Calage sur bornes minimales et choix des bornes de calage.

Neuvième colloque francophone sur les Sondages

Emmanuel Gros – Antoine Rebecq

14 octobre 2016

INSEE - DMCSI - Division Sondages

Sommaire

Introduction

Notations

- Soit U une population finie de taille N. On s'intéresse à diverses variables Y_1,...,Y_p dont on souhaite estimer les totaux sur U :

$$T_{Y_j} = \sum_{i \in U} y_{j,i}$$

- s : échantillon de taille n tiré dans U, selon un plan de sondage $p(s)$ conduisant à des probabilités d'inclusion π_i. L'estimateur d'Horvitz-Thompson s'écrit :

$$\hat{T}_{Y\pi} = \sum_{i \in s} \frac{y_i}{\pi_i}$$

L'estimateur par calage

- Soit $\mathbf{X} = (X_1, ..., X_J)'$ un vecteur de J variables auxiliaires dont on connaît les totaux sur la population. On préfère souvent l'estimateur par calage (Deville et Särndal (1992)) :

$$\hat{T}_{Yc} = \sum_{i \in s} w_i y_i$$

où les w_i minimisent la « distance » aux $\frac{1}{\pi_i}$ tout en vérifiant les J équations de calage :

$$\sum_{i \in s} w_i \mathbf{x}_i = T_{\mathbf{X}}$$

Calage sur marges – formalisation du problème

- On définit une fonction G telle que $G(\frac{w_i}{d_i})$ mesure la « distance » entre w_i et d_i :
  - $G(1) = 0$
  - G est positive et convexe : $G(\frac{w_i}{d_i})$ est d'autant plus élevée que le rapport $\frac{w_i}{d_i}$ est éloigné de 1

- Les poids calés w_i sont alors solutions du problème d'optimisation suivant :

$$\begin{cases} \min\limits_{w_i} \sum\limits_{i \in s} d_i G(\frac{w_i}{d_i}) \\ \text{s.c. } \sum_{i \in s} w_i \mathbf{x}_i = T_{\mathbf{x}} \end{cases}$$

Calage sur marges – résolution

- Solution :

$$w_i = d_i F(\mathbf{x}_i' \hat{\lambda})$$

où est appelée fonction de calage ($F = (G')^{-1}$).

- Le facteur de calage $g_i = \dfrac{w_i}{d_i} = F(\mathbf{x}_i' \hat{\lambda})$ dépend (entre autres) du choix de la distance G.

Méthodes de calage

- Méthode linéaire : $F(u) = 1 + u$
- Méthode exponentielle ou raking ratio : $F(u) = \exp(u) > 0$

L et U sont deux constantes telles que $L < 1 < U$.

- Méthode logit : il s'agit d'une méthode exponentielle bornée : $F(u) = \dfrac{L(U-1) + U(1-L)\exp(Au)}{(U-1) + (1-L)\exp(Au)}$ et alors $L < F(u) < U$

Choix de la méthode de calage

- Pas de critère théorique permettant de déterminer la « bonne » méthode de calage.
- En pratique, on privilégie en général les méthodes bornées :
  - permettent d'éviter les poids négatifs ou inférieurs à 1 ;
  - évitent l'apparition d'unités influentes (notamment estimations sur domaines)
  - Pas de garantie de convergence : pour des L et U choisis trop proches de 1, le calage n'est plus possible.

Calage sur bornes minimales

Calage sur bornes minimales

- Objectif : déterminer les bornes L^* et U^* conduisant au calage le plus « serré » possible, *i.e.* minimisant l'étendue $U^* - L^*$.
- Solution : résolution numérique via un programme d'optimisation.

Problèmatique

- Les poids calés doivent vérifier les contraintes de calage :

$$\widetilde{\mathbf{X}}_s' \mathbf{g} = T_{\mathbf{X}}$$

(où $\mathbf{g}$ est le vecteur des facteurs de calage et $\widetilde{\mathbf{X}}_s = (d_i x_{j,i})_{1 \leq i \leq n, 1 \leq j \leq J}$)

- Les bornes L^* et U^* correspondent au vecteur $\mathbf{g}^*$ qui minimise l'étendue des rapports de poids tout en respectant les contraintes de calage $\widetilde{\mathbf{X}}_s' \mathbf{g}^* = T_{\mathbf{X}}$.

Formalisation du problème

- Cela correspond au programme d'optimisation suivant :

$$\begin{cases} \min\limits_{\mathbf{g}\in\mathbb{R}^n} \left(\max\limits_{i\in[[1,n]]} g_i - \min\limits_{j\in[[1,n]]} g_j \right) \\ \text{s. c. } \widetilde{\mathrm{X}}_s' \mathbf{g} = T_{\mathrm{X}} \end{cases}$$

- Ce programme est un programme linéaire, qui peut donc être résolu numériquement via l'algorithme du simplexe.
- Ce programme est de taille $n(n-1) \times (n+1)$, donc en $\mathcal{O}(n^3) \Rightarrow$ problèmes de mémoire et de temps de calcul.

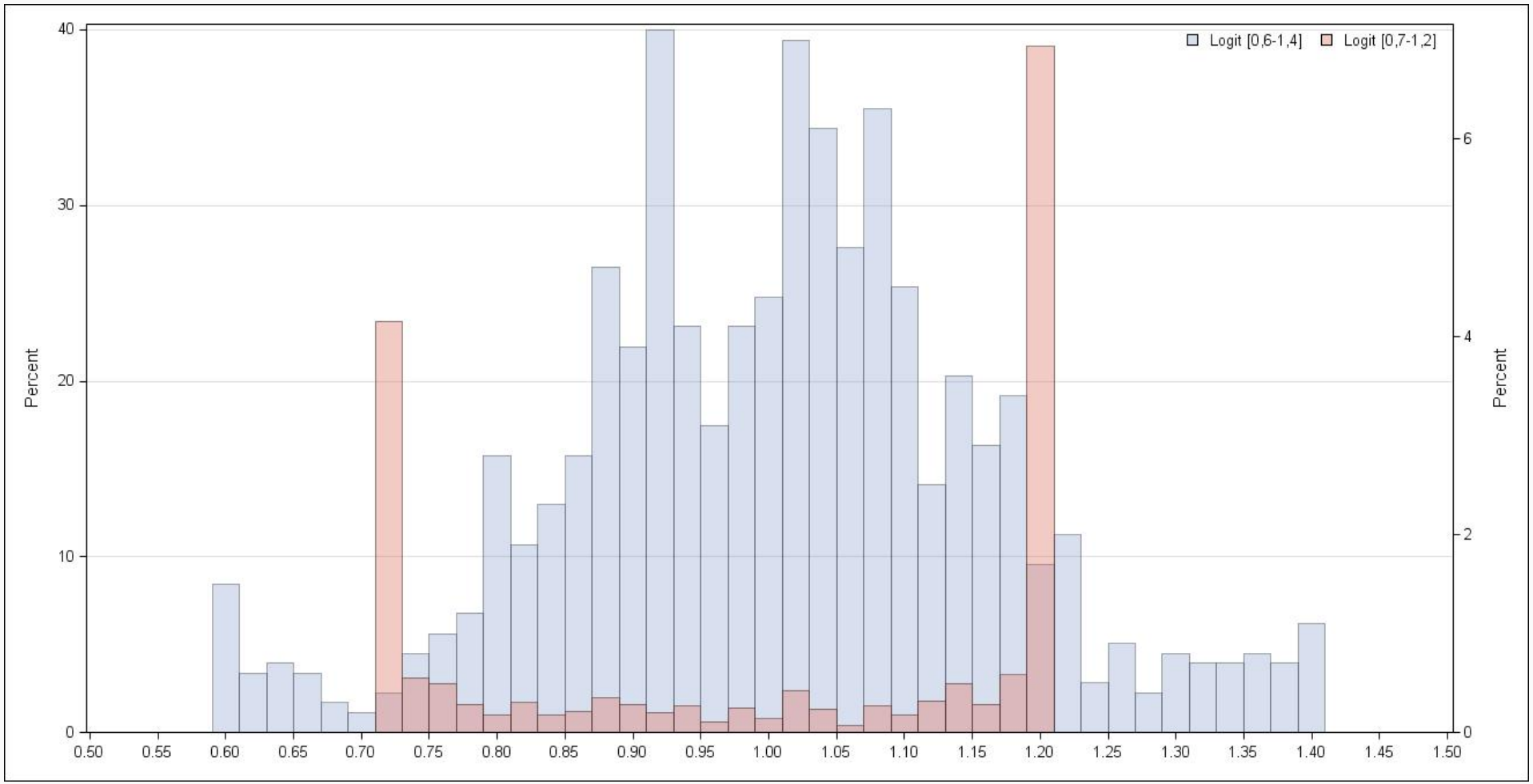
Reformulation du problème en $\mathcal{O}(n^2)$

- On peut montrer que le programme précédent est équivalent au programme suivant :

$$\begin{cases} \min\limits_{\mathbf{g}\in\mathbb{R}^n, a\in\mathbb{R}} \max\limits_{i\in[[1,n]]} |g_i - a| \\ \text{s. c. } \widetilde{\mathbf{X}}_s'\mathbf{g} = T_{\mathbf{X}} \end{cases}$$

- Ce programme est encore un programme linéaire, qui peut donc également être résolu numériquement via l'algorithme du simplexe...
- ...mais de dimension $2n \times (n+2)$, *i.e.* en $\mathcal{O}(n^2) \Rightarrow$ Beaucoup plus efficace.
- implémenté dans la fonction calibration du package R Icarus

Exemple de choix des bornes de calage



Choix des bornes de calage

Problématique

- Pratique à l'Insee : on préconise en général l'utilisation de méthodes de calage bornées, avec des bornes :
  - suffisamment serrées pour éviter les déformations de poids extrêmes...
  - ... mais sans être trop proches des bornes L^* et U^* pour éviter les accumulations trop importantes de rapports de poids à ces bornes.
- Mais il n'y a aucune justification théorique à cette pratique :
  - asymptotiquement, toutes les méthodes sont équivalentes ;
  - à distance finie, on ne dispose pas de formules permettant d'évaluer l'impact du choix des bornes sur les estimateurs.
- Étude par simulations pour essayer de quantifier cet impact.

Cadre de l'étude (1)

On s'appuie sur les données issues d'Esane 2013 dans le commerce hors exhaustif (~ les unités de plus de 20 salariés ou plus de 38 M€ de CA) : les liasses fiscales des entreprises fournissent de très nombreuses variables disponibles sur l'ensemble des unités du champ → variables auxiliaires pour le calage & calcul de biais et d'EQM.

⇒ Procédure retenue :

- on tire K=40 000 échantillons de 1 000 unités par SAS stratifié par APE⊗[tranches de taille] avec allocation proportionnelle ;
- on cale chaque échantillon selon trois scénarios de calage – raking ratio, logit [0,5-2] et logit minimal – sur les marges suivantes :
  - structures par secteur, tranches d'effectif, ZEAT ;
  - totaux de chiffre d'affaires, valeur ajoutée, actif total et passif total.

Cadre de l'étude (2)

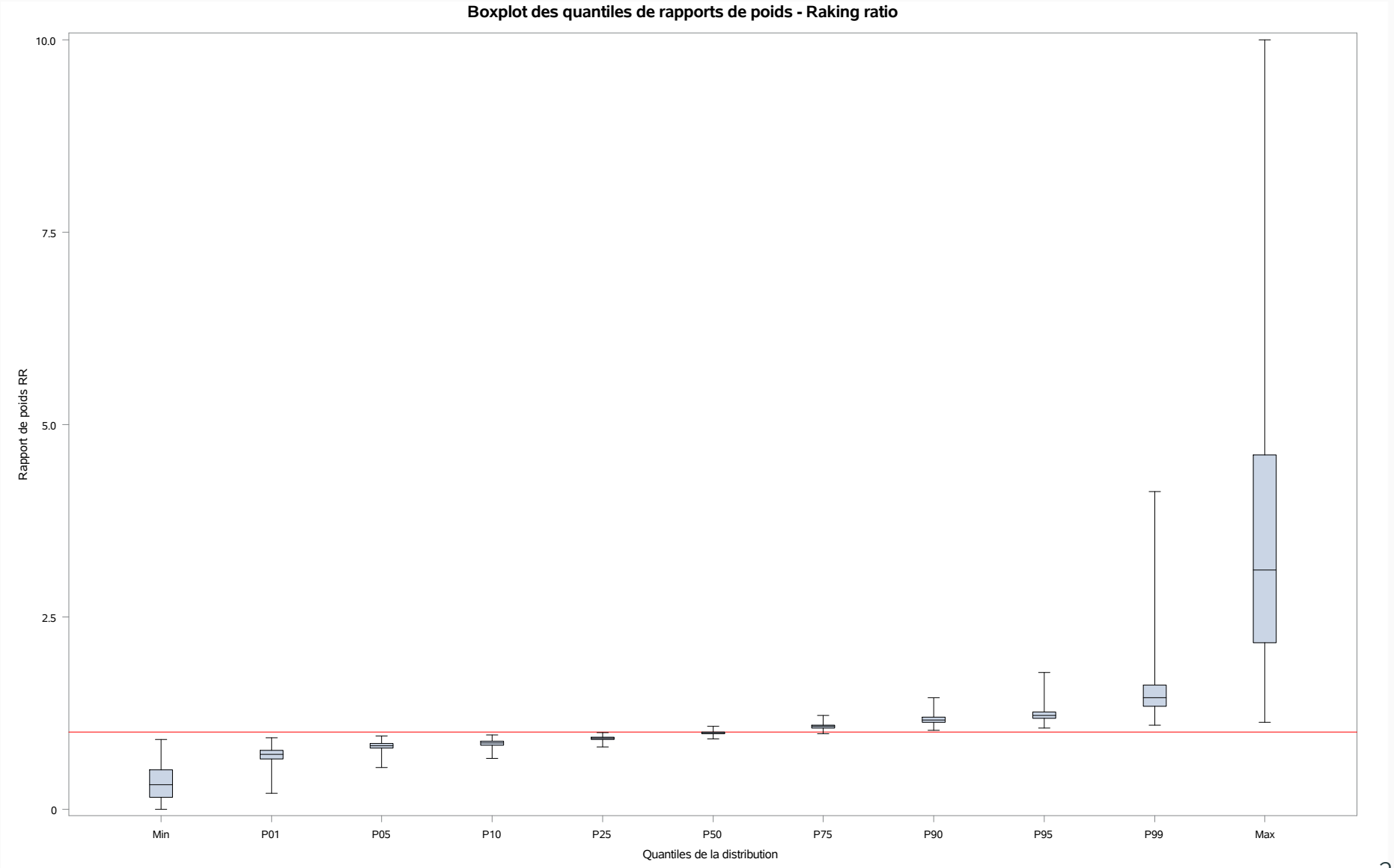
- Pour une variable d'intérêt Y donnée, on calcule les estimateurs calés pour chaque échantillon de chaque scénario de calage :
  - $\hat{T}^k_{Y,w^{RR}} = \sum_{i \in s_k} w_i^{RR} y_i, k = 1, ..., K$
  - $\hat{T}^k_{Y,w^{logit\ [0,5-2]}} = \sum_{i \in s_k} w_i^{logit\ [0,5-2]} y_i, k = 1, ..., K$
  - $\hat{T}^k_{Y,w^{logit\ min}} = \sum_{i \in s_k} w_i^{logit\ min} y_i, k = 1, ..., K$
- On évalue enfin la qualité des estimateurs à l'aide de leurs biais relatifs et racines carrée des EQM relatives Monte-Carlo :
  - Biais relatif absolu (en %) :

$$RB(\hat{T}_{Y,w}) = \frac{100}{K} \sum_{k=1}^{K} \frac{|\hat{T}^k_{Y,w} - T_Y|}{T_Y}$$

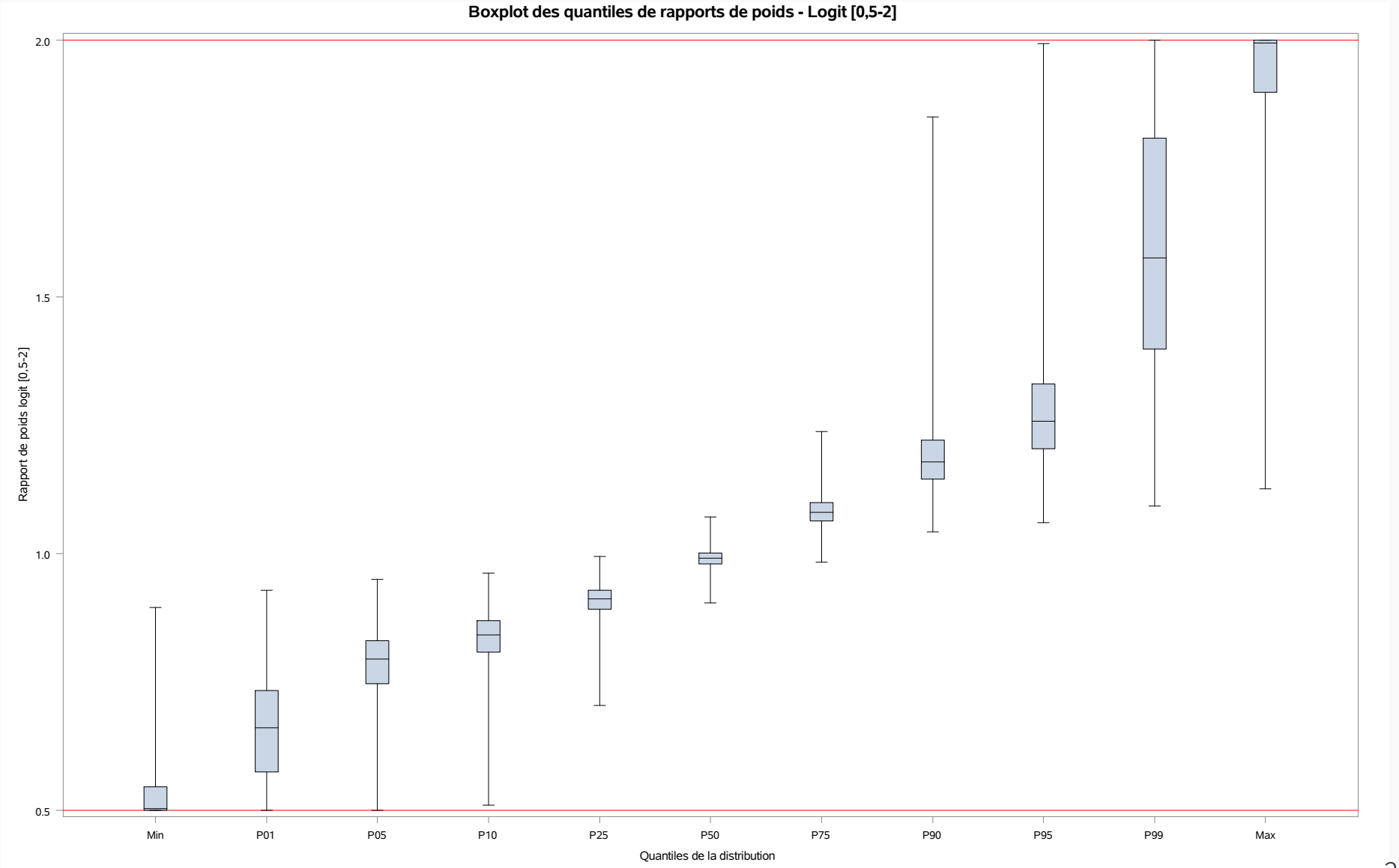
  - Racine carrée de l'EQM relative (en %) :

$$RRMSE(\hat{T}_{Y,w}) = 100 \frac{\sqrt{K^{-1} \times \sum_{k=1}^{K} (\hat{T}^k_{Y,w} - T_Y)^2}}{T_Y}$$

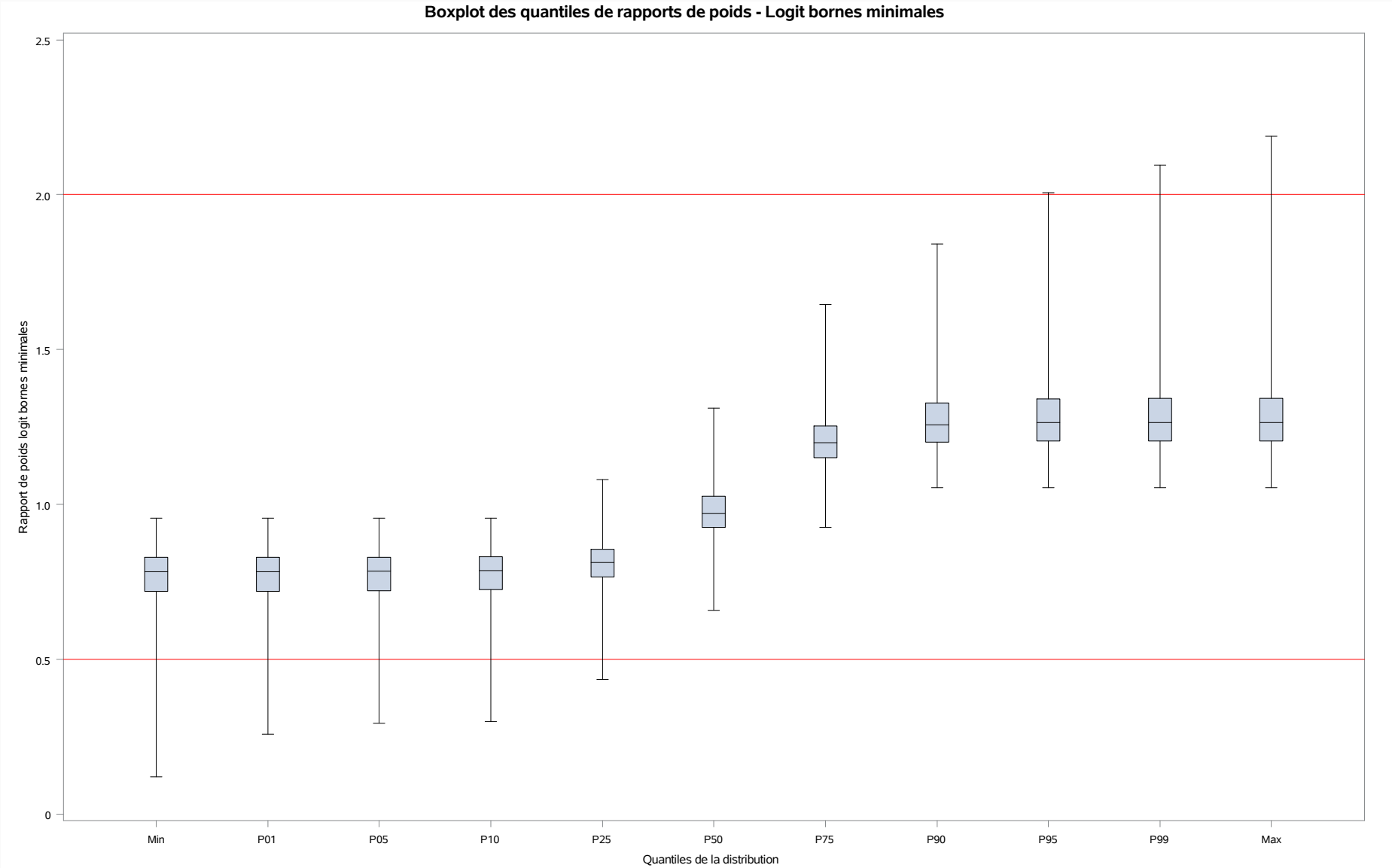
Distribution des rapports de poids – Raking ratio



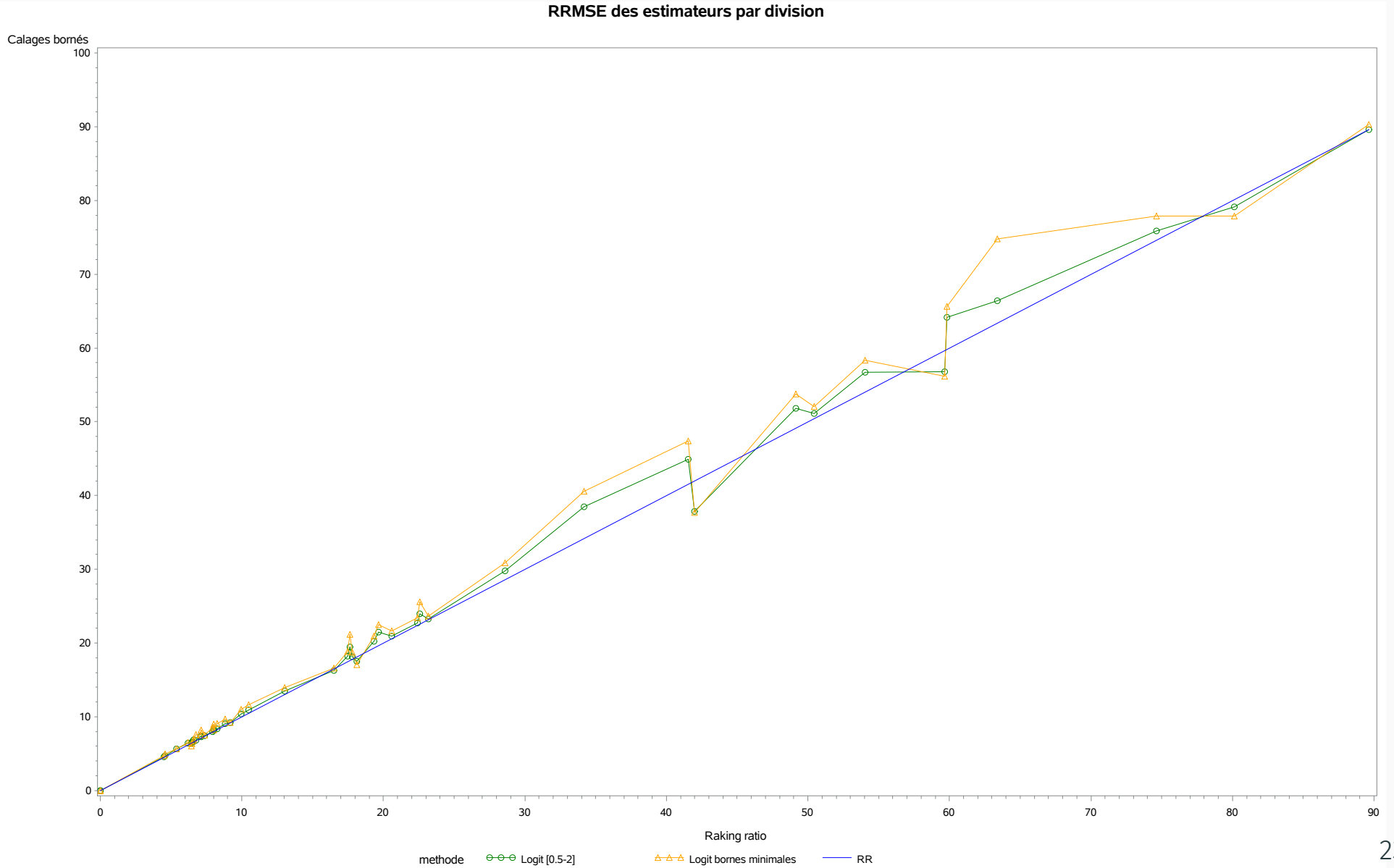
Distribution des rapports de poids – Logit [0,5-2]



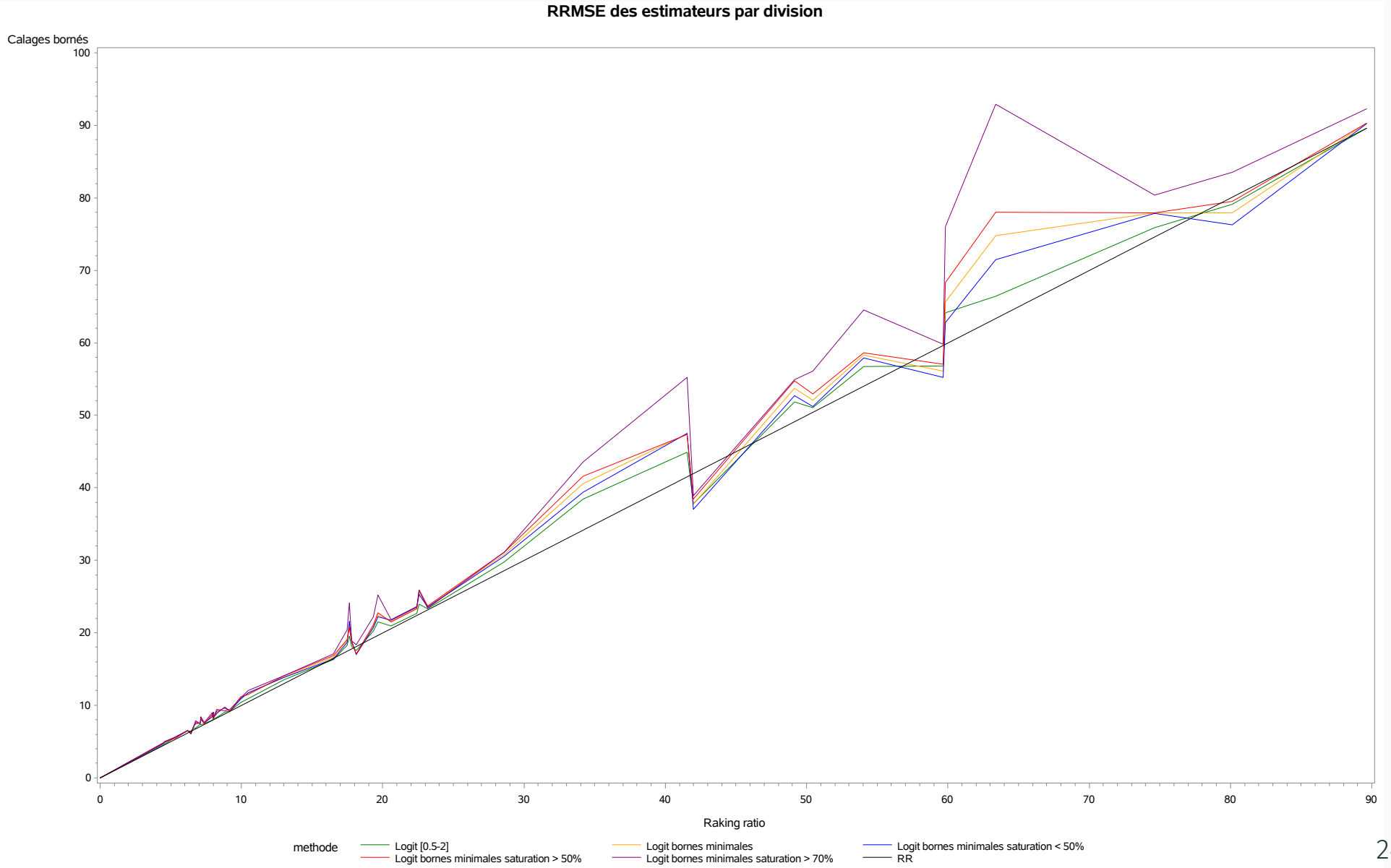
Distribution des rapports de poids – Logit bornes minimales



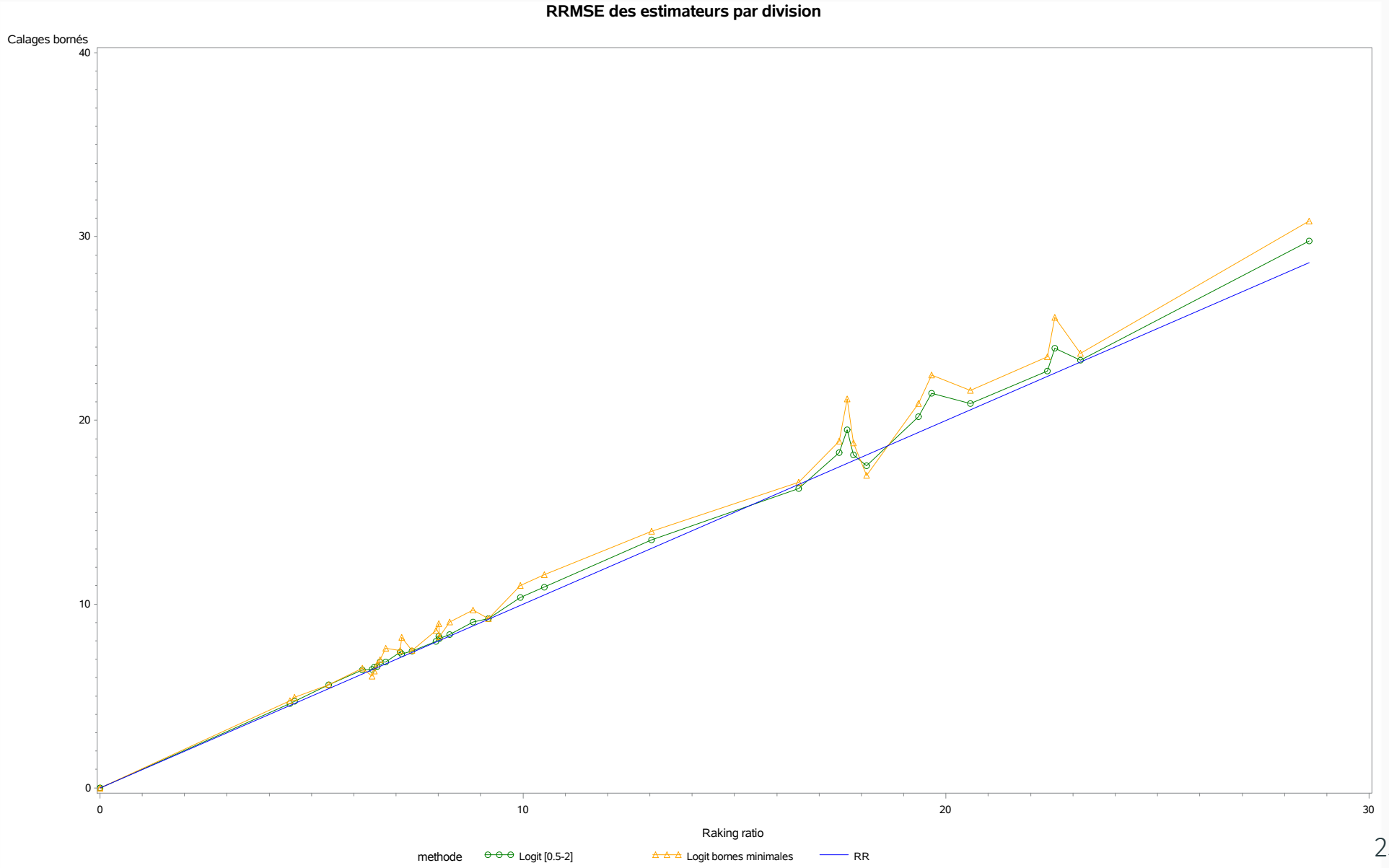
RRMSE des estimateurs par division (1)



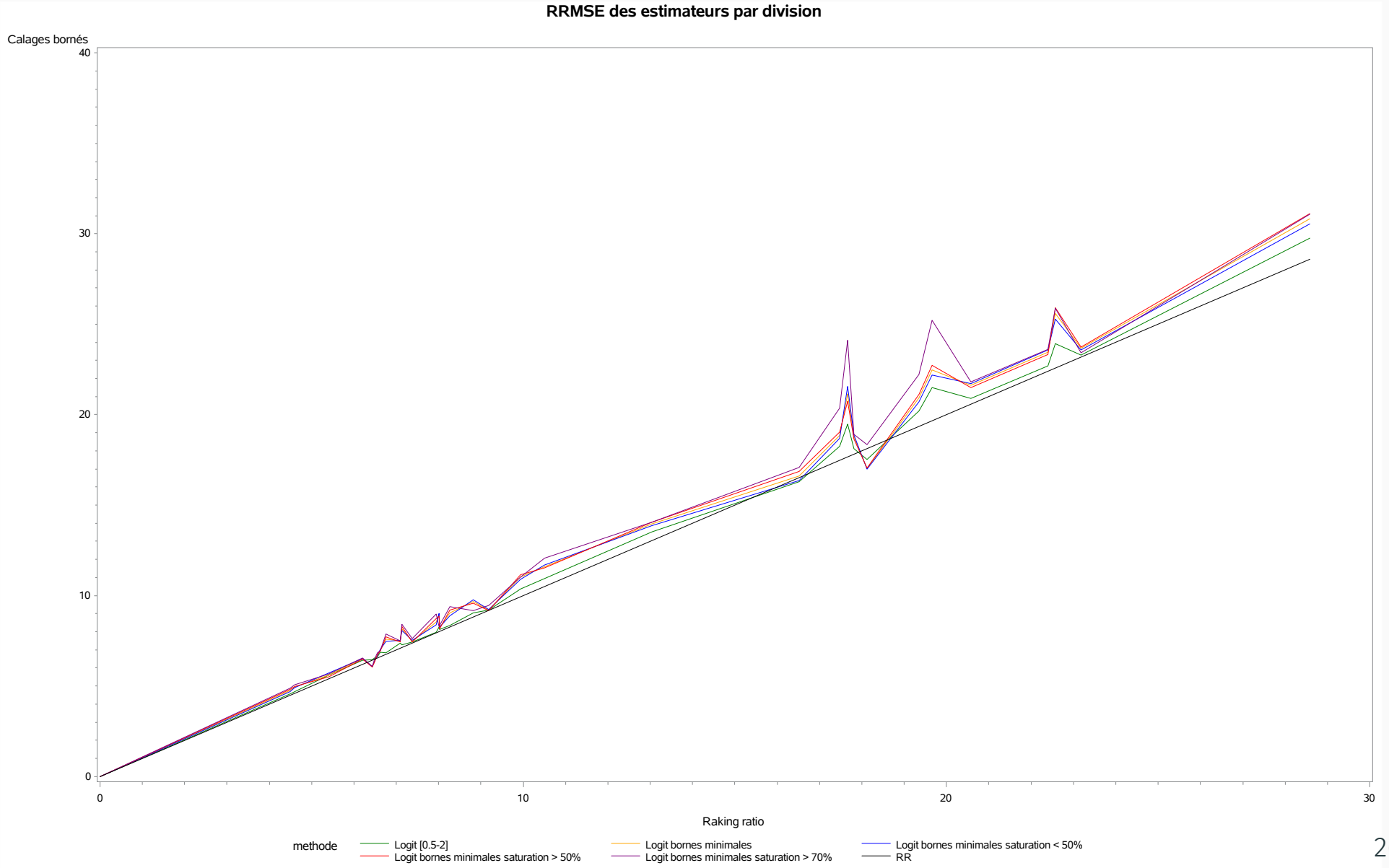
RRMSE des estimateurs par division (2)



RRMSE des estimateurs par division (3)



RRMSE des estimateurs par division (4)

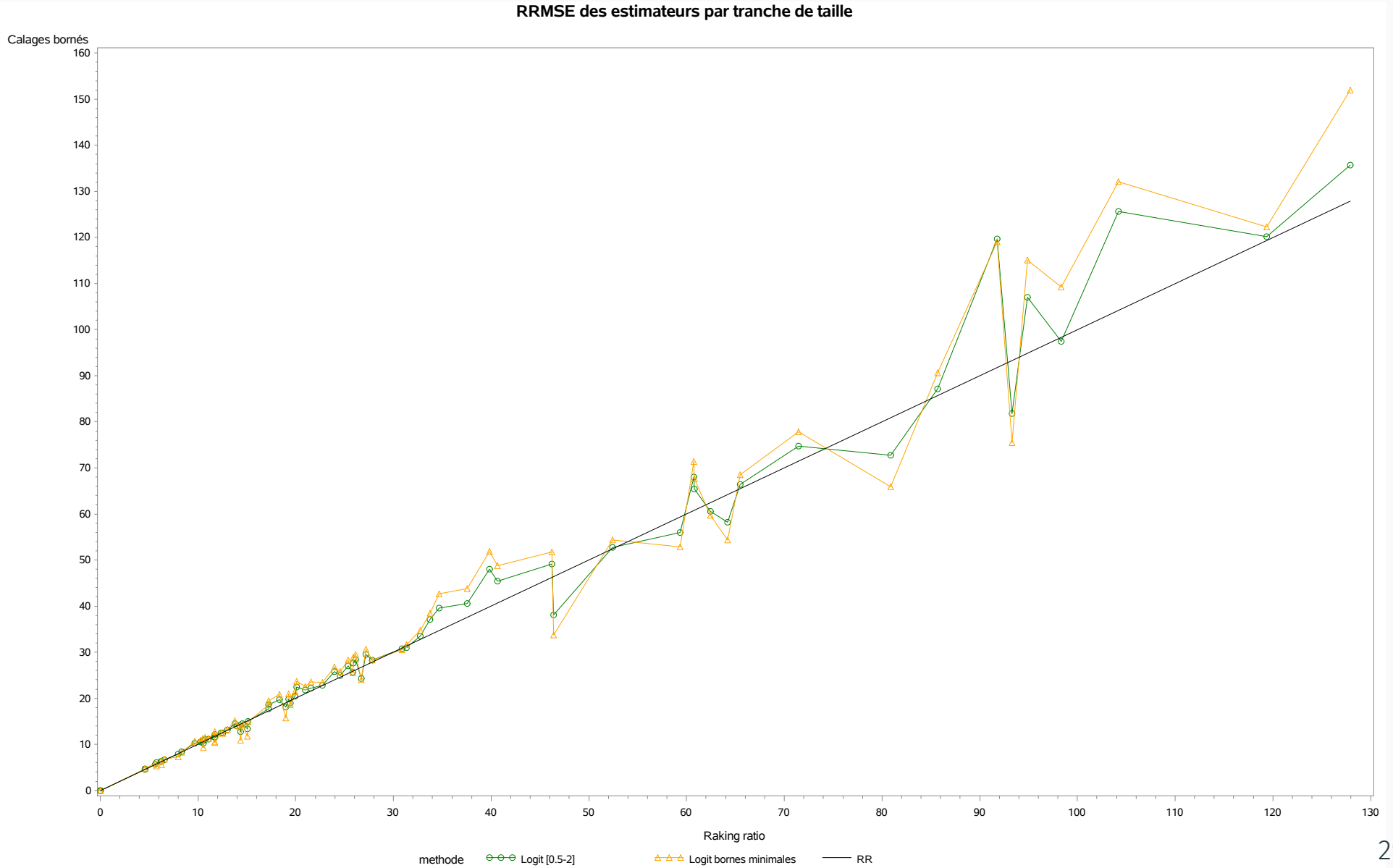


Conclusions

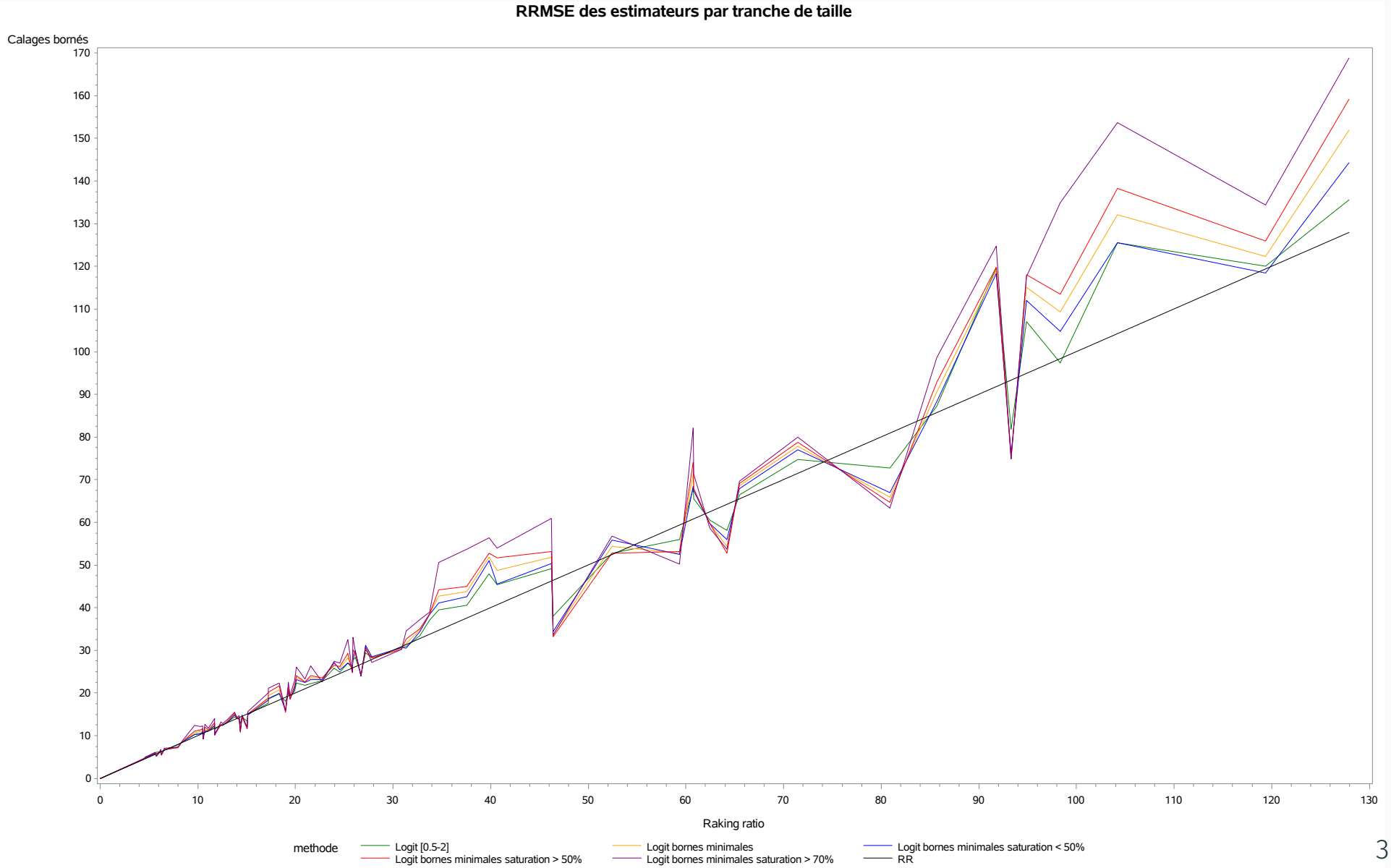
- Le choix de bornes de calage « serrées » semble conduire à des estimateurs moins efficaces, en particulier lorsqu'on observe de fortes accumulations de rapports de poids aux bornes.
- on observe des résultats similaires avec d'autres scénarios de simulations :
  - échantillons tirés selon le PdS de l'ESA, avec génération de non réponse puis calage direct $\rightarrow$ on observe de plus dans ce cas des problèmes importants en termes de biais ;
  - échantillons tirés selon le PdS de l'ESA, avec génération de non réponse puis correction de la non-réponse par GRH et calage post-CNR $\rightarrow$ résultats similaires à ceux présentés ici.
- Ces résultats semblent valider la pratique à l'Insee en termes de choix des bornes de calage.
- Résultats à analyser plus en détails, et à compléter via d'autres simulations (sur données d'enquêtes ménages par exemple).

Merci de votre attention !

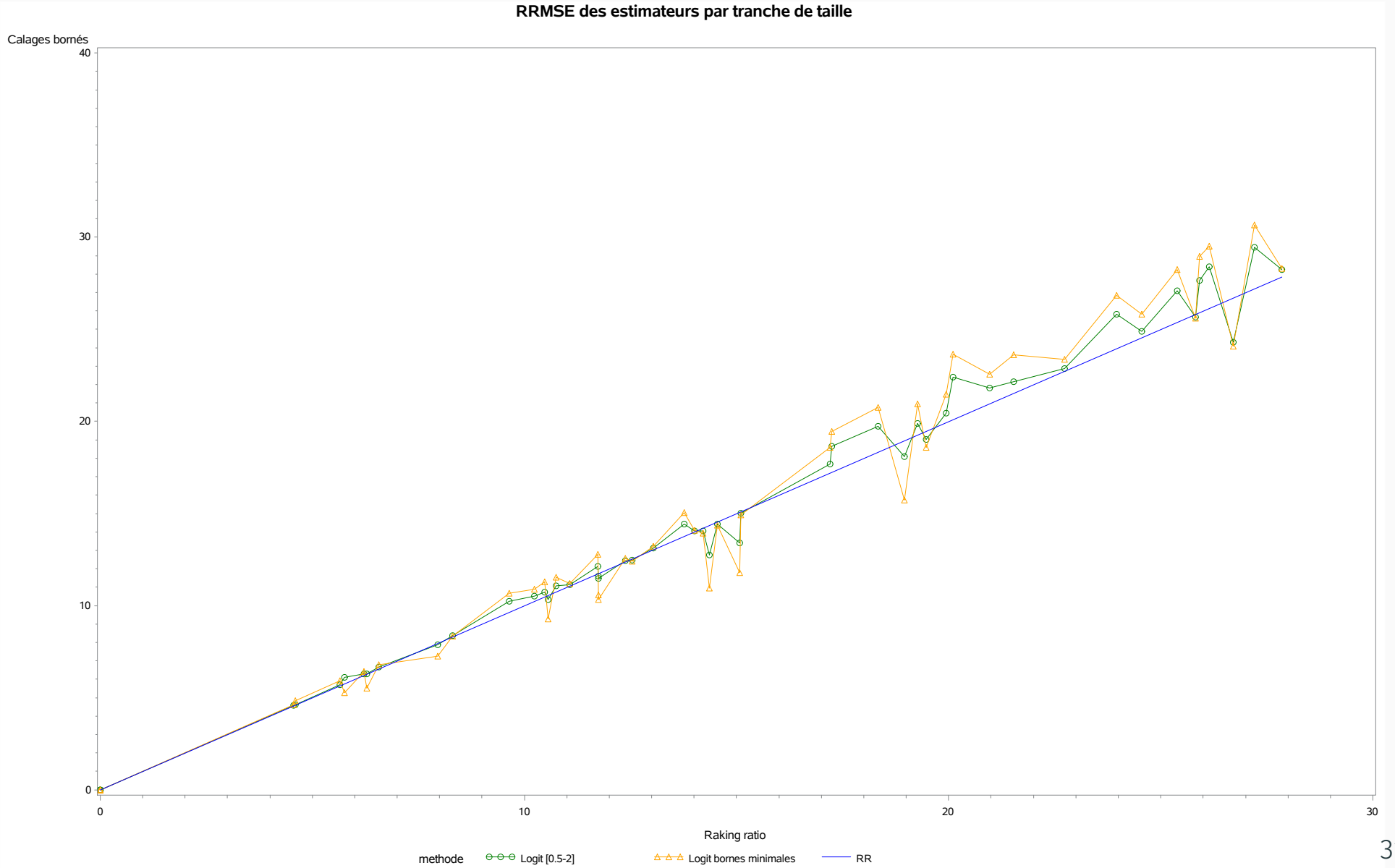
RRMSE des estimateurs par tranches d'effectif (1)



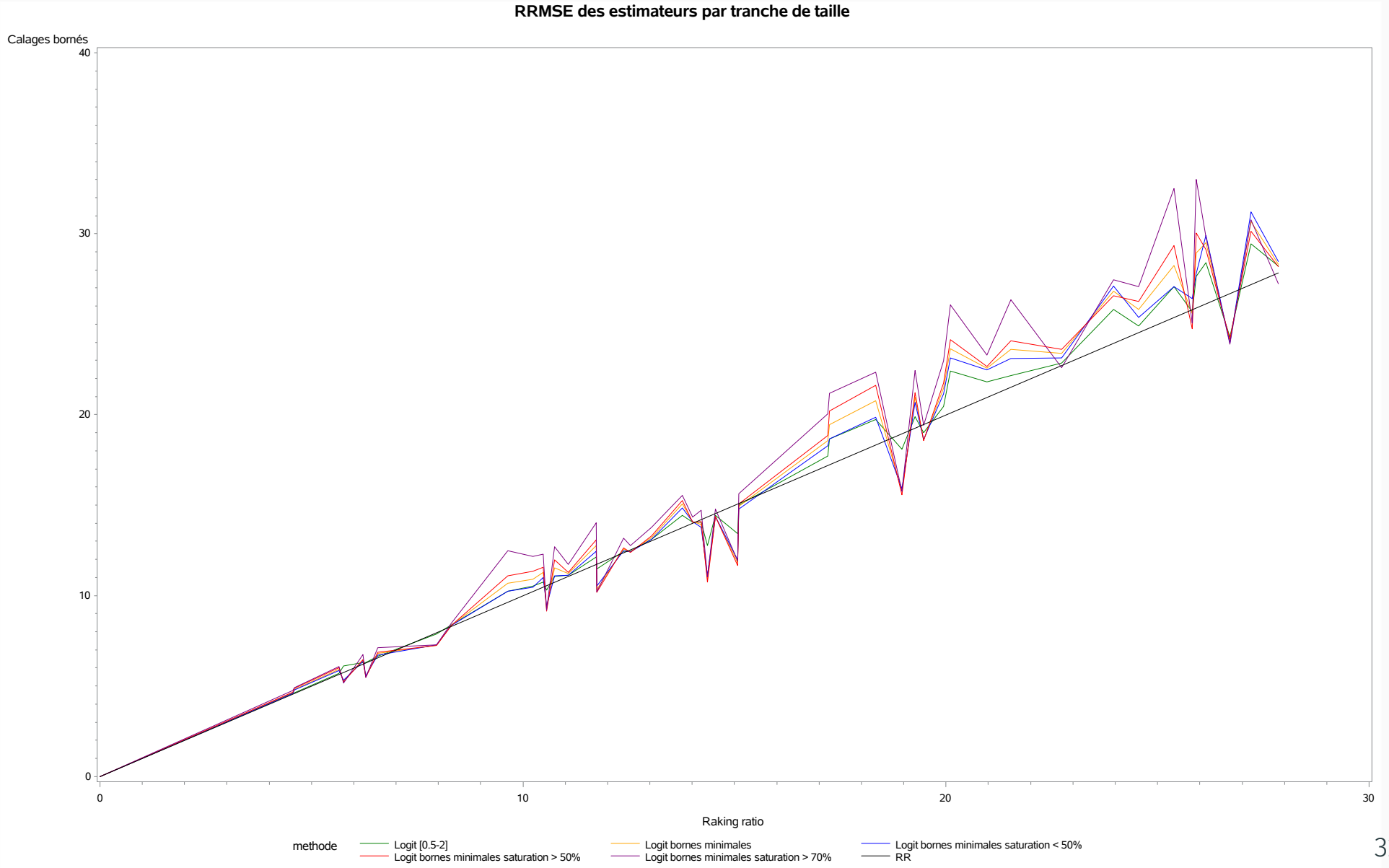
RRMSE des estimateurs par tranches d'effectif (2)



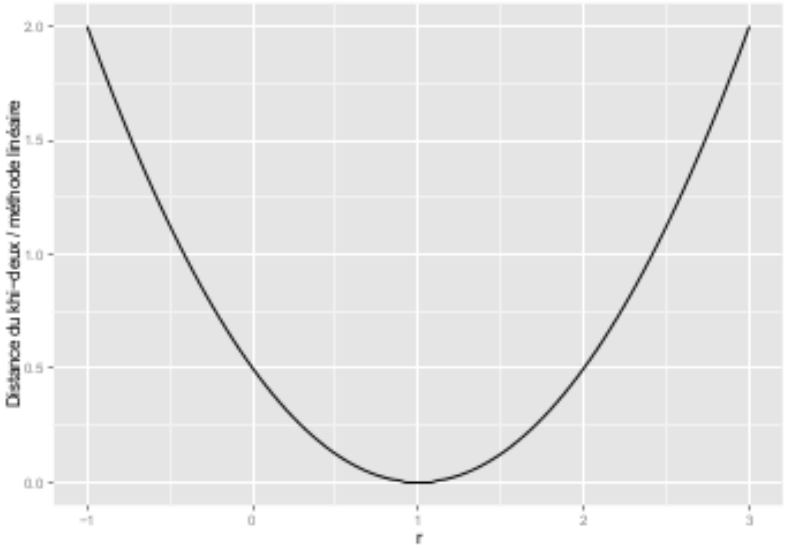
RRMSE des estimateurs par tranches d'effectif (3)



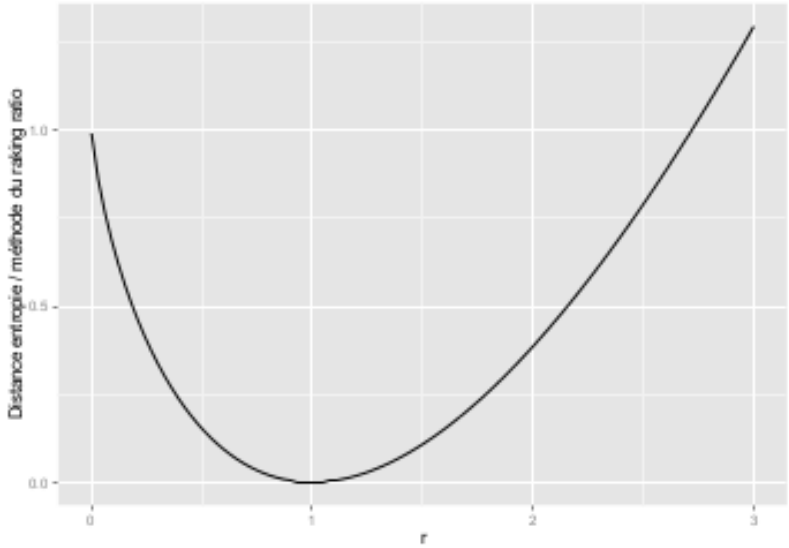
RRMSE des estimateurs par tranches d'effectif (4)



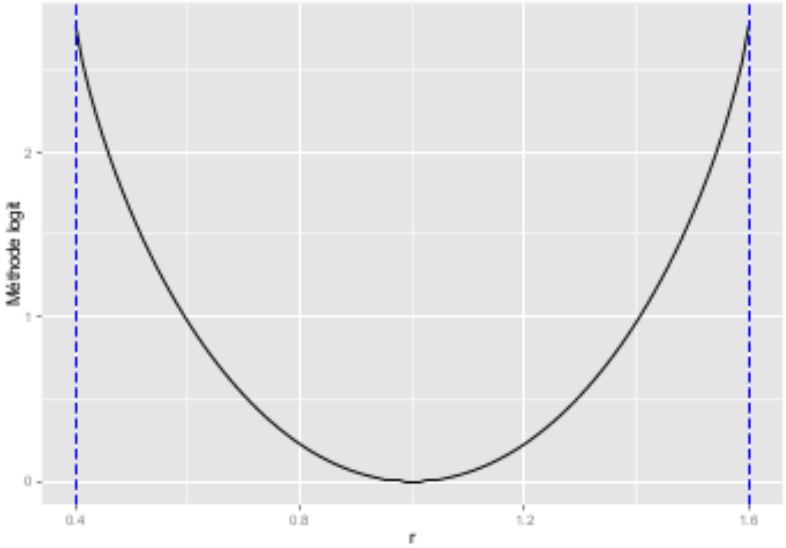
Méthodes de calage : $G(r)$ *versus* r



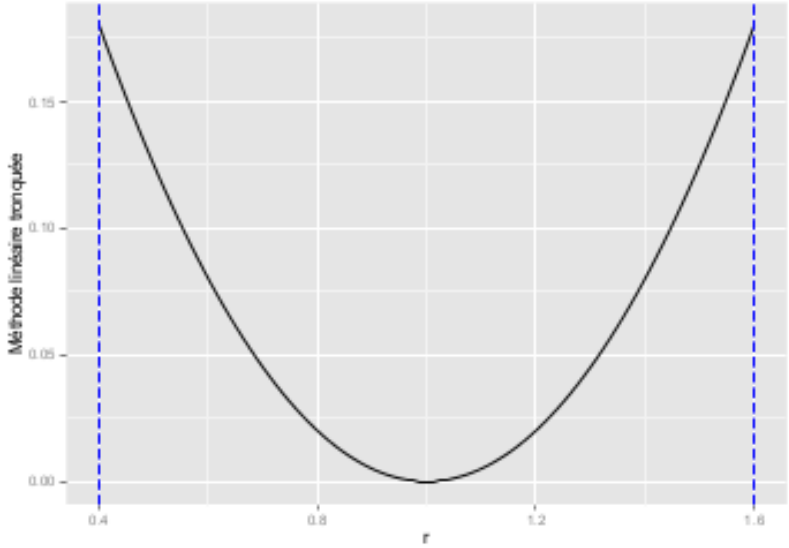
(a) Linéaire



(b) Raking ratio

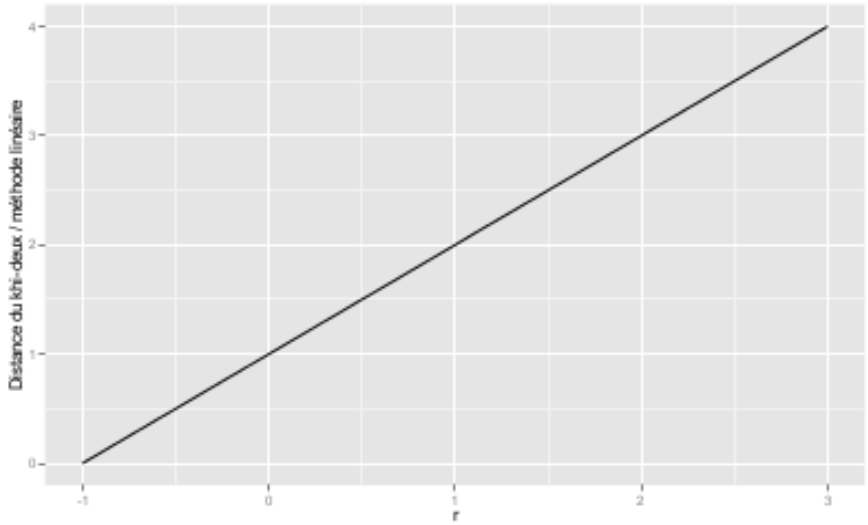


(c) Logit

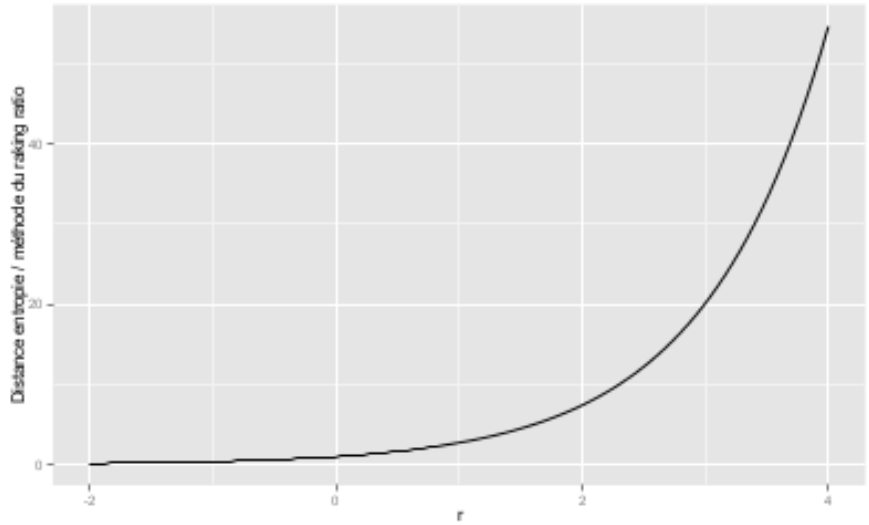


(d) Linéaire tronquée

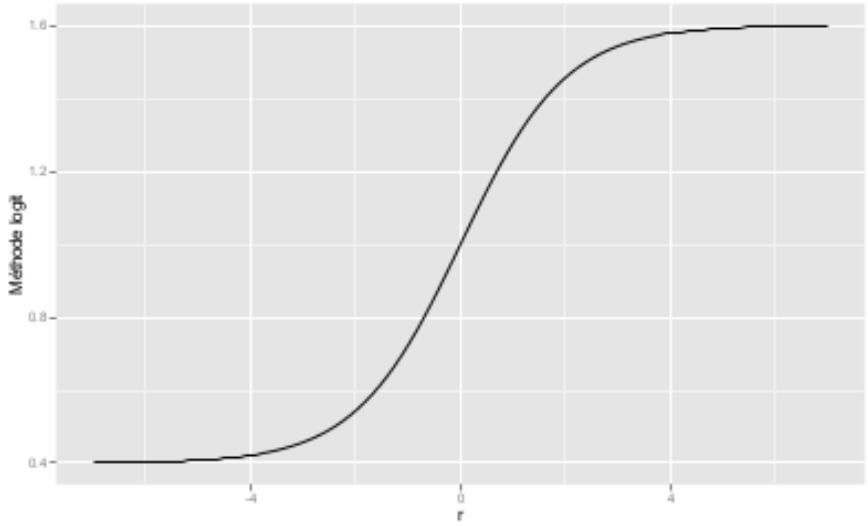
Méthodes de calage : $F(u)$ *versus* u



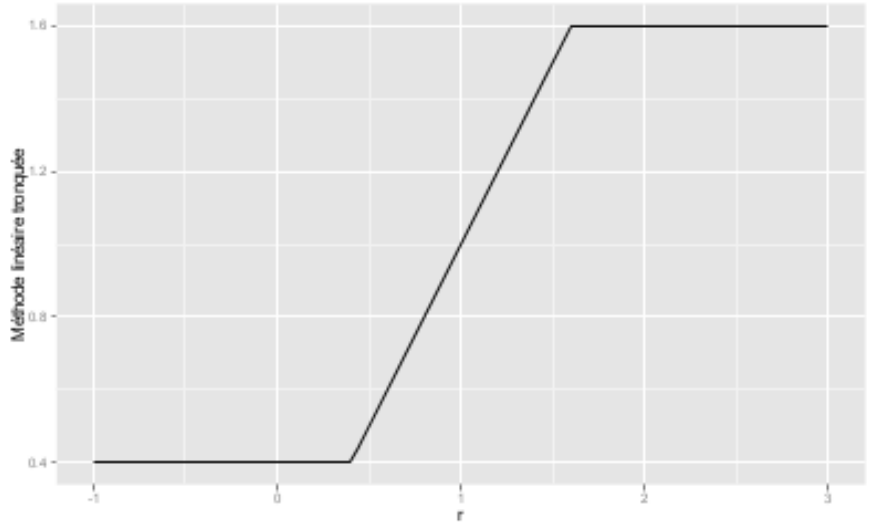
(a) Linéaire



(b) Raking ratio



(c) Logit



(d) Linéaire tronquée