Utilisation d'une approche conditionnelle pour traiter les outliers

Éric Lesage

Laboratoire de statistique d'enquête
Crest-Ensai

6 novembre 2012

7^e Colloque Francophone sur les Sondages
Ensai, Bruz

en collaboration avec François Coquet, Crest-Ensai et Irmar

Plan de l'exposé

Outlier pour la variable d'intérêt y

- Le plan de sondage est noté $\mathbb{P}(s)$
  - $\mathcal{S}$ est l'ensemble des échantillons possibles a priori (au moment de l'échantillonnage)
  - $I_k(s)$ l'indicatrice d'appartenance de l'unité k à l'échantillon s
  - La probabilité d'inclusion de l'unité k (a priori)

$$\pi_k = \mathbb{P}([k \in s]) = \mathbb{E}(I_k(s))$$

  - $d_k = \frac{1}{\pi_k}$ le poids d'échantillonnage

- L'outlier
  - La variable d'intérêt y est par exemple le chiffre d'affaire de l'année n
  - Dans l'échantillon s_0 enquêté, l'unité 1 possède une valeur extrême :

$$y_1 = 11\ 695$$

Outlier pour la variable d'intérêt y

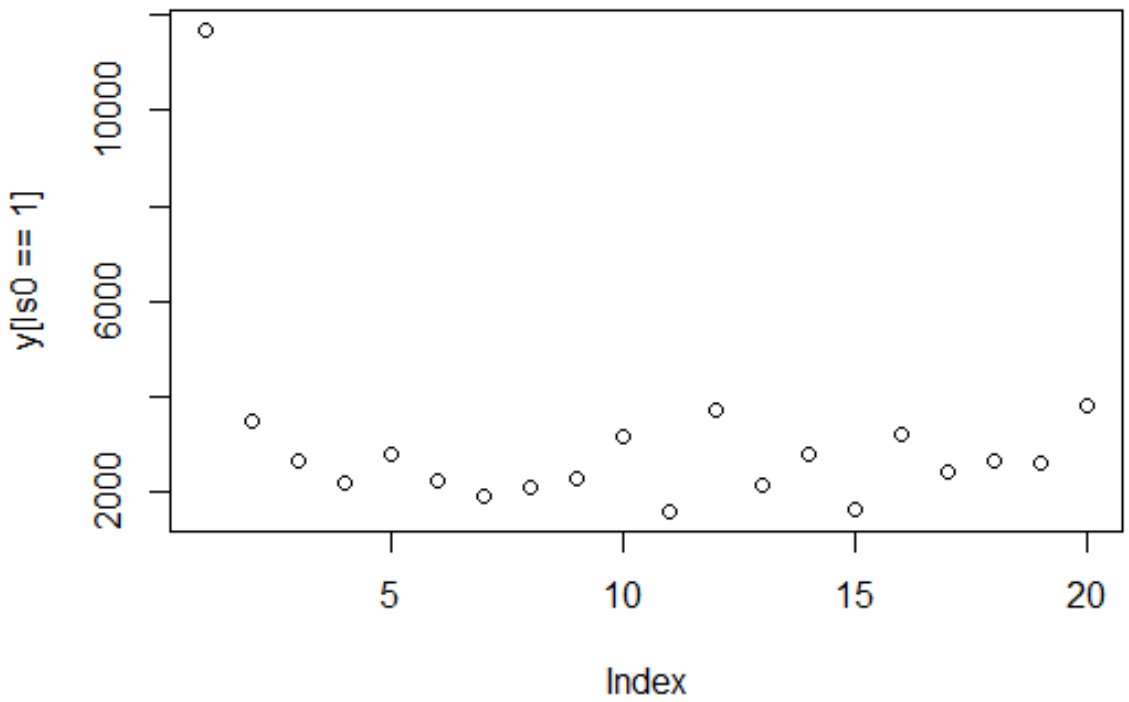


Figure: y_k pour les unités de l'échantillon s_0

Le problème causé par la présence d'un outlier

- Pour un sondage aléatoire simple de fraction $f = 0.2$, le poids de l'outlier est $d_1 = 5$.
  I.e. que l'outlier représente 4 autres unités semblables
- Le praticien a le sentiment que son estimateur HT est entaché d'une forte erreur d'échantillonnage et il souhaiterait réduire le poids de l'outlier
- La variance de l'estimateur Horvitz-Thompson de la moyenne μ_y est importante
- L'outlier a une influence très forte sur l'estimateur Horvitz-Thompson de la moyenne μ_y (Haziza et al., 2011)

Les solutions pour traiter la présence d'un outlier

1. Traiter l'outlier comme une unité non-substituable et imposer un poids de 1 (méthode empirique)
2. Winsorizer les poids : on diminue le poids de façon à réduire l'influence de l'outlier
3. Utiliser une information auxiliaire sensible à la présence de l'outlier dans le cadre d'une approche conditionnelle...
   C'est ce que nous allons voir

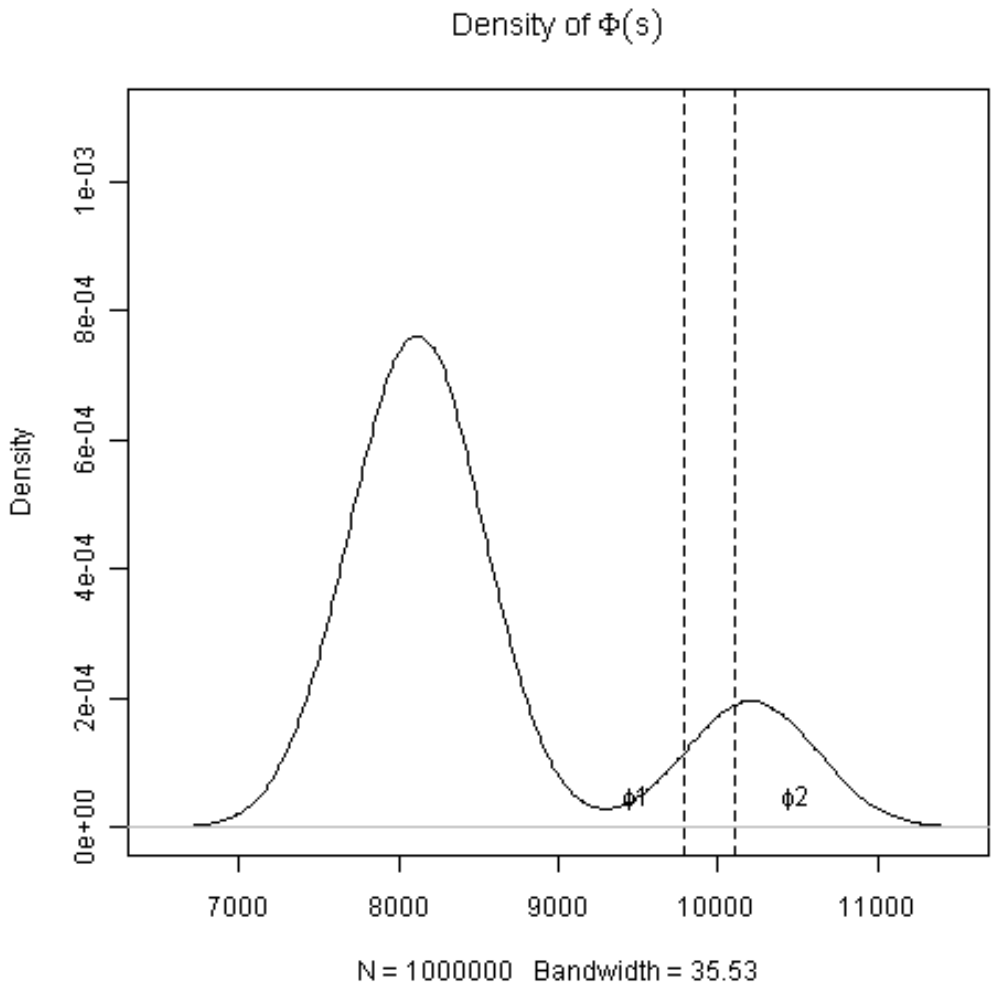
Une information auxiliaire complète liée à la variable d'intérêt

- Une variable auxiliaire x_k est connue pour toutes les unités de la base de sondage après l'échantillonnage
- Par exemple, les effectifs de l'entreprise ou son chiffre d'affaire de l'année *n-1*
- L'unité 1 est également un outlier pour la variable x
- On connaît parfaitement la distribution (Monte Carlo) de la statistique

$$\phi(s) = \hat{\mu}_{x,\pi}$$

sous le plan de sondage

La distribution de $\hat{\mu}_{x,\pi}$ fournit une information importante sur l'hétérogénéité de l'ensemble $\mathcal{S}$



Un sous-ensemble $\mathcal{S}_\varphi$ adapté à l'inférence

- A chaque intervalle $[\varphi_1, \varphi_2]$ de valeurs de $\hat{\mu}_{x,\pi}$ correspond un sous-ensemble de $\mathcal{S}$

$$\mathcal{S}_\varphi = \{s \in \mathcal{S};\ \hat{\mu}_{x,HT}(s) \in [\varphi_1, \varphi_2]\}$$

- On ne peut pas ignorer au moment de l'estimation la valeur $\hat{\mu}_{x,\pi}(s_0)$ qui caractérise notre échantillon enquêté s_0
- On réduit alors l'inférence au sous-ensemble "adapté" $\mathcal{S}_\varphi$ (Rao, 1985) correspondant à un intervalle contenant $\hat{\mu}_{x,\pi}(s_0)$
- On utilise une inférence conditionnelle à $\mathcal{S}_\varphi$
  (on regarde la distribution de notre estimateur pour des échantillons similaires à s_0)

Un exemple pour se convaincre...

- On réalise un sondage Bernoullien fraction de sondage f
- Au moment de l'estimation on utilise la taille de l'échantillon n et on réduit l'inférence à l'ensemble des échantillons de taille n
- Les poids des unités échantillonnées ne sont plus $1/f$ mais N/n

Poids conditionnels

- La probabilité d'inclusion conditionnelle de l'unité k

$$\pi_k^{\varphi} = \mathbb{P}\left([k \in s] \mid \mathcal{S}_{\varphi}\right) = \mathbb{E}\left(I_k(s) \mid \mathcal{S}_{\varphi}\right)$$

- On obtient un nouveau jeu de poids conditionnels pour les unités de l'échantillon

$$w_k = \frac{1}{\pi_k^{\varphi}}$$

- L'estimateur pondéré conditionnel (Tillé, 1998 et 1999) qui est un estimateur HT conditionnel

$$\hat{\mu}_{y,CHT} = \frac{1}{N} \sum_{k \in s} \frac{1}{\pi_k^{\varphi}} y_k$$

Estimateur conditionnellement sans biais

- $\hat{\mu}_{y,CHT}$ est conditionnellement sans biais

$$\forall y \qquad \mathbb{E}(\hat{\mu}_{y,CHT} \mid \hat{\mu}_{x,HT}) = \mu_y,$$

(Si $\forall k \in \mathcal{U}$, $\pi_k^{\varphi} \neq 0$)

- Pour la variance, on a

$$\mathbb{V}(\hat{\mu}_{y,CHT}) = \mathbb{E}\left(\mathbb{V}\left(\hat{\mu}_{y,CHT} \mid \hat{\mu}_{x,HT}\right)\right) + 0$$

- Et on prend comme estimateur de variance, l'estimateur de la variance conditionnelle :

$$v(\hat{\mu}_{y,CHT}) = \sum_{k,l \in s} \frac{1}{\pi_{k,l}^{\varphi}} \frac{y_k}{\pi_k^{\varphi}} \frac{y_l}{\pi_l^{\varphi}} (\pi_{k,l}^{\varphi} - \pi_k^{\varphi}\pi_l^{\varphi})$$

où $\pi_{k,l}^{\varphi} = \mathbb{E}\left(I_k I_l \mid \mathcal{S}_{\varphi}\right)$ est la probabilité d'inclusion double conditionnelle des unités k et l

Calcul des probabilités d'inclusion conditionnelles π_k^φ

Différentes approches :

1. Approximation par des lois normales des distributions conditionnelles et non-conditionnelle de $\hat{\mu}_{x,HT}$ (Tillé, 1999)
   Estimateur conditionnel équivalent à l'estimateur par la régression (estimateur linéaire optimal)
2. Calcul direct lorsqu'on connaît la loi $\mathbb{P}(s \mid \mathcal{S}_\varphi)$ (ou une méthode de calcul de π_k^φ)
   - SAS + conditionnement sur les tailles des sous-échantillons n_h dans les strates
     Estimateur post-stratifié
   - Tirage poissionnien conditionnel + conditionnement sur les tailles des sous-échantillons n_h dans les strates
     Estimateur HT d'un tirage Poissionnien stratifié
3. Estimation des π_k^φ par méthodes Monte-Carlo (Thompson(2008) et Fattorini(2006)) : $\hat{\pi}_k^\varphi$

L'estimateur conditionnel Monte Carlo

- Estimateur ponctuel

$$\hat{\mu}_{y,MC}(s) = \frac{1}{N} \sum_{k \in s} \frac{y_k}{\hat{\pi}_k^{\varphi}(s)}$$

où $\hat{\pi}_k^{\varphi}(s)$ est estimé par $K = 10^6$ itérations du mécanisme d'échantillonnage

$$\hat{\pi}_k^{\varphi}(s) = \frac{\text{nb d'échant. qui appartiennent à } \mathcal{S}_{\varphi} \text{ et qui contiennent k}}{\text{nombre d'échantillons qui appartiennent à } \mathcal{S}_{\varphi}}$$

- Estimateur de variance

$$v(\hat{\mu}_{y,CHT}) = \sum_{k,l \in s} \frac{1}{\hat{\pi}_{k,l}^{\varphi}} \frac{y_k}{\hat{\pi}_k^{\varphi}} \frac{y_l}{\hat{\pi}_l^{\varphi}} \left(\hat{\pi}_{k,l}^{\varphi} - \hat{\pi}_k^{\varphi} \hat{\pi}_l^{\varphi} \right)$$

où $\hat{\pi}_{k,l}^{\varphi}$ est l'estimateur Monte Carlo de la probabilité d'inclusion double conditionnelle des unités k et l

Un exemple sur données simulées

- Une population $\mathcal{U}$ de taille $N = 100$
- La variable auxiliaire est connue au moment de l'estimation
  - Pour $k \neq 1$
    $$x_k \sim \mathcal{N}(8\ 000, (2\ 000)^2)$$
  - Pour l'unité $k = 1$
    $$x_1 = 50\ 000$$
- La variable d'intérêt y_k est liée à x_k

$$y_k = 1000 + 0.2\, x_k + u_k,$$

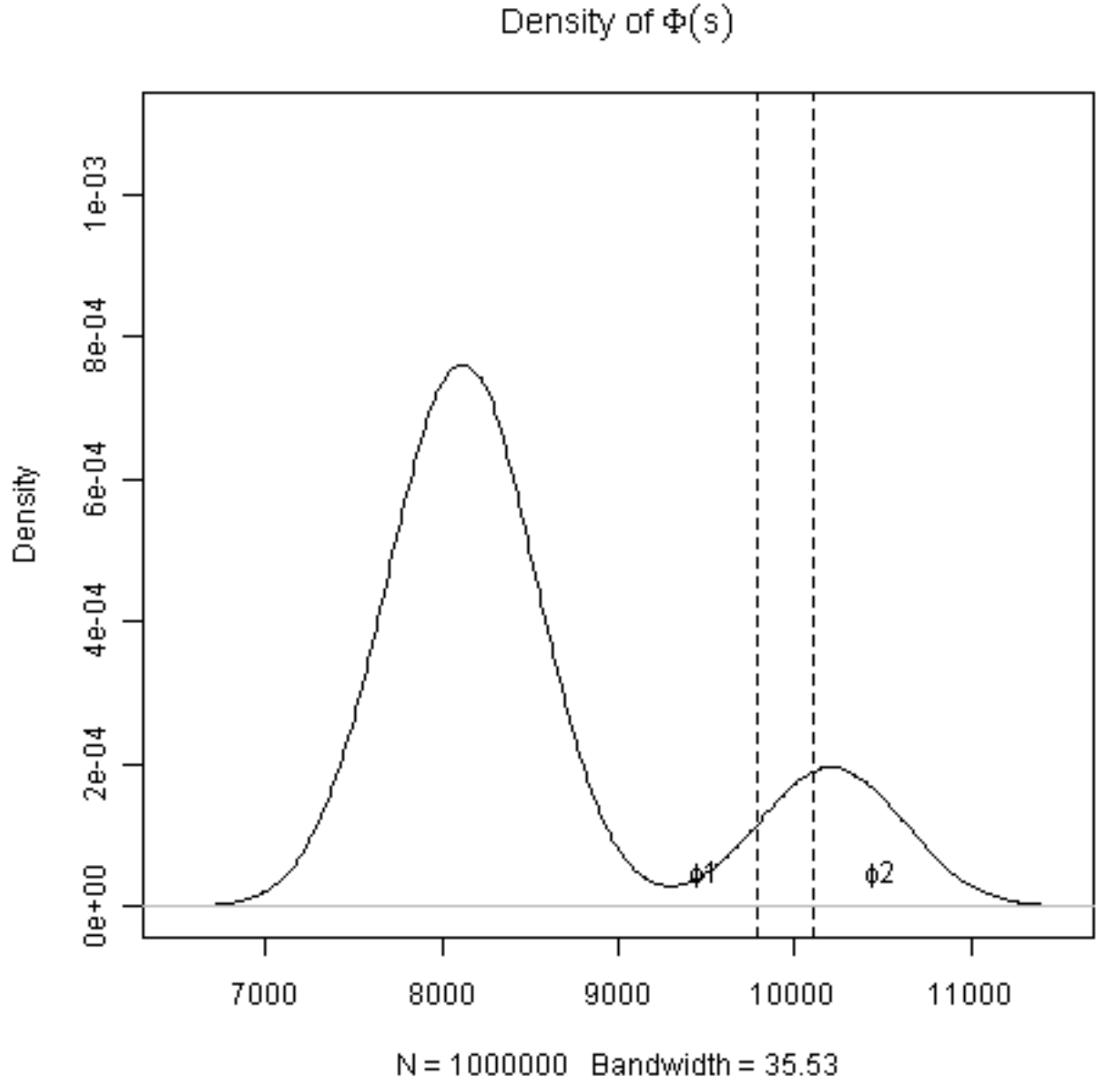
où $u_k \sim \mathcal{N}(0, (500)^2)$

Plan de sondage

- Sondage aléatoire simple sans remise de taille $n = 20$
- La probabilité d'inclusion initiale vaut

$$\pi_k = 0.2$$

  et $d_k = 5$
- L'échantillon tiré s_0 contient l'unité $k = 1$ (outlier)
- On obtient
  - $\hat{\mu}_{x,HT}(s_0) = 9\ 970$,
    ce qui représente 17% de plus que la vraie valeur $\mu_x = 8\ 531$
  - $\hat{\mu}_{y,HT}(s_0) = 3\ 039$
    ce qui représente 13% de plus que la vraie valeur $\mu_y = 2\ 695$


Density of Φ(s)
Density
0e+00
2e-04
4e-04
6e-04
8e-04
1e-03
7000
8000
9000
10000
11000
ϕ1
ϕ2
N = 1000000 Bandwidth = 35.53

Choix de $\mathcal{S}_\varphi$

On passe à l'estimateur pondéré conditionnel...

- L'ensemble des échantillons "réalistes" dans notre approche conditionnelle est :

$$\mathcal{S}_\varphi = \{s \in \mathcal{S},\ \hat{\mu}_{x,HT} \in [9\ 793, 10\ 110]\}$$

- Ce sous-ensemble d'échantillons est assez important

$$\mathbb{P}\left([\hat{\mu}_{x,HT} \in [9\ 793, 10\ 110]]\right) = \alpha = 5\%$$

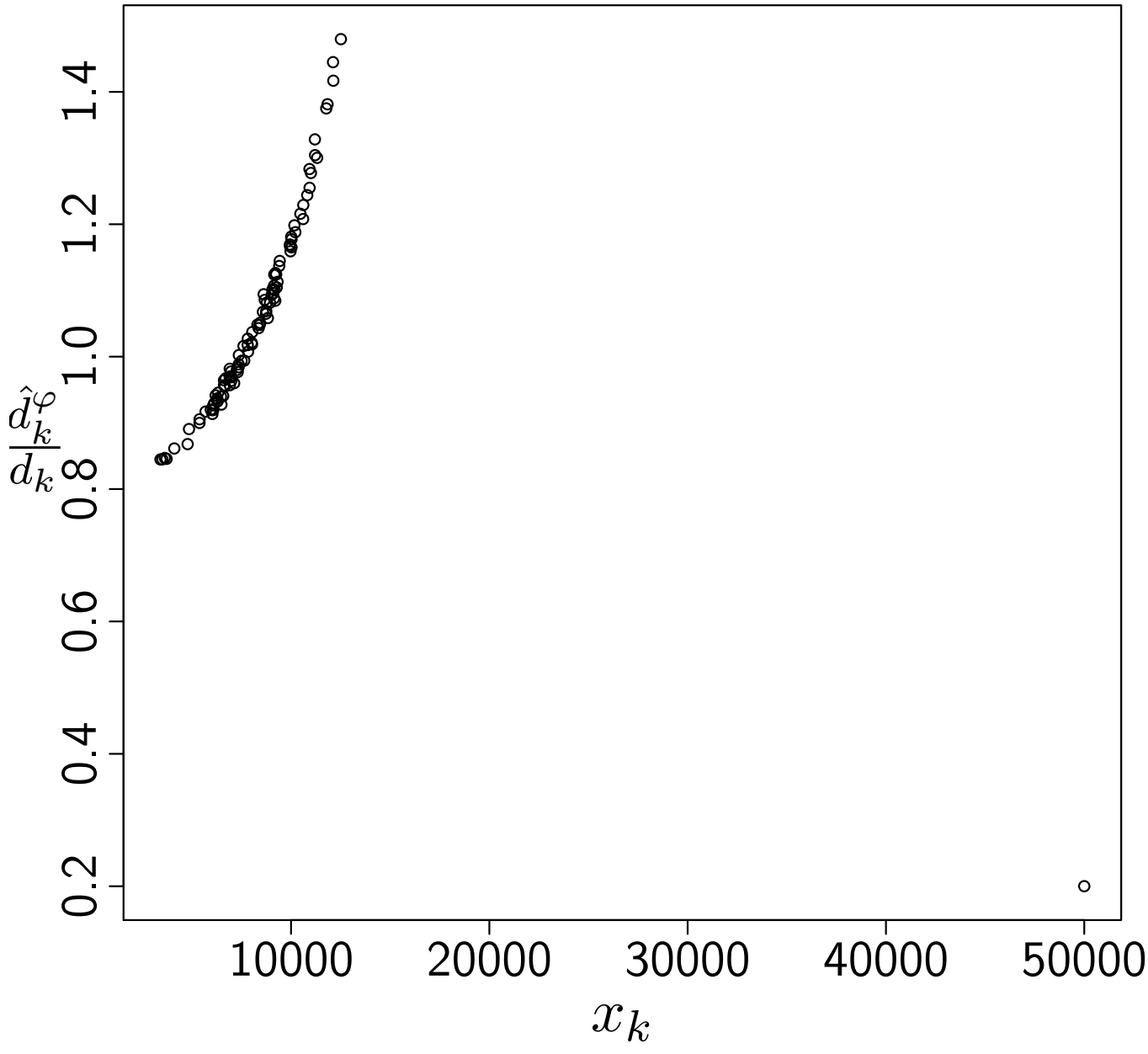
Probabilités d'inclusion conditionnelles estimées $\hat{\pi}_k^{\varphi}$

- On réalise $K = 10^6$ échantillonnages pour estimer les probabilités d'inclusion $\hat{\pi}_k^{\varphi}$
- $49\ 782$ (4.98%) échantillons simulés tombent dans $\mathcal{S}_{\varphi}$
- Poids conditionnel de l'unité 1 (outlier)
  - Parmi eux, $49\ 767$ échantillons contiennent l'outlier, ce qui donne une probabilité d'inclusion estimée de l'outlier de

$$\hat{\pi}_1^{\varphi} = 0.9997$$

  - Le poids de l'unité 1 passe de $d_1 = \frac{1}{0.2} = 5$ à $\hat{d}_1^{\varphi} = 1.0003$
- Les poids conditionnels des autres unités de s_0 sont plus comparables à leurs poids initiaux $d_k = 5$ (voir le graphique)

Sampling Weights Corrections



Estimateur pondéré conditionnel

- L'estimateur conditionnel Monte Carlo de μ_y donne une meilleure estimation de $\mu_y = 2\ 695$:

$$\hat{\mu}_{y,MC}(s_0) = 2\ 671$$

à comparer à $\hat{\mu}_{y,HT}(s_0) = 3\ 039$

Bibliographie

- Fattorini L. (2006). Applying the Horvitz-Thompson criterion in complexe designs : A computer-intensive perspective for estimating inclusion probabilities. Biometrika, 93, 269-278.
- Beaumont, J.-F., Haziza, D. and Ruiz-Gazen, A. (2011). A unified approach to robust estimation in finite population sampling. In revision for Biometrika.
- Rao, J.N.K. (1985). Conditional inference in survey sampling. Survey Methodology, 11, 15-31.
- Thompson M., E. and Wu C. (2008). Simulation-based Randomized Systematic PPS Sampling Under Substitution of Units. Survey Methodology, 34, 3-10.
- Tillé, Y. (1998). Estimation in surveys using conditional inclusion probabilities : Simple random sampling. International Statistical Review, 66, 303-322.
- Tillé, Y. (1999). Estimation in surveys using conditional inclusion probabilities : comlex design. Survey Methodology, 25, 57-66.

Merci de votre attention.