Bootstrap avec remise pour l'estimation de variance dans les enquêtes auprès de ménages
Principes et exemples

Guillaume Chauvet
Travail joint avec Pascal Bessonneau, Gwennaëlle Brilhault et Cédric Garcia

ENSAI/IRMAR-INED-Université Gustave Eiffel

30/03/2022

Le bootstrap, pour quoi faire?

Dans le cadre d'une enquête, nous aimerions assortir les estimations produites d'une mesure de précision (estimation de variance, coefficient de variation, intervalle de confiance).

C'est généralement difficile, mais possible si nous avons toute l'information nécessaire (Delta et Paliod, 2021). Cependant, l'utilisateur de données d'enquête a souvent une connaissance limitée du plan de sondage et de la procédure d'estimation.

Même si un calcul de précision est possible et programmé sous forme de logiciel, l'utilisateur de données d'enquête n'a pas forcément les connaissances nécessaires pour le paramétrer.

Le bootstrap, pour quoi faire?

Technique computationnelle (Efron, 1979) qui permet (en théorie) d'estimer la distribution d'un estimateur, en reproduisant de façon répétée le mécanisme de sélection et la procédure d'estimation utilisée.

Dans le cadre des enquêtes, l'objectif est souvent plus simplement de produire un estimateur de variance. Le bootstrap consiste alors :

- à répliquer B fois la création des poids d'extrapolation par rééchantillonnage + réestimation,
- à obtenir ainsi un jeu de B poids bootstrap,
- à les utiliser pour calculer les statistiques bootstrappées.
  Leur dispersion est alors utilisée comme estimateur de variance.

Procédure très simple d'un point de vue utilisateur.
Du point de vue du bootstrappeur, c'est un peu plus compliqué.

Le bootstrap, comment faire?

Devant la difficulté d'étendre de façon naturelle le bootstrap au cas d'enquêtes complexes, les auteurs ont été pragmatiques.

Nous cherchons des procédures qui garantissent :

$$\begin{aligned} E^*(\hat{t}^*_{y\pi}) &= \hat{t}_{y\pi}, \\ V^*(\hat{t}^*_{y\pi}) &\simeq \hat{V}(\hat{t}_{y\pi}). \end{aligned}$$

Autrement dit, on cherche à reproduire sous la loi de rééchantillonnage l'estimateur du total et son estimateur de variance.
On peut montrer que dans le cas d'une fonction lisse de totaux $\theta = f(t_y)$, on reproduit approximativement son estimateur de variance par linéarisation (e.g., Rao et Wu, 1988).

Le bootstrap, comment faire?

La plupart des méthodes de bootstrap proposées pour les enquêtes visent à vérifier ces deux égalités. On trouve plusieurs approches :

1. Appliquer le bootstrap i.i.d., et recaler les poids bootstrap obtenus sur $\hat{t}_{y\pi}$ et $\hat{V}(\hat{t}_{y\pi})$: méthode du rescaled bootstrap (Rao, Wu et Yue, 1992).
2. Trouver une méthode de rééchantillonnage conduisant directement à des poids bootstrap respectant ces deux équations (e.g., Gross, 1980; Sitter, 1992; Antal et Tillé, 2011).
3. Générer directement les poids bootstrap selon une distribution avec des moments appropriés (e.g., Bertail et Combris, 1997; Beaumont et Patak, 2012).

Quelques idées reçues

"Le" bootstrap en Sondages

Quelques idées reçues

"Le" bootstrap en Sondages

Il n'y a pas une mais de nombreuses méthodes différentes (et des variantes).

Les propriétés sont à apprécier au cas par cas.

Quelques idées reçues

"Le" bootstrap en Sondages

Il n'y a pas une mais de nombreuses méthodes différentes (et des variantes).
Les propriétés sont à apprécier au cas par cas.

Le bootstrap permet d'estimer la distribution de l'estimateur

Quelques idées reçues

"Le" bootstrap en Sondages

Il n'y a pas une mais de nombreuses méthodes différentes (et des variantes).
Les propriétés sont à apprécier au cas par cas.

Le bootstrap permet d'estimer la distribution de l'estimateur

Non démontré pour la plupart des méthodes de bootstrap en Sondages.
Ce sont essentiellement des méthodes d'estimation de variance pour des fonctions de totaux.

Quelques idées reçues

"Le" bootstrap en Sondages
Il n'y a pas une mais de nombreuses méthodes différentes (et des variantes).
Les propriétés sont à apprécier au cas par cas.

Le bootstrap permet d'estimer la distribution de l'estimateur
Non démontré pour la plupart des méthodes de bootstrap en Sondages.
Ce sont essentiellement des méthodes d'estimation de variance pour des fonctions de totaux.

Le bootstrap permet d'estimer la variance pour un quantile

Quelques idées reçues

"Le" bootstrap en Sondages
Il n'y a pas une mais de nombreuses méthodes différentes (et des variantes).
Les propriétés sont à apprécier au cas par cas.

Le bootstrap permet d'estimer la distribution de l'estimateur
Non démontré pour la plupart des méthodes de bootstrap en Sondages.
Ce sont essentiellement des méthodes d'estimation de variance pour des fonctions de totaux.

Le bootstrap permet d'estimer la variance pour un quantile
Non démontré pour la plupart des méthodes de bootstrap en Sondages.
Dans le contexte Sondages, démontré par Shao et Chen (1998) pour la méthode de bootstrap avec remise, sous un tirage avec remise des unités + hypothèses de modèle sur la variable d'intérêt.

Bootstrap avec remise des unités primaires

Basé sur Bessonneau, Brilhault, Chauvet et Garcia (2021),
With-replacement bootstrap variance estimation for household surveys: Principles, examples and implementation.
Survey Methodology 47(2), pp. 313-347.

Bootstrap avec remise des unités primaires

Cas particulier de la méthode de Rao, Wu et Yue (1992), mais qui semble la plus utilisée en pratique (Rust and Rao, 1996; Yeo et al., 1999).
Supposons S sélectionné selon un plan à plusieurs degrés, avec sélection d'un échantillon S_I de n_I unités primaires (UP).

La méthode de bootstrap procéde ainsi :

1. On sélectionne un échantillon avec remise et à probabilités égales de $n_I - 1$ unités dans S_I.
2. On donne à l'UP u_i le poids bootstrap de sondage :

$$d_{Ii}^{*} = \frac{n_I}{n_I - 1} \times \{\text{Nb de rééchantillonnages de } u_i\} \times d_{Ii}$$

3. On reproduit la chaîne d'estimation selon les deux principes suivants :
   - Les étapes d'échant./NR post 1er degré ne sont pas bootstrappées.
   - Les étapes d'estim. (correction de la NR, calage) sont bootstrappées.

Bootstrap avec remise des unités primaires

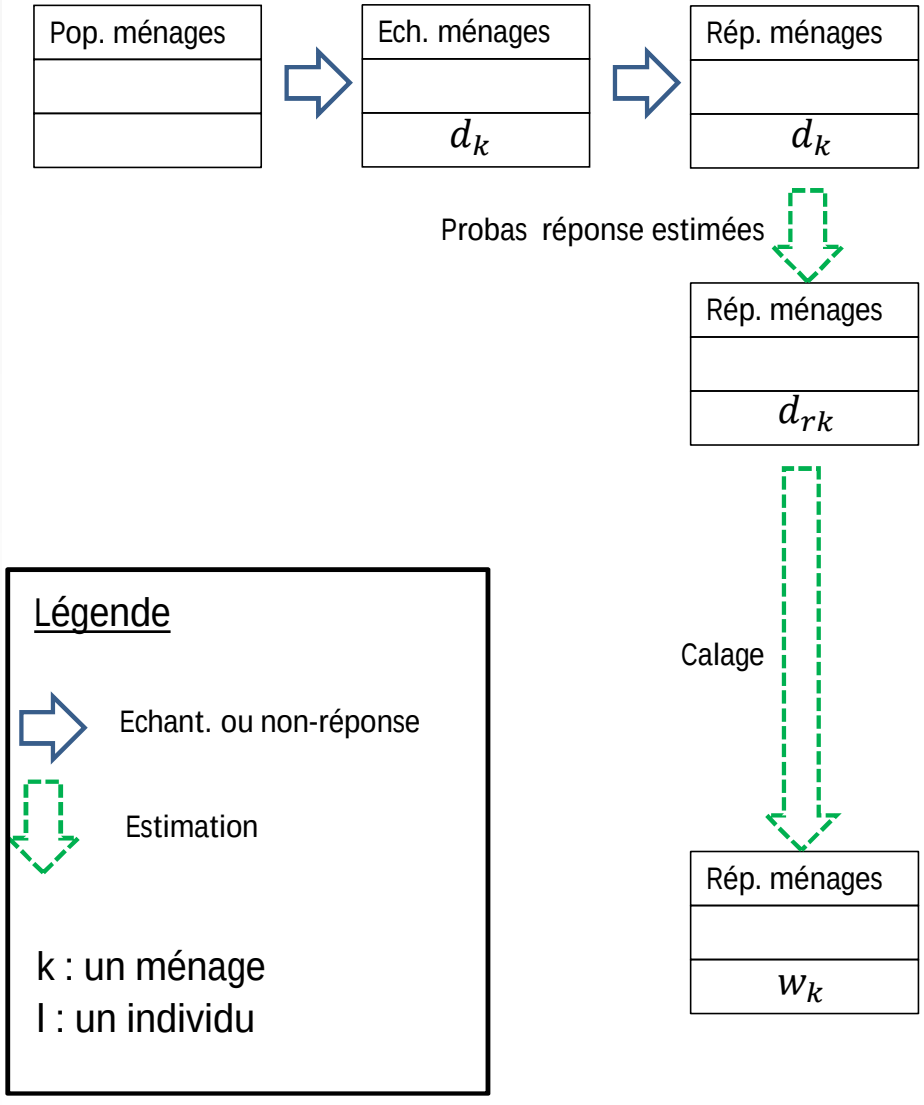
Cette méthode donne un estimateur sans biais de la variance si les UP sont sélectionnées avec remise: cf Bessonneau et al. (2021), où nous expliquons quel est l'estimateur de variance de référence que l'on cherche à reproduire.

Quand les UP sont sélectionnées sans remise, cette méthode :

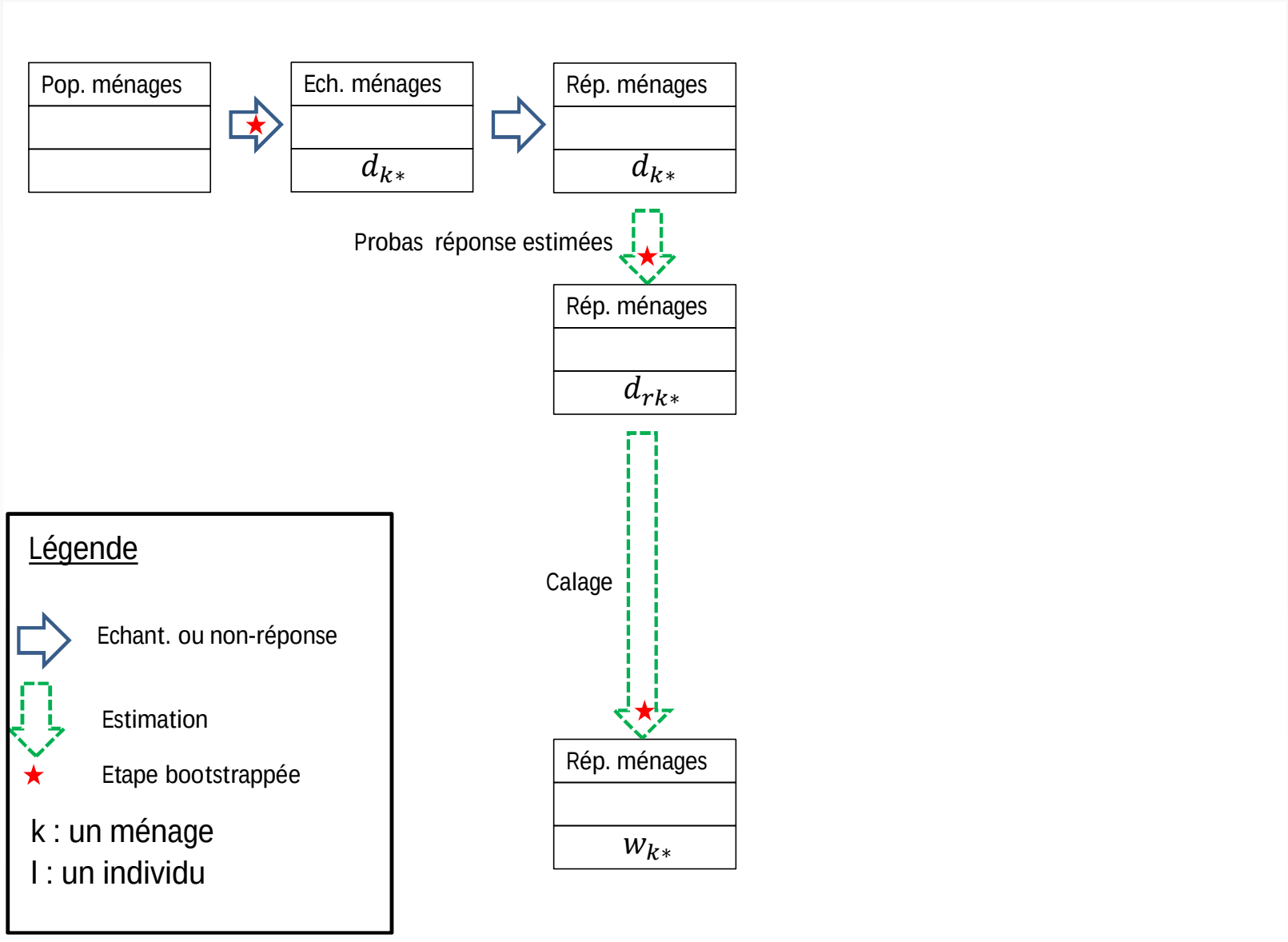
- surestime la variance du premier degré,
- estime correctement la variance due aux étapes suivantes de tirage et de non-réponse.

Pour estimer correctement la variance due au premier degré, une correction des poids bootstrap est possible (e.g., Kleim et Bélanger, 2007).
Mais cette correction n'est possible que pour un SRS-WOR stratifié au premier degré, et elle conduit à sous-estimer la variance due aux étapes suivantes de tirage.

Echantillon de ménages



Etape de bootstrap sur les ménages



Echantillon de ménages : exemple

$d_A = d_B = d_C = d_D = 4 \qquad d_E = d_F = d_G = d_H = d_I = d_J = 16$

$\hat{p}_1 = \frac{3}{4}$ $\qquad$ $\hat{p}_2 = \frac{4}{6}$

$d_{rA} = \frac{16}{3} \qquad d_{rD} = 6 \qquad d_{rE} = d_{rH} = d_{rI} = 24 \qquad d_{rF} = d_{rJ} = \frac{64}{3}$

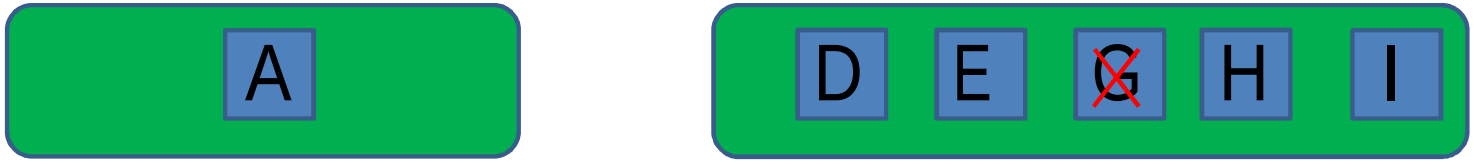
$x_A = (1,1) \quad x_D = (1,0) \quad x_E = (1,1) \quad x_F = (1,0) \quad x_H = (1,1) \quad x_I = (1,0) \quad x_J = (1,1)$

$X_{men} = (100\ ,\ 60) \qquad \hat{X}_{r,men} = (126\ ,\ 74.67)$

$w_A = 4.29 \quad w_D = 4.68 \quad w_E = w_H = 19.29 \quad w_F = 16.62 \quad w_I = 18.70 \quad w_J = 17.14$

Etape de bootstrap sur les ménages : exemple

$d_{A*} = \frac{40}{3}$ $d_{D*} = \frac{40}{9}$ $d_{E*} = d_{H*} = d_{I*} = \frac{160}{9}$ $d_{G*} = \frac{320}{9}$



$\hat{p}_{1*} = 1$

$\hat{p}_{2*} = \frac{4}{5}$

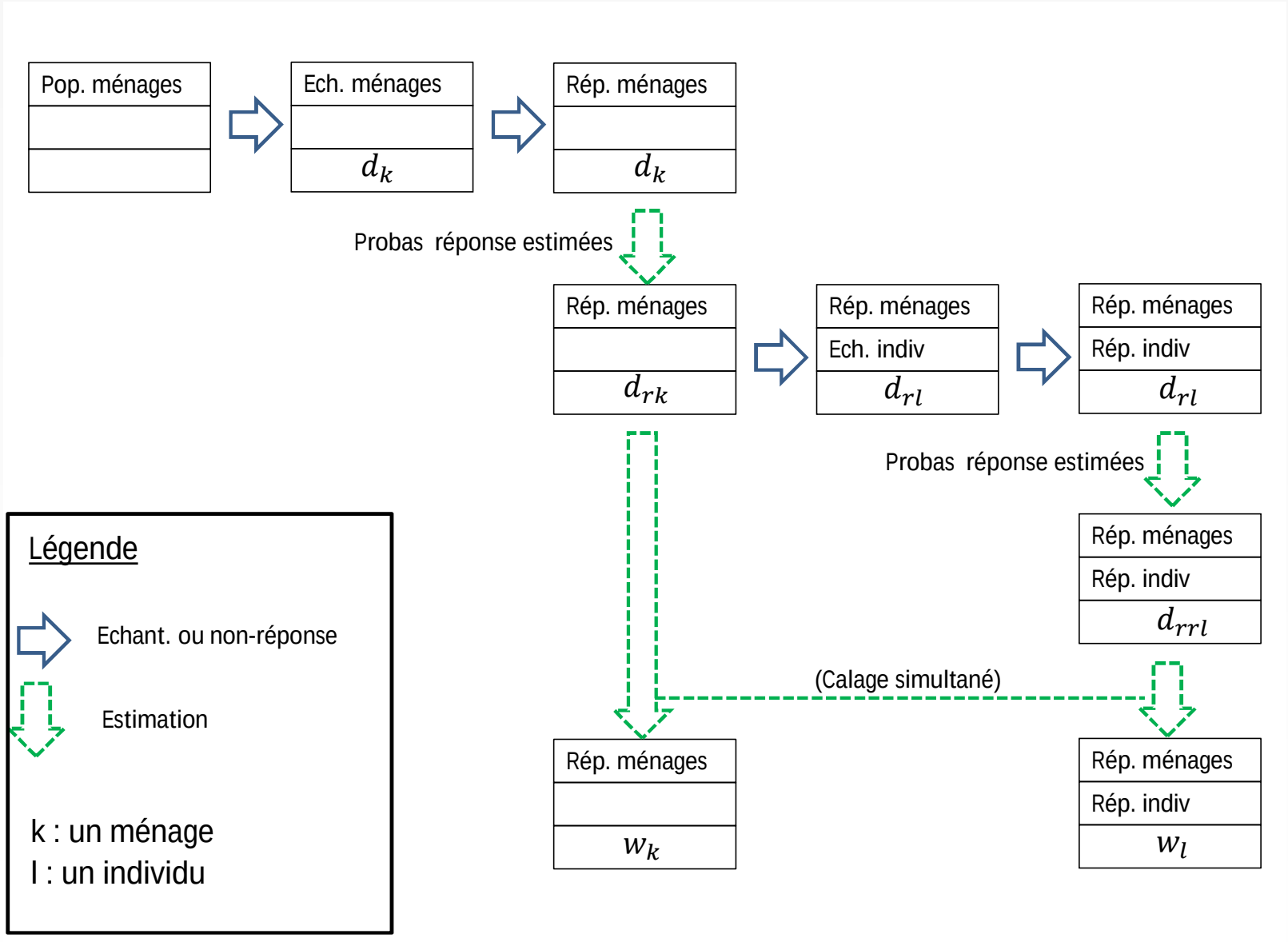
$d_{rA*} = \frac{40}{3}$ $d_{rD*} = \frac{50}{9}$ $d_{rE*} = d_{rH*} = d_{rI*} = \frac{200}{9}$

$x_A = (1,1)$ $x_D = (1,0)$ $x_E = (1,1)$ $x_H = (1,1)$ $x_I = (1,0)$

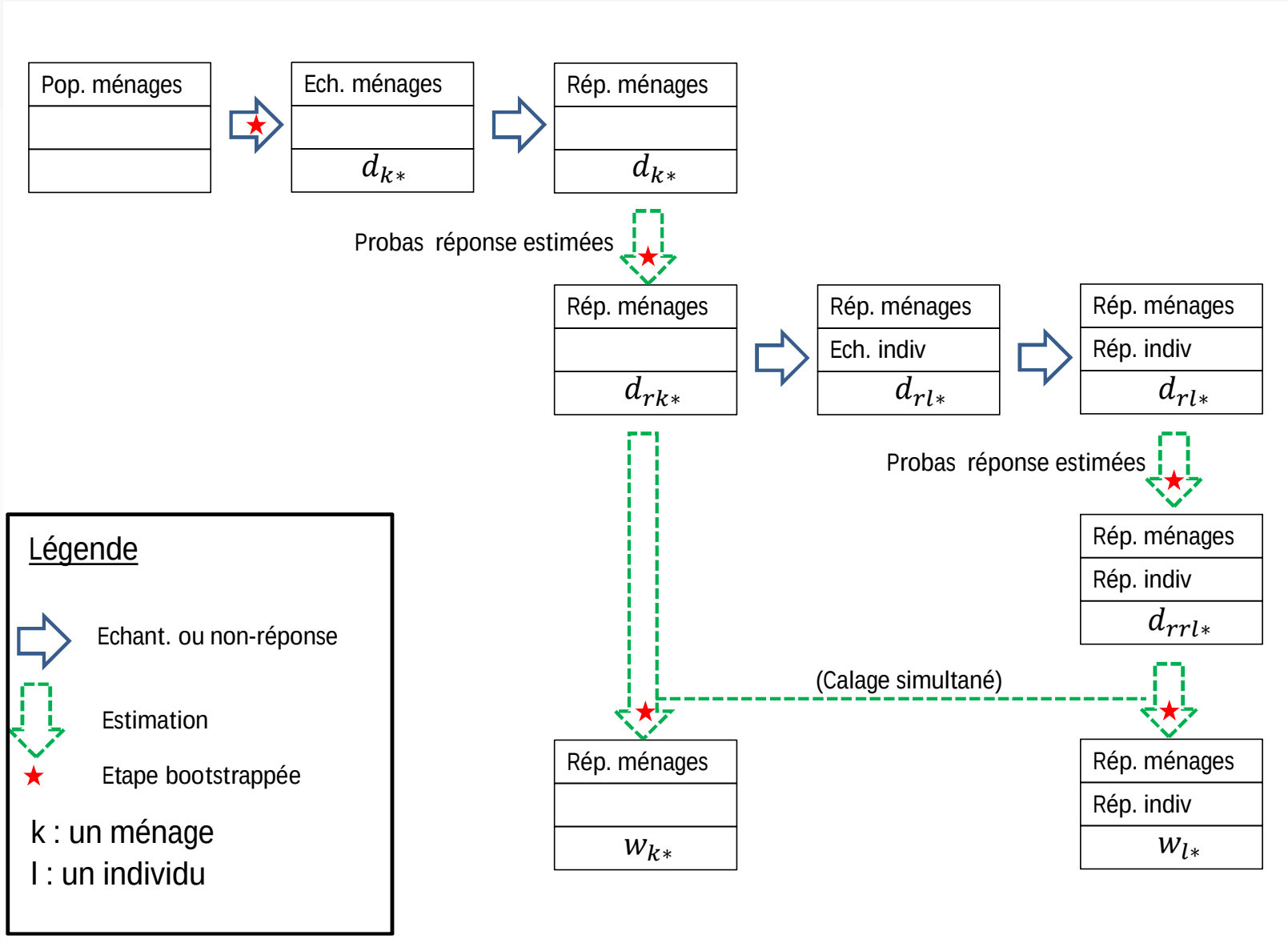
$X_{men} = (100\ ,\ 60)$ $\hat{X}_{r,men*}$=(85.56 , 57.78)

$w_{A*} = 13.85$ $w_{D*} = 8.00$ $w_{E*} = w_{H*} = 23.08$ $w_{I*} = 32.00$

Echantillon d'individus



Etape de bootstrap sur les individus



Travaux en cours / travaux futurs

Travaux en cours / travaux futurs

Des programmes SAS et R paramétrés sont disponibles sur demande, permettant d'appliquer la méthode. Appliqué par le LISER (merci Maria) sur l'enquête sur le racisme et la discrimination au Luxembourg. Résultats très proches des estimations de variance analytiques :)

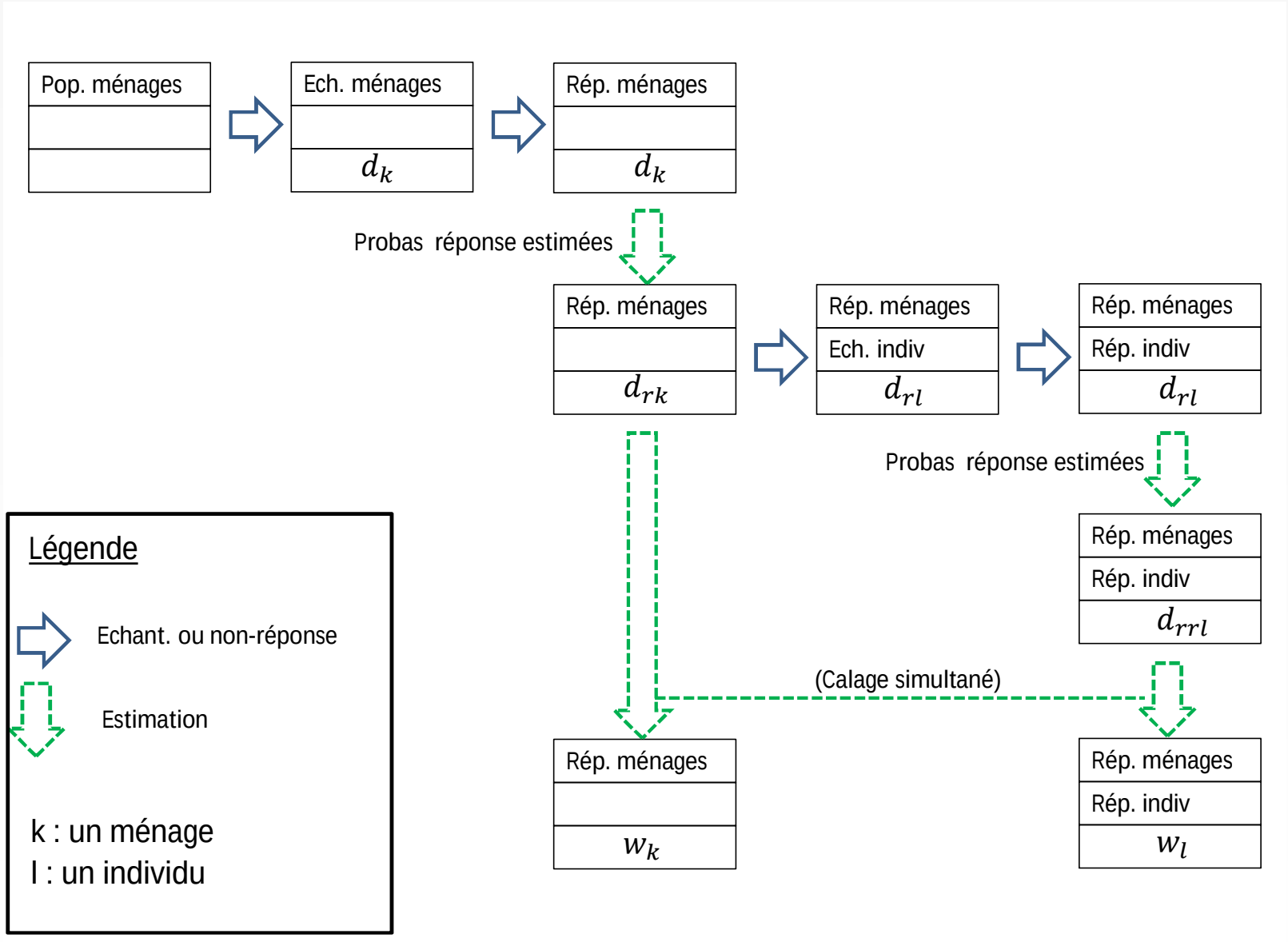
Travail en cours avec la division Sondages pour l'enquête Histoire de Vie et Patrimoine (enquête répétée avec échantillons rotatifs, partage des poids, calage sur sources multiples ... : miam miam).

Application dans le cas d'un plan produit, type celui utilisé par l'Enquête Longitudinale Française depuis l'Enfance (ELFE) : cf présentation de Jean Rubin.

Merci de votre attention :)

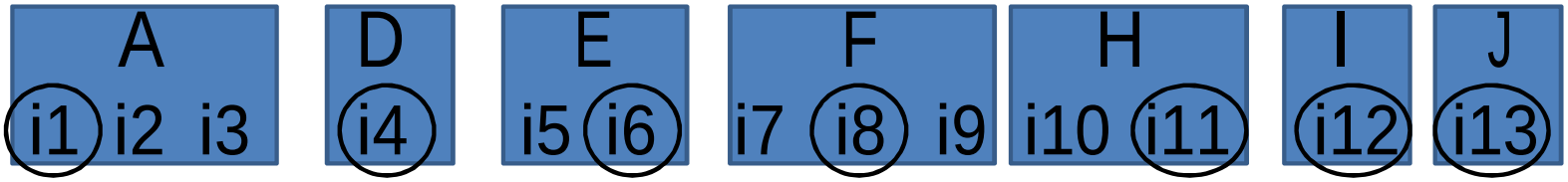
Exemple pour un échantillon d'individus

Echantillon d'individus



Echantillon d'individus : exemple

$d_{rA} = {}^{16}/_{3}$ $d_{rD} = 6$ $d_{rE} = d_{rH} = d_{rI} = 24$ $d_{rF} = d_{rJ} = {}^{64}/_{3}$



$d_{r1} = 16$ $d_{r4} = 6$ $d_{r6} = 48$ $d_{r8} = 64$ $d_{r11} = 48$ $d_{r12} = 24$ $d_{r13} = {}^{64}/_{3}$

$\hat{p}_1 = {}^{3}/_{4}$ $\hat{p}_2 = {}^{1}/_{3}$

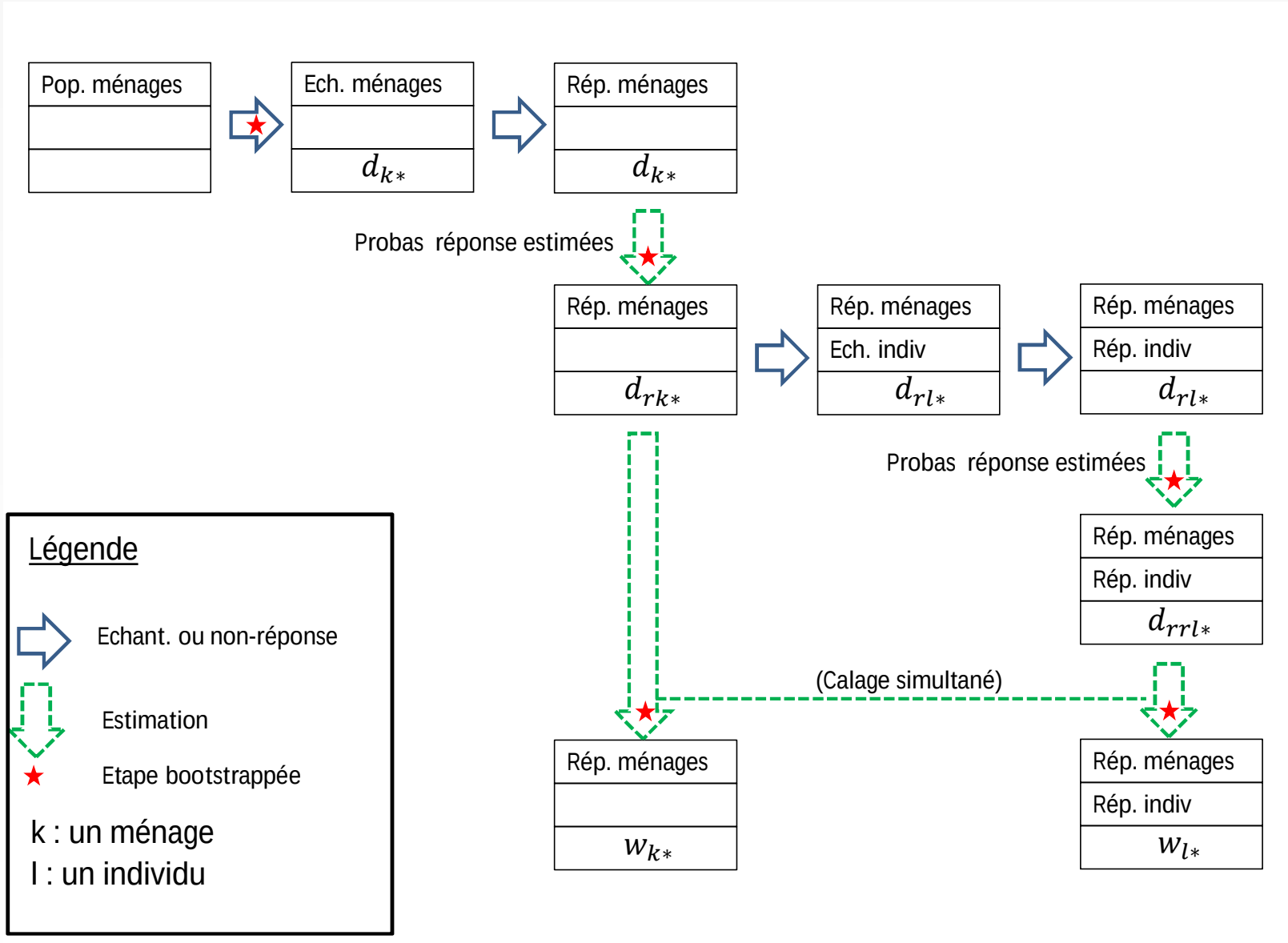
$d_{rr1} = 21.33$ $d_{rr6} = 64$ $d_{rr8} = 85.33$ $d_{rr12} = 72$

$z_1 = (1,3)$ $z_6 = (1,2)$ $z_8 = (1,3)$ $z_{12} = (1,1)$

$Z_{ind} = (200\ ,\ 450)$ $\hat{Z}_{rr,ind} = (242.7\ ,\ 520.0)$

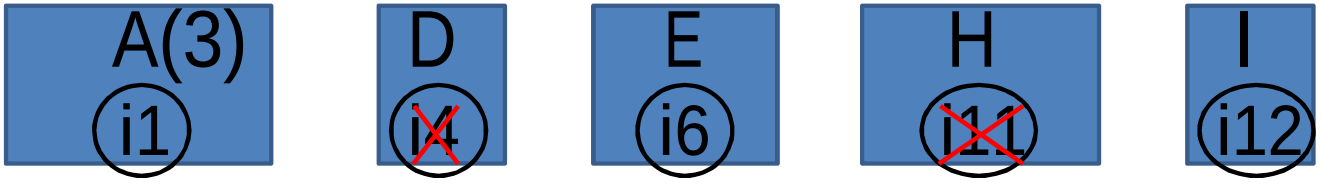
$w_1 = 19.84$ $w_6 = 51.62$ $w_8 = 79.35$ $w_{13} = 49.19$

Etape de bootstrap sur les individus



Etape de bootstrap sur les individus : exemple

d_{rA*} = 40/3 d_{rD*} = 50/9 d_{rE*} = 200/9 d_{rH*} = 200/9 d_{rI*} = 200/9



d_{r1*} = 40 d_{r4*} = 50/9 d_{r6*} = 400/9 d_{r11*} = 200/3 d_{r12*} = 200/9

$\hat{p}_{1*}$ = 2/3 $\hat{p}_{2*}$ = 1/2

d_{rr1*} = 60 d_{rr6*} = 66.67 d_{rr12*} = 44.44

z_1 = (1,3) z_6 = (1,2) z_{12} = (1,1)

Z_{ind}=(200 , 450) $\hat{Z}_{rr,ind*}$=(171.1, 357.8)

w_{1*} = 86.87 w_{6*} = 76.05 w_{12*} = 36.97