ESTIMATEURS DE VARIANCE ISSUS D'UN PLAN PRODUIT pour l'enquête Elfe

Hélène JUILLARD [(1)], Guillaume CHAUVET [(2)] et Anne RUIZ-GAZEN [(3)]

[(1)]Ined [(2)]Crest/Ensai [(3)]TSE

Journées de méthodologie statistique - Insee - avril 2015

Plan

Plan

Elfe

Etude Longitudinale Française depuis l'Enfance : cohorte de 18 000 enfants

Cohorte démarrée en 2011, organisée par une unité Ined-Inserm-EFS.

POPULATION
Nourrissons nés en 2011 dans l'une des 544 maternités métropolitaines.

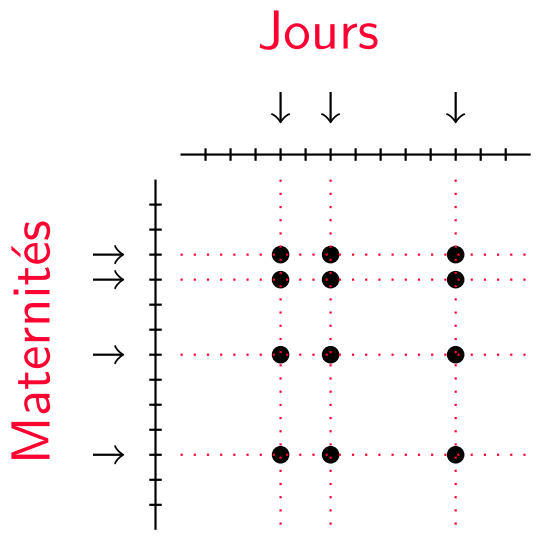
THEMES
Leur **santé**, leur **alimentation**, leur **lieu** d'habitation, leur **scolarité** ainsi que leur **vie familiale** et **sociale**.

Le premier temps d'enquête s'est déroulé en maternité quelques jours après la naissance de l'enfant.

Plan

Echantillonnage produit

Croisement indépendant de deux échantillonnages



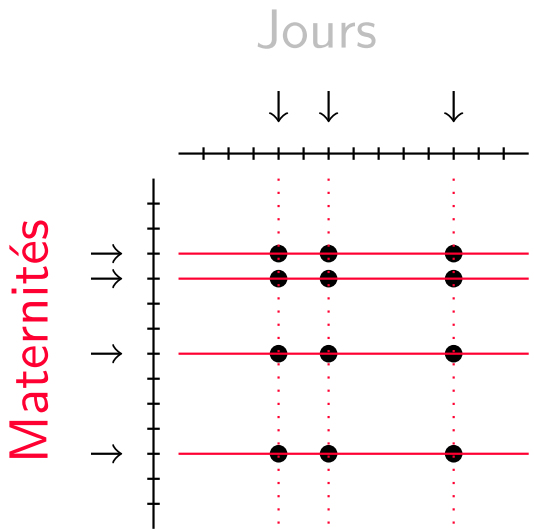
PLAN DE SONDAGE
Croisement indépendant de 2 échantillons

échantillon de jours
$\times$
échantillon de maternités

tous les nourrissons nés dans ces maternités durant ces jours

Echantillonnage produit

Croisement indépendant de deux échantillonnages



MATERNITES i dans U_M

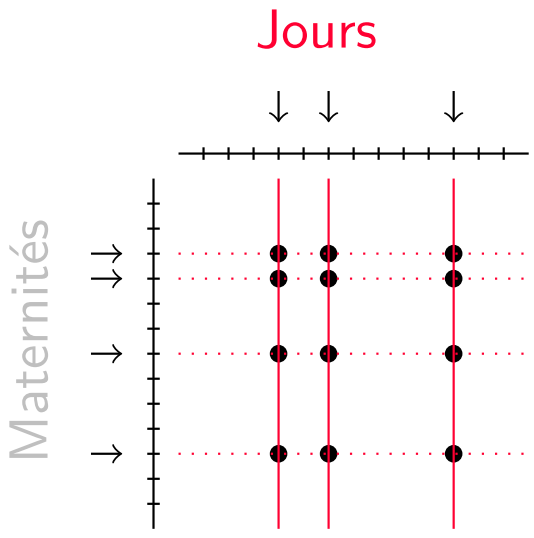
Plan stratifié (selon la taille des maternités) STSI

Strates $g = 1$ à 5
Echantillons S_{Mg}

$$Pr(i \in S_{Mg}) = \frac{n_{Mg}}{N_{Mg}}$$

Echantillonnage produit

Croisement indépendant de deux échantillonnages



JOURS k dans U_D

Modélisation par un plan stratifié (selon la saison) STSI

Strates $h = 1$ à 4
Echantillons S_{Dh}

$$Pr(k \in S_{Dh}) = \frac{n_{Dh}}{N_{Dh}}$$

Echantillonnage produit

Estimation d'un total : le cas STSI $\times$ STSI

La variable d'étude Y prend la valeur Y_{ik} pour $i \in U_M$ et $k \in U_D$.

Exemple : Y_{ik} = le nombre de naissances par césarienne dans la maternité i le jour k.

Echantillonnage produit

Estimation d'un total : le cas STSI $\times$ STSI

La variable d'étude Y prend la valeur Y_{ik} pour $i \in U_M$ et $k \in U_D$.

Le total $t_Y = \sum_{i \in U_M} \sum_{k \in U_D} Y_{ik}$ est estimé sans biais par

$$\begin{aligned}
\hat{t}_Y &= \sum_{g=1}^{G} \frac{N_{Mg}}{n_{Mg}} \sum_{h=1}^{H} \frac{N_{Dh}}{n_{Dh}} \sum_{i \in S_{Mg}} \sum_{k \in S_{Dh}} Y_{ik} \\
&= \sum_{g=1}^{G} \frac{N_{Mg}}{n_{Mg}} \sum_{i \in S_{Mg}} \hat{Y}_{i\bullet} \\
&= \sum_{h=1}^{H} \frac{N_{Dh}}{n_{Dh}} \sum_{k \in S_{Dh}} \hat{Y}_{\bullet k}.
\end{aligned}$$

Echantillonnage produit

Estimation d'un total : le cas STSI $\times$ STSI

La variable d'étude Y prend la valeur Y_{ik} pour $i \in U_M$ et $k \in U_D$.

Le total $t_Y = \sum_{i \in U_M} \sum_{k \in U_D} Y_{ik}$ est estimé sans biais par

$$\hat{t}_Y = \sum_{g=1}^{G} \frac{N_{Mg}}{n_{Mg}} \sum_{h=1}^{H} \frac{N_{Dh}}{n_{Dh}} \sum_{i \in S_{Mg}} \sum_{k \in S_{Dh}} Y_{ik}.$$

Un estimateur sans biais de la variance de $\hat{t}_Y$ peut se décomposer ainsi :

$$\hat{\mathbf{V}}_{\mathrm{HT}}\left(\hat{t}_Y\right) = \hat{\mathbf{V}}_D\left(\hat{t}_Y\right) + \hat{\mathbf{V}}_M\left(\hat{t}_Y\right) - \hat{\mathbf{V}}_E\left(\hat{t}_Y\right) \qquad (1)$$

où $\hat{\mathbf{V}}_D\left(\hat{t}_Y\right)$ inter-jours, $\hat{\mathbf{V}}_M\left(\hat{t}_Y\right)$ inter-maternités et $\hat{\mathbf{V}}_E\left(\hat{t}_Y\right)$ résiduel.

Echantillonnage produit

Estimation de variance : le cas STSI $\times$ STSI

$$\hat{\mathbf{V}}_{\mathrm{HT}}\left(\hat{t}_Y\right) = \hat{\mathbf{V}}_D\left(\hat{t}_Y\right) + \hat{\mathbf{V}}_M\left(\hat{t}_Y\right) - \hat{\mathbf{V}}_E\left(\hat{t}_Y\right)$$

où

$$\hat{\mathbf{V}}_D\left(\hat{t}_Y\right) = \sum_{h=1}^{H} N_{Dh}^2 \left(\frac{1}{n_{Dh}} - \frac{1}{N_{Dh}}\right) \frac{1}{n_{Dh}-1} \sum_{k \in S_{Dh}} (\hat{Y}_{\bullet k} - \frac{1}{n_{Dh}} \sum_{l \in S_{Dh}} \hat{Y}_{\bullet l})^2$$

inter-jours

$$\hat{\mathbf{V}}_M\left(\hat{t}_Y\right) = \sum_{g=1}^{G} N_{Mg}^2 \left(\frac{1}{n_{Mg}} - \frac{1}{N_{Mg}}\right) \frac{1}{n_{Mg}-1} \sum_{i \in S_{Mg}} (\hat{Y}_{i\bullet} - \frac{1}{n_{Mg}} \sum_{j \in S_{Mg}} \hat{Y}_{j\bullet})^2$$

inter-maternités

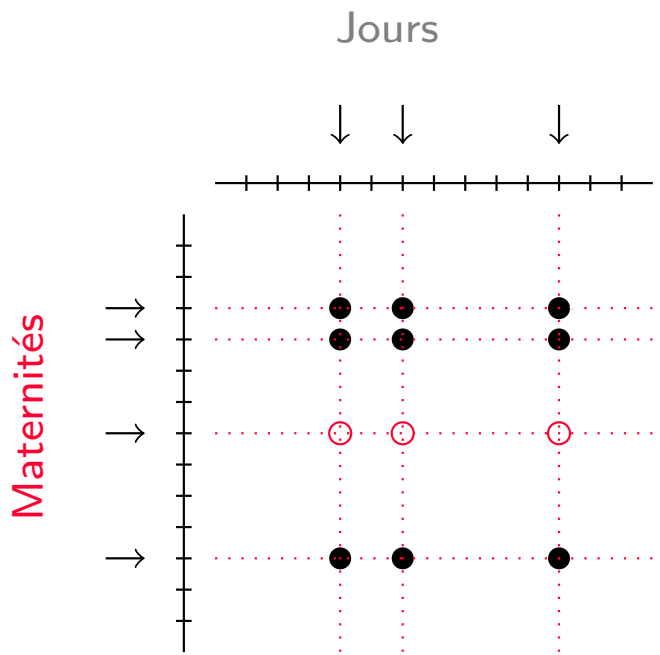
$$\hat{\mathbf{V}}_E\left(\hat{t}_Y\right) = \sum_{g=1}^{G} N_{Mg}^2 \left(\frac{1}{n_{Mg}} - \frac{1}{N_{Mg}}\right) \sum_{h=1}^{H} N_{Dh}^2 \left(\frac{1}{n_{Dh}} - \frac{1}{N_{Dh}}\right) \frac{1}{(n_{Mg}-1)(n_{Dh}-1)}$$
$$\sum_{i \in S_{Mg}} \sum_{k \in S_{Dh}} (Y_{ik} - \bar{Y}_{\bullet k,g} - \bar{Y}_{i\bullet,h} + \bar{Y}_{\bullet\bullet,gh})^2$$

résiduel

Plan

Non-réponse

Maternités, jours et nourrissons



MATERNITES

320 maternités participantes parmi les 349 sélectionnées

7 % de non-réponse

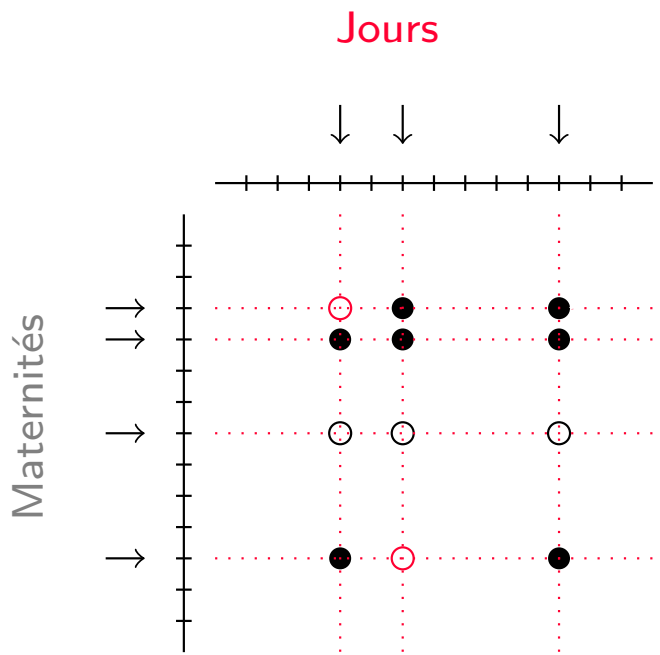
JOURS

7741 maternités × jours d'enquête parmi les 8000 attendus

3 % de non-réponse

Non-réponse

Maternités, jours et nourrissons



MATERNITES

320 maternités participantes parmi les 349 sélectionnées

7 % de non-réponse

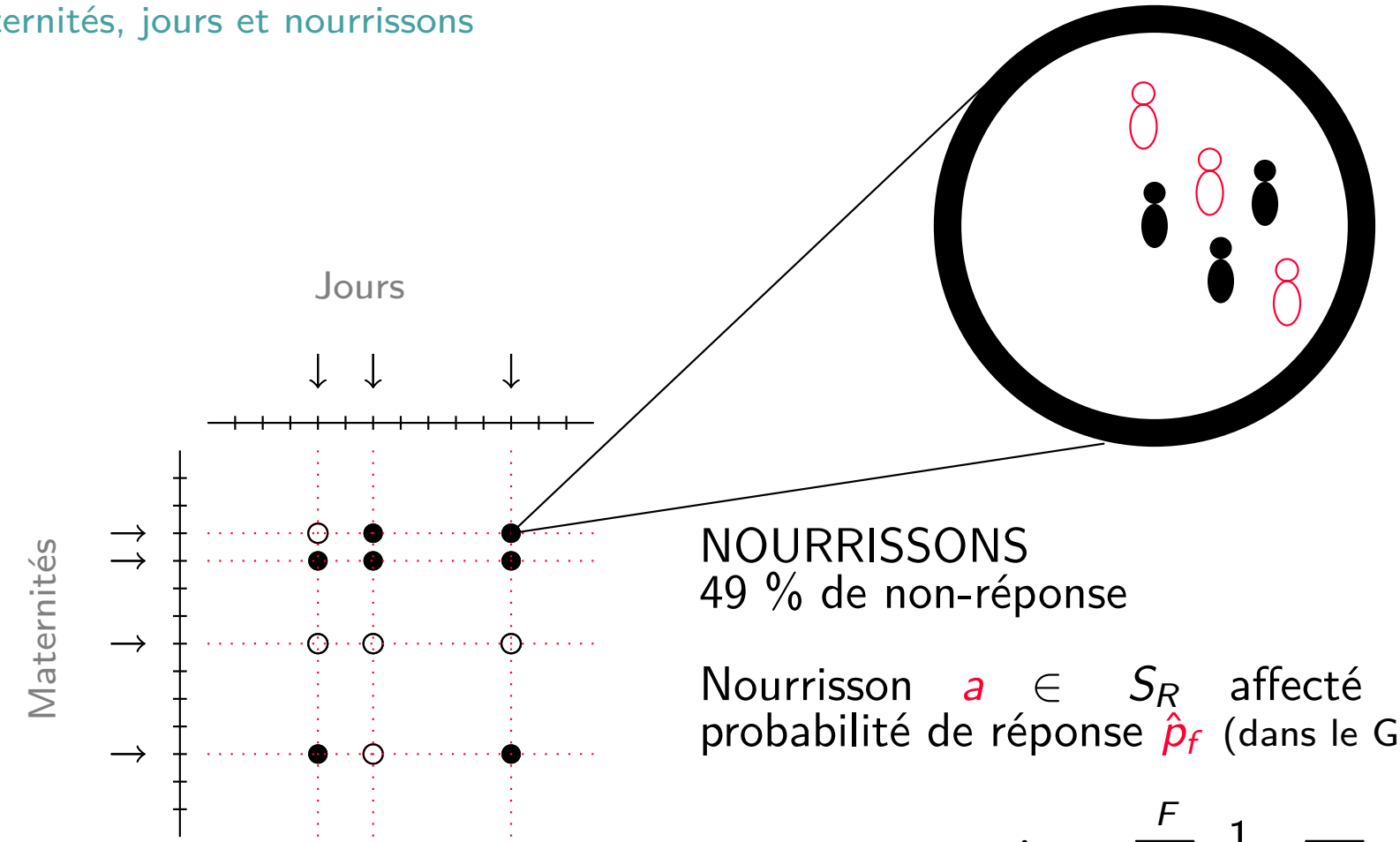
JOURS

7741 maternités $\times$ jours d'enquête parmi les 8000 attendus

3 % de non-réponse

Non-réponse

Maternités, jours et nourrissons



NOURRISSONS
49 % de non-réponse

Nourrisson $a \in S_R$ affecté d'une probabilité de réponse $\hat{p}_f$ (dans le GRH f)

$$Y_{ik} \text{ estimé par } \hat{Y}_{ik} = \sum_{f=1}^{F} \frac{1}{\hat{p}_f} \sum_{a \in S_{R_{ik_f}}} y_a$$

$\hat{t}_Y$ estimé par $\hat{t}_{Y\star}$

Non-réponse

Estimation de la variance : le cas STSI $\times$ STSI avec prise en compte de la non-réponse au niveau nourrisson

$$\hat{\mathbf{V}}\left(\hat{t}_{Y\star}\right) = \hat{\mathbf{V}}_{\text{ech}}^{\text{NR}}\left(\hat{t}_{Y\star}\right) + \hat{\mathbf{V}}_{\text{NR}}\left(\hat{t}_{Y\star}\right) \quad (2)$$

où

$$\hat{\mathbf{V}}_{\text{ech}}^{\text{NR}}\left(\hat{t}_{Y\star}\right) = \left[\hat{\mathbf{V}}_{D}^{\text{NR}}\left(\hat{t}_{Y\star}\right) + \hat{\mathbf{V}}_{M}^{\text{NR}}\left(\hat{t}_{Y\star}\right) - \hat{\mathbf{V}}_{E}^{\text{NR}}\left(\hat{t}_{Y\star}\right)\right] - \hat{\mathbf{V}}_{\text{echC}}^{\text{NR}}\left(\hat{t}_{Y\star}\right)$$

estime la variance d'échantillonnage

$$\hat{\mathbf{V}}_{\text{NR}}\left(\hat{t}_{Y\star}\right) = \sum_{g=1}^{G} \frac{N_{Mg}^2}{n_{Mg}^2} \sum_{h=1}^{H} \frac{N_{Dh}^2}{n_{Dh}^2} \sum_{f=1}^{F} \frac{1-\hat{p}_f}{\hat{p}_f^2} \sum_{a \in S_{Rf,gh}} \left(y_a - \frac{1}{n_{Rf,gh}} \sum_{b \in S_{Rf,gh}} y_b \right)^2$$

estime la variance de non-réponse

Plan

Estimateurs simplifiés

Procédures logicielles déjà existantes

$$\hat{\mathbf{V}}\left(\hat{t}_{Y\star}\right) = \left[\hat{\mathbf{V}}_D^{\mathrm{NR}}\left(\hat{t}_{Y\star}\right) + \hat{\mathbf{V}}_M^{\mathrm{NR}}\left(\hat{t}_{Y\star}\right) - \hat{\mathbf{V}}_E^{\mathrm{NR}}\left(\hat{t}_{Y\star}\right)\right] - \left[\hat{\mathbf{V}}_{\mathrm{echC}}^{\mathrm{NR}}\left(\hat{t}_{Y\star}\right) - \hat{\mathbf{V}}_{\mathrm{NR}}\left(\hat{t}_{Y\star}\right)\right]$$

- l'estimateur sans biais n'est a priori programmé dans aucun logiciel ;
- l'estimateur sans biais peut prendre des valeurs négatives

$$\hat{\mathbf{V}}_{\mathrm{SIMP1}} \equiv \hat{\mathbf{V}}_M^{\mathrm{NR}}\left(\hat{t}_{Y\star}\right) = \sum_{g=1}^{G} N_{Mg}^2 \left(\frac{1}{n_{Mg}} - \frac{1}{N_{Mg}}\right) s^2_{\hat{\hat{Y}}_{\circ\bullet,g}}$$

inter-maternités

$$\hat{\mathbf{V}}_{\mathrm{SIMP2}} \equiv \hat{\mathbf{V}}_D^{\mathrm{NR}}\left(\hat{t}_{Y\star}\right) = \sum_{h=1}^{H} N_{Dh}^2 \left(\frac{1}{n_{Dh}} - \frac{1}{N_{Dh}}\right) s^2_{\hat{\hat{Y}}_{\bullet\circ,h}}$$

inter-jours

$$\hat{\mathbf{V}}_{\mathrm{SIMP3}} \equiv \hat{\mathbf{V}}_M^{\mathrm{NR}}\left(\hat{t}_{Y\star}\right) + \hat{\mathbf{V}}_D^{\mathrm{NR}}\left(\hat{t}_{Y\star}\right)$$

$\longrightarrow$ SAS / R / Stata

Estimateurs simplifiés

Illustration sur données Elfe

Modélisation STSI × STSI, NR	Nombre de naissances	Nombre de nourrissons nés sous césarienne	Nombre de nourrissons ayant une mère suivie par sage-femme
$\hat{t}_{Y\star}$	753342	73644	97775
$\hat{\mathbf{V}}\left(\hat{t}_{Y\star}\right)$	$2.9\ 10^8$	$7.1\ 10^7$	$1.9\ 10^7$
$\hat{\mathbf{V}}_{\text{SIMP1}}\left(\hat{t}_{Y\star}\right)$ (ER)	$8.9\ 10^7$ (-69 %)	$3.7\ 10^6$ (-95 %)	$1.4\ 10^7$ (-29 %)
$\hat{\mathbf{V}}_{\text{SIMP2}}\left(\hat{t}_{Y\star}\right)$ (ER)	$2.4\ 10^8$ (-16 %)	$7.0\ 10^7$ (-02 %)	$8.7\ 10^6$ (-55 %)
$\hat{\mathbf{V}}_{\text{SIMP3}}\left(\hat{t}_{Y\star}\right)$ (ER)	$3.3\ 10^8$ (**14** %)	$7.3\ 10^7$ (**3.1** %)	$2.2\ 10^7$ (**15** %)

$$ER = \frac{\hat{\mathbf{V}}_{\text{SIMP}}\left(\hat{t}_{Y\star}\right) - \hat{\mathbf{V}}\left(\hat{t}_{Y\star}\right)}{\hat{\mathbf{V}}\left(\hat{t}_{Y\star}\right)}$$

Conclusion

- Estimateur de variance issu d'un plan produit
- Estimateur de variance avec non-réponse
- Estimateur simplifié pour l'enquête Elfe

A venir :

- Calage
- Longitudinal

Références

Kim, J. K. and Kim, J. J. (2007). Nonresponse Weighting Adjustment Using Estimated Response Probability. The Canadian Journal of Statistics, 35, 501-514.

Ohlsson, E. (1996). Cross-Classified Sampling. Journal of Official Statistics, 12, No.3, 241-251.

R Core Team (2012). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Särndal, C.-E., Swensson, B. and Wretman, J.H. (1992). Model Assisted Survey Sampling. Springer-Verlag.

SAS Institute (2010). *SAS/STAT® 9.22 User's Guide*. Cary: SAS Institute.

StataCorp. 2013. *Stata: Release 13*. Statistical Software. College Station, TX: StataCorp LP.

Vos, J. W. E. (1964). Sampling in space and time. Review of the International Statistical Institute, 32, No.3, 226-241.