Discussion sur la repondération pour la non-réponse: Quand les probabilités de réponse sont estimées par calage ou par maximum de vraisemblance.

Caren Hasler

Institut de Statistique, Université de Neuchâtel

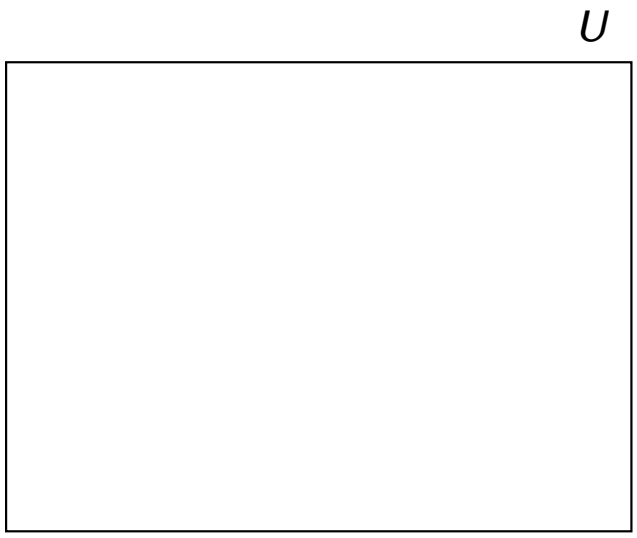
SFdS Séminaire en ligne sur les sondages
20 avril 2023

Aperçu

Aperçu

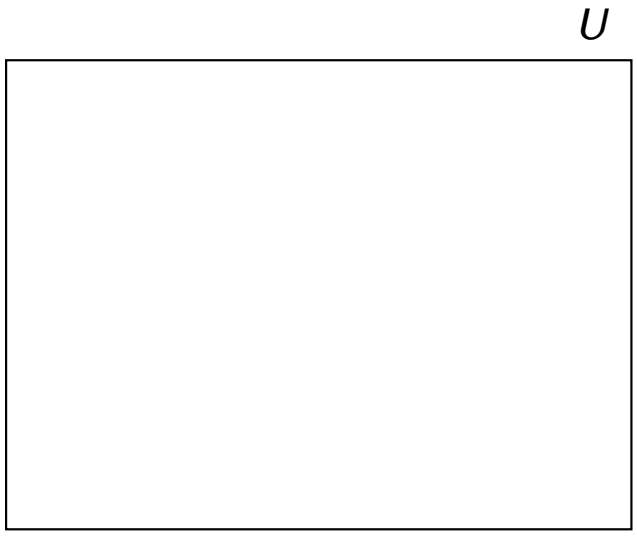
Notions utiles

- Population finie
  $U = \{1, 2, \ldots, i, \ldots, N\}$ de taille N.



Notions utiles

- Population finie
  $U = \{1, 2, \ldots, i, \ldots, N\}$ de taille N.
- Vecteur de q variables auxiliaires
  $\mathbf{x}_i = (1, x_{i2}, \ldots, x_{iq})$ attaché à une unité i, première variable constante.

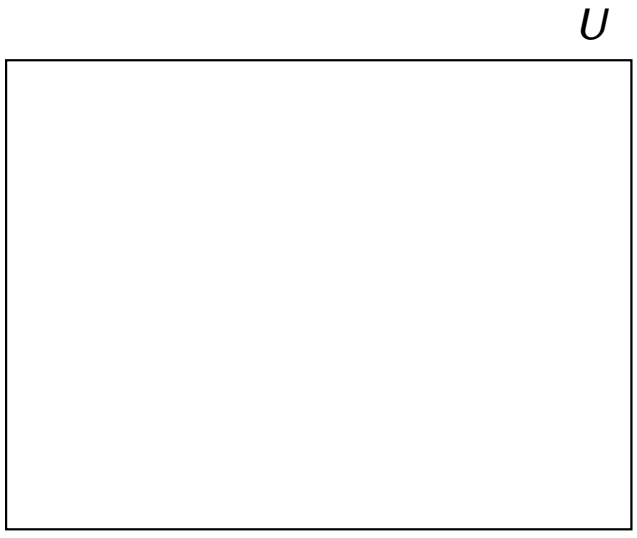


Notions utiles

- Population finie
  $U = \{1, 2, \ldots, i, \ldots, N\}$ de taille N.
- Vecteur de q variables auxiliaires
  $\mathbf{x}_i = (1, x_{i2}, \ldots, x_{iq})$ attaché à une unité i, première variable constante.
- Paramètre d'intérêt:

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

pour une variable d'intérêt y.

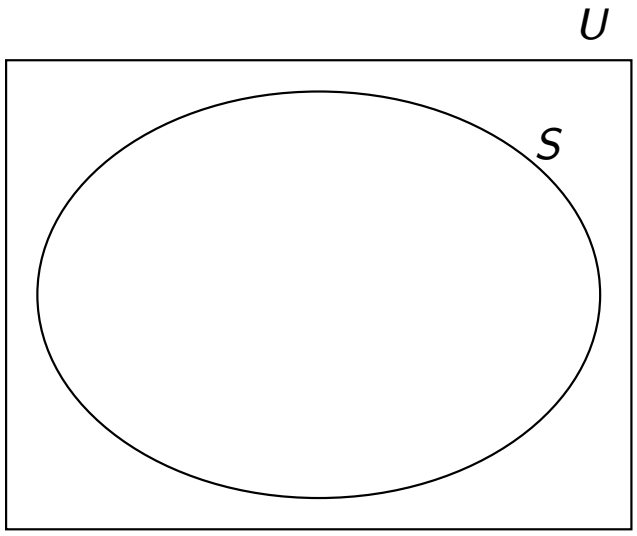


Notions utiles

- Population finie
  $U = \{1, 2, \ldots, i, \ldots, N\}$ de taille N.
- Vecteur de q variables auxiliaires
  $\mathbf{x}_i = (1, x_{i2}, \ldots, x_{iq})$ attaché à une unité i, première variable constante.
- Paramètre d'intérêt:

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

  pour une variable d'intérêt y.
- Echantillon S de taille n sélectionné dans U avec un plan de sondage $p(.)$

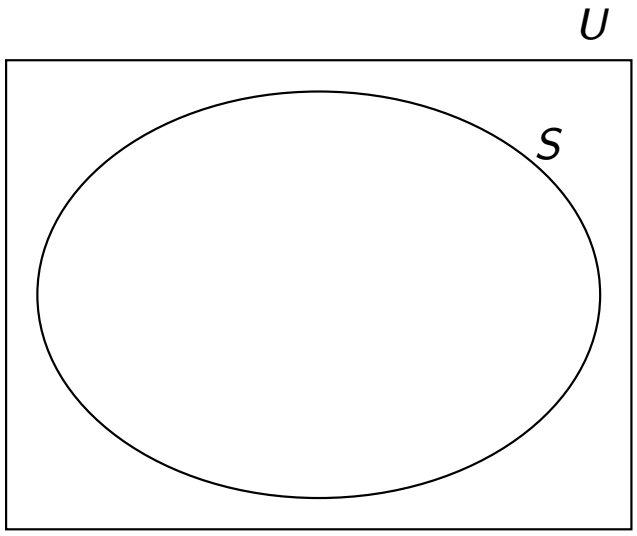


Notions utiles

- Population finie
  $U = \{1, 2, \ldots, i, \ldots, N\}$ de taille N.
- Vecteur de q variables auxiliaires
  $\mathbf{x}_i = (1, x_{i2}, \ldots, x_{iq})$ attaché à une unité i, première variable constante.
- Paramètre d'intérêt:

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

  pour une variable d'intérêt y.
- Echantillon S de taille n sélectionné dans U avec un plan de sondage $p(.)$
  - $\pi_i = \sum_{S; S \ni i} p(S)$,

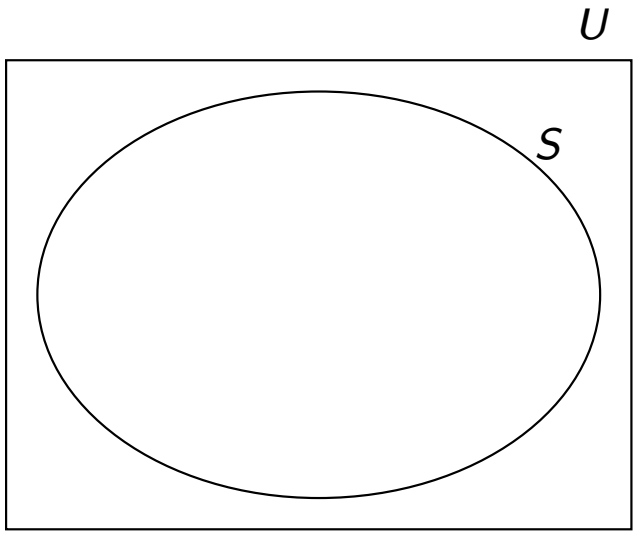


Notions utiles

- Population finie
  $U = \{1, 2, \ldots, i, \ldots, N\}$ de taille N.
- Vecteur de q variables auxiliaires
  $\mathbf{x}_i = (1, x_{i2}, \ldots, x_{iq})$ attaché à une unité i, première variable constante.
- Paramètre d'intérêt:

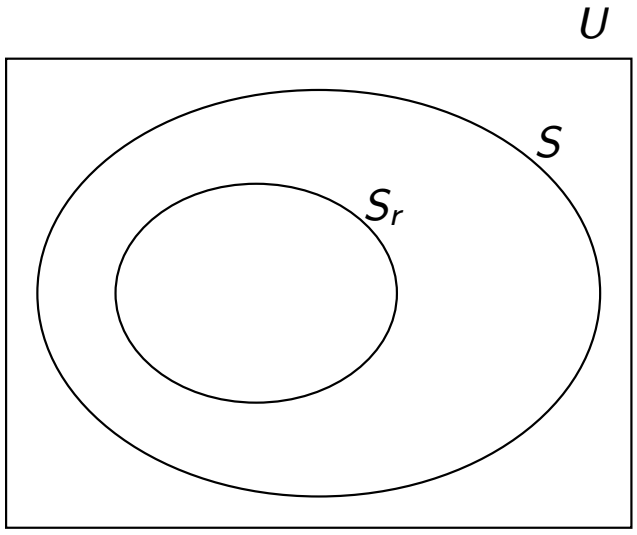
$$Y = \sum_{i \in U} y_i$$

  pour une variable d'intérêt y.
- Echantillon S de taille n sélectionné dans U avec un plan de sondage $p(.)$
  - $\pi_i = \sum_{S; S \ni i} p(S)$,
  - $\pi_i > 0$ for all $i \in U$.



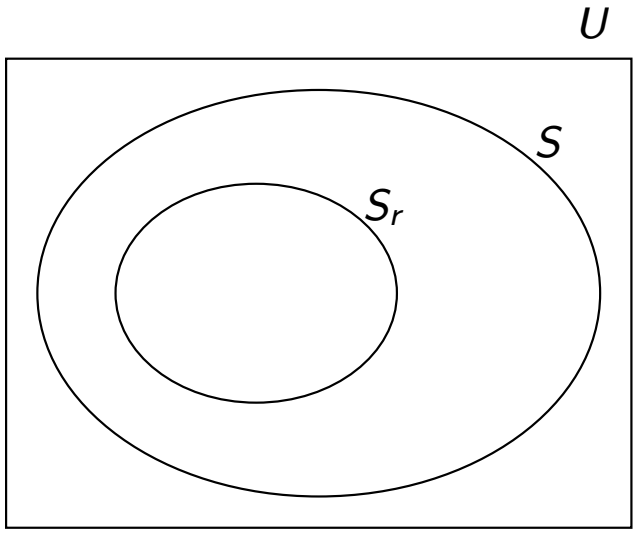
Notions utiles

- Non-réponse:
  - y_i observé pour $i \in S_r \subset S$
    S_r: *répondants*



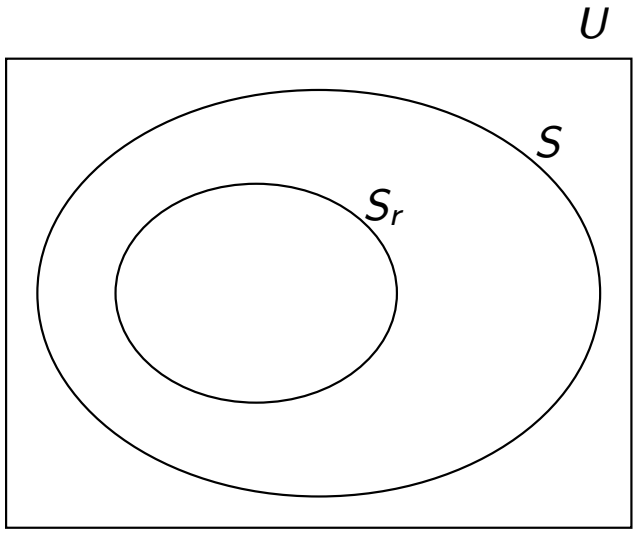
Notions utiles

- Non-réponse:
  - y_i observé pour $i \in S_r \subset S$
    S_r: *répondants*
- Echantillon S_r sélectionné dans S avec distribution $q(.|S)$
  - $p_i = \Pr(i \in S_r | i \in S)$



Notions utiles

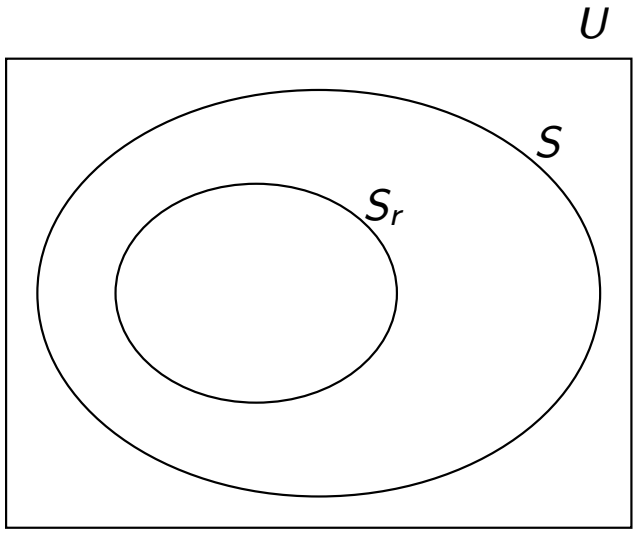
- Non-réponse:
  - y_i observé pour $i \in S_r \subset S$
    S_r: *répondants*
- Echantillon S_r sélectionné dans S avec distribution $q(.|S)$
  - $p_i = \Pr(i \in S_r | i \in S)$



- Conditions:
  - Données manquantes aléatoirement, (MAR) see [Rubin, 1976]:
    $Pr(i \in S_r | i \in S, \mathbf{x}_i, y_i) = Pr(i \in S_r | i \in S, \mathbf{x}_i)$,

Notions utiles

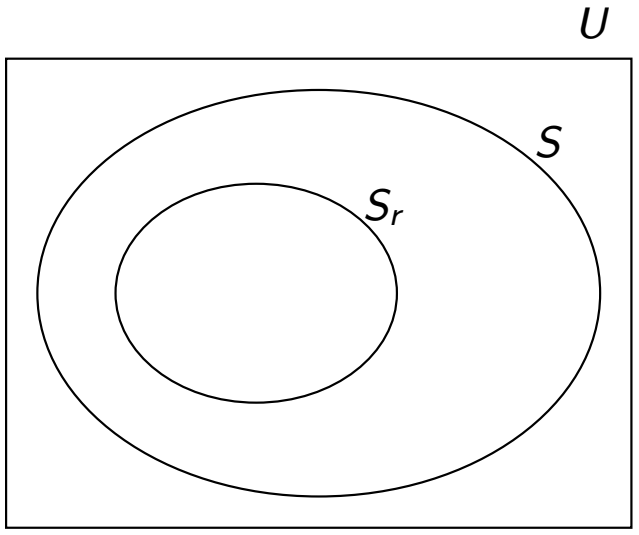
- Non-réponse:
  - y_i observé pour $i \in S_r \subset S$
    S_r: *répondants*
- Echantillon S_r sélectionné dans S avec distribution $q(.|S)$
  - $p_i = \Pr(i \in S_r | i \in S)$



- Conditions:
  - Données manquantes aléatoirement, (MAR) see [Rubin, 1976]:
    $Pr(i \in S_r | i \in S, \mathbf{x}_i, y_i) = Pr(i \in S_r | i \in S, \mathbf{x}_i)$,
  - Unités répondent indépendamment les unes des autres:
    $Pr(i, j \in S_r | i, j \in S) = p_i p_j$,

Notions utiles

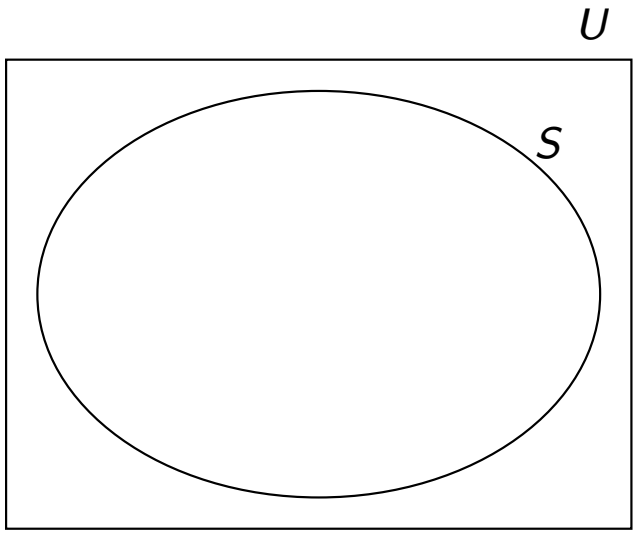
- Non-réponse:
  - y_i observé pour $i \in S_r \subset S$
    S_r: *répondants*
- Echantillon S_r sélectionné dans S avec distribution $q(.|S)$
  - $p_i = \Pr(i \in S_r | i \in S)$



- Conditions:
  - Données manquantes aléatoirement, (MAR) see [Rubin, 1976]:
    $Pr(i \in S_r | i \in S, \mathbf{x}_i, y_i) = Pr(i \in S_r | i \in S, \mathbf{x}_i)$,
  - Unités répondent indépendamment les unes des autres:
    $Pr(i, j \in S_r | i, j \in S) = p_i p_j$,
  - Les probabilités de réponses admettent une borne inférieure:
    il existe $c > 0$ tel que $p_i > c$ pour tout $i \in S$.

Estimateur du total ajusté pour la non-réponse

- But: estimer $Y = \sum_{i \in U} y_i$

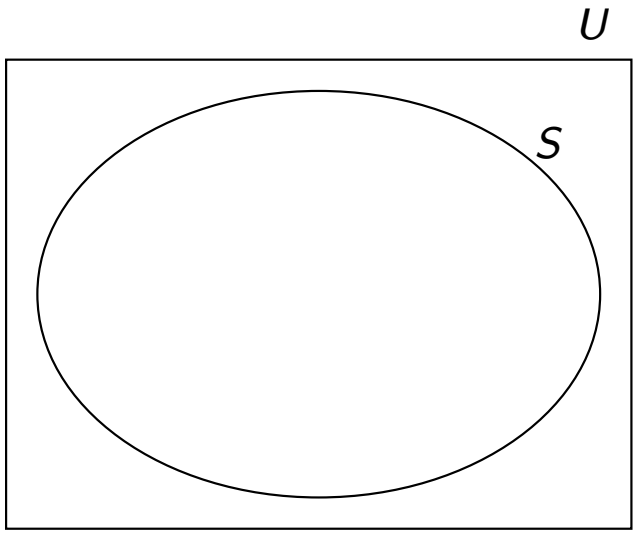


Estimateur du total ajusté pour la non-réponse

- But: estimer $Y = \sum_{i \in U} y_i$
- Estimateur par expansion ou d'Horvitz-Thompson (HT) [Horvitz and Thompson, 1952]

$$\widehat{Y} = \sum_{i \in S} \frac{y_i}{\pi_i}$$

Sans biais pour Y si $\pi_i > 0$ pour tout $i \in U$.

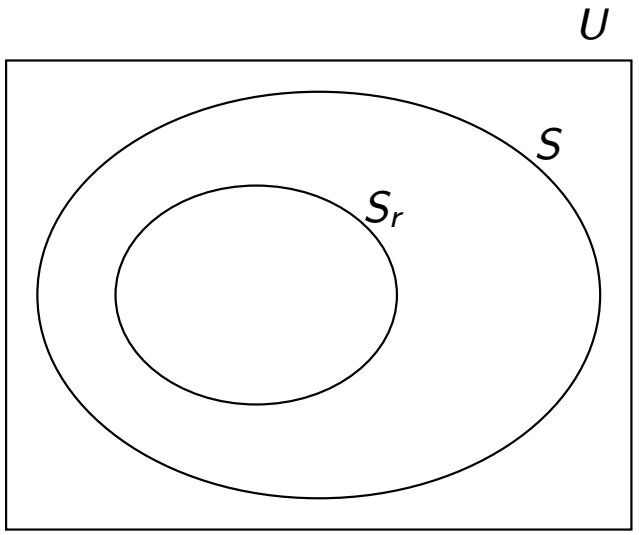


Estimateur du total ajusté pour la non-réponse

- Estimateur pour le deux-phases (ou estimateur par double expansion)

$$\widehat{Y}_p = \sum_{i \in S_r} \frac{y_i}{\pi_i p_i}$$

Sans biais pour Y si $\pi_i, p_i > 0$ pour tout $i \in S$.



Estimateur du total ajusté pour la non-réponse

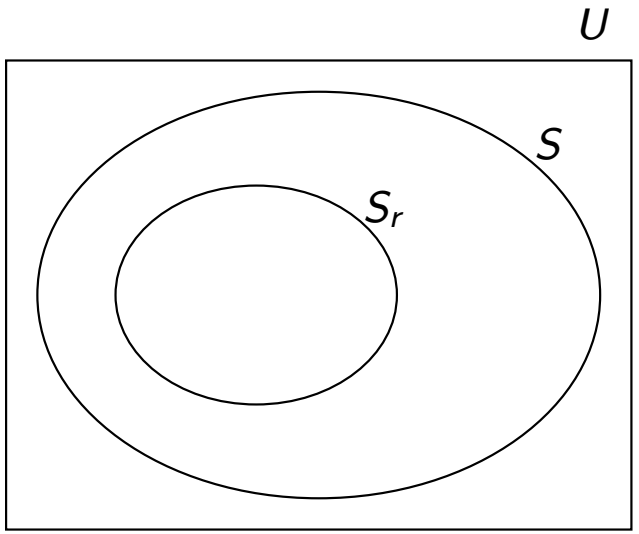
- Estimateur pour le deux-phases (ou estimateur par double expansion)

$$\widehat{Y}_p = \sum_{i \in S_r} \frac{y_i}{\pi_i p_i}$$

Sans biais pour Y si $\pi_i, p_i > 0$ pour tout $i \in S$.

- Estimateur par double expansion empirique (DEE)

$$\widehat{Y}_{\widehat{p}} = \sum_{i \in S_r} \frac{y_i}{\pi_i \widehat{p}_i}$$



Estimateur du total ajusté pour la non-réponse

- Estimateur pour le deux-phases (ou estimateur par double expansion)

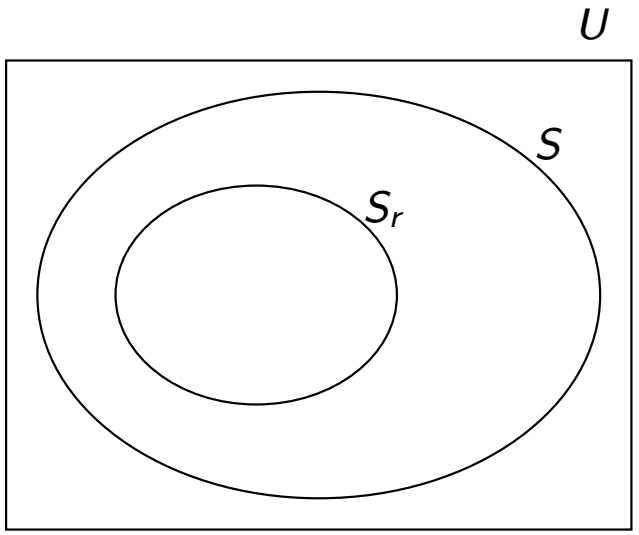
$$\widehat{Y}_p = \sum_{i \in S_r} \frac{y_i}{\pi_i p_i}$$

  Sans biais pour Y si $\pi_i, p_i > 0$ pour tout $i \in S$.

- Estimateur par double expansion empirique (DEE)

$$\widehat{Y}_{\widehat{p}} = \sum_{i \in S_r} \frac{y_i}{\pi_i \widehat{p}_i}$$

- Correction du biais de non-réponse



Aperçu

Approches

1. **Maximum de vraisemblance**
   - [Kim and Kim, 2007]: asymptotique du DEE pour modèle de réponse géneral
   - [Beaumont, 2005]: efficacité du DEE avec modèle de réponse logistique

Approches

1. **Maximum de vraisemblance**
   - [Kim and Kim, 2007]: asymptotique du DEE pour modèle de réponse géneral
   - [Beaumont, 2005]: efficacité du DEE avec modèle de réponse logistique
2. **Calage Direct**: [Särndal and Lundström, 2005]
   - [Deville and Dupont, 1993], [Dupont, 1993], [Lundström and Särndal, 1999]: estimateur ponctuel et estimateur de la variance DEE
   - [Iannacchione et al., 1991]: étude empirique DEE, calage au niveau de S
   - [Kim and Riddles, 2012]: asymptotique pour calage au niveau de S

Approches

1. **Maximum de vraisemblance**
   - [Kim and Kim, 2007]: asymptotique du DEE pour modèle de réponse géneral
   - [Beaumont, 2005]: efficacité du DEE avec modèle de réponse logistique
2. **Calage Direct**: [Särndal and Lundström, 2005]
   - [Deville and Dupont, 1993], [Dupont, 1993], [Lundström and Särndal, 1999]: estimateur ponctuel et estimateur de la variance DEE
   - [Iannacchione et al., 1991]: étude empirique DEE, calage au niveau de S
   - [Kim and Riddles, 2012]: asymptotique pour calage au niveau de S
3. **Autres approches robustes**:
   - Méthode en deux étapes: mle suivi de calage [Haziza and Lesage, 2016]
   - Méthode du score: mle puis partition en classes homogènes [Haziza and Beaumont, 2007]

Approches

1. **Maximum de vraisemblance**
   - [Kim and Kim, 2007]: asymptotique du DEE pour modèle de réponse géneral
   - [Beaumont, 2005]: efficacité du DEE avec modèle de réponse logistique
2. **Calage Direct**: [Särndal and Lundström, 2005]
   - [Deville and Dupont, 1993], [Dupont, 1993], [Lundström and Särndal, 1999]: estimateur ponctuel et estimateur de la variance DEE
   - [Iannacchione et al., 1991]: étude empirique DEE, calage au niveau de S
   - [Kim and Riddles, 2012]: asymptotique pour calage au niveau de S
3. **Autres approches robustes**:
   - Méthode en deux étapes: mle suivi de calage [Haziza and Lesage, 2016]
   - Méthode du score: mle puis partition en classes homogènes [Haziza and Beaumont, 2007]

1. **Travail**:
   - Cadre théorique commun pour les deux approches
   - Asymptotique pour calage au niveau de U et S
   - Double robustesse
   - Comparaison et discussion

1. **Travail**:
   - Cadre théorique commun pour les deux approches
   - Asymptotique pour calage au niveau de U et S
   - Double robustesse
   - Comparaison et discussion
2. **Aujourd'hui**:
   - Intuition, modèles sous-jacents
   - Simulations

Probabilités de réponse: modèle de régression logistique

$$\boxed{\widehat{Y}_{\widehat{p}} = \sum_{i \in S_r} \frac{1}{\pi_i} \frac{1}{\widehat{p}_i} y_i}$$

- Modèle de régression logistique

$$p_i = p(\mathbf{x}_i; \boldsymbol{\lambda}) = \frac{\exp\left(\mathbf{x}_i^\top \boldsymbol{\lambda}\right)}{1 + \exp\left(\mathbf{x}_i^\top \boldsymbol{\lambda}\right)} = \frac{1}{1 + \exp\left(-\mathbf{x}_i^\top \boldsymbol{\lambda}\right)}$$

Probabilités de réponse: modèle de régression logistique

$$\boxed{\widehat{Y}_{\widehat{p}} = \sum_{i \in S_r} \frac{1}{\pi_i} \frac{1}{\widehat{p}_i} y_i}$$

- Modèle de régression logistique

$$p_i = p(\mathbf{x}_i; \boldsymbol{\lambda}) = \frac{\exp\left(\mathbf{x}_i^\top \boldsymbol{\lambda}\right)}{1 + \exp\left(\mathbf{x}_i^\top \boldsymbol{\lambda}\right)} = \frac{1}{1 + \exp\left(-\mathbf{x}_i^\top \boldsymbol{\lambda}\right)}$$

- Probabilités de réponse: $0 < \widehat{p}_i < 1$

Probabilités de réponse: modèle de régression logistique

$$\boxed{\widehat{Y}_{\widehat{p}} = \sum_{i \in S_r} \frac{1}{\pi_i} \frac{1}{\widehat{p}_i} y_i}$$

- Modèle de régression logistique

$$p_i = p(\mathbf{x}_i; \boldsymbol{\lambda}) = \frac{\exp\left(\mathbf{x}_i^\top \boldsymbol{\lambda}\right)}{1 + \exp\left(\mathbf{x}_i^\top \boldsymbol{\lambda}\right)} = \frac{1}{1 + \exp\left(-\mathbf{x}_i^\top \boldsymbol{\lambda}\right)}$$

- Probabilités de réponse: $0 < \widehat{p}_i < 1$
- Poids de correction de la non-réponse: $\dfrac{1}{\widehat{p}_i} > 1$

Probabilités de réponse: Maximum de vraisemblance et Calage

Modèle de régression logistique

$$\widehat{p}_i = p(\mathbf{x}_i; \widehat{\boldsymbol{\lambda}}) = \frac{\exp\left(\mathbf{x}_i^\top \widehat{\boldsymbol{\lambda}}\right)}{1 + \exp\left(\mathbf{x}_i^\top \widehat{\boldsymbol{\lambda}}\right)} = \frac{1}{1 + \exp\left(-\mathbf{x}_i^\top \widehat{\boldsymbol{\lambda}}\right)}$$

où $\widehat{\boldsymbol{\lambda}}$ est solution d'une équation estimante

Probabilités de réponse: Maximum de vraisemblance et Calage

Modèle de régression logistique

$$\widehat{p}_i = p(\mathbf{x}_i; \widehat{\boldsymbol{\lambda}}) = \frac{\exp\left(\mathbf{x}_i^\top \widehat{\boldsymbol{\lambda}}\right)}{1 + \exp\left(\mathbf{x}_i^\top \widehat{\boldsymbol{\lambda}}\right)} = \frac{1}{1 + \exp\left(-\mathbf{x}_i^\top \widehat{\boldsymbol{\lambda}}\right)}$$

où $\widehat{\boldsymbol{\lambda}}$ est solution d'une équation estimante

$$\text{MLE:} \qquad \sum_{i \in S_r} k_i \mathbf{x}_i = \sum_{i \in S} k_i \widehat{p}_i \mathbf{x}_i \qquad k_i = 1, 1/\pi_i, \ldots$$

$$\text{Calage sur } S\text{:} \qquad \sum_{i \in S_r} \frac{\mathbf{x}_i}{\pi_i \widehat{p}_i} = \sum_{i \in S} \frac{\mathbf{x}_i}{\pi_i}$$

$$\text{Calage sur } U\text{:} \qquad \sum_{i \in S_r} \frac{\mathbf{x}_i}{\pi_i \widehat{p}_i} = \sum_{i \in U} \mathbf{x}_i$$

Différences

- **Information nécessaire sur x**

$$\text{MLE:} \sum_{i \in S_r} k_i \mathbf{x}_i = \sum_{i \in S} k_i \widehat{p}_i \mathbf{x}_i$$

$$\text{Calage sur } S\text{:} \sum_{i \in S_r} \frac{\mathbf{x}_i}{\pi_i \widehat{p}_i} = \sum_{i \in S} \frac{\mathbf{x}_i}{\pi_i}$$

$$\text{Calage sur } U\text{:} \sum_{i \in S_r} \frac{\mathbf{x}_i}{\pi_i \widehat{p}_i} = \sum_{i \in U} \mathbf{x}_i$$

Différences

$$\text{MLE: } \sum_{i \in S_r} k_i \mathbf{x}_i = \sum_{i \in S} k_i \widehat{p}_i \mathbf{x}_i$$

$$\text{Calage sur } S\text{: } \sum_{i \in S_r} \frac{\mathbf{x}_i}{\pi_i \widehat{p}_i} = \sum_{i \in S} \frac{\mathbf{x}_i}{\pi_i}$$

$$\text{Calage sur } U\text{: } \sum_{i \in S_r} \frac{\mathbf{x}_i}{\pi_i \widehat{p}_i} = \sum_{i \in U} \mathbf{x}_i$$

- **Information nécessaire sur x**
- **Consistence**: calage, consistence avec des totaux connus

Différences

$$\text{MLE:} \sum_{i \in S_r} k_i \mathbf{x}_i = \sum_{i \in S} k_i \widehat{p}_i \mathbf{x}_i$$

$$\text{Calage sur } S\text{:} \sum_{i \in S_r} \frac{\mathbf{x}_i}{\pi_i \widehat{p}_i} = \sum_{i \in S} \frac{\mathbf{x}_i}{\pi_i}$$

$$\text{Calage sur } U\text{:} \sum_{i \in S_r} \frac{\mathbf{x}_i}{\pi_i \widehat{p}_i} = \sum_{i \in U} \mathbf{x}_i$$

- **Information nécessaire sur x**
- **Consistence**: calage, consistence avec des totaux connus
- **Esprit**: MLE centré sur les probabilités de réponse, calage sur l'estimation de totaux

Différences

$$\text{MLE:} \sum_{i \in S_r} k_i \mathbf{x}_i = \sum_{i \in S} k_i \widehat{p}_i \mathbf{x}_i$$

$$\text{Calage sur } S\text{:} \sum_{i \in S_r} \frac{\mathbf{x}_i}{\pi_i \widehat{p}_i} = \sum_{i \in S} \frac{\mathbf{x}_i}{\pi_i}$$

$$\text{Calage sur } U\text{:} \sum_{i \in S_r} \frac{\mathbf{x}_i}{\pi_i \widehat{p}_i} = \sum_{i \in U} \mathbf{x}_i$$

- **Information nécessaire sur x**
- **Consistence**: calage, consistence avec des totaux connus
- **Esprit**: MLE centré sur les probabilités de réponse, calage sur l'estimation de totaux
- **Correction du biais**: MLE et calage sur S corrigent biais dû à la non-réponse, calage sur U biais dû à la non-réponse et à l'échantillonnage

Aperçu

Modèles

1. **Modèle de non-réponse (NR)**: modèle pour les probabilités de réponse
   Exemple: modéle de régression logistique:

$$p_i = p(\mathbf{x}_i; \boldsymbol{\lambda}) = \frac{\exp\left(\mathbf{x}_i{}^\top \lambda\right)}{1 + \exp\left(\mathbf{x}_i{}^\top \boldsymbol{\lambda}\right)} = \frac{1}{1 + \exp\left(-\mathbf{x}_i{}^\top \boldsymbol{\lambda}\right)}$$

Modèles

1. **Modèle de non-réponse (NR)**: modèle pour les probabilités de réponse
   Exemple: modéle de régression logistique:

$$p_i = p(\mathbf{x}_i; \boldsymbol{\lambda}) = \frac{\exp\left(\mathbf{x}_i{}^\top \lambda\right)}{1 + \exp\left(\mathbf{x}_i{}^\top \boldsymbol{\lambda}\right)} = \frac{1}{1 + \exp\left(-\mathbf{x}_i{}^\top \boldsymbol{\lambda}\right)}$$

2. **Modèle de superpopulation (SUP)**: modèle pour le lien entre $\mathbf{x}$ et y:

Modèles

1. **Modèle de non-réponse (NR)**: modèle pour les probabilités de réponse
   Exemple: modéle de régression logistique:

$$p_i = p(\mathbf{x}_i; \boldsymbol{\lambda}) = \frac{\exp\left(\mathbf{x}_i{}^\top \lambda\right)}{1 + \exp\left(\mathbf{x}_i{}^\top \boldsymbol{\lambda}\right)} = \frac{1}{1 + \exp\left(-\mathbf{x}_i{}^\top \boldsymbol{\lambda}\right)}$$

2. **Modèle de superpopulation (SUP)**: modèle pour le lien entre $\mathbf{x}$ et y:
   - MLE:

$$y_i = k_i \pi_i p_i \mathbf{x}_i{}^\top \beta + \varepsilon_i$$

Modèles

1. **Modèle de non-réponse (NR)**: modèle pour les probabilités de réponse
   Exemple: modéle de régression logistique:

$$p_i = p(\mathbf{x}_i; \boldsymbol{\lambda}) = \frac{\exp\left(\mathbf{x}_i{}^\top \lambda\right)}{1 + \exp\left(\mathbf{x}_i{}^\top \boldsymbol{\lambda}\right)} = \frac{1}{1 + \exp\left(-\mathbf{x}_i{}^\top \boldsymbol{\lambda}\right)}$$

2. **Modèle de superpopulation (SUP)**: modèle pour le lien entre $\mathbf{x}$ et y:
   - MLE:
   $$y_i = k_i \pi_i p_i \mathbf{x}_i{}^\top \boldsymbol{\beta} + \varepsilon_i$$
   - Calage:
   $$y_i = \mathbf{x}_i{}^\top \boldsymbol{\beta} + \varepsilon_i$$

Asymptotique sous NR

1. $\widehat{Y}_{\widehat{p}}$ asymptotiquement équivalent à un estimateur sans-biais

Asymptotique sous NR

1. $\widehat{Y}_{\widehat{p}}$ asymptotiquement équivalent à un estimateur sans-biais
2. $\widehat{Y}_{\widehat{p}}$ asymptotiquement équivalent à un estimateur au moins autant efficace que $\widehat{Y}_p$

Asymptotique sous NR

1. $\widehat{Y}_{\widehat{p}}$ asymptotiquement équivalent à un estimateur sans-biais
2. $\widehat{Y}_{\widehat{p}}$ asymptotiquement équivalent à un estimateur au moins autant efficace que $\widehat{Y}_p$

   Intuition: estimation comme lissage des probabilités de réponse

Asymptotique sous NR

1. $\widehat{Y}_{\widehat{p}}$ asymptotiquement équivalent à un estimateur sans-biais
2. $\widehat{Y}_{\widehat{p}}$ asymptotiquement équivalent à un estimateur au moins autant efficace que $\widehat{Y}_p$

   Intuition: estimation comme lissage des probabilités de réponse

   Montré par [Kim and Kim, 2007] et [Beaumont, 2005] pour MLE

Asymptotique sous NR

1. $\widehat{Y}_{\widehat{p}}$ asymptotiquement équivalent à un estimateur sans-biais
2. $\widehat{Y}_{\widehat{p}}$ asymptotiquement équivalent à un estimateur au moins autant efficace que $\widehat{Y}_p$

   Intuition: estimation comme lissage des probabilités de réponse

   Montré par [Kim and Kim, 2007] et [Beaumont, 2005] pour MLE
3. Plus la relation du SUP est forte, plus $\widehat{Y}_{\widehat{p}}$ est efficace

Asymptotique sous NR

1. $\widehat{Y}_{\widehat{p}}$ asymptotiquement équivalent à un estimateur sans-biais
2. $\widehat{Y}_{\widehat{p}}$ asymptotiquement équivalent à un estimateur au moins autant efficace que $\widehat{Y}_p$

   Intuition: estimation comme lissage des probabilités de réponse

   Montré par [Kim and Kim, 2007] et [Beaumont, 2005] pour MLE
3. Plus la relation du SUP est forte, plus $\widehat{Y}_{\widehat{p}}$ est efficace
4. Gain en efficacité avec calage sur U vs calage sur S

Asymptotique sous NR

1. $\widehat{Y}_{\widehat{p}}$ asymptotiquement équivalent à un estimateur sans-biais
2. $\widehat{Y}_{\widehat{p}}$ asymptotiquement équivalent à un estimateur au moins autant efficace que $\widehat{Y}_p$

   Intuition: estimation comme lissage des probabilités de réponse

   Montré par [Kim and Kim, 2007] et [Beaumont, 2005] pour MLE
3. Plus la relation du SUP est forte, plus $\widehat{Y}_{\widehat{p}}$ est efficace
4. Gain en efficacité avec calage sur U vs calage sur S

   Intuition: calage sur S corrige l'erreur due à la non-réponse, calage sur U corrige l'erreur due à la non-réponse et à l'échantillonnage

$$\text{Calage sur } S: \sum_{i \in S_r} \frac{\mathbf{x}_i}{\pi_i \widehat{p}_i} = \sum_{i \in S} \frac{\mathbf{x}_i}{\pi_i}$$

$$\text{Calage sur } U: \sum_{i \in S_r} \frac{\mathbf{x}_i}{\pi_i \widehat{p}_i} = \sum_{i \in U} \mathbf{x}_i$$

Asymptotique sous SUP - Double robustesse

- $\widehat{Y}_{\widehat{p}}$ asymptotiquement équivalent à un estimateur sans-biais quand:

Asymptotique sous SUP - Double robustesse

- $\widehat{Y}_{\widehat{p}}$ asymptotiquement équivalent à un estimateur sans-biais quand:
  - $\widehat{p}_i$ obtenus par calage

Asymptotique sous SUP - Double robustesse

- $\widehat{Y}_{\widehat{p}}$ asymptotiquement équivalent à un estimateur sans-biais quand:
  - $\widehat{p}_i$ obtenus par calage
  - $y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i$

Asymptotique sous SUP - Double robustesse

- $\widehat{Y}_{\widehat{p}}$ asymptotiquement équivalent à un estimateur sans-biais quand:
  - $\widehat{p}_i$ obtenus par calage
  - $y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i$
  - NR pas forcément verifié

Asymptotique sous SUP - Double robustesse

- $\widehat{Y}_{\widehat{p}}$ asymptotiquement équivalent à un estimateur sans-biais quand:
  - $\widehat{p}_i$ obtenus par calage
  - $y_i = \mathbf{x}_i^\top \beta + \varepsilon_i$
  - NR pas forcément verifié
- Double robustesse de $\widehat{Y}_{\widehat{p}}$ obtenu par calage

Asymptotique sous SUP - Double robustesse

- $\widehat{Y}_{\widehat{p}}$ asymptotiquement équivalent à un estimateur sans-biais quand:
  - $\widehat{p}_i$ obtenus par calage
  - $y_i = \mathbf{x}_i^\top \beta + \varepsilon_i$
  - NR pas forcément verifié
- Double robustesse de $\widehat{Y}_{\widehat{p}}$ obtenu par calage

  Voire aussi [Kott, 2006], [Kott and Liao, 2012] et [Haziza and Lesage, 2016]

Asymptotique sous SUP - Double robustesse

- $\widehat{Y}_{\widehat{p}}$ asymptotiquement équivalent à un estimateur sans-biais quand:
  - $\widehat{p}_i$ obtenus par calage
  - $y_i = \mathbf{x}_i^\top \beta + \varepsilon_i$
  - NR pas forcément verifié
- Double robustesse de $\widehat{Y}_{\widehat{p}}$ obtenu par calage

  Voire aussi [Kott, 2006], [Kott and Liao, 2012] et [Haziza and Lesage, 2016]
- Double robustesse avec MLE? Pas évident.

Aperçu

Problèmes convergence et poids extrêmes

$$\text{MLE:} \sum_{i \in S_r} k_i \mathbf{x}_i = \sum_{i \in S} k_i \widehat{p}_i \mathbf{x}_i$$

$$\text{Calage sur } S\text{:} \sum_{i \in S_r} \frac{\mathbf{x}_i}{\pi_i \widehat{p}_i} = \sum_{i \in S} \frac{\mathbf{x}_i}{\pi_i}$$

$$\text{Calage sur } U\text{:} \sum_{i \in S_r} \frac{\mathbf{x}_i}{\pi_i \widehat{p}_i} = \sum_{i \in U} \mathbf{x}_i$$

- Méthode numérique pour résoudre les équations estimantes

Problèmes convergence et poids extrêmes

$$\text{MLE:} \sum_{i \in S_r} k_i \mathbf{x}_i = \sum_{i \in S} k_i \widehat{p}_i \mathbf{x}_i$$

$$\text{Calage sur } S\text{:} \sum_{i \in S_r} \frac{\mathbf{x}_i}{\pi_i \widehat{p}_i} = \sum_{i \in S} \frac{\mathbf{x}_i}{\pi_i}$$

$$\text{Calage sur } U\text{:} \sum_{i \in S_r} \frac{\mathbf{x}_i}{\pi_i \widehat{p}_i} = \sum_{i \in U} \mathbf{x}_i$$

- Méthode numérique pour résoudre les équations estimantes
- Problèmes de convergence ou de poids extrêmes

Problèmes convergence et poids extrêmes

$$\text{MLE:} \sum_{i \in S_r} k_i \mathbf{x}_i = \sum_{i \in S} k_i \widehat{p}_i \mathbf{x}_i$$

$$\text{Calage sur } S\text{:} \sum_{i \in S_r} \frac{\mathbf{x}_i}{\pi_i \widehat{p}_i} = \sum_{i \in S} \frac{\mathbf{x}_i}{\pi_i}$$

$$\text{Calage sur } U\text{:} \sum_{i \in S_r} \frac{\mathbf{x}_i}{\pi_i \widehat{p}_i} = \sum_{i \in U} \mathbf{x}_i$$

- Méthode numérique pour résoudre les équations estimantes
- Problèmes de convergence ou de poids extrêmes
- Plus fréquent avec calage sur U que calage sur S

Problèmes convergence et poids extrêmes

$$\text{MLE:} \sum_{i \in S_r} k_i \mathbf{x}_i = \sum_{i \in S} k_i \widehat{p}_i \mathbf{x}_i$$

$$\text{Calage sur } S\text{:} \sum_{i \in S_r} \frac{\mathbf{x}_i}{\pi_i \widehat{p}_i} = \sum_{i \in S} \frac{\mathbf{x}_i}{\pi_i}$$

$$\text{Calage sur } U\text{:} \sum_{i \in S_r} \frac{\mathbf{x}_i}{\pi_i \widehat{p}_i} = \sum_{i \in U} \mathbf{x}_i$$

- Méthode numérique pour résoudre les équations estimantes
- Problèmes de convergence ou de poids extrêmes
- Plus fréquent avec calage sur U que calage sur S

  Intuition: calage sur U corrige erreur de non-réponse et d'échantillonnage

Problèmes convergence et poids extrêmes

$$\text{MLE:} \sum_{i \in S_r} k_i \mathbf{x}_i = \sum_{i \in S} k_i \widehat{p}_i \mathbf{x}_i$$

$$\text{Calage sur } S\text{:} \sum_{i \in S_r} \frac{\mathbf{x}_i}{\pi_i \widehat{p}_i} = \sum_{i \in S} \frac{\mathbf{x}_i}{\pi_i}$$

$$\text{Calage sur } U\text{:} \sum_{i \in S_r} \frac{\mathbf{x}_i}{\pi_i \widehat{p}_i} = \sum_{i \in U} \mathbf{x}_i$$

- Méthode numérique pour résoudre les équations estimantes
- Problèmes de convergence ou de poids extrêmes
- Plus fréquent avec calage sur U que calage sur S

  Intuition: calage sur U corrige erreur de non-réponse et d'échantillonnage
- Presque inexistant avec MLE

Aperçu

Scenarios et simulations

- $N = 2000$, $n = 200$

Scenarios et simulations

- $N = 2000$, $n = 200$
- $\mathbf{x}_i = (1, x_i)$, x_i uniform(0,100)

Scenarios et simulations

- $N = 2000$, $n = 200$
- $\mathbf{x}_i = (1, x_i)$, x_i uniform(0,100)
- **Cinq populations**:

Scenarios et simulations

- $N = 2000$, $n = 200$
- $\mathbf{x}_i = (1, x_i)$, x_i uniform(0,100)
- **Cinq populations**:
  1. $y_{1i} = 1000 + 20x_i + \varepsilon_{1i}$, ε_i N(0,750) $r^2 = 0.6$

Scenarios et simulations

- $N = 2000$, $n = 200$
- $\mathbf{x}_i = (1, x_i)$, x_i uniform(0,100)
- **Cinq populations**:
  1. $y_{1i} = 1000 + 20x_i + \varepsilon_{1i}$, ε_i N(0,750) $\quad r^2 = 0.6$
  2. $y_{2i} = 1000 + \varepsilon_{2i}$, ε_{2i} N(0,750) $\quad r^2 \approx 0$

Scenarios et simulations

- $N = 2000$, $n = 200$
- $\mathbf{x}_i = (1, x_i)$, x_i uniform(0,100)
- **Cinq populations**:
  1. $y_{1i} = 1000 + 20x_i + \varepsilon_{1i}$, ε_i N(0,750) $r^2 = 0.6$
  2. $y_{2i} = 1000 + \varepsilon_{2i}$, ε_{2i} N(0,750) $r^2 \approx 0$
  3. $y_{3i} = 1000 + 20x_i + \varepsilon_i$, ε_{3i} N(0,50) $r^2 = 0.99$

Scenarios et simulations

- $N = 2000$, $n = 200$
- $\mathbf{x}_i = (1, x_i)$, x_i uniform(0,100)
- **Cinq populations**:
  1. $y_{1i} = 1000 + 20x_i + \varepsilon_{1i}$, ε_i N(0,750) $r^2 = 0.6$
  2. $y_{2i} = 1000 + \varepsilon_{2i}$, ε_{2i} N(0,750) $r^2 \approx 0$
  3. $y_{3i} = 1000 + 20x_i + \varepsilon_i$, ε_{3i} N(0,50) $r^2 = 0.99$
  4. $y_{4i} = 1500 + 500\exp(-10 + 0.1x)$, ε_i N(0, 100)

Scenarios et simulations

- $N = 2000$, $n = 200$
- $\mathbf{x}_i = (1, x_i)$, x_i uniform(0,100)
- **Cinq populations**:
  1. $y_{1i} = 1000 + 20x_i + \varepsilon_{1i}$, ε_i N(0,750) $r^2 = 0.6$
  2. $y_{2i} = 1000 + \varepsilon_{2i}$, ε_{2i} N(0,750) $r^2 \approx 0$
  3. $y_{3i} = 1000 + 20x_i + \varepsilon_i$, ε_{3i} N(0,50) $r^2 = 0.99$
  4. $y_{4i} = 1500 + 500\exp(-10 + 0.1x)$, ε_i N(0, 100)
  5. y_{5i} Bernoulli(ϕ_i), ϕ_i vaut 0.2 si $x_i \in [0, 75]$ et 0.8 si $x_i \in]75, 100]$

Scenarios et simulations

- $N = 2000$, $n = 200$
- $\mathbf{x}_i = (1, x_i)$, x_i uniform(0,100)
- **Cinq populations**:
  1. $y_{1i} = 1000 + 20x_i + \varepsilon_{1i}$, ε_i N(0,750) $r^2 = 0.6$
  2. $y_{2i} = 1000 + \varepsilon_{2i}$, ε_{2i} N(0,750) $r^2 \approx 0$
  3. $y_{3i} = 1000 + 20x_i + \varepsilon_i$, ε_{3i} N(0,50) $r^2 = 0.99$
  4. $y_{4i} = 1500 + 500\exp(-10 + 0.1x)$, ε_i N(0, 100)
  5. y_{5i} Bernoulli(ϕ_i), ϕ_i vaut 0.2 si $x_i \in [0, 75]$ et 0.8 si $x_i \in]75, 100]$
- **Trois plans de sondages**: SRS, Poisson, Stratified

Scenarios et simulations

- $N = 2000$, $n = 200$
- $\mathbf{x}_i = (1, x_i)$, x_i uniform(0,100)
- **Cinq populations**:
  1. $y_{1i} = 1000 + 20x_i + \varepsilon_{1i}$, ε_i N(0,750) $r^2 = 0.6$
  2. $y_{2i} = 1000 + \varepsilon_{2i}$, ε_{2i} N(0,750) $r^2 \approx 0$
  3. $y_{3i} = 1000 + 20x_i + \varepsilon_i$, ε_{3i} N(0,50) $r^2 = 0.99$
  4. $y_{4i} = 1500 + 500\exp(-10 + 0.1x)$, ε_i N(0, 100)
  5. y_{5i} Bernoulli(ϕ_i), ϕ_i vaut 0.2 si $x_i \in [0, 75]$ et 0.8 si $x_i \in]75, 100]$
- **Trois plans de sondages**: SRS, Poisson, Stratified
- **Deux mechanismes de non-réponse**: p logistique, p non logistique
  Indicatrices de réponse Bernoulli(p_i)

Scenarios et simulations

- $N = 2000$, $n = 200$
- $\mathbf{x}_i = (1, x_i)$, x_i uniform(0,100)
- **Cinq populations**:
  1. $y_{1i} = 1000 + 20x_i + \varepsilon_{1i}$, ε_i N(0,750) $r^2 = 0.6$
  2. $y_{2i} = 1000 + \varepsilon_{2i}$, ε_{2i} N(0,750) $r^2 \approx 0$
  3. $y_{3i} = 1000 + 20x_i + \varepsilon_i$, ε_{3i} N(0,50) $r^2 = 0.99$
  4. $y_{4i} = 1500 + 500\exp(-10 + 0.1x)$, ε_i N(0, 100)
  5. y_{5i} Bernoulli(ϕ_i), ϕ_i vaut 0.2 si $x_i \in [0, 75]$ et 0.8 si $x_i \in]75, 100]$
- **Trois plans de sondages**: SRS, Poisson, Stratified
- **Deux mechanismes de non-réponse**: p logistique, p non logistique
  Indicatrices de réponse Bernoulli(p_i)

  $\rightarrow$ 45 scénarios

Scenarios et simulations

- $N = 2000$, $n = 200$
- $\mathbf{x}_i = (1, x_i)$, x_i uniform(0,100)
- **Cinq populations**:
  1. $y_{1i} = 1000 + 20x_i + \varepsilon_{1i}$, ε_i N(0,750) $r^2 = 0.6$
  2. $y_{2i} = 1000 + \varepsilon_{2i}$, ε_{2i} N(0,750) $r^2 \approx 0$
  3. $y_{3i} = 1000 + 20x_i + \varepsilon_i$, ε_{3i} N(0,50) $r^2 = 0.99$
  4. $y_{4i} = 1500 + 500\exp(-10 + 0.1x)$, ε_i N(0, 100)
  5. y_{5i} Bernoulli(ϕ_i), ϕ_i vaut 0.2 si $x_i \in [0, 75]$ et 0.8 si $x_i \in]75, 100]$
- **Trois plans de sondages**: SRS, Poisson, Stratified
- **Deux mechanismes de non-réponse**: p logistique, p non logistique
  Indicatrices de réponse Bernoulli(p_i)

  $\rightarrow$ 45 scénarios
- Taux de réponse moyen sur U: 0.5

Scenarios et simulations

- $N = 2000$, $n = 200$
- $\mathbf{x}_i = (1, x_i)$, x_i uniform(0,100)
- **Cinq populations**:
  1. $y_{1i} = 1000 + 20x_i + \varepsilon_{1i}$, ε_i N(0,750) $r^2 = 0.6$
  2. $y_{2i} = 1000 + \varepsilon_{2i}$, ε_{2i} N(0,750) $r^2 \approx 0$
  3. $y_{3i} = 1000 + 20x_i + \varepsilon_i$, ε_{3i} N(0,50) $r^2 = 0.99$
  4. $y_{4i} = 1500 + 500\exp(-10 + 0.1x)$, ε_i N(0, 100)
  5. y_{5i} Bernoulli(ϕ_i), ϕ_i vaut 0.2 si $x_i \in [0, 75]$ et 0.8 si $x_i \in]75, 100]$
- **Trois plans de sondages**: SRS, Poisson, Stratified
- **Deux mechanismes de non-réponse**: p logistique, p non logistique
  Indicatrices de réponse Bernoulli(p_i)

  $\rightarrow$ 45 scénarios
- Taux de réponse moyen sur U: 0.5
- 10000 simulations

Estimateurs

1. $\widehat{Y}$ (HT): estimateur d'Horvitz-Thompson, indisponible en pratique, point de comparaison

Estimateurs

1. $\widehat{Y}$ (HT): estimateur d'Horvitz-Thompson, indisponible en pratique, point de comparaison
2. $\widehat{Y}_p$ (p): 2 phases, vraies p_i, indisponible en pratique, point de comparaison

Estimateurs

1. $\widehat{Y}$ (HT): estimateur d'Horvitz-Thompson, indisponible en pratique, point de comparaison
2. $\widehat{Y}_p$ (p): 2 phases, vraies p_i, indisponible en pratique, point de comparaison
3. $\widehat{Y}_{\text{naïf}}$ (naïf): $N n_r^{-1} \sum_{i \in S_r} y_i$

Estimateurs

1. $\widehat{Y}$ (HT): estimateur d'Horvitz-Thompson, indisponible en pratique, point de comparaison
2. $\widehat{Y}_p$ (p): 2 phases, vraies p_i, indisponible en pratique, point de comparaison
3. $\widehat{Y}_{\text{naïf}}$ (naïf): $Nn_r^{-1} \sum_{i \in S_r} y_i$
4. $\widehat{Y}^{mle}_{\widehat{p}}$ (mle): probabilités de réponse estimées via MLE

Estimateurs

1. $\widehat{Y}$ (HT): estimateur d'Horvitz-Thompson, indisponible en pratique, point de comparaison
2. $\widehat{Y}_p$ (p): 2 phases, vraies p_i, indisponible en pratique, point de comparaison
3. $\widehat{Y}_{\text{naïf}}$ (naïf): $Nn_r^{-1} \sum_{i \in S_r} y_i$
4. $\widehat{Y}^{mle}_{\widehat{p}}$ (mle): probabilités de réponse estimées via MLE
5. $\widehat{Y}^{cal,U}_{\widehat{p}}$ (cal$U$): probabilités de réponse estimées via calage sur U

Estimateurs

1. $\widehat{Y}$ (HT): estimateur d'Horvitz-Thompson, indisponible en pratique, point de comparaison
2. $\widehat{Y}_p$ (p): 2 phases, vraies p_i, indisponible en pratique, point de comparaison
3. $\widehat{Y}_{\text{naïf}}$ (naïf): $Nn_r^{-1} \sum_{i \in S_r} y_i$
4. $\widehat{Y}_{\widehat{p}}^{mle}$ (mle): probabilités de réponse estimées via MLE
5. $\widehat{Y}_{\widehat{p}}^{cal,U}$ (cal$U$): probabilités de réponse estimées via calage sur U
6. $\widehat{Y}_{\widehat{p}}^{cal,S}$ (cal$S$): probabilités de réponse estimées via calage sur S

Mesures de comparaison

- Biais relatif empirique

$$\text{RB} = \frac{B}{Y},$$

où B est le biais de Monte-Carlo (moyenne estimateur sur simulations - vrai total Y)

Mesures de comparaison

- Biais relatif empirique

$$\text{RB} = \frac{B}{Y},$$

où B est le biais de Monte-Carlo (moyenne estimateur sur simulations - vrai total Y)

- RRVAR empirique

$$\text{RRVAR} = \frac{(\text{VAR})^{1/2}}{Y},$$

où VAR est la variance empirique (variance estimateur sur simulations)

Mesures de comparaison

- Biais relatif empirique

$$\text{RB} = \frac{B}{Y},$$

où B est le biais de Monte-Carlo (moyenne estimateur sur simulations - vrai total Y)

- RRVAR empirique

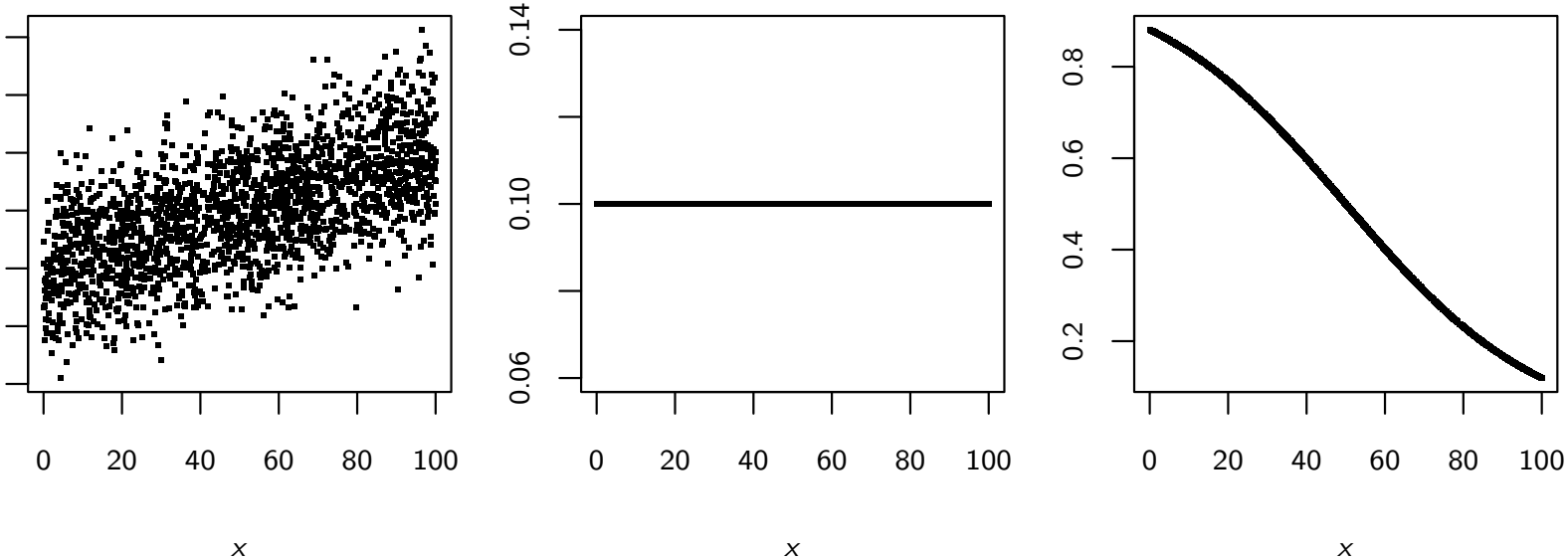
$$\text{RRVAR} = \frac{(\text{VAR})^{1/2}}{Y},$$

où VAR est la variance empirique (variance estimateur sur simulations)

- Taux de convergence des méthodes appliquées pour estimer p (# simulations convergence / # total simulations)

Scenario 1

- Population 1: $y_{1i} = 1000 + 20x_i + \varepsilon_{1i}$, ε_i N(0,750) $r^2 = 0.6$
- SRS: $\pi_i = n/N = 0.1$
- NR correct: méchanisme de réponse logistique



Population, Probabilités d'inclusion, Probabilités de réponse

Scenario 1: résultats

	RB	RRVAR
HT	-1.26×10^{-5}	3.20×10^{-2}
p	1.03×10^{-3}	1.33×10^{-1}
naïf	-1.43×10^{-1}	4.41×10^{-2}
mle	1.51×10^{-3}	8.11×10^{-2}
calU	-2.89×10^{-4}	4.46×10^{-2}
calS	-2.22×10^{-4}	4.88×10^{-2}

- RB de $\widehat{Y}_{\widehat{p}}$ proche de RB de $\widehat{Y}_p$ (sans biais) pour MLE, calage sur S et U

Scenario 1: résultats

	RB	RRVAR
HT	-1.26×10^{-5}	3.20×10^{-2}
p	1.03×10^{-3}	1.33×10^{-1}
naïf	-1.43×10^{-1}	4.41×10^{-2}
mle	1.51×10^{-3}	8.11×10^{-2}
calU	-2.89×10^{-4}	4.46×10^{-2}
calS	-2.22×10^{-4}	4.88×10^{-2}

- RB de $\widehat{Y}_{\widehat{p}}$ proche de RB de $\widehat{Y}_p$ (sans biais) pour MLE, calage sur S et U
- RRVAR de $\widehat{Y}_{\widehat{p}}$ plus petit que RRVAR de $\widehat{Y}_p$ pour MLE, calage sur S et U

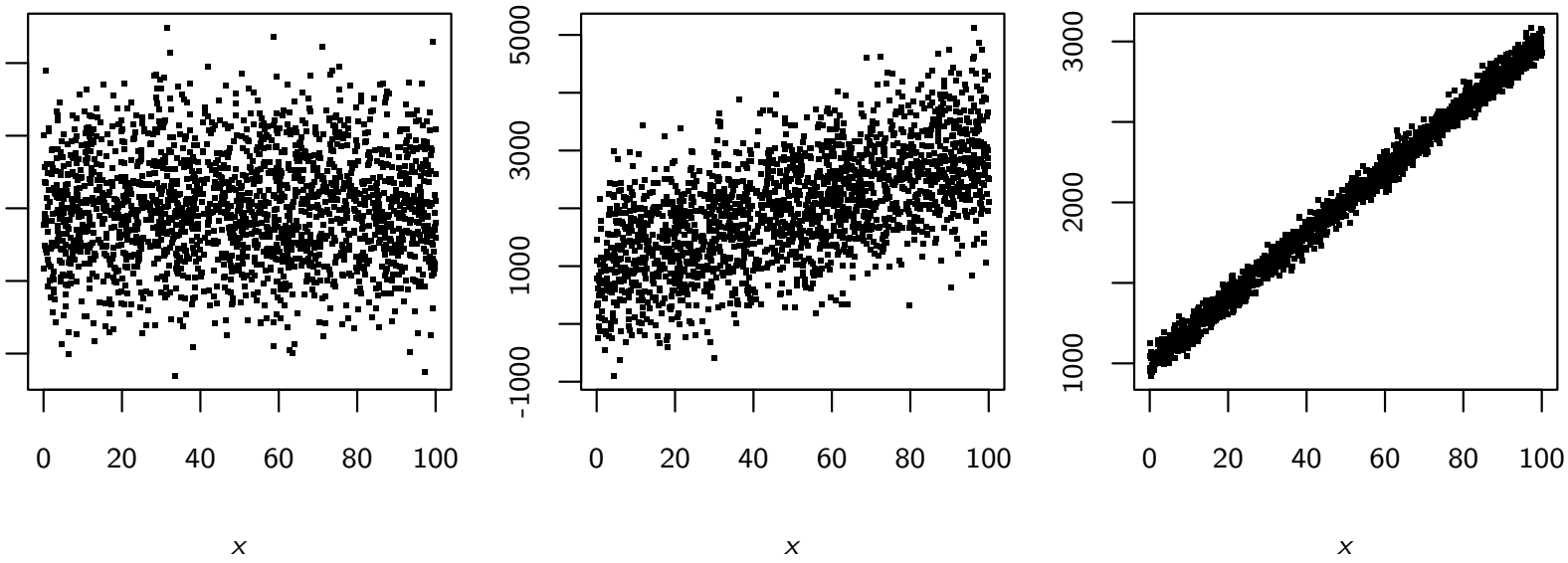
Scenario 1: résultats

	RB	RRVAR
HT	-1.26×10^{-5}	3.20×10^{-2}
p	1.03×10^{-3}	1.33×10^{-1}
naïf	-1.43×10^{-1}	4.41×10^{-2}
mle	1.51×10^{-3}	8.11×10^{-2}
calU	-2.89×10^{-4}	4.46×10^{-2}
calS	-2.22×10^{-4}	4.88×10^{-2}

- RB de $\widehat{Y}_{\widehat{p}}$ proche de RB de $\widehat{Y}_p$ (sans biais) pour MLE, calage sur S et U
- RRVAR de $\widehat{Y}_{\widehat{p}}$ plus petit que RRVAR de $\widehat{Y}_p$ pour MLE, calage sur S et U
- Naïf a le plus grand RB

Scenario 2

- Population 1 à 3, $r^2 = 0, 0.6, 0.99$
- Relation linéaire devient de plus en plus forte
- SRS: $\pi_i = n/N = 0.1$
- NR correct: méchanisme de réponse logistique



Population, Probabilités d'inclusion, Probabilités de réponse

Scenario 2: résultats RRVAR

	1	2	3
mle	9.58×10^{-2}	8.11×10^{-2}	6.54×10^{-2}
calU	9.11×10^{-2}	4.46×10^{-2}	3.58×10^{-3}
calS	8.99×10^{-2}	4.88×10^{-2}	1.96×10^{-2}

- Plus la relation linéaire est forte, plus RRVAR diminue avec calage sur U et S

Scenario 2: résultats RRVAR

	1	2	3
mle	9.58×10^{-2}	8.11×10^{-2}	6.54×10^{-2}
calU	9.11×10^{-2}	4.46×10^{-2}	3.58×10^{-3}
calS	8.99×10^{-2}	4.88×10^{-2}	1.96×10^{-2}

- Plus la relation linéaire est forte, plus RRVAR diminue avec calage sur U et S
- Plus flagrant avec calage sur U

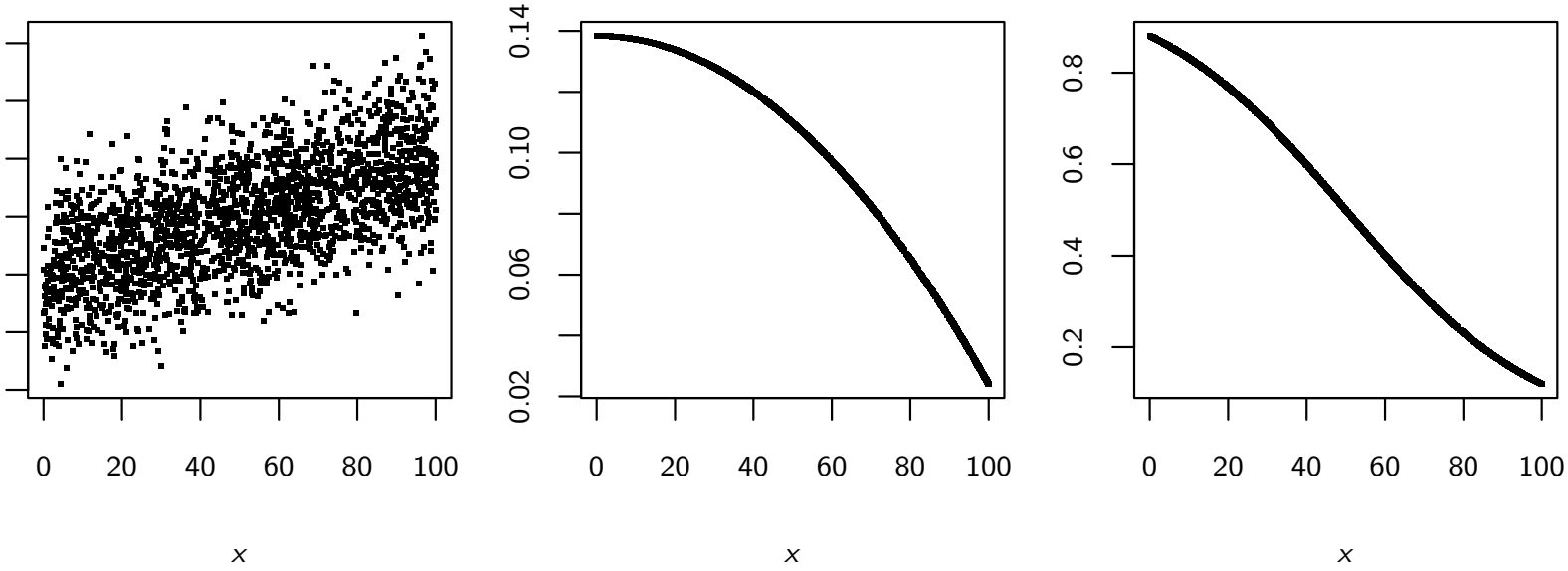
Scenario 2: résultats RRVAR

	1	2	3
mle	9.58×10^{-2}	8.11×10^{-2}	6.54×10^{-2}
calU	9.11×10^{-2}	4.46×10^{-2}	3.58×10^{-3}
calS	8.99×10^{-2}	4.88×10^{-2}	1.96×10^{-2}

- Plus la relation linéaire est forte, plus RRVAR diminue avec calage sur U et S
- Plus flagrant avec calage sur U
- RRVAR plus petite avec calage sur U que calage sur S (sauf si pas de relation linéaire)

Scenario 3

- Population 1: $y_{1i} = 1000 + 20x_i + \varepsilon_{1i}$, ε_i N(0,750) $r^2 = 0.6$
- Poisson: $\pi_i \propto 110^2 - x_i^2$
- NR correct: méchanisme de réponse logistique



Population, Probabilités d'inclusion, Probabilités de réponse

Scenario 3: résultats

	RRVAR	rate cal/estim eqn
mle	9.35×10^{-2}	0.9999
calU	4.47×10^{-2}	0.2133
calS	9.71×10^{-2}	0.1591

- RRVAR plus petite avec calage sur U que calage sur S

Scenario 3: résultats

	RRVAR	rate cal/estim eqn
mle	9.35×10^{-2}	0.9999
calU	4.47×10^{-2}	0.2133
calS	9.71×10^{-2}	0.1591

- RRVAR plus petite avec calage sur U que calage sur S
- Très faible taux de convergence pour calage sur U et calage sur S

Scenario 3: résultats

	RRVAR	rate cal/estim eqn
mle	9.35×10^{-2}	0.9999
calU	4.47×10^{-2}	0.2133
calS	9.71×10^{-2}	0.1591

- RRVAR plus petite avec calage sur U que calage sur S
- Très faible taux de convergence pour calage sur U et calage sur S
- Presque que de très petites valeurs de x dans les répondants

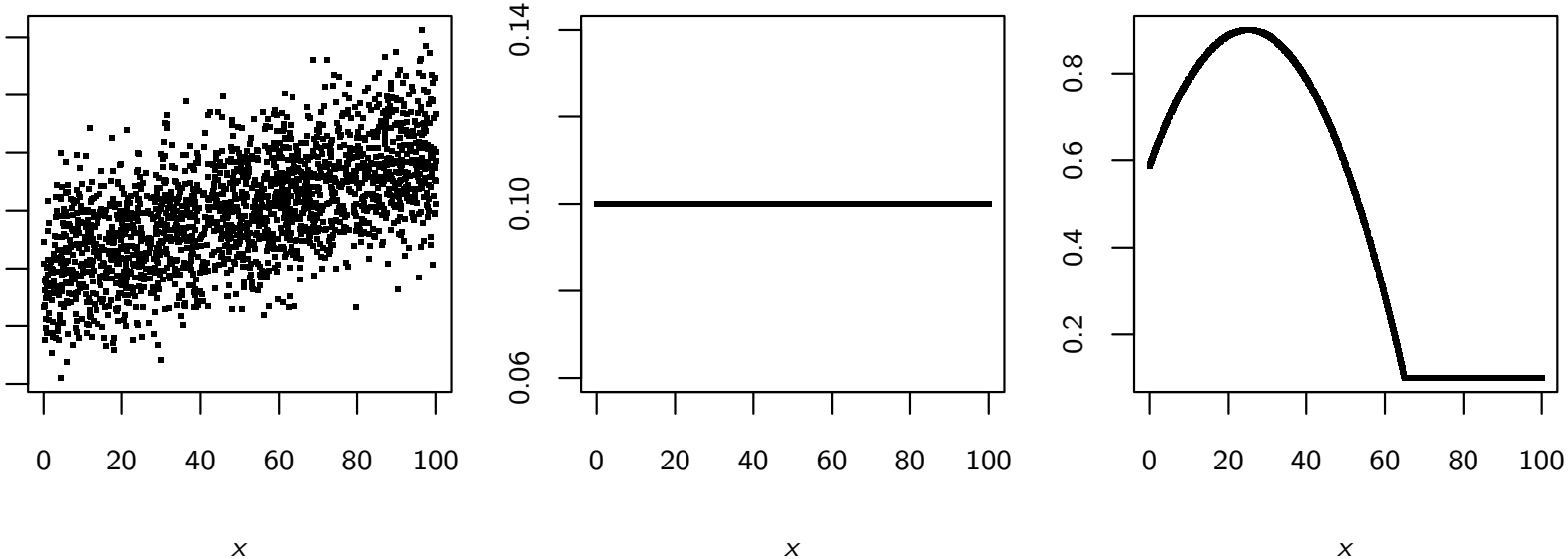
Scenario 3: résultats

	RRVAR	rate cal/estim eqn
mle	9.35×10^{-2}	0.9999
calU	4.47×10^{-2}	0.2133
calS	9.71×10^{-2}	0.1591

- RRVAR plus petite avec calage sur U que calage sur S
- Très faible taux de convergence pour calage sur U et calage sur S
- Presque que de très petites valeurs de x dans les répondants
- MLE converge toujours (ou presque)

Scenario 4

- Population 1: $y_{1i} = 1000 + 20x_i + \varepsilon_{1i}$, ε_i N(0,750) $r^2 = 0.6$
- SRS: $\pi_i = n/N = 0.1$
- NR incorrect: méchanisme de réponse non-logistique



Population, Probabilités d'inclusion, Probabilités de réponse

Scenario 4: résultats

	RB	RRVAR
HT	-1.40×10^{-4}	3.20×10^{-2}
p	3.06×10^{-3}	1.81×10^{-1}
naïf	-1.76×10^{-1}	4.29×10^{-2}
mle	-1.13×10^{-1}	8.90×10^{-2}
calU	2.30×10^{-3}	5.33×10^{-2}
calS	2.73×10^{-3}	5.72×10^{-2}

- RB de $\widehat{Y}_{\widehat{p}}$ proche de RB de $\widehat{Y}_p$ (sans biais) calage sur S et U

Scenario 4: résultats

	RB	RRVAR
HT	-1.40×10^{-4}	3.20×10^{-2}
p	3.06×10^{-3}	1.81×10^{-1}
naïf	-1.76×10^{-1}	4.29×10^{-2}
mle	-1.13×10^{-1}	8.90×10^{-2}
calU	2.30×10^{-3}	5.33×10^{-2}
calS	2.73×10^{-3}	5.72×10^{-2}

- RB de $\widehat{Y}_{\widehat{p}}$ proche de RB de $\widehat{Y}_p$ (sans biais) calage sur S et U
- RB plus grand avec MLE

Scenario 4: résultats

	RB	RRVAR
HT	-1.40×10^{-4}	3.20×10^{-2}
p	3.06×10^{-3}	1.81×10^{-1}
naïf	-1.76×10^{-1}	4.29×10^{-2}
mle	-1.13×10^{-1}	8.90×10^{-2}
calU	2.30×10^{-3}	5.33×10^{-2}
calS	2.73×10^{-3}	5.72×10^{-2}

- RB de $\widehat{Y}_{\widehat{p}}$ proche de RB de $\widehat{Y}_p$ (sans biais) calage sur S et U
- RB plus grand avec MLE
- NR pas correctement spécifié, SUP linéaire correctement spécifié

Aperçu

Discussion

- Propriétés asymptotiques de l'estimateur par double expansion empirique quand les probabilités de réponse estimées par MLE, calage sur S et calage sur U

Discussion

- Propriétés asymptotiques de l'estimateur par double expansion empirique quand les probabilités de réponse estimées par MLE, calage sur S et calage sur U
- Conclusions principales (illustrées par simulation dans cette présentation):
  - Asymptotiquement sans biais
  - Au moins autant efficace que l'estimateur par double expansion (avec vraies probabilités de réponse)
  - Doublement robuste avec calage, double robustesse avec MLE pas évidente

Discussion

- Propriétés asymptotiques de l'estimateur par double expansion empirique quand les probabilités de réponse estimées par MLE, calage sur S et calage sur U
- Conclusions principales (illustrées par simulation dans cette présentation):
  - Asymptotiquement sans biais
  - Au moins autant efficace que l'estimateur par double expansion (avec vraies probabilités de réponse)
  - Doublement robuste avec calage, double robustesse avec MLE pas évidente
- Problémes de convergence (ou poids extrêmes) avec calage
- Beaucoup moins fréquent avec MLE

Discussion

- **Pratique: choix de la méthode peut provenir de**
  - Niveau de connaissance de **x**
  - Besoin de consistance entre totaux estimés et totaux connus pour certains **x**

Discussion

- **Pratique: choix de la méthode peut provenir de**
  - Niveau de connaissance de **x**
  - Besoin de consistance entre totaux estimés et totaux connus pour certains **x**
- **Pratique: approches plus robustes**
  - Méthode en deux étapes: mle suivi de calage [Haziza and Lesage, 2016]
  - Méthode du score: mle puis partition en classes homogènes [Haziza and Beaumont, 2007]

Discussion

- **Pratique: choix de la méthode peut provenir de**
  - Niveau de connaissance de **x**
  - Besoin de consistance entre totaux estimés et totaux connus pour certains **x**
- **Pratique: approches plus robustes**
  - Méthode en deux étapes: mle suivi de calage [Haziza and Lesage, 2016]
  - Méthode du score: mle puis partition en classes homogènes [Haziza and Beaumont, 2007]
- Contrôle du biais de non-réponse, consistence entre totaux estimés et totaux connus, reduit problème convergence ou de poids extrêmes

Discussion

- **Pratique: choix de la méthode peut provenir de**
  - Niveau de connaissance de **x**
  - Besoin de consistance entre totaux estimés et totaux connus pour certains **x**
- **Pratique: approches plus robustes**
  - Méthode en deux étapes: mle suivi de calage [Haziza and Lesage, 2016]
  - Méthode du score: mle puis partition en classes homogènes [Haziza and Beaumont, 2007]
- Contrôle du biais de non-réponse, consistence entre totaux estimés et totaux connus, reduit problème convergence ou de poids extrêmes
- Pas de théorie (à ma connaissance)

Remerciements

Pr. Yves Tillé, Dr. Lionel Qualité et collègues de
l'institut de statistique de l'Université de Neuchâtel

Office Fédéral de la Statistique

Les points de vues exprimées dans cette
présentation sont celles de l'auteur uniquement

Beaumont, J.-F. (2005).
Calibrated imputation in surveys under a quasi-model-assisted approach.
Journal of the Royal Statistical Society. Series B, 67:445–458.

Deville, J.-C. and Dupont, F. (1993).
Non-réponse: principes et méthodes.
In *Actes des Journées de Méthodologie Statistique*, pages 53–70, INSEE, Paris.

Dupont, F. (1993).
Calage et redressement de la non-réponse totale : validité de la pratique courante de redressement et comparaison des méthodes alternatives pour l'enquête sur la consommation alimentaire de 1989.
In *Actes des Journées de Méthodologie Statistique*, pages 9–42, INSEE, Paris.

Haziza, D. and Beaumont, J.-F. (2007).
On the construction of imputation classes in surveys.
International Statistical Review, 75(1):25–43.

Haziza, D. and Lesage, E. (2016).
A discussion of weighting procedures for unit nonresponse.
Journal of Official Statistics, 32(1):129–145.

Horvitz, D. G. and Thompson, D. J. (1952).
A generalization of sampling without replacement from a finite universe.
Journal of the American Statistical Association, 47:663–685.

Iannacchione, V. G., Milne, J. G., and Folsom, R. E. (1991).
Response probability weight adjustments using logistic regression.
In *In Proceedings of the Survey Research Methods Section, American Statistical Association*, pages 637–642.

Kim, J. K. and Kim, J. (2007).
Nonresponse weighting adjustment using estimated response probability.
The Canadian Journal of Statistics / La Revue Canadienne de Statistique, 35(4):501–514.

Kim, J. K. and Riddles, M. K. (2012).
Some theory for propensity-score-adjustment estimators in survey sampling.

Survey Methodology, 38(2):157–165.

Kott, P. S. (2006).

Using calibration weighting to adjust for nonresponse and coverage errors.
Survey Methodology, 32(2):133–142.

Kott, P. S. and Liao, D. (2012).

Providing double protection for unit nonresponse with a nonlinear calibration-weighting routine.
Survey Research Methods, 6(2):105–111.

Lundström, S. and Särndal, C.-E. (1999).

Calibration as a standard method for treatment of nonresponse.
Journal of Official Statistics, 15:305–327.

Rubin, D. B. (1976).

Inference and missing data.
Biometrika, 63:581–592.

Särndal, C.-E. and Lundström, S. (2005).

Estimation in surveys with nonresponse.
Wiley, New York.