
Insee
Mesurer pour comprendre

# Le package R Icarus

## Calage sur marges en R

Antoine Rebecq
INSEE

# Le package R icarus

1. L'écosystème R - calage sur marges

2. Calage sur marges avec Icarus

3. Variantes du calage sur marges

# L'écosystème R - calage sur marges

# R

- Logiciel et langage de programmation dédié aux statistiques
- Logiciel libre (*"free as in freedom"*)
- Communauté de développeurs : le CRAN (https://cran.r-project.org/) regroupe les différents packages (lien vers la taskview "Official statistics and survey methodology" (https://cran.r-project.org/web/views/OfficialStatistics.html))
- Langage avec des fonctions structurées : facilite la réutilisation et l'échange du code
- *git* permet de travailler à plusieurs sur un même projet et de suivre les modifications
- Beaucoup de packages ont leur code source hébergé sur Github (https://github.com)

# Le calage sur marges

Le calage permet de proposer un estimateur avec de bonnes propriétés statistiques vérifiant les contraintes de calage :

- Cohérence entre les sources statistiques
- Réduction de variance pour certaines estimations
- Méthode adaptée pour les échantillons non probabilistes

*Praticiens du calage* : statistique officielle et non officielle, recherche en épidémiologie, biologie, sociologie, etc.

# Implémentations du calage sur marges

- En SAS :
  - Calmar (Sautory (1993), disponible en téléchargement sur le site de l'INSEE (http://www.insee.fr/fr/methodes/default.asp?page=outils/calmar/accueil_calmar.htm))
  - Calmar 2 (Le Guennec and Sautory (2002))
- En R :
  - package sampling (Tillé and Matei (2015))
  - package survey (Lumley (2016))

Objectif de Icarus (Icarus CAle et Redresse les Unités en Sondages) : proposer un package avec des fonctionnalités spécifiquement dédiées au calage, avec une interface proche de Calmar.

# Calage sur marges avec Icarus

# Installation du package Icarus

Pour installer le package :

```
install.packages("icarus")
```

Ou pour avoir la dernière version à jour du package (nécessite le package devtools)

```
install.packages("devtools")
library(devtools)
install_github("haroine/icarus")
```

Page github (https://github.com/haroine/icarus) du package Icarus : code, historique, wiki (https://github.com/haroine/icarus/wiki), issues (https://github.com/haroine/icarus/issues)

# Calage sur marges simple

On considère une population de taille 300 (salariés d'une entreprise). On veut mesurer la fréquentation mensuelle des cinémas par les salariés. On dispose des structures pour les variables catégorielles et des totaux pour les variables quantitatives suivantes :

- le nombre d'employés par département (2 positions)
- le niveau hiérarchique du salarié dans l'entreprise (3 positions)
- le sexe (2 positions)
- le salaire (variable quantitative)

```
library(icarus)

N <- 300 ## Taille de la population
```

```
data_ex2 ## Données d'enquête
```

| | ID | SERVICE | CATEG | SEXE | SALAIRE | CINEMA | POIDS |
|---|---|---|---|---|---|---|---|
| **1** | a01 | 1 | 1 | 1 | 1000 | 1 | 10 |
| **2** | a02 | 1 | 2 | 2 | 1100 | 2 | 10 |
| **3** | a03 | 2 | 2 | 2 | 1500 | 4 | 10 |
| **4** | a04 | 2 | 3 | 1 | 2300 | 15 | 10 |
| **5** | a05 | 2 | 1 | 1 | 1000 | 2 | 10 |
| **6** | a06 | 1 | 1 | 2 | 500 | 3 | 10 |
| **7** | a07 | 2 | 2 | 2 | 1000 | 1 | 10 |
| **8** | b01 | 1 | 3 | 2 | 2000 | 0 | 20 |
| **9** | b02 | 1 | 1 | 1 | 2100 | 0 | 20 |
| **10** | b03 | 2 | 2 | 1 | 2000 | 3 | 20 |
| **11** | b04 | 2 | 1 | 2 | 3200 | 6 | 20 |
| **12** | b05 | 1 | 1 | 2 | 1800 | 0 | 20 |
| **13** | b06 | 1 | 2 | 1 | 2800 | 0 | 20 |
| **14** | b07 | 1 | 3 | 1 | 1100 | 1 | 20 |
| **15** | b08 | 2 | 1 | 2 | 2500 | 1 | 20 |

La table des marges de calage possède un format proche de celle de Calmar. On indique :

- le nom de la variable dans la table échantillon
- le nombre de modalités de la variable
- les marges correspondantes

```
mar1 <- c("categ",3,80,90,60)
mar2 <- c("sexe",2,140,90,0)
mar3 <- c("service",2,100,130,0)
mar4 <- c("salaire",0,470000,0,0)
margins <- rbind(mar1, mar2, mar3, mar4)
```

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| **mar1** | categ | 3 | 80 | 90 | 60 |
| **mar2** | sexe | 2 | 140 | 90 | 0 |
| **mar3** | service | 2 | 100 | 130 | 0 |
| **mar4** | salaire | 0 | 470000 | 0 | 0 |

## Calage avec la méthode du raking ratio

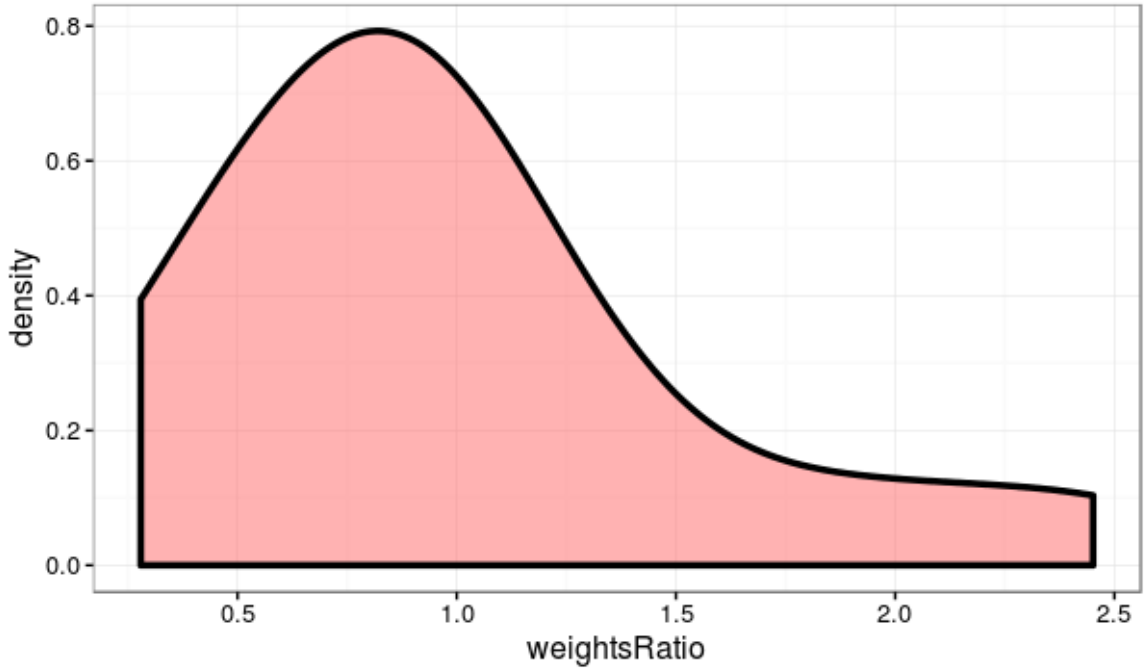
```
wCalesRaking <- calibration(data=data_ex2, marginMatrix=margins,
                            colWeights="poids", method="raking",
                            popTotal = 230)
```

```
##
## ################### Summary of before/after weight ratios ###################
## Calibration method : raking
## Mean : 0.9852
##     0%     1%    10%    25%    50%    75%    90%    99%   100%
## 0.2792 0.2980 0.4528 0.6389 0.8833 1.0450 1.7080 2.3755 2.4514
##
## ##################### Comparison Margins Before/After calibration ####################
## $Total
## Before calibration  After Calibration             Margin
##                230                 230                230
##
##
##
##
##
```

```
##  $categ
##    Before calibration After Calibration Margin
##  1              47.83             34.78  34.78
##  2              30.43             39.13  39.13
##  3              21.74             26.09  26.09
##
##  $sexe
##    Before calibration After Calibration Margin
##  1              47.83             60.87  60.87
##  2              52.17             39.13  39.13
##
##  $service
##    Before calibration After Calibration Margin
##  1              56.52             43.48  43.48
##  2              43.48             56.52  56.52
##
##  $salaire
##  Before calibration  After Calibration             Margin
##              434000             470000             470000
```

## Sorties :

- Quantiles de la distribution des rapports de poids (ou *facteurs de calage*) $g_k = \frac{w_k}{d_k}$
- Stats sur les estimateurs des variables de calage (variables catégorielles exprimées en pourcentages par défaut)
- Disponible dans l'output ou en TeX à l'aide de la commande *marginStats*
- Sortie graphique : graphe de la densité des facteurs de calage



## Calage avec la méthode logit

```
wCalesLogit <- calibration(data=data_ex2, marginMatrix=margins,
                           colWeights="poids", method="logit",
                           bounds=c(0.4,2.2), popTotal = 230)
```

```
##
## ################### Summary of before/after weight ratios ###################
## Calibration method : logit
##   L bound : 0.4
##   U bound : 2.2
## Mean : 0.9701
##     0%     1%    10%    25%    50%    75%    90%    99%   100%
## 0.4026 0.4042 0.4193 0.5186 0.8564 1.1952 1.7807 2.1475 2.1707
##
## ################### Comparison Margins Before/After calibration ###################
## $Total
## Before calibration  After Calibration             Margin
##                230                 230                230
##
## $categ
##   Before calibration After Calibration Margin
## 1              47.83             34.78  34.78
## 2              30.43             30.13  30.13
```

# Fonctionnalités

- Utilisation de l'inverse généralisée
- Facteur d'échelle en présence de non-réponse
- Possibilités d'indiquer les marges catégorielles en pourcentages

# Variantes du calage sur marges

# Le calage pénalisé

```
wLogitSerres <- calibration(data=data_ex2, marginMatrix=margins,
                            colWeights="poids", method="logit",
                            bounds=c(0.6,2.0), popTotal = 230)
```

```
## Warning in calibAlgorithm(Xs, d, total, q, inverseDistance,
## updateParameters, : No convergence
```

Le ***calage pénalisé*** (Bocci and Beaumont (2008)) permet de relâcher la contrainte d'exactitude sur les estimateurs pour les variables de calage. Un paramètre supplémentaire apparaît : le *gap*, qui spécifie l'étendue de la distribution des poids souhaitée (similaire à la méthode logit).

Utilité :

- lorsque l'on a beaucoup de contraintes de calage
- lorsque la convergence du calage est difficile
- lorsque l'on ne veut pas accorder trop d'importance à une "marge" de calage (variable de contrôle)

Applications à l'INSEE : présentation donnée le 15 mars (http://www.insee.fr/fr/insee-statistique-publique/connaitre/colloques/sms/sms-150316-rebecq.pdf)

```
costs <- c(1,1,1,Inf)

wPenalise <- calibration(data=data_ex2, marginMatrix=margins,
                                    colWeights="poids", costs=costs,
                                    gap=1.4, popTotal=230)
```

```
## ################### Summary of before/after weight ratios ###################
## Calibration method : linear
## Mean : 0.9661
##     0%     1%    10%    25%    50%    75%    90%    99%   100%
## 0.1950 0.2435 0.5774 0.7713 0.9579 1.1268 1.4663 1.5818 1.5949
##
## ################### Comparison Margins Before/After calibration ###################
## Careful, calibration may not be exact
## $Total
## Before calibration  After Calibration             Margin
##                230                 230                230
##
## $categ
##   Before calibration After Calibration Margin
## 1              47.83             40.34  34.78
## 2              30.43             36.55  39.13
## 3              21.74             23.11  26.09
##
##
##
##
##
##
```

```
## $sexe
##   Before calibration After Calibration Margin
## 1              47.83             56.04  60.87
## 2              52.17             43.96  39.13
##
## $service
##   Before calibration After Calibration Margin
## 1              56.52             47.89  43.48
## 2              43.48             52.11  56.52
##
## $salaire
## Before calibration  After Calibration             Margin
##             434000             470000             470000
```

# Le calage sur bornes minimales

```
wCalMin <- calibration(data=data_ex2, marginMatrix=margins,
                       colWeights="poids", method="min", popTotal=230)
```

```
## Solution found for calibration on minimal bounds:
## L = 0.334476843910806
## U = 1.65866209262436
##
## ################### Summary of before/after weight ratios ###################
## Calibration method : min
## Mean : 0.9693
##     0%     1%    10%    25%    50%    75%    90%    99%   100%
## 0.3344 0.3344 0.3344 0.3345 0.6451 1.6587 1.6587 1.6587 1.6587
##
## ################### Comparison Margins Before/After calibration ###################
## $Total
## Before calibration  After Calibration             Margin
##                230                 230                230
##
## $categ
##   Before calibration After Calibration Margin
## 1              47.83             34.78  34.78
```

# Merci !

Lien vers les slides : http://nc233.com/icarus (http://nc233.com/icarus)

## Références

[1] J. Bocci and C. Beaumont. "Another look at ridge calibration". In: *Metron* 66.1 (2008), pp. 5-20.

[2] J. Le Guennec and O. Sautory. "Calmar 2: Une nouvelle version de la macro calmar de redressement d'échantillon par calage". In: *Journées de Méthodologie Statistique, Paris. INSEE* (2002).

[3] T. Lumley. *survey: analysis of complex survey samples*. R package version 3.31-2. 2016.

[4] O. Sautory. "La macro Calmar. Redressement d'un échantillon par calage sur marges". In: *Document F9310, DSDS, INSEE* (1993).

[5] Y. Tillé and A. Matei. *sampling: Survey Sampling*. R package version 2.7. 2015. .