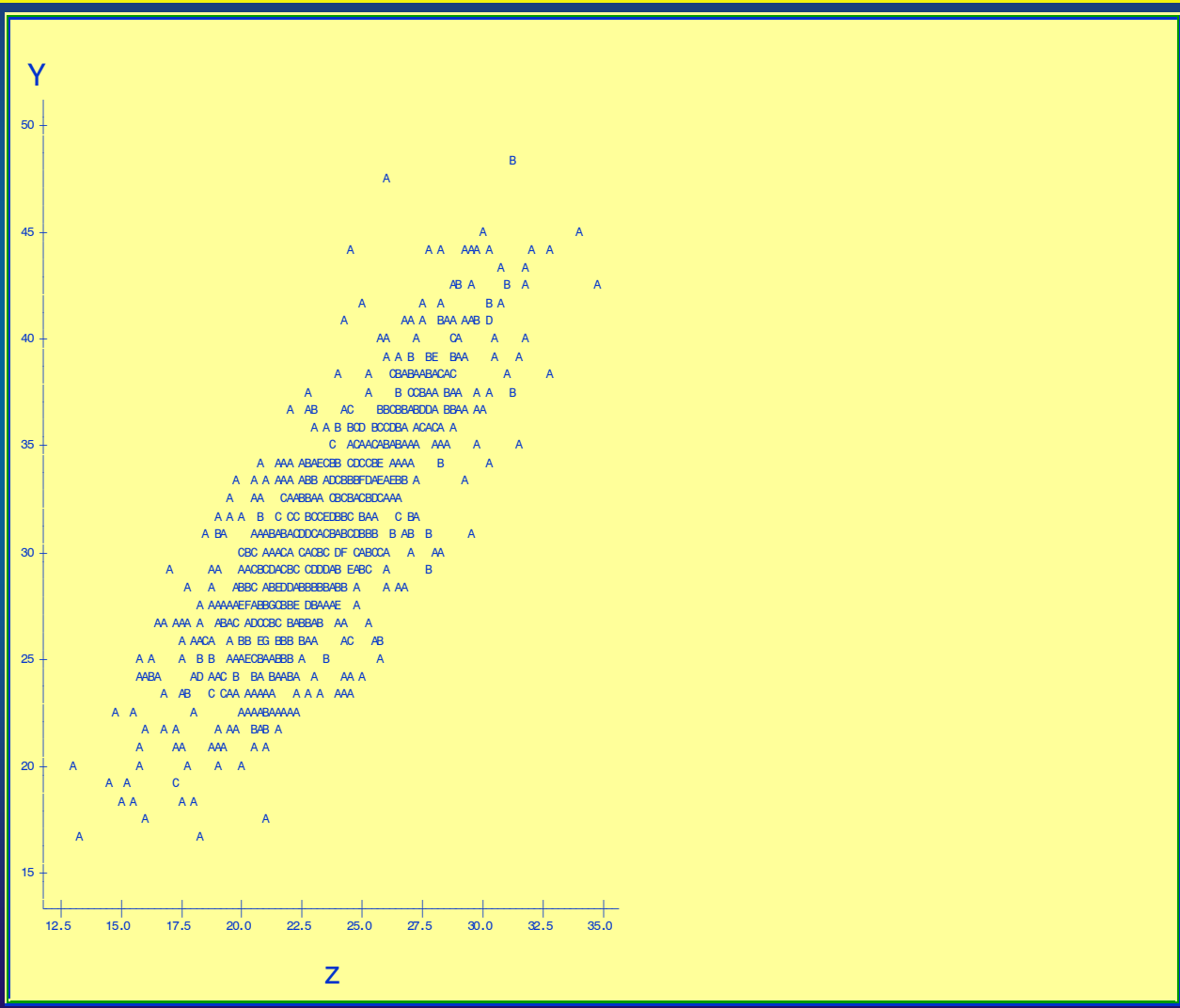
- Nous avons une population de taille $N = 1000$ comprenant deux variables y et z tel que $\rho_{yz} = 0.81$
- Nous tirons $R = 10000$ EASSR de taille $n = 100$
- Dans chaque échantillon, on génère la non-réponse de telle sorte que p_i dépend de z_i et que le taux de réponse soit 70 %

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 DE 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 DE SIMULATION

Résultats

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

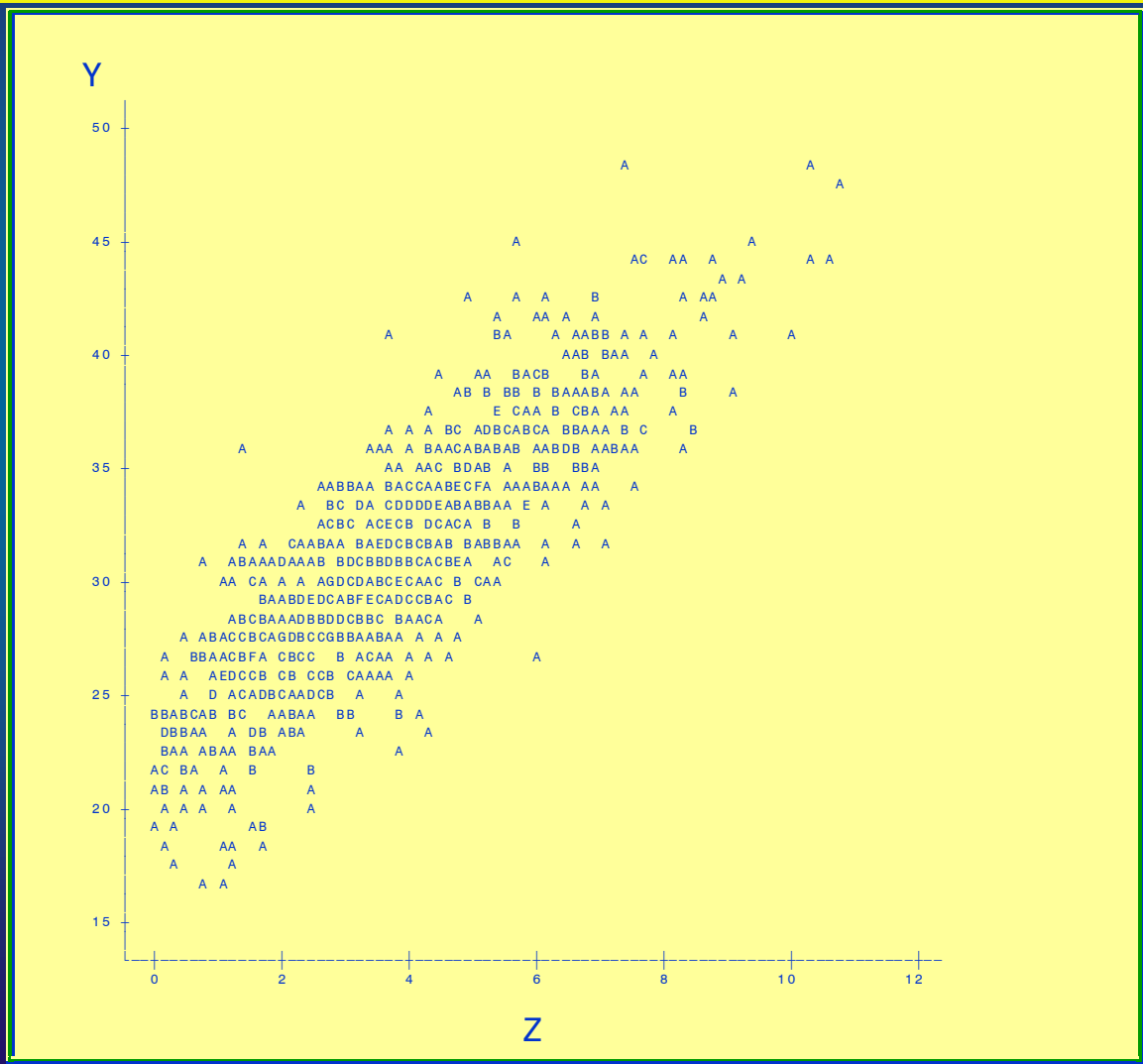
	Moyenne	Ratio	Régression
Biais Relatif (%)	3.99	0.038	-0.098
EQM	1.94	0.31	0.32

ÉTUDES DE 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 DE SIMULATION



ÉTUDES DE 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 DE 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

BIAIS QUAND LES HYPOTHÈSES NE SONT PAS VALIDES

- 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 non-réponse et/ou du modèle d'imputation peut mener à des estimations considérablement biaisées

CLASSES D'IMPUTATION

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

$$\bar{Y} = \frac{1}{N}\sum_{P} y_i$$

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

JUSTIFICATION THÉORIQUE

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

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

- 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 hot-deck aléatoire, $\bar{y}_{I,1}$ est biaisé. Le biais est donné par

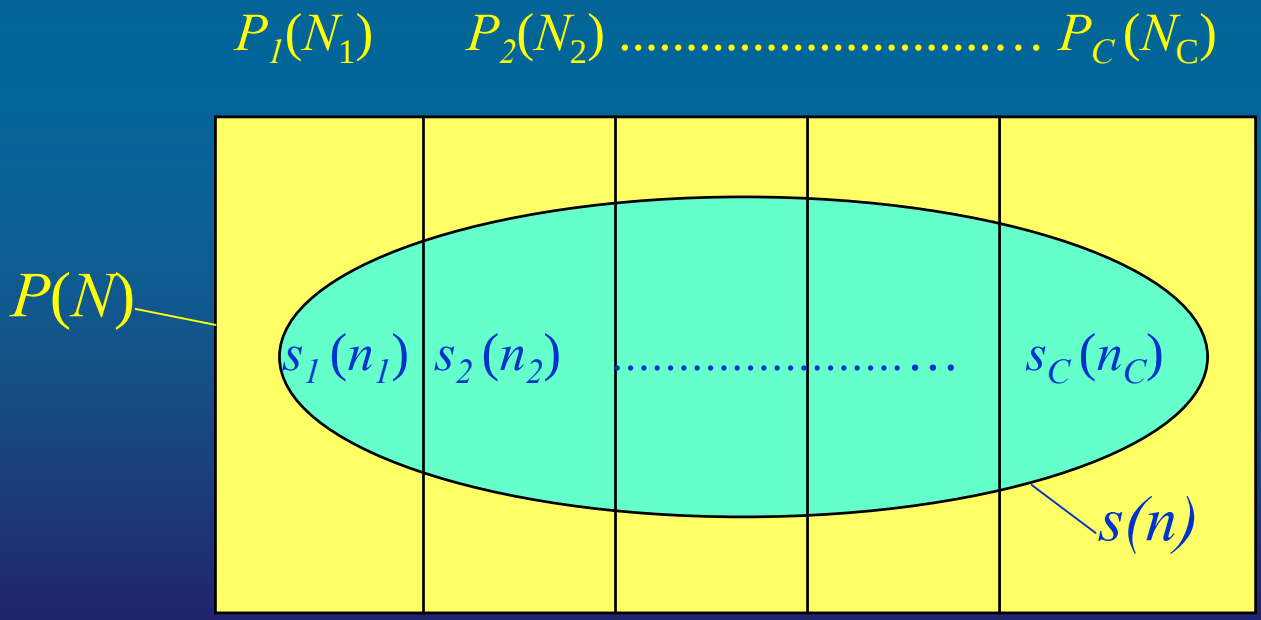
$$Biais\,(\bar{y}_{I,1}) = E\left(\bar{y}_{I,1} - \bar{Y}\right) = E_p E_r\left(\bar{y}_{I,1} - \bar{Y}\right)$$
$$= \frac{1}{N\bar{P}} \sum_P (p_i - \bar{P})(y_i - \bar{Y})$$

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

- L’objectif sera donc de créer des classes en partitionnant la population.



JUSTIFICATION THÉORIQUE

- À l'intérieur de chacune des classes, on utilise l'imputation par hot-deck aléatoire, ce qui mène à l'estimateur

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

où

$$w'_v = \sum_{s_v} w_i \Big/ \sum_{s} w_i, \quad \text{et} \quad \bar{y}_v = \frac{1}{\sum_{s_v} w_i} \left[\sum_{s_{r_v}} w_i y_i + \sum_{s_{m_v}} w_i y_i^* \right].$$

JUSTIFICATION THÉORIQUE

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

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

où $\bar{P}_v = \frac{1}{N_v}\sum_{P_v} p_i$ 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
- Pour obtenir des classes homogènes, il faudra donc effectuer un bon travail de modélisation

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 $\hat{p}_i$

ESTIMATION DE LA VARIANCE

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):

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

estimation

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

ESTIMATION DE LA VARIANCE

- L'approche deux-phases

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

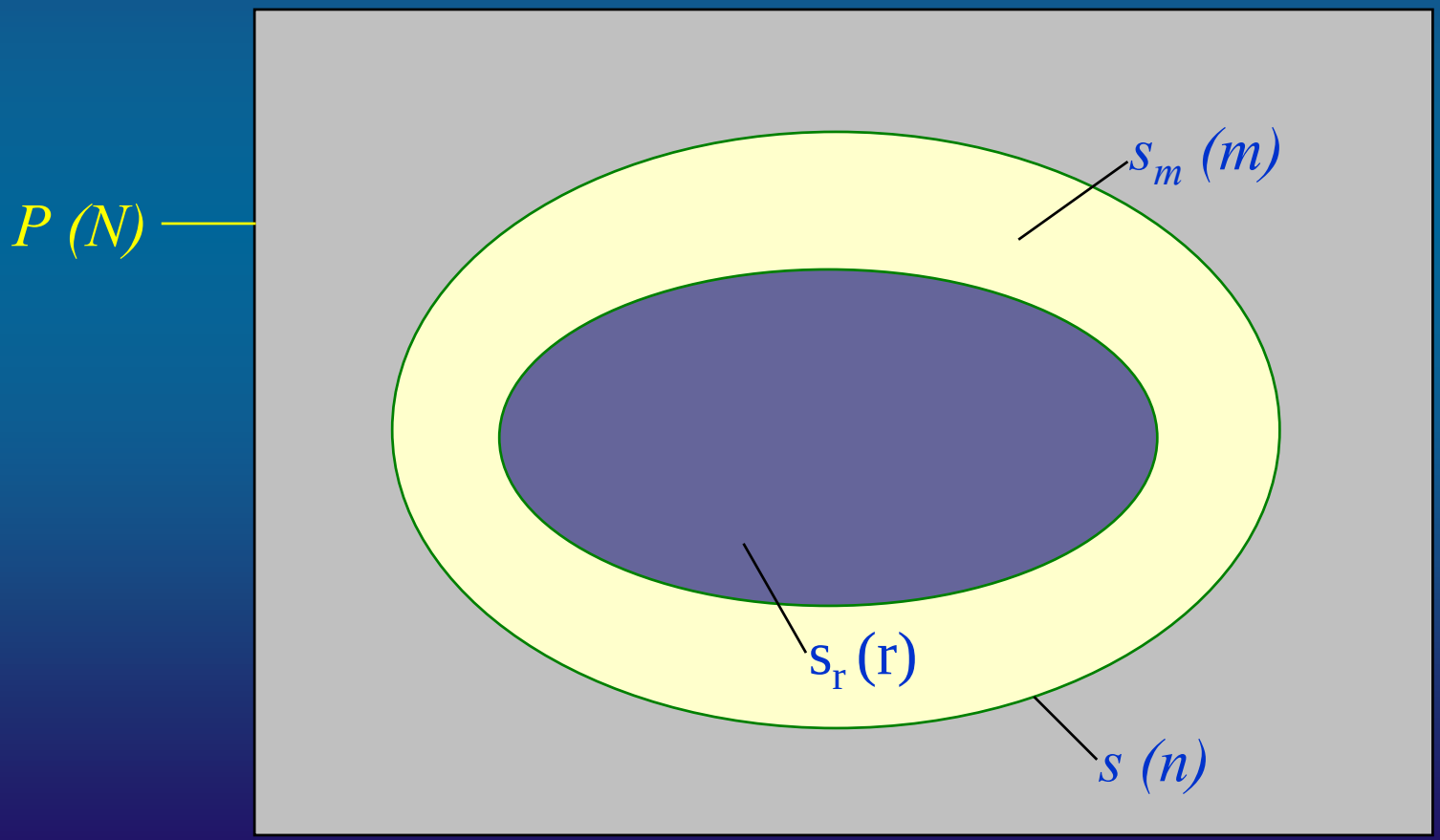
- L'approche renversée (Fay, 1991)

$$P \longrightarrow (P_r, P_m) \longrightarrow s = (s_r, s_m)$$

- Toutes les méthodes supposent que l'estimateur imputé est approximativement sans biais

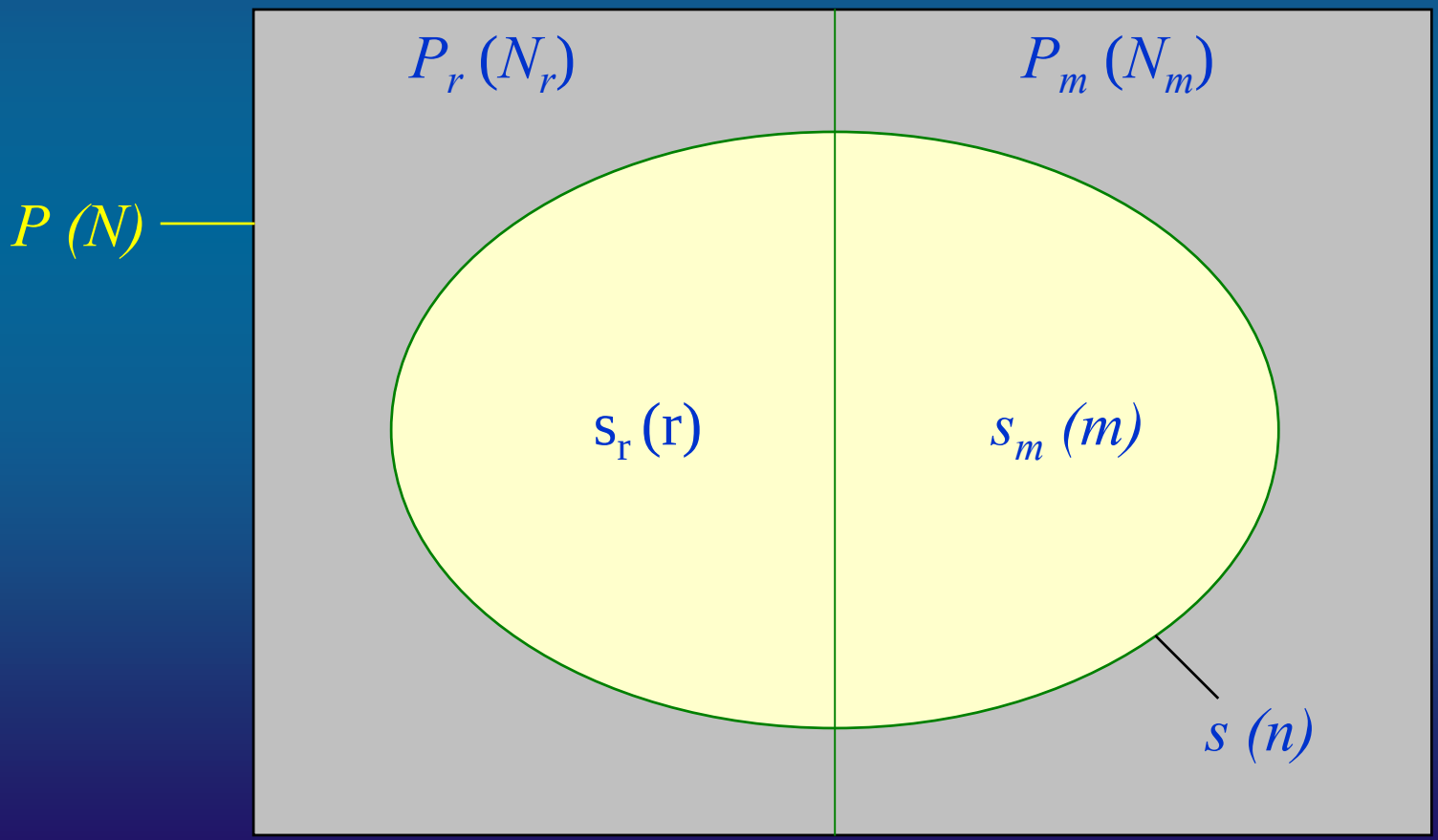
ESTIMATION DE LA VARIANCE

L'approche deux-phases

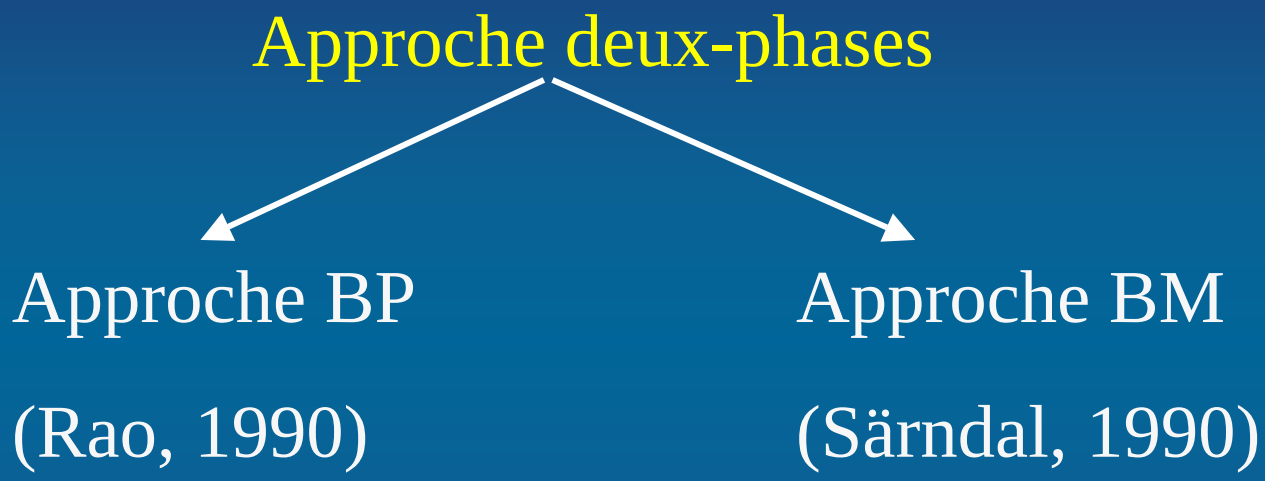


ESTIMATION DE LA VARIANCE

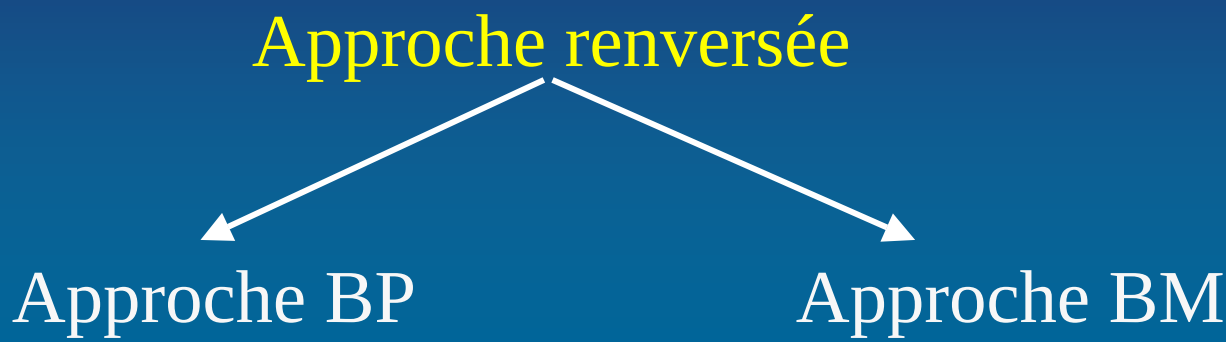
L'approche renversée



ESTIMATION DE LA VARIANCE



ESTIMATION DE LA VARIANCE



(Shao et Steel, 1999)

DEUX-PHASES: APPROCHE BP

- Cette approche est due à Rao (1990) et Rao et Sitter (1995).
- Cette approche a été développée sous l'approche BP.
- Considérons le cas d'un échantillon aléatoire simple, *s*, de taille *n* tiré sans remise d'une population *P* de taille *N*. Dans le cas d'imputation par la moyenne, l'estimateur imputé d'une moyenne est

$$\bar{y}_I = \bar{y}_r = \frac{1}{r} \sum_{i \in s_r} y_i$$

DEUX-PHASES: APPROCHE BP

- La variance de l'estimateur imputé est donnée par

$$V(\bar{y}_I) = V_p E_r(\bar{y}_r \mid s,r) + E_p V_r(\bar{y}_r \mid s,r)$$

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

- Un estimateur correct de la variance est donné par

$$v_{cor}(\bar{y}_I) = \left(\frac{1}{r} - \frac{1}{N}\right) s_{yr}^2$$

DEUX-PHASES: APPROCHE BP

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

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

- On a

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

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

DEUX-PHASES: APPROCHE BM

- Cette approche est due à Särndal (1990)
- Cette approche a été développée sous l'approche BM.
- Basée sur la décomposition:

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

DEUX-PHASES: APPROCHE BM

- La variance de l'estimateur imputé est donnée par

$$V(\bar{y}_I - \bar{Y}) = E(\bar{y}_I - \bar{Y})^2 = E_r E_p E_m (\bar{y}_I - \bar{Y})^2$$

$$= V_{éch} + V_{imp} + V_{mix}$$

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

DEUX-PHASES: APPROCHE BM

- Deville et Särndal (1994) ont généralisé la méthode de Särndal (1990) au cas de plans de sondage arbitraires
- Dans le cas d'enquêtes à plusieurs degrés, devrait-on considérer des modèles à effets aléatoires?
- La méthode peut être généralisée au cas de données multivariées
- La méthode peut être généralisée au cas d'estimateurs non-linéaires (ex: ratio)

L'APPROCHE RENVERSÉE

Approche deux-phases:

$$V(\bar{y}_I) = V_p E_r(\bar{y}_I \mid s,r) + E_p V_r(\bar{y}_I \mid s,r)$$

Approche renversée:

$$V(\bar{y}_I) = E_r V_p(\bar{y}_I \mid a_i) + V_r E_p(\bar{y}_I \mid a_i)$$

L'APPROCHE RENVERSÉE

On estime chacune des composantes séparément

i. Estimation de $V_1 = E_r V_p(\bar{y}_I - \bar{Y} \mid a_i)$

ii. Estimation de $V_2 = V_r E_p(\bar{y}_I - \bar{Y} \mid a_i)$

iii. Lorsque la fraction de sondage n/N est négligeable, la composante V_2 est négligeable par rapport à V_1

iv. 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

$$\bar{y}_I = \bar{y}_r = \frac{1}{r}\sum_{i\in s_r} y_i = \frac{\sum_{i\in s} a_i y_i}{\sum_{i\in s} a_i}$$

- 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,

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

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: $\theta = \bar{Y}$ et imputation par la moyenne

$$v_J(\bar{y}_I) = \frac{r-1}{n-1}\frac{s_r^2}{n} = v_{inc}(\bar{y}_I)$$

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: $y_j^* \rightarrow y_j^* + \bar{y}_r(j) - \bar{y}_I$

$$v_{JRS}(\bar{y}_I) = \frac{n-1}{n} \sum_{j \in s} \left(\bar{y}_{I(j)}^a - \bar{y}_I \right)^2$$

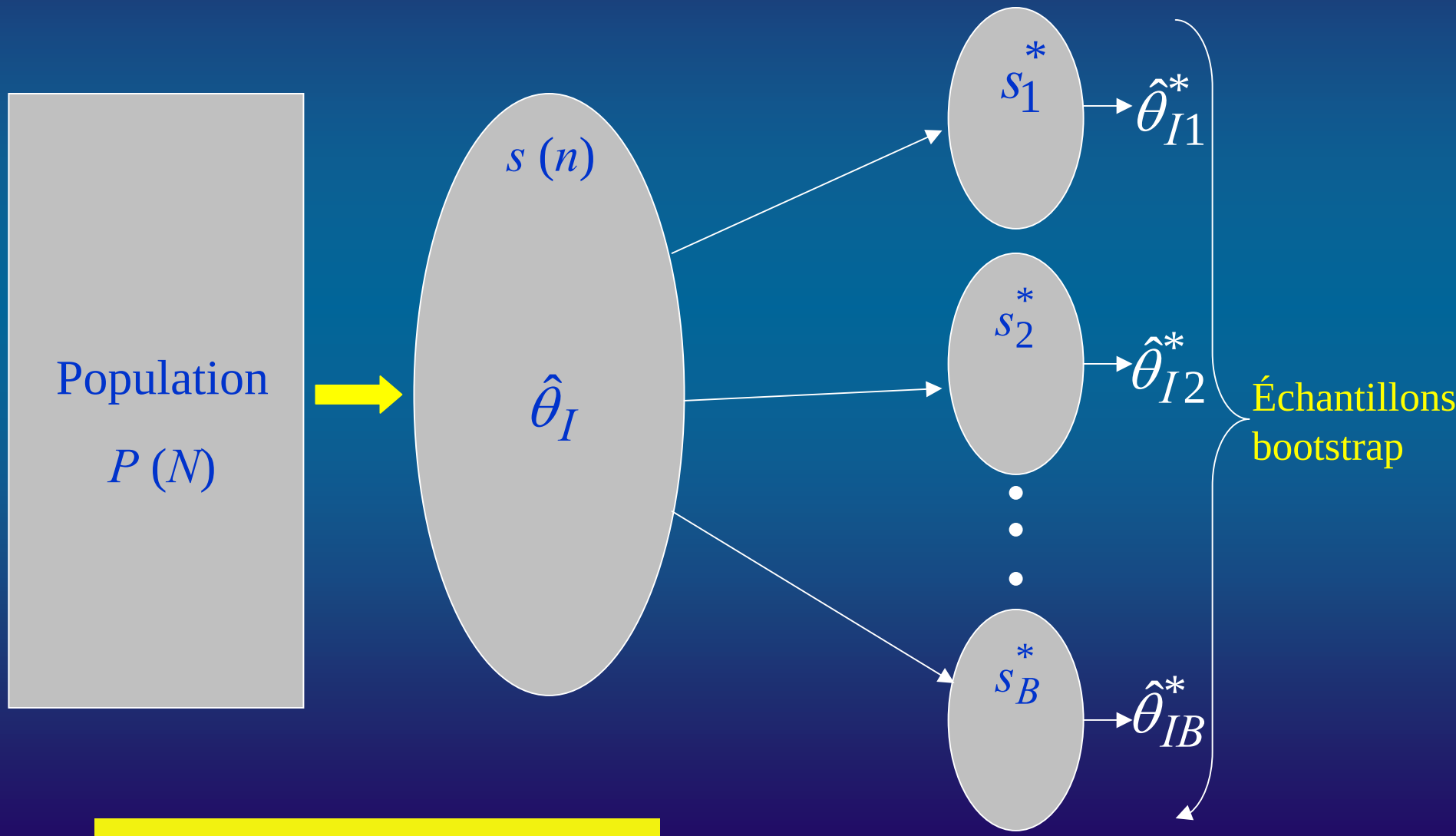
$$= \frac{s_r^2}{r} = v_{cor}(\bar{y}_I)$$

- Le jackknife de Rao-Shao est un estimateur de V_1 dans l'approche renversée

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

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

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

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

LE JACKKNIFE & LE BOOTSTRAP

- Les méthodes peuvent être appliquées à plusieurs méthodes d'imputation
- Les méthodes peuvent être utilisées pour des plans complexes
- Le bootstrap peut être utilisée pour des fonctions non-lisses de totaux (ex: quantiles), mais pas le jackknife
- Supposent que les unités ont été tirées avec remise ou que la fraction de sondage est négligeable

DISTORSION DES RELATIONS

PARAMÈTRES COMPLEXES

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

- Coefficient de régression: $\mathbf{B}_N = \left(\sum_{i \in P} \mathbf{x_i x_i'} \right)^{-1} \sum_{i \in P} \mathbf{x_i} y_i$

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

DOMAINES

- Des estimations au niveau des domaines sont souvent requises

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

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 pop. $\bar{Y} = \$555$

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 $\bar{y}_{dI}$

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

DOMAINES

- Sous un mécanisme de réponse uniforme et imputation par la moyenne, $y_i^* = \bar{y}_r$, le biais de l'estimateur imputé $\bar{y}_{dI}$ est donné par

$$Biais(\bar{y}_{dI}) = (1-p)(\bar{Y} - \bar{Y}_d)$$

- Haziza et Rao (2001) ont proposé un estimateur ajusté qui est approximativement sans biais sous les approches BP et BM

COEFFICIENT DE CORRÉLATION

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

- 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 (2002a) ont proposé un estimateur ajusté après imputation "traditionnelle"

IMPUTATION PONDÉRÉE VS NON-PONDÉRÉE

- En pratique, on utilise souvent des méthodes d'imputation non-pondérées et ce, même dans les cas de plans à probabilités inégales
  - Le stade d'imputation précède généralement le stade d'estimation si bien que les poids de sondage ne sont pas disponibles dans le fichier au stade de l'imputation
  - Certains systèmes de vérification et d'imputation n'ont pas tous l'option d'utiliser des méthodes d'imputation pondérées

IMPUTATION PONDÉRÉE VS NON-PONDÉRÉE

- L'utilisation de méthodes non-pondérées mène généralement à des estimateurs imputés biaisés

- Exemple: On tire un échantillon aléatoire s, de taille n, à l'aide d'un plan de sondage arbitraire $p(.)$. Un estimateur imputé de $\bar{Y}$ est donné par

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

où $w_i = 1/\pi_i$

IMPUTATION PONDÉRÉE VS NON-PONDÉRÉE

(i) Imputation par hot-deck aléatoire pondéré (IP)

$$y_i^* = y_j \text{ pour } j \in s_r \text{ tel que } P\left(y_i^* = y_j\right) = w_j \Big/ \sum_{i \in s_r} w_i$$

(ii) Imputation par hot-deck aléatoire non-pondéré (INP)

$$y_i^* = y_j \text{ pour } j \in s_r \text{ tel que } P\left(y_i^* = y_j\right) = 1/r$$

IMPUTATION PONDÉRÉE VS NON-PONDÉRÉE

- Dans le cas d' IP, l'estimateur imputé $\bar{y}_I$ est approximativement sans biais sous l'approche BP

$$E(\bar{y}_I) \approx \bar{Y}$$

- Dans le cas d' INP, l'estimateur imputé $\bar{y}_I$ est biaisé sous l' approche BP

$$\mathrm{BR}(\bar{y}_I) = E(\bar{y}_I - \bar{Y})/\bar{Y} \approx (1-p) C_\pi C_y \rho_{\pi y}$$

IMPUTATION PONDÉRÉE VS NON-PONDÉRÉE

- Avec $C_y > 0$, le biais relatif est donc 0 si
  - $p = 1$ (cas de pleine réponse)

  ou
  - $C_\pi = 0$ (cas d'un plan à probabilités égales)

  ou
  - $\rho_{\pi y} = 0$
- Haziza et Rao (2002b) ont proposé un estimateur ajusté qui est approximativement sans biais sous les approches BP et BM

CONCLUSIONS

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