Miscellanées sur le calage

Les différentes méthodes de calage et le choix des « bornes » de calage

Emmanuel Gros

15 mars 2016

INSEE - DMCSI - Division Sondages

Sommaire

Théorie du calage sur marges

Contexte et 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 :

$$\pi_i = \sum_{s \in S, s \ni i} p(s)$$

- $d_i = \frac{1}{\pi_i}$: poids de sondage des unités $\Rightarrow$ pondération appropriée en l'absence d'erreurs non dues à l'échantillonnage : non-réponse, erreurs de mesures, défaut de couverture, etc.

L'estimateur d'Horvitz-Thompson

- Les poids de sondage permettent de construire, pour toute variable d'intérêt Y, l'estimateur d'**Horvitz-Thompson** :

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

- Propriétés de l'estimateur d'Horvitz-Thompson :
  - estimateur sans biais (pourvu que $\pi_i > 0$ pour tout i) :

$$E(\hat{T}_{Y\pi}) = T_Y$$

    pour toute variable Y ;

  - précision de l'estimation fonction :
    - de la dispersion de la variable d'intérêt considérée ;
    - de la corrélation entre la variable d'intérêt considérée et les poids de sondage.

Information auxiliaire

- Soit $\mathbf{X} = (X_1, ..., X_J)'$ un vecteur de J variables auxiliaires
  $\Rightarrow$ exemples : sexe, diplôme, revenu, secteur d'activité, chiffre d'affaires, etc.
- On suppose $\mathbf{x}_i$ connu pour toutes les unités de l'échantillon ; on suppose par ailleurs connu le vecteur des totaux de ces variables sur la population :

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

- Si l'estimateur d'Horvitz-Thompson est sans biais, il ne garantit pas une estimation exacte des totaux des variables auxiliaires :

$$\text{on a } E(\hat{T}_{\mathbf{X}\pi}) = T_{\mathbf{X}}, \text{ mais en général } \hat{T}_{\mathbf{X}\pi} \neq T_{\mathbf{X}}$$

- Solution : modifier les pondérations initiales pour assurer la cohérence avec les informations auxiliaires.

Calage sur marges – problématique

- Objectif : prendre en compte l'information auxiliaire disponible en modifiant les poids initiaux de façon à améliorer l'estimateur d'Horvitz-Thompson $\Rightarrow$ on cherche de nouveaux poids w_i qui soient :
  - « les plus proches possibles » des poids de sondage d_i...
  - ... tout en vérifiant les J équations de calage :

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

- Articles fondateurs : Deville et Särndal (1992); Deville, Särndal et Sautory (1993).
- Pour toute variable d'intérêt Y, l'estimateur calé est défini par :

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

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 $F(.)$ est appelée fonction de calage (dépend de G).

- Le facteur de calage $F(\mathbf{x}_i' \hat{\lambda})$ dépend :
  - de $F(.)$, donc du choix de la distance G ;
  - des caractéristiques $\mathbf{x}_i$ de l'unité i ;
  - de $\hat{\lambda}$ qui peut être vu comme une mesure du déséquilibre entre $\hat{T}_{\mathbf{x}\pi}$ et $T_{\mathbf{x}}$.

Méthodes de calage (1) – méthodes linéaire et exponentielle

On indique les fonctions $G(r)$ (où $r = \frac{w_i}{d_i}$) et $F(u)$ (où $u = \mathbf{x}_i' \hat{\lambda}$).

- Méthode linéaire : distance du chi-deux généralisée, fonction de calage linéaire

$$G(r) = \frac{1}{2}(r-1)^2 \rightarrow F(u) = 1 + u$$

- Méthode exponentielle ou raking ratio : distance de type « entropie », fonction de calage exponentielle

$$G(r) = r\log(r) - r + 1 \rightarrow F(u) = \exp(u)$$

Méthodes de calage (2) – méthodes bornées

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

- Méthode logit : il s'agit d'une méthode exponentielle bornée

$$\begin{cases} G(r) &= \frac{1}{A}\left[(r-L)\log\frac{r-L}{1-L} + (U-R)\log\frac{U-r}{U-1}\right] \text{ si } r \in]L;U[\\ &= +\infty \text{ sinon ; avec } A = \frac{U-L}{(1-L)(U-1)} \end{cases}$$

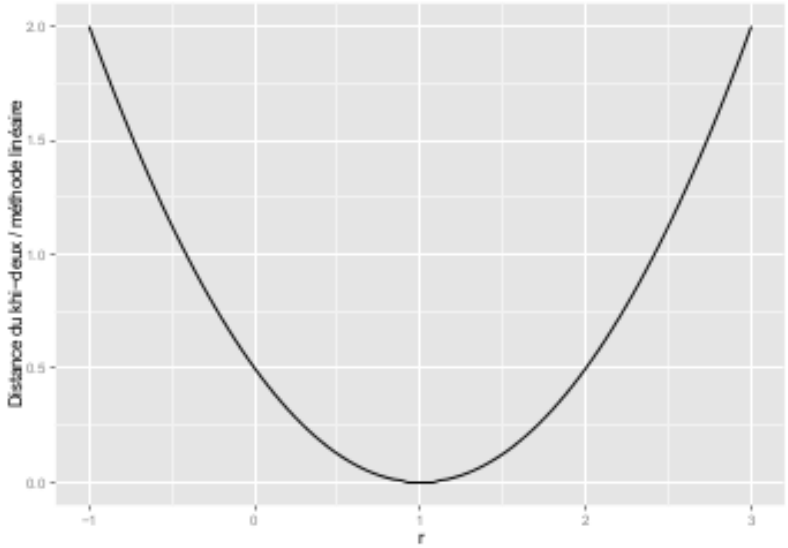
$$\rightarrow F(u) = \frac{L(U-1) + U(1-L)\exp(Au)}{(U-1) + (1-L)\exp(Au)} \text{ et alors } F(u) \in]L;U[$$

- Méthode linéaire tronquée :

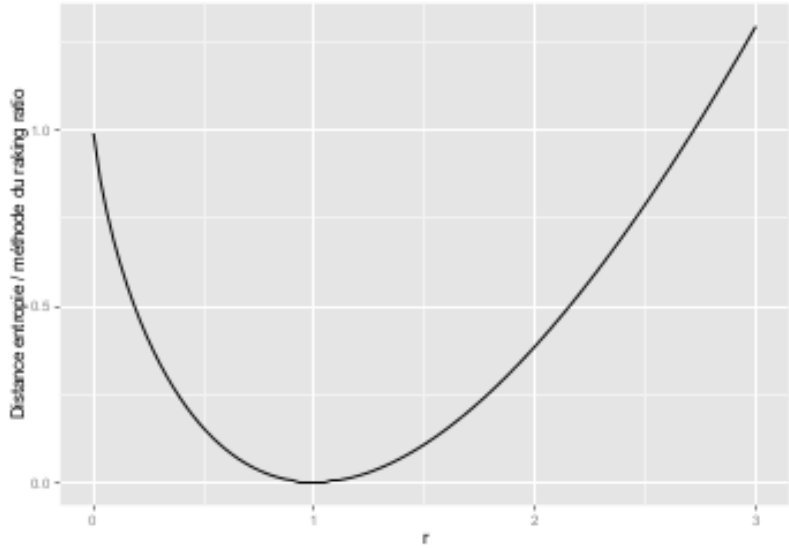
$$\begin{cases} G(r) &= \frac{1}{2}(r-1)^2 \text{ si } r \in [L;U] \\ &= +\infty \text{ sinon} \end{cases}$$

$$\rightarrow F(u) = 1 + u \in [L;U]$$

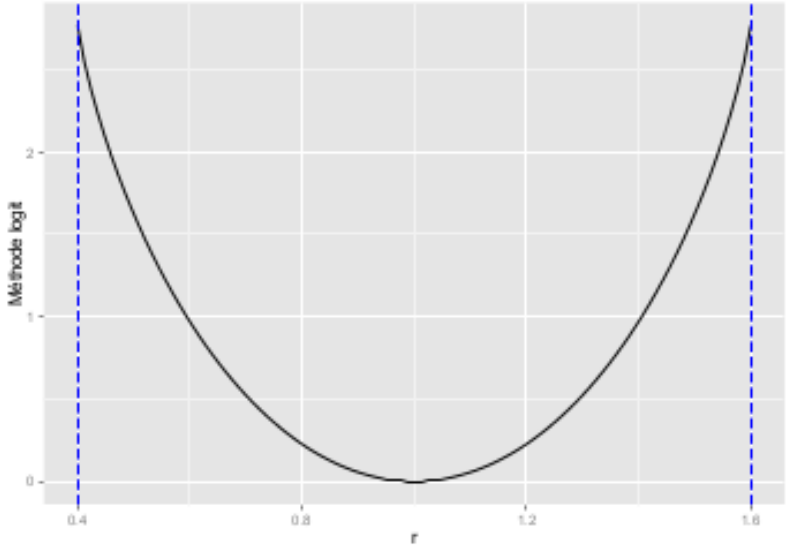
Méthodes de calage : $G(r)$ *versus* r



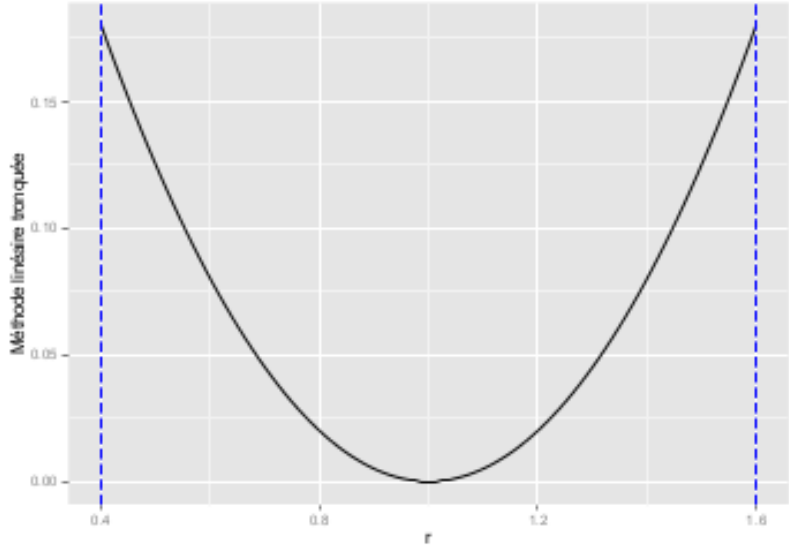
(a) Linéaire



(b) Raking ratio

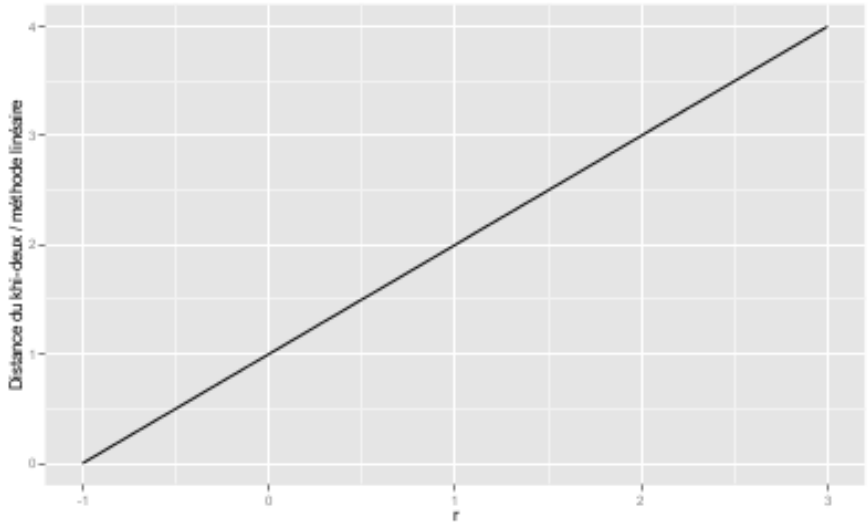


(c) Logit

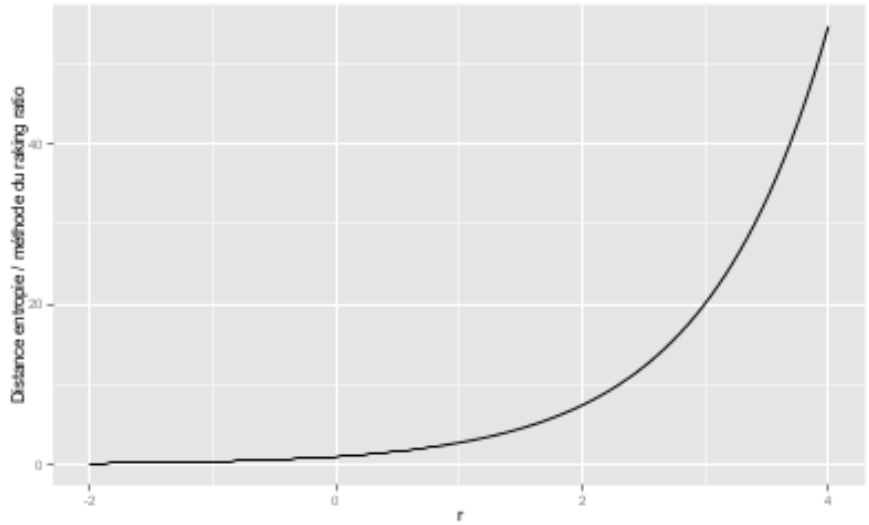


(d) Linéaire tronquée

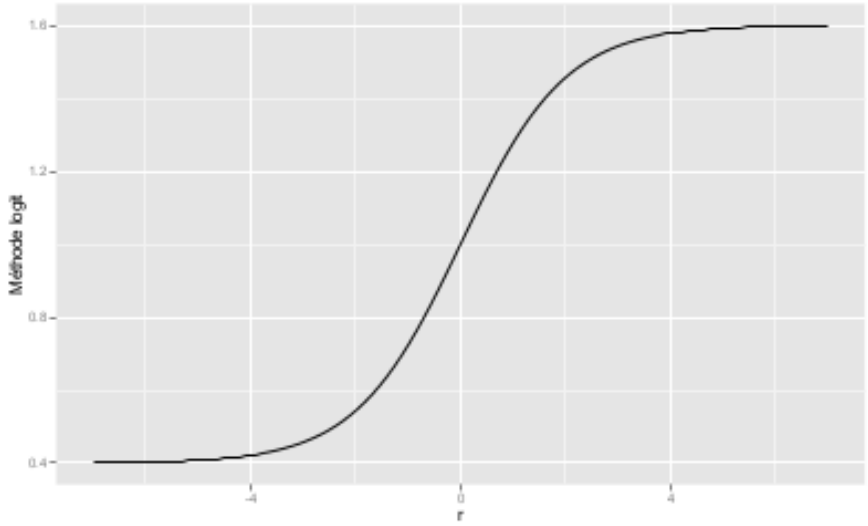
Méthodes de calage : $F(u)$ *versus* u



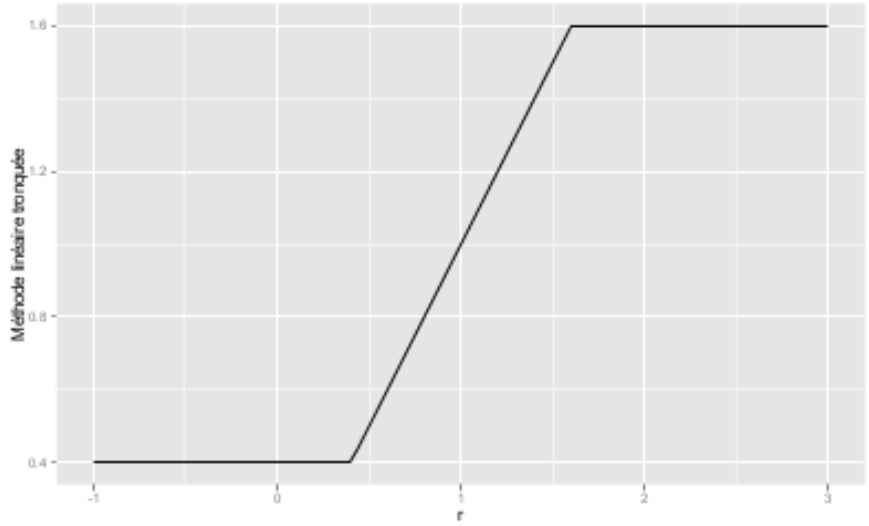
(a) Linéaire



(b) Raking ratio



(c) Logit



(d) Linéaire tronquée

Propriétés de l'estimateur calé

- La méthode linéaire permet de retrouver l'estimateur par la régression généralisée de Y sur les variables auxiliaires X_1 à X_J...

- ... et toutes les méthodes sont **asymptotiquement équivalentes**
  $\rightarrow$ les propriétés de l'estimateur calé sont donc celles de l'estimateur par la régression généralisée :

  - asymptotiquement sans biais pour toute variable Y ;
  - la variance de l'estimateur est fonction des résidus de la régression de Y sur les $X_1, ..., X_J$, et non plus de la variable d'intérêt Y directement $\Rightarrow$ amélioration de la précision des estimations pour toutes variables d'intérêt bien corrélées aux variables de calage.

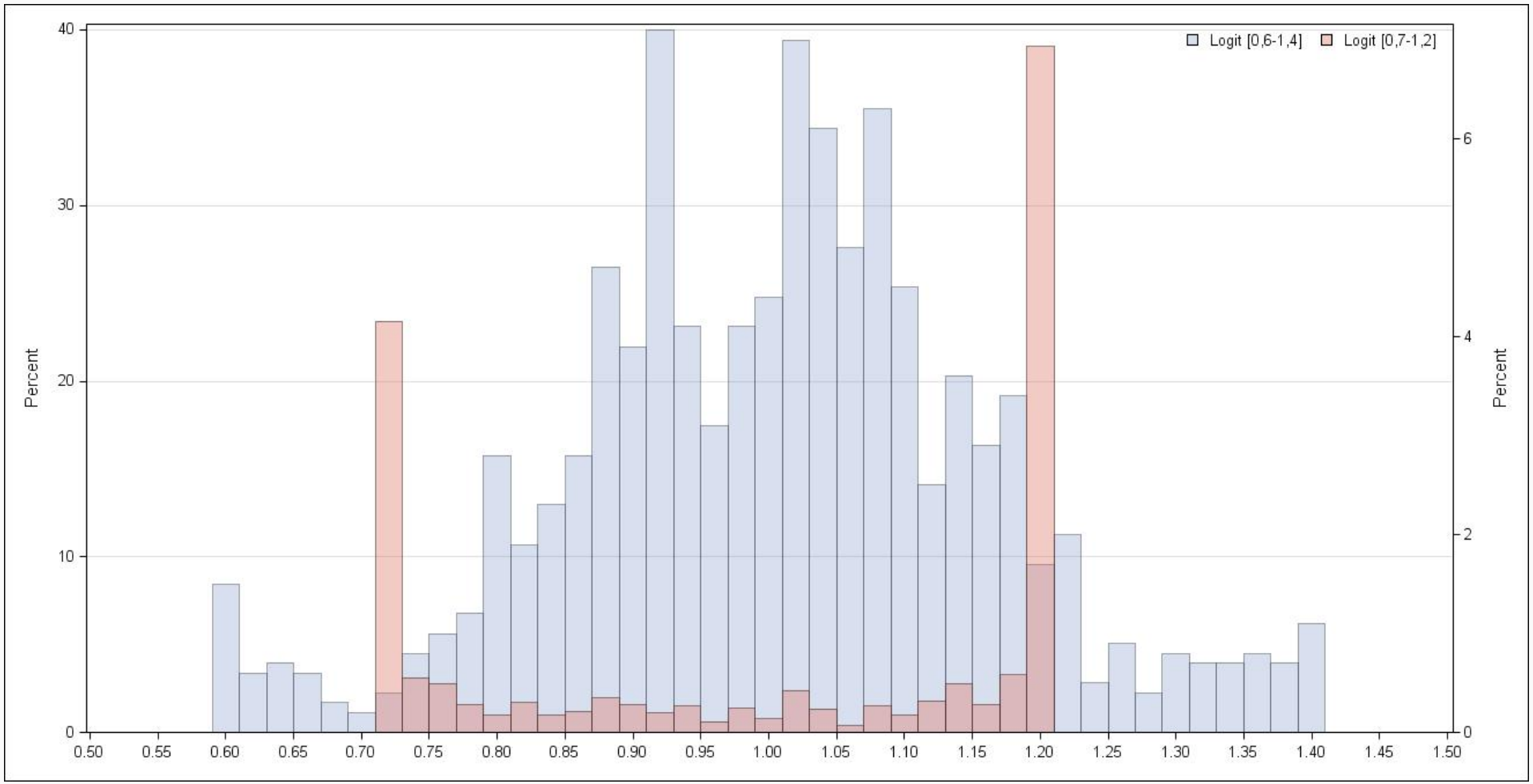
Comparaison des méthodes de calage

- La méthode linéaire converge systématiquement, mais peut conduire à des poids négatifs ou très élevés.
- La méthode exponentielle converge quasi-systématiquement, garantit des poids positifs, mais peut conduire à des poids inférieurs à 1 ou très élevés.
- Les méthodes bornées permettent un contrôle des déformations maximales de poids induites par le calage via les paramètres L et U, mais ne convergent pas systématiquement :
  - il existe en général pour L une valeur $L_{max} < 1$ et pour U une valeur $U_{min} > 1$ telles qu'au delà de ces valeurs, le calage n'est plus possible ;
  - en général, lorsque l'on retient des valeurs de L et U proches de L_{max} et U_{min}, on observe des accumulations de rapports de poids à ces bornes.

Choix de la méthode de calage

- Il n'existe pas de critères théoriques sans ambiguïté permettant de déterminer la « bonne » méthode de calage.
- En pratique, on privilégie en général les méthodes bornées :
  - pour des raisons « cosmétiques » : elles permettent d'éviter les poids négatifs ou inférieurs à 1 ;
  - pour des raisons de robustesse des estimations par sous-population : elles permettent d'éviter l'apparition de poids extrêmes, susceptibles de créer des unités influentes conduisant à des estimations peu robustes sur certaines sous-populations.
- Le choix des bornes L et U résulte d'un arbitrage : on souhaite d'une part contrôler suffisamment les déformations de poids maximales induites par le calage et d'autre part éviter les accumulations trop importantes de rapports de poids aux bornes.

Choix des bornes de calage



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 $U^* - L^*$.
- Il n'existe pas de formule analytique permettant de calculer ces bornes L^* et U^* en fonction des données – échantillon, poids initiaux, variables de calage et marges – impliquées dans le calage $\rightarrow$ détermination empirique des bornes minimales.
- Solution : résolution numérique via un programme d'optimisation :
  - problématique déjà envisagée en 2001 par Camille Vanderhoeft dans une version simplifiée ;
  - présentation de Monique Graf au 8^{eme} colloque francophone sur les Sondages à Dijon en 2014 ;
  - une version plus efficace implémentée dans le package R Icarus.

Problèmatique

- Les poids calés doivent vérifier les contraintes de calage :

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

- En notant :
  - $\widetilde{\mathbf{X}}_s = (d_i x_{j,i})_{1 \leq i \leq n, 1 \leq j \leq J}$,
  - $g_i = \dfrac{w_i}{d_i}$ le rapport de poids pour l'unité i, et $\mathbf{g}$ le vecteur des rapports de poids,

  ces équations de calage se ré-écrivent sous la forme :

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

- 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 :

$$\begin{cases} \min\limits_{\mathbf{g}\in\mathbb{R}^n} \left(\max\limits_{i\in[[1,n]]} g_i - \min\limits_{j\in[[1,n]]} g_j \right) \\ \qquad \text{s. c. } \widetilde{X}_s' \mathbf{g} = T_X \ ; \ \mathbf{g} \geq 0 \end{cases}$$

- Ce programme est un programme linéaire, qui peut donc être résolu numériquement via l'algorithme du simplexe.

- Problème : 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 :

$$\begin{cases} \min\limits_{\mathbf{g}\in\mathbb{R}^n, a\in\mathbb{R}} \max\limits_{i\in[[1,n]]} |g_i - a| \\ \text{s. c. } \widetilde{\mathbf{X}}_s'\mathbf{g} = T_{\mathbf{X}} \ ; \ \mathbf{g} \geq 0 \end{cases}$$

- 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

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)

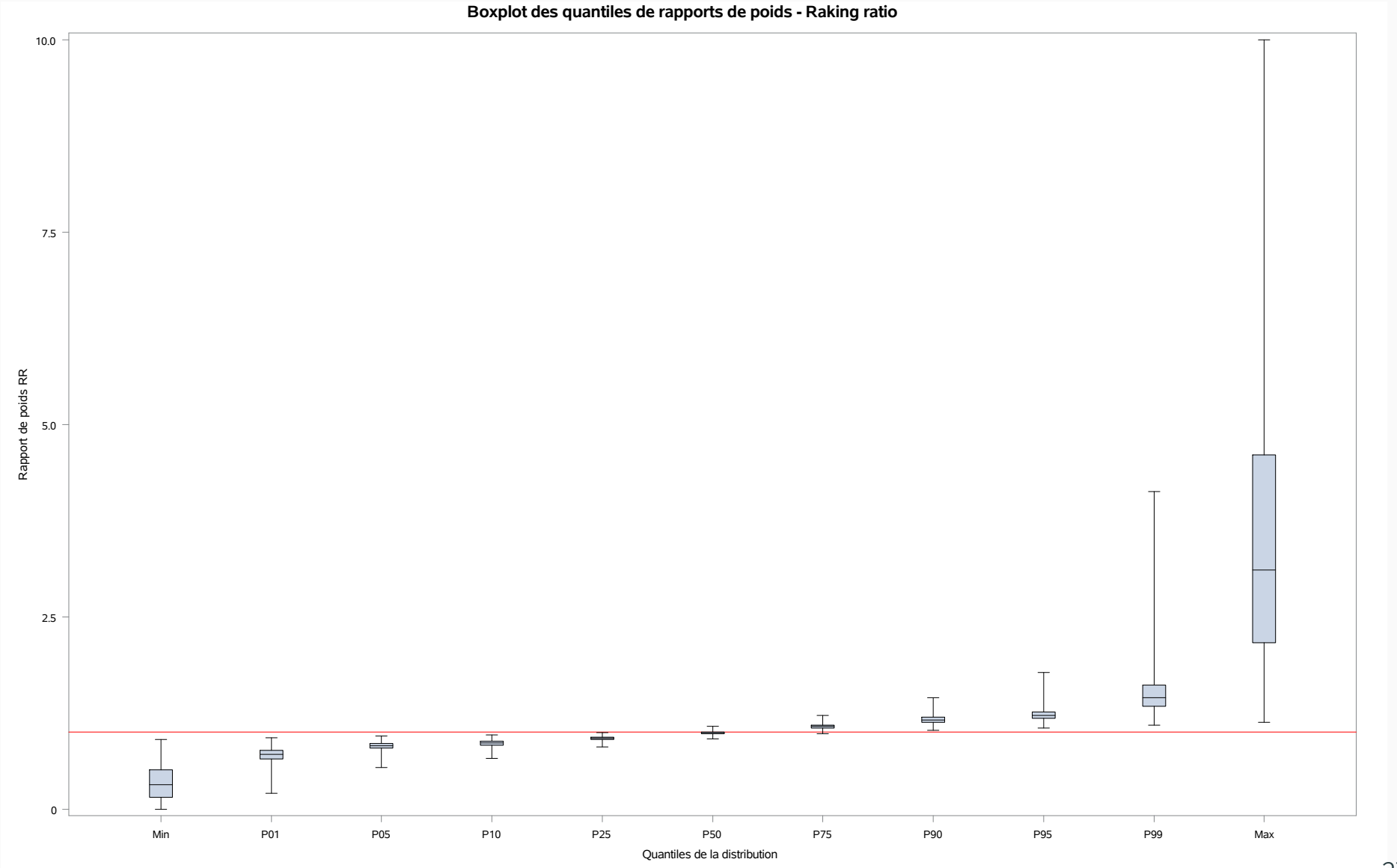
- Pour une variable d'intérêt Y donnée, on calcule les estimateurs calés pour chaque échantillon de chaque scénario de calage :
  - $\hat{T}^k_{Y,w^{RR}} = \sum_{i \in s_k} w_i^{RR} y_i, k = 1, ..., K$
  - $\hat{T}^k_{Y,w^{logit\ [0,5-2]}} = \sum_{i \in s_k} w_i^{logit\ [0,5-2]} y_i, k = 1, ..., K$
  - $\hat{T}^k_{Y,w^{logit\ min}} = \sum_{i \in s_k} w_i^{logit\ min} y_i, k = 1, ..., K$
- 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 %) :

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

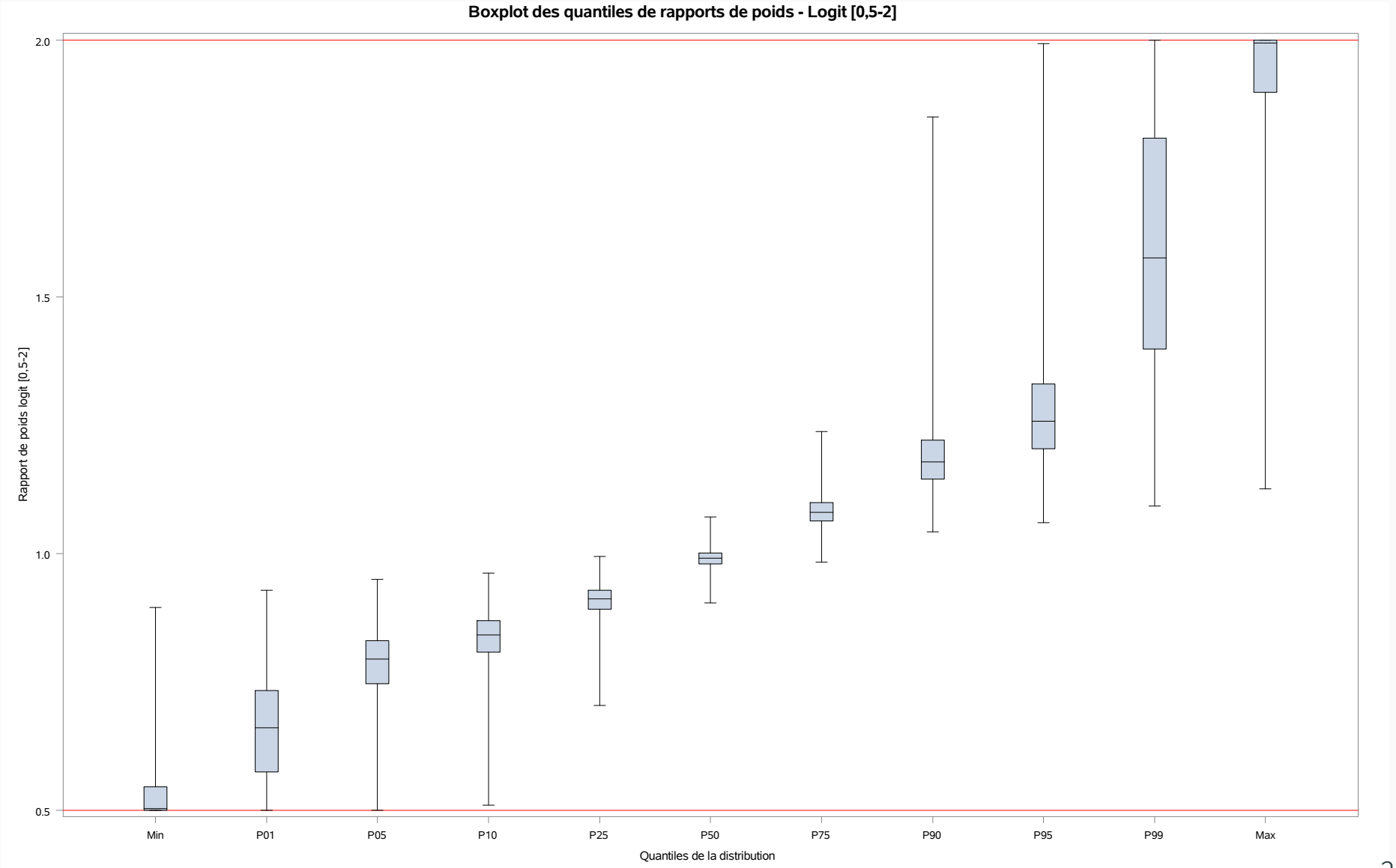
  - Racine carrée de l'EQM relative (en %) :

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

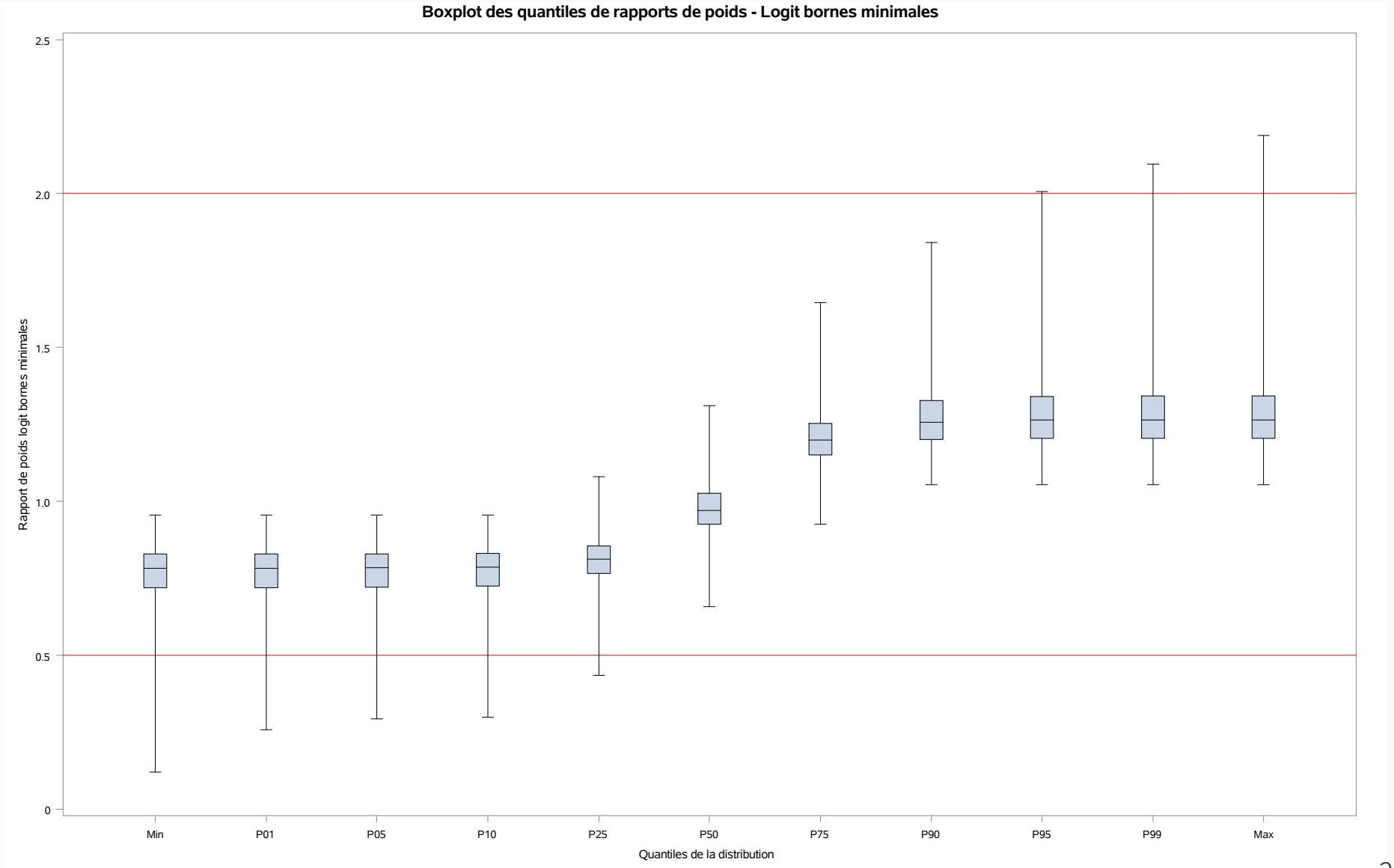
Distribution des rapports de poids – Raking ratio



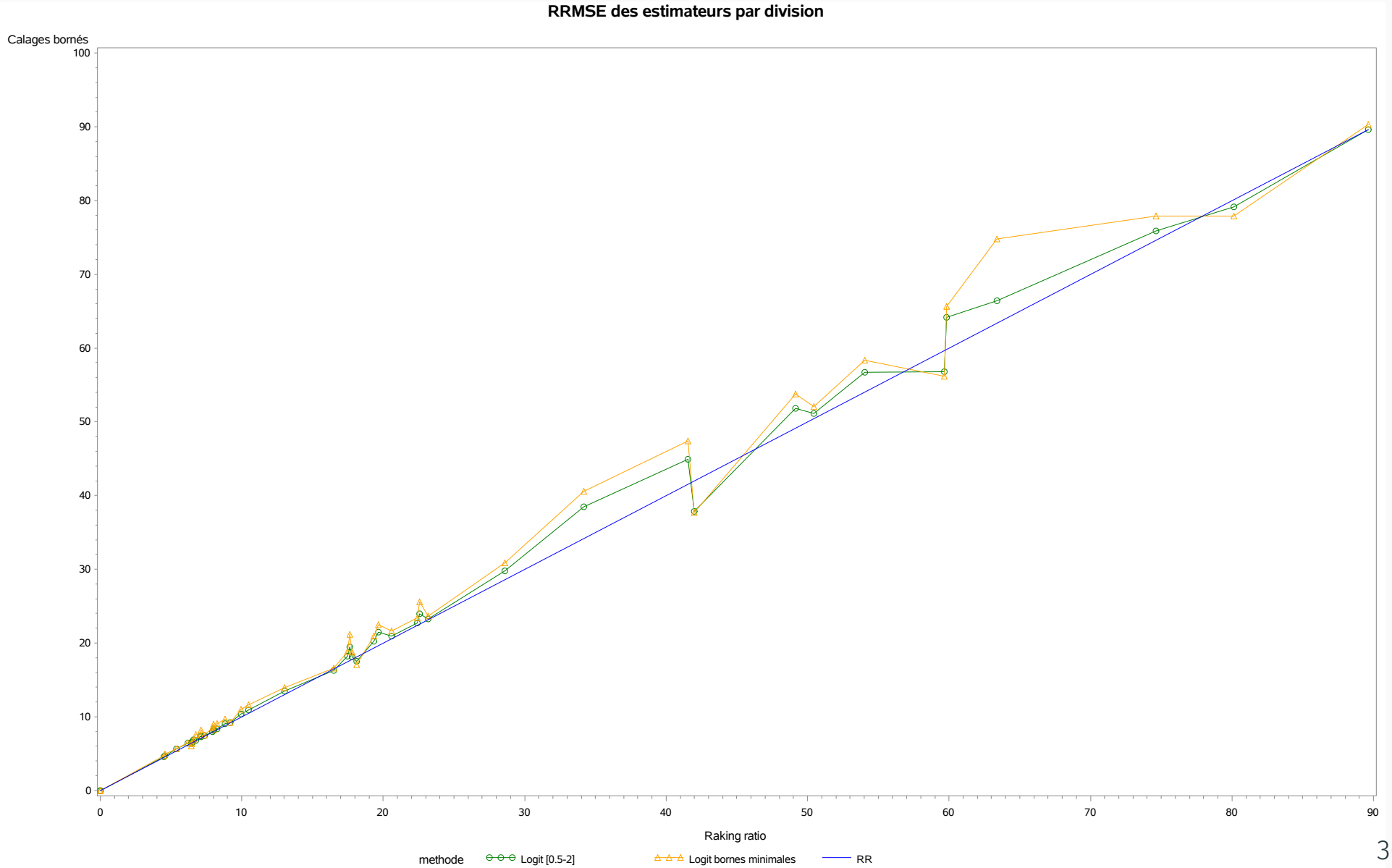
Distribution des rapports de poids – Logit [0,5-2]



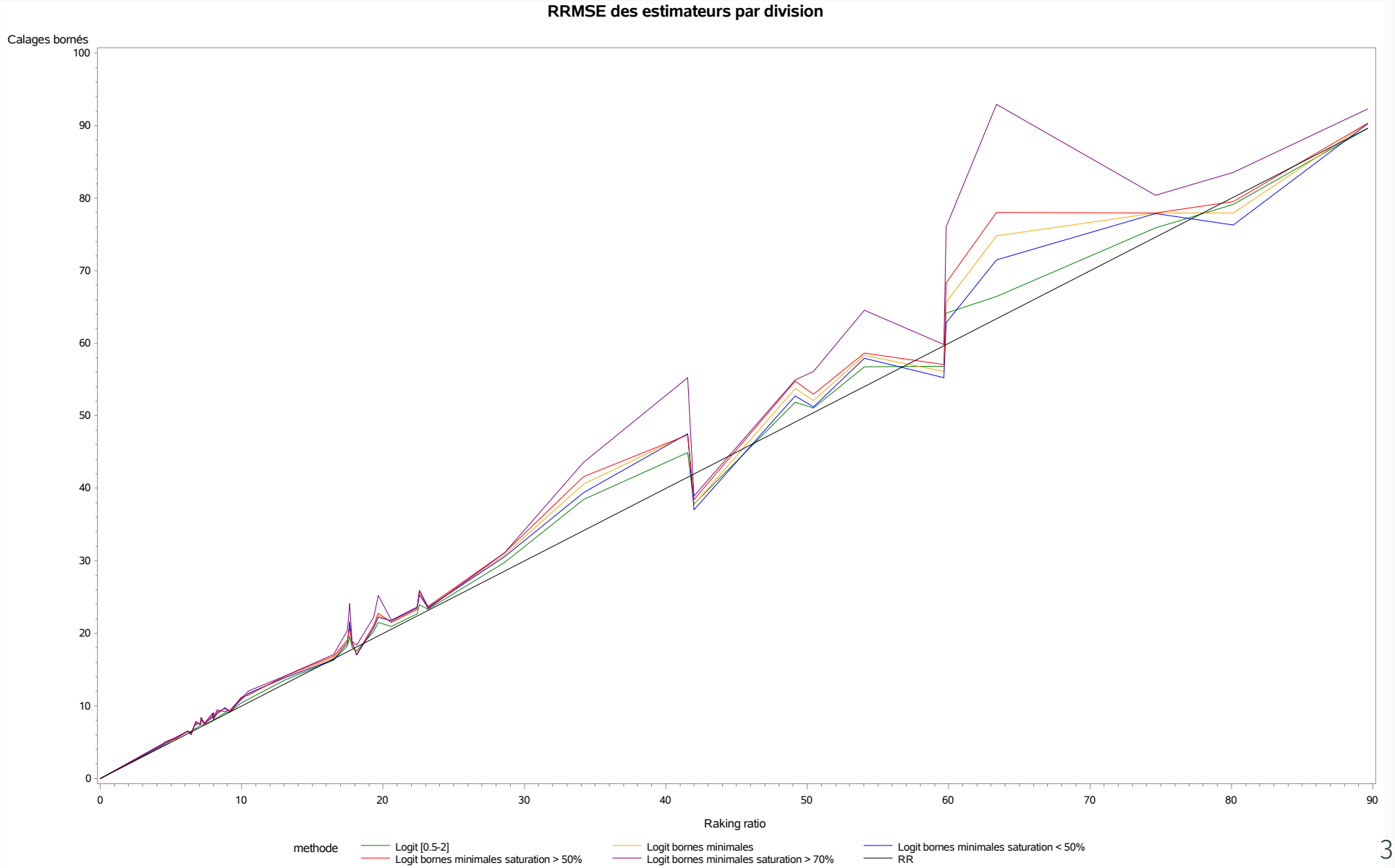
Distribution des rapports de poids – Logit bornes minimales



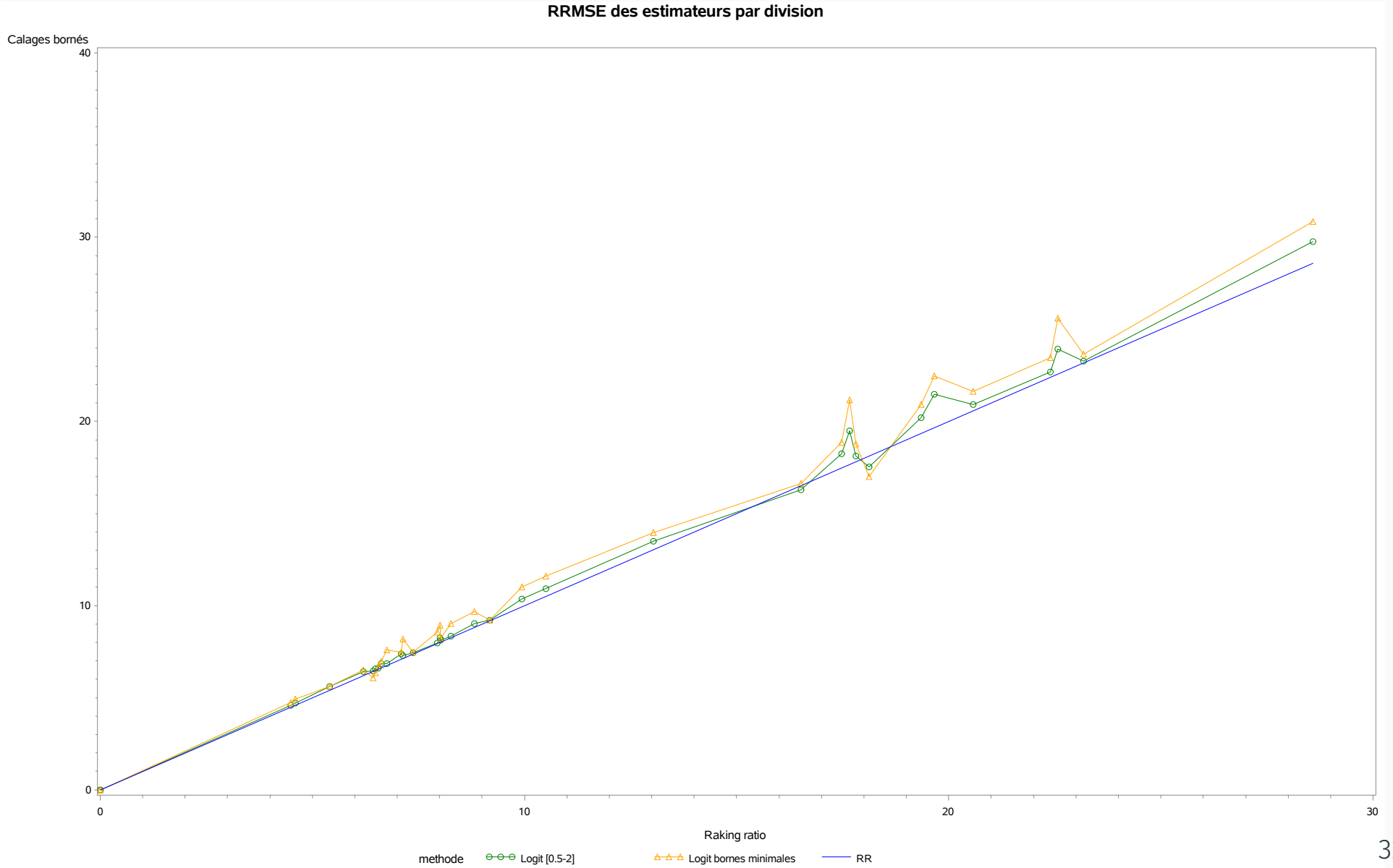
RRMSE des estimateurs par division (1)



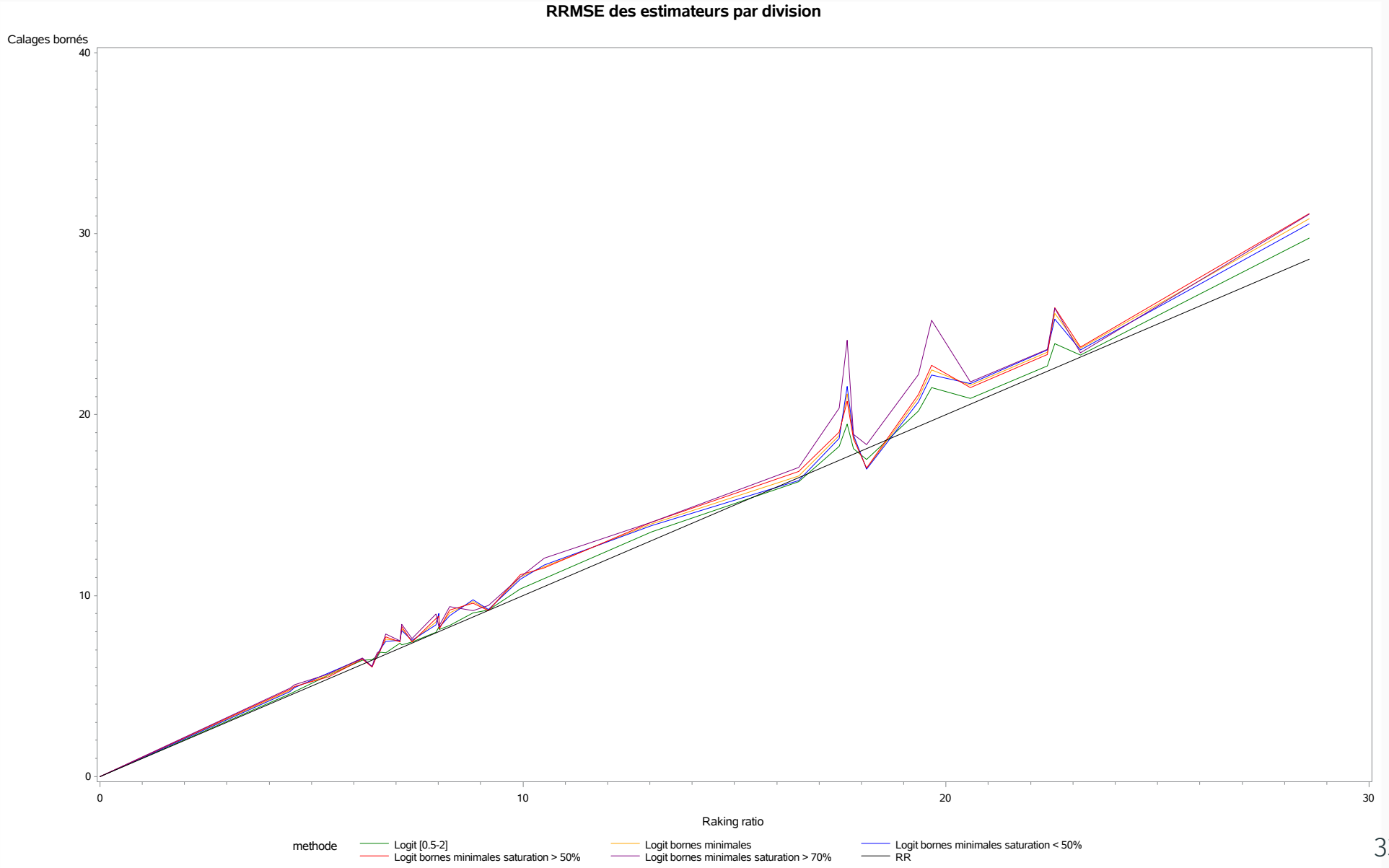
RRMSE des estimateurs par division (2)



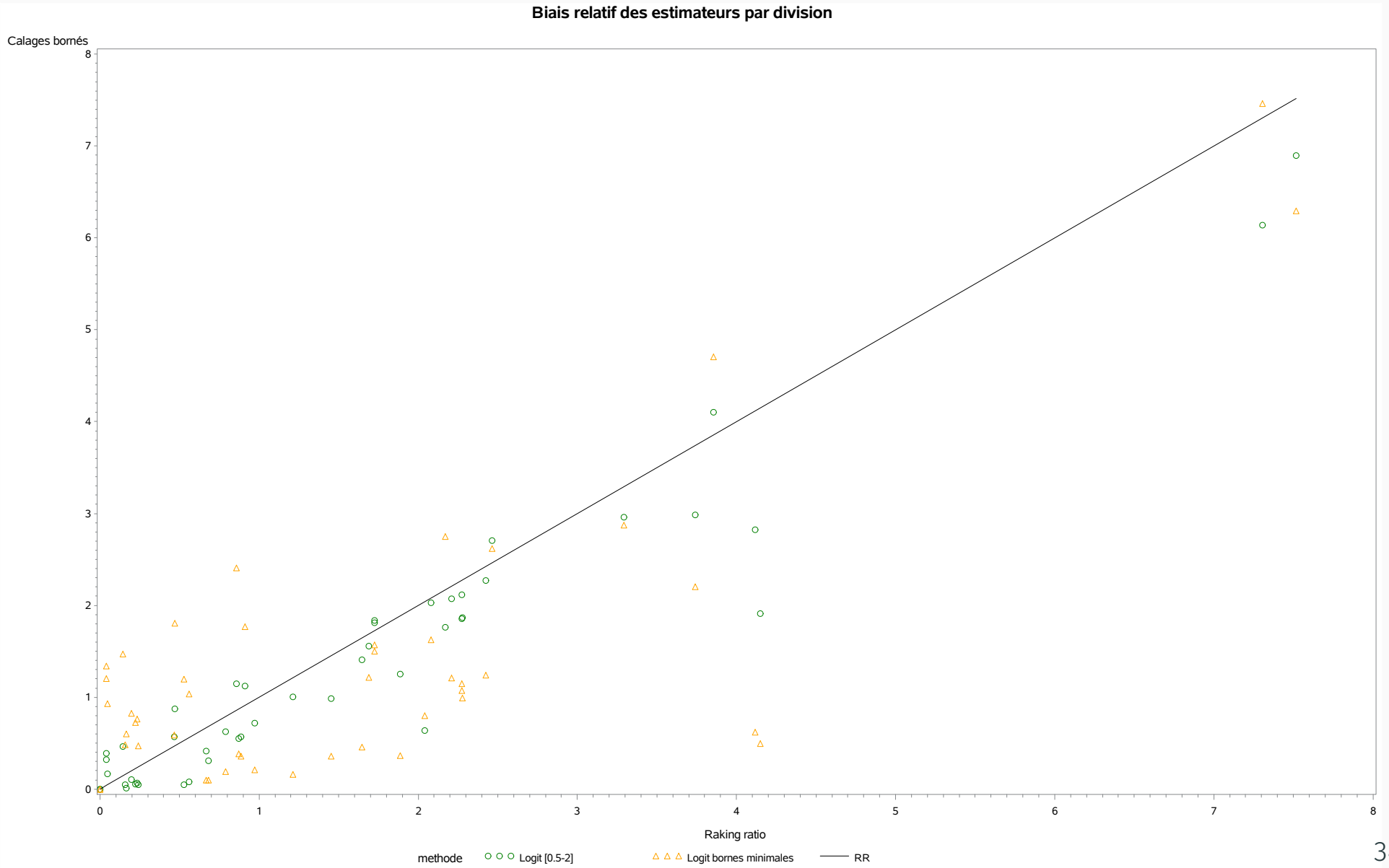
RRMSE des estimateurs par division (3)



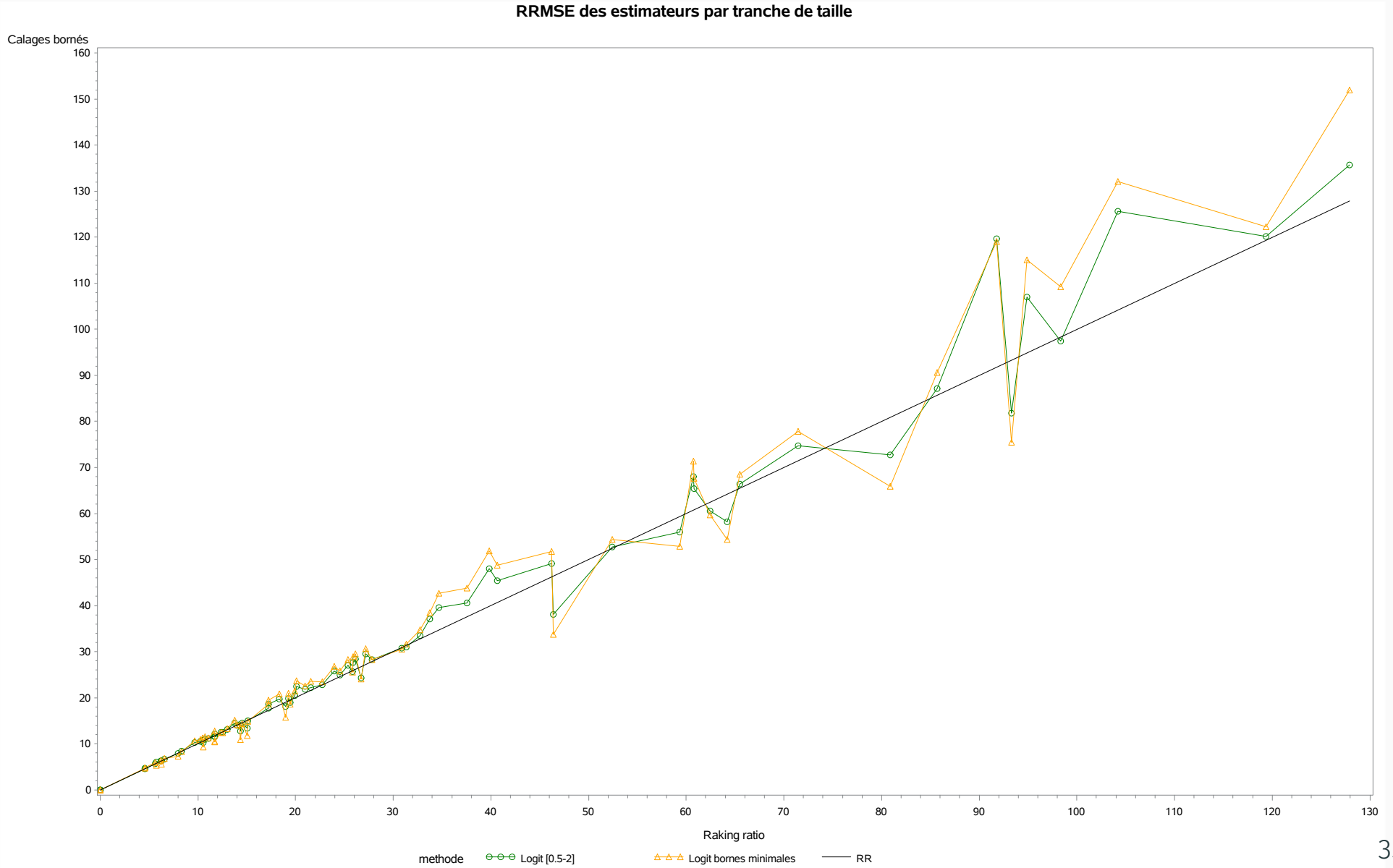
RRMSE des estimateurs par division (4)



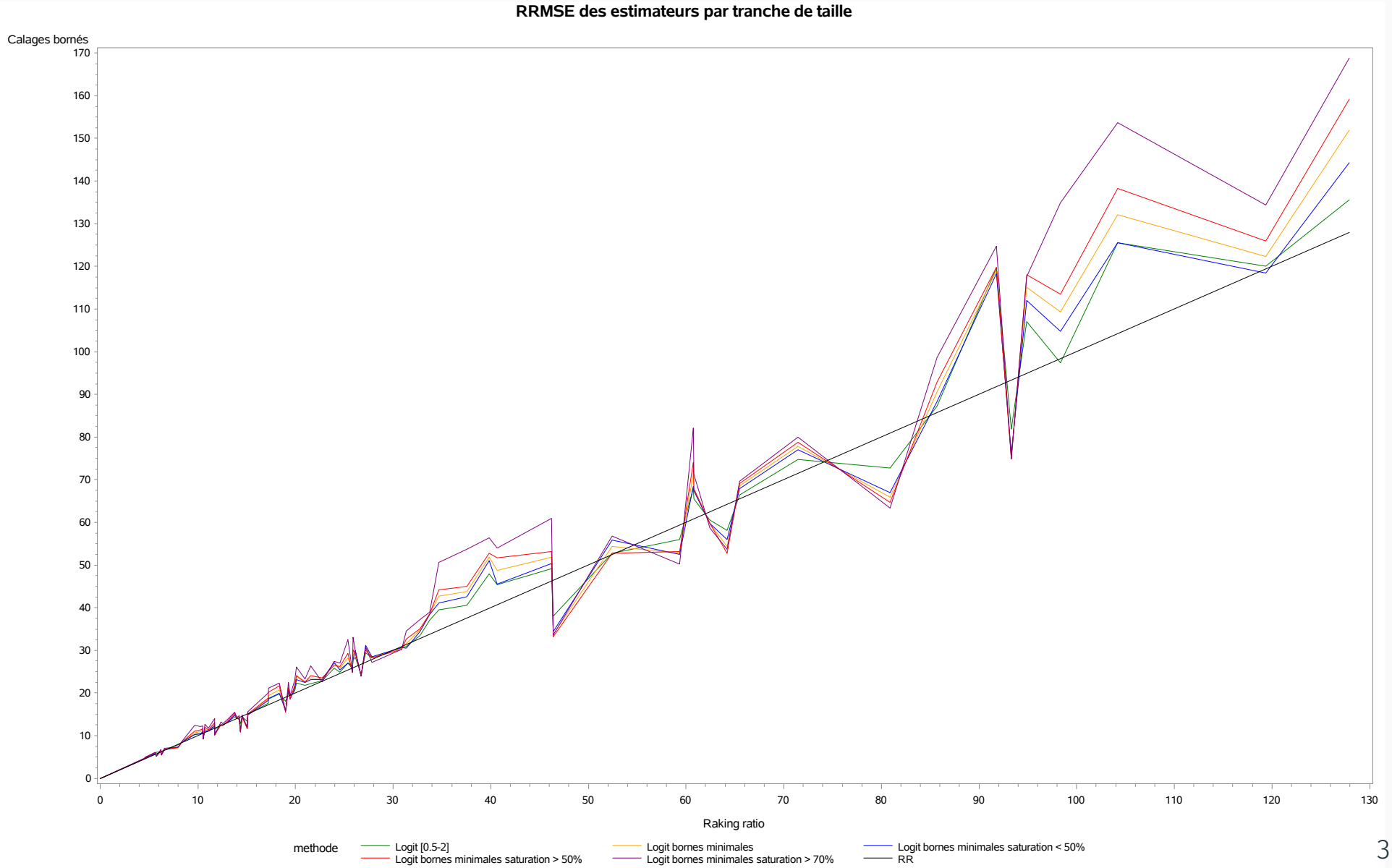
Biais relatif absolu des estimateurs par division



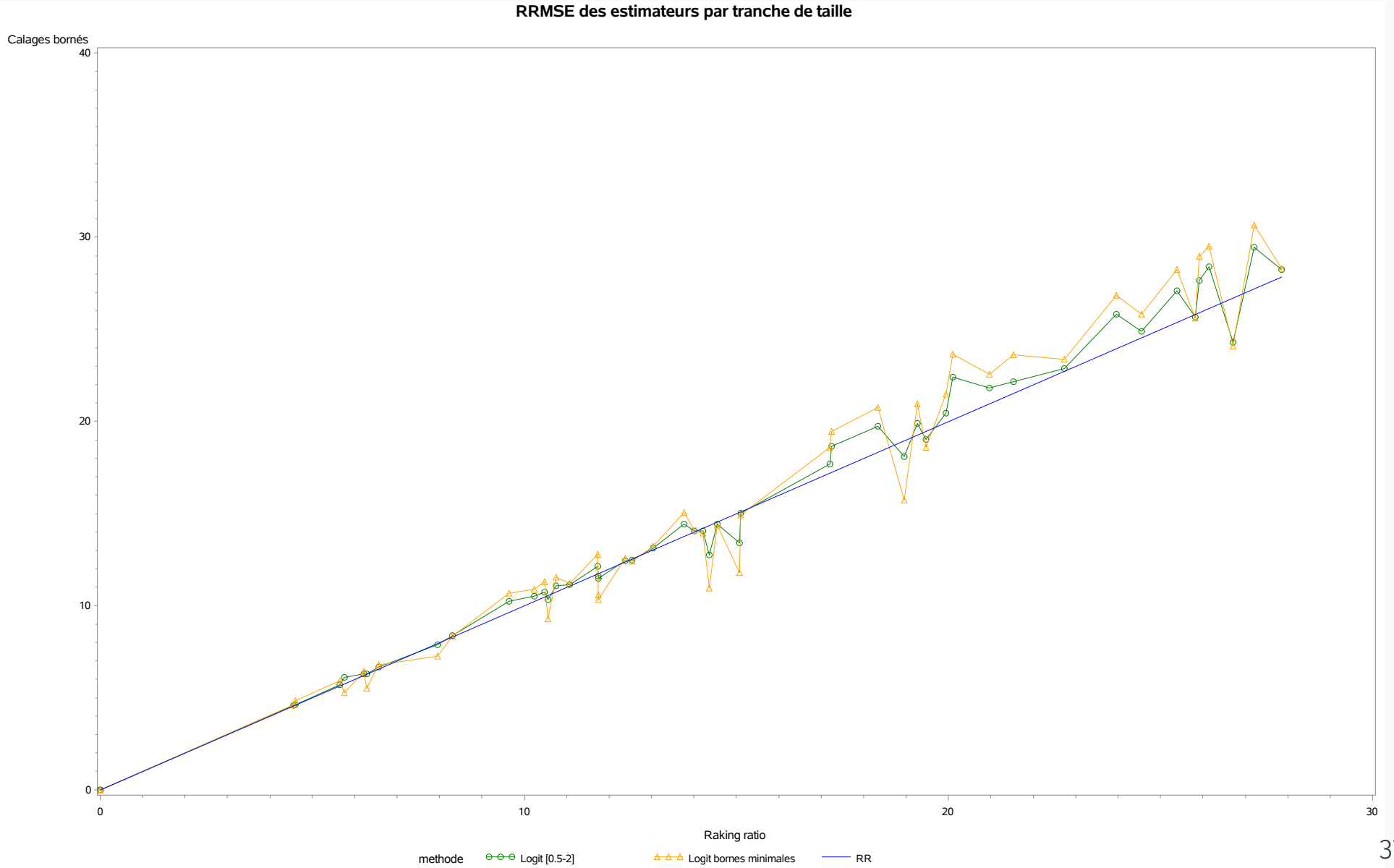
RRMSE des estimateurs par tranches d'effectif (1)



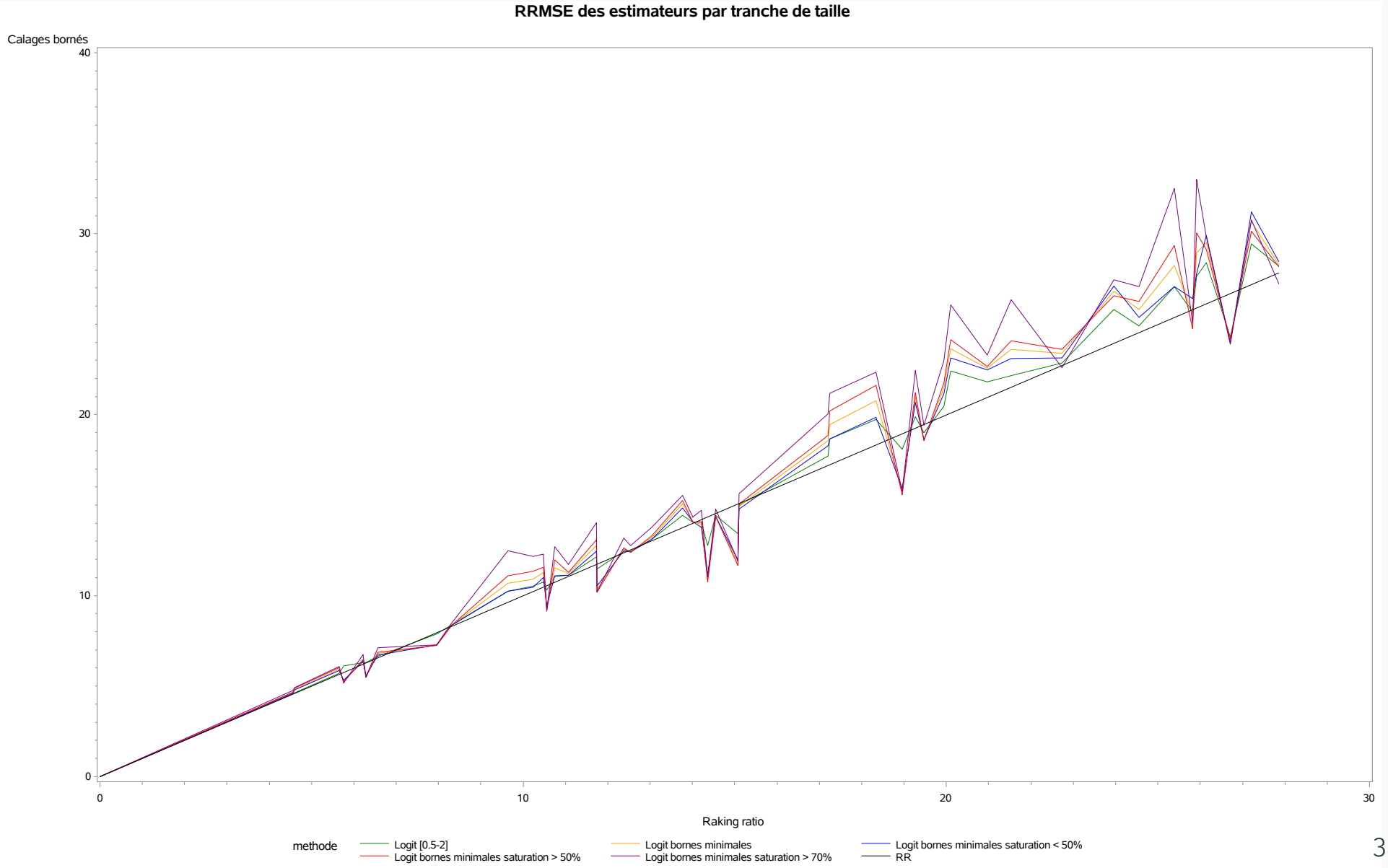
RRMSE des estimateurs par tranches d'effectif (2)



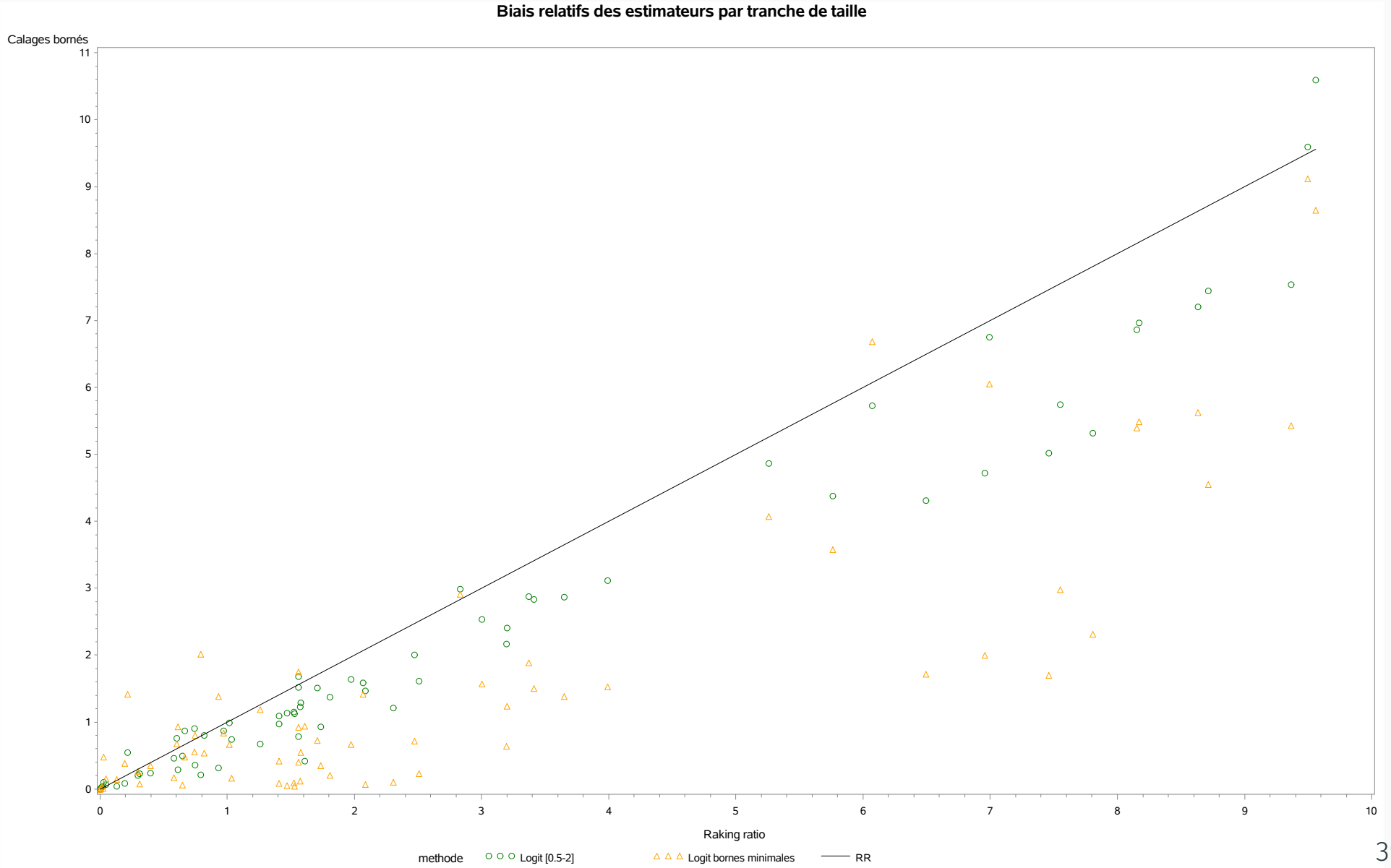
RRMSE des estimateurs par tranches d'effectif (3)



RRMSE des estimateurs par tranches d'effectif (4)



Biais relatif absolu des estimateurs par tranches d'effectif



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 !