
1431
Université
de Poitiers

Projet Analyses Statistiques des Sondages

Rédigé par Kodzo LIMA
Chargé de l'UE : M.Slaoui YOUSRI

Mars 2023

Table des matières

1 Introduction

Dans l'étude du module Analyses Statistiques des Sondages,il nous est demlandé dans la partie projet du module de faire un rapport sur trois fonctions dans les packages "sampling" et "tydiverse" en faisant sortir les parties mathématiques,leurs rôles,les méthodes et l'algorithme ou la pratique sur R et les références utilisées

dans le rapport ridigé avec Latex.

2 Estimateurs Par Calage (Calibration estimator and its variance estimation)

La technique d'estimation par calage a été introduite par Deville et Särndal en 1992. Le calage est une méthode d'ajustement des poids d'un estimateur linéaire pondéré inital dans le but d'obtenir un nouvel estimateur linéaire qui estime parfaitement un certain nombre de totaux connues sur la population U (qu'on appellera totaux de calage ou totaux de contrôle).

2.1 Rôle de l'Estimation par Calage

L'idée est d'utiliser des informations auxiliaires pour obtenir une meilleure estimation d'une statistique démographique. La méthode du calage permet de construire une classe d'estimateurs pondérés linéaires, appelés estimateurs par calage.

2.2 La Méthode de calage

* calculer des poids qui tiennent compte de l'information auxiliaire spécifiée et sont soumis à des contraintes précisées par une ou plusieurs équations de calage.

* utiliser ces poids pour calculer des estimations linéairement pondérées des totaux et d'autres paramètres de population finie, c'est à dire multiplier la valeur de la variable par le poids et faire la somme sur un ensemble d'unités observées.

* Se fixer l'objectif d'obtenir des estimations presque sans biais sous le plan de sondage, à condition qu'il n'y ait pas d'erreur de non réponse ni d'autres erreurs non dues à l'échantillonnage.

2.3 Partie Mathématique

Nous allons présenter la méthode du calage dans le cas classique où l'on cale sur des totaux tx connus sur la population U et où l'estimateur initial est l'estimateur par expansion. Les estimateurs par calage sont de la forme :

$$\hat{t}_y, CAL = \sum_{k \in S} w_k y_k \ ,$$

où les poids de calage w_k , $k \in S$, vérifient les équations de calage

$$\sum_{k \in S} w_k x_k = t_x \ ,$$

où x est un vecteur de variables auxiliaires dont on connaît les valeurs pour l'échantillon s et le total tx.

2.4 Calage par minimisation de la distance entre les poids d'échantillonnage et les poids de calage (minimum distance method)

Les poids de calage sont obtenus par résolution du programme d'optimisation suivant : $\min(w_k, k \in S)$ $\sum_{k\in S} G_k(w_k, d_k)$ sous les contraintes : $\sum_{k\in S} w_k x_k = t_x$, où $G_k(w, d)$ est une pseudo-distance qui mesure la distance entre les poids d'échantillonnage et les poids de calage

On suppose que la fonction $G_k(w, d)$ est dérivable par rapport à w, strictement convexe, avec une dérivée partielle $g_k(w, d)$ $=\partial G_k(w, d)/\partial w$ continue et telle que $g_k(d, d) = 0$. Habituellement la fonction de distance est choisie de telle sorte que $g_k(w, d) = g(w/d)/q_k$, où q_k est une pondération judicieusement choisie , $g()$ une fonction continue, dérivable en 1, strictement croissante, telle que $g(1) = 0$ et $g(1) = 1$. On note $F(u) = g^1$(u),la fonction réciproque de $g()$ que l'on nommera fonction de calage par la suite. On suppose en outre que $F"(0) < \infty$. On résout le programme d'optimisation avec un lagrangien et on obtient les poids de calage $w_k = d_k F(q_k \hat{\lambda}^T x_k)$ où $\hat{\lambda}$ est solution (s'il en existe) des équations de calage :

$$\sum_{k\in S} d_k F(q_k \hat{\lambda}^T x_k) = \sum_{l\in U} x_l$$

On peut trouver de nombreux exemples de fonction de distance dans Deville et Särndal (1992). On citera ici, à titre d'exemple, trois distances qui correspondent à trois fonctions de calage particulièrement intéressantes.

* La première est la distance du χ^2 :

$$G_k(w_k, d_k) = (w_k d_k)^2/(2q_k d_k)$$

Cette distance correspond à la fonction de calage dite linéaire F(u) = 1 + u et les poids de calage obtenus sont appelés poids de calage linéaires et valent :

$$w_k = d_k(1 + q_k \lambda^T x_k)$$

L'estimateur obtenu avec la fonction de calage linéaire est égal à l'estimateur GREG linéaire t_y le modèle d'ajustement linéaire (Särndal et al. ,1992) où les

poids c_k valent q^1.

$$\hat{t}_y, AM = \hat{B}_\pi^T t_x + \sum_{k \in U} d_k (y_k - \hat{B}_\pi^T x_k)$$

où

$$\hat{B}_\pi^T = (\sum_{k \in U} c^- 1_k d_k x_k x_k^T)^- 1 \sum_{k \in U} c^- 1 d_k x_k y_k$$

* La deuxième distance

$$G_k(w_k, d_k) = w_k log(w_k/d_k) w_k + d_k$$

qui correspond à la fonction de calage exponentielle F (u) = exp(u).

Les poids de calage exponentiels qui ont l'avantage d'être positifs.

$$w_k = d_k exp(q_k \lambda^T x_k)$$

En outre, la fonction de calage exponentielle est égale à sa fonction dérivée $h(u) = F(u) = exp(u)$.

* Troisième distance

IL s'agit dela distance dite logit qui permet d'obtenir des poids bornés

$$w_k = d_k \frac{L(U-1) + U(1-L) exp(A q_k \lambda^T x_k)}{(U-1) + (1-L) exp(A q_k \lambda^T x_k)}$$

où $A = (U-L)/(1-L)(U-1)$, et L et U sont les bornes inférieure et supérieure de la fonction de calage.

Par ailleurs on utilise la variance approchée de l'estimateur GREG pour construire un estimateur de variance des estimateurs par calage. On rappelle que la variance approchée de l'estimateur GREG qui vaut :

$$Var(\hat{t}_y, GREG) = (\pi_{k,l} - \pi_k \pi_l) d_k E_k d_l E_l$$

où $E_k = y_k B^T x_k$ et $B = (\sum_{k \in U} q_k x_k x_k{}^T)^- 1 \sum_{k \in U} q_k x_k y_k$ (Ici U est la population).
Les E_k sont les résidus de l'ajustement par les moindres carrés de y_k en fonction de x_k sur U, pondérés par q_k. L'estimateur par calage sera donc d'autant plus précis que l'ajustement linéaire sera bon.

L'estimateur de variance proposée pour les estimateurs par calage est :

$$\hat{V}(\hat{t}_{y,CAL}) = \sum_{k\in S}\sum_{l\in S}\frac{\pi_{k,l}-\pi_k\pi_l}{\pi_{k,l}}w-ke-kw-le-l$$

où $e_k = y_k\hat{B}_w^T x_k$ et

$$\hat{B}_w = (\sum_{k\in S} w_k q_k x_k {x_k}^T)^{-}1\sum_{k\in S} w_k q_k x_k y_k$$

.

2.5 La pratique de calage sur R

Après intallation des packages : "sampling","survey"et "icarus" ,nous allons dès à présent faire du calage
Nous allons utiliser dans cette partie les jeux de données suivantes tirés de `http://nc233.com/icarus/#11` de l'INSEE qui a inspiré la rédaction de cette partie

id	service	categ	sexe	salaire	cinema	poids
a0	1	1	1	1000	1	10
a01	1	2	2	1100	2	10
a02	2	2	2	1500	4	10
a03	2	3	1	2300	15	10
a04	2	1	1	1000	2	
a05	1	1	2	500	3	
a06	2	2	2	1000	1	10
a07	1	3	1	2000	0	20
b01	1	1	1	2100	0	20
b02	2	2	2	2000	3	20
b03	2	1	2	3200	6	20
b04	1	1	1	1800	0	20
b05	1	2	1	2800	0	20
b06	1	3	2	1100	1	20
b07	2	1	2	2500	1	20

* Calage sur marges simple

```
library(icarus)
N <- 300 ## Taille de la population
PS ## Données d'enquête
ar1 <- 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)
print(margins)
```

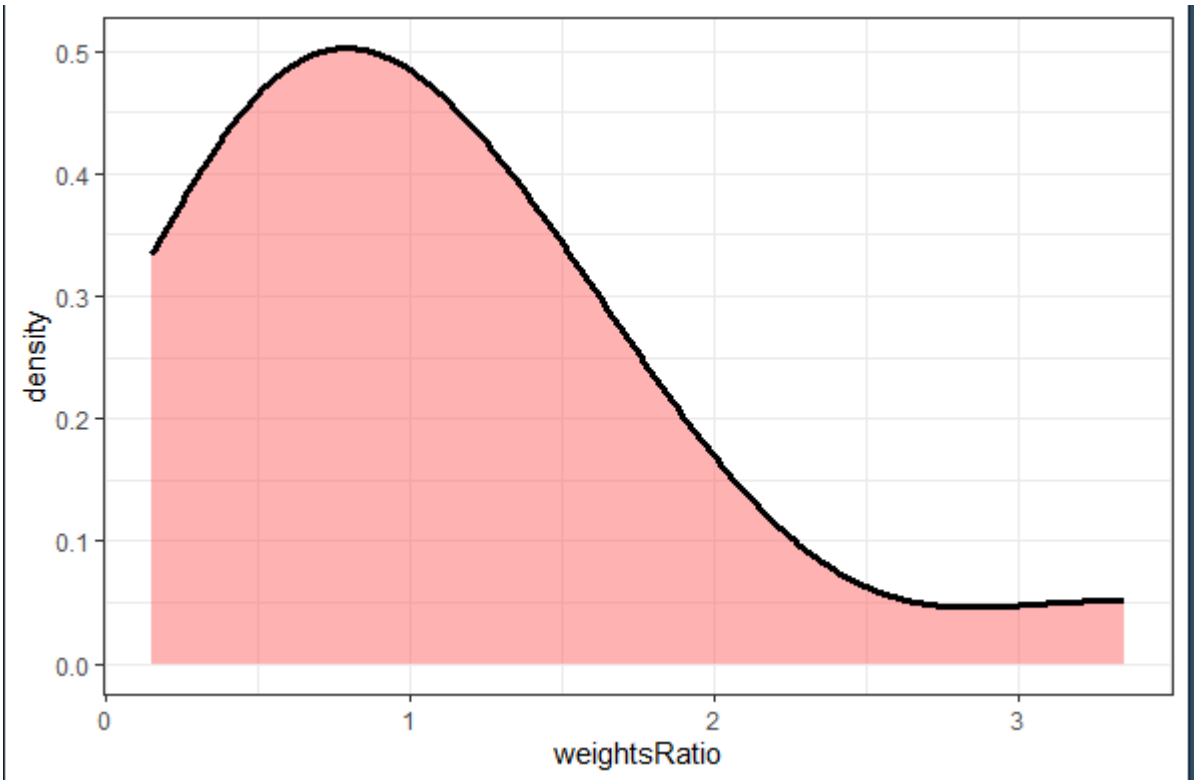
```
      [,1]       [,2] [,3]     [,4]  [,5]
mar1 "categ"    "3"  "80"     "90"  "60"
mar2 "sexe"     "2"  "140"    "90"  "0"
mar3 "service"  "2"  "100"    "130" "0"
mar4 "salaire"  "0"  "470000" "0"   "0"
```

* Calage avec la méthode du raking ration

```
##Calage avec la méthode du raking ratio
wCalesRaking <- calibration(data=PS, marginMatrix=margins,
colWeights="poids", method="raking",popTotal = 230)
```

```
################### Summary of before/after weight ratios ###################
Calibration method : raking
Mean : 1.053
    0%     1%    10%    25%    50%    75%    90%    99%   100%
0.1564 0.1720 0.3156 0.6193 0.7466 1.3014 1.7459 3.1242 3.3471

################### 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
```

Les résultats sortis par R (Rstudio) sont les Quantiles de la distribution des rapports de poids (ou facteurs de calage) $g_k = w_k/d_k$ et les Statatistiques sur les estimateurs des variables de calage (variables catégorielles exprimées en pourcentages par défaut)



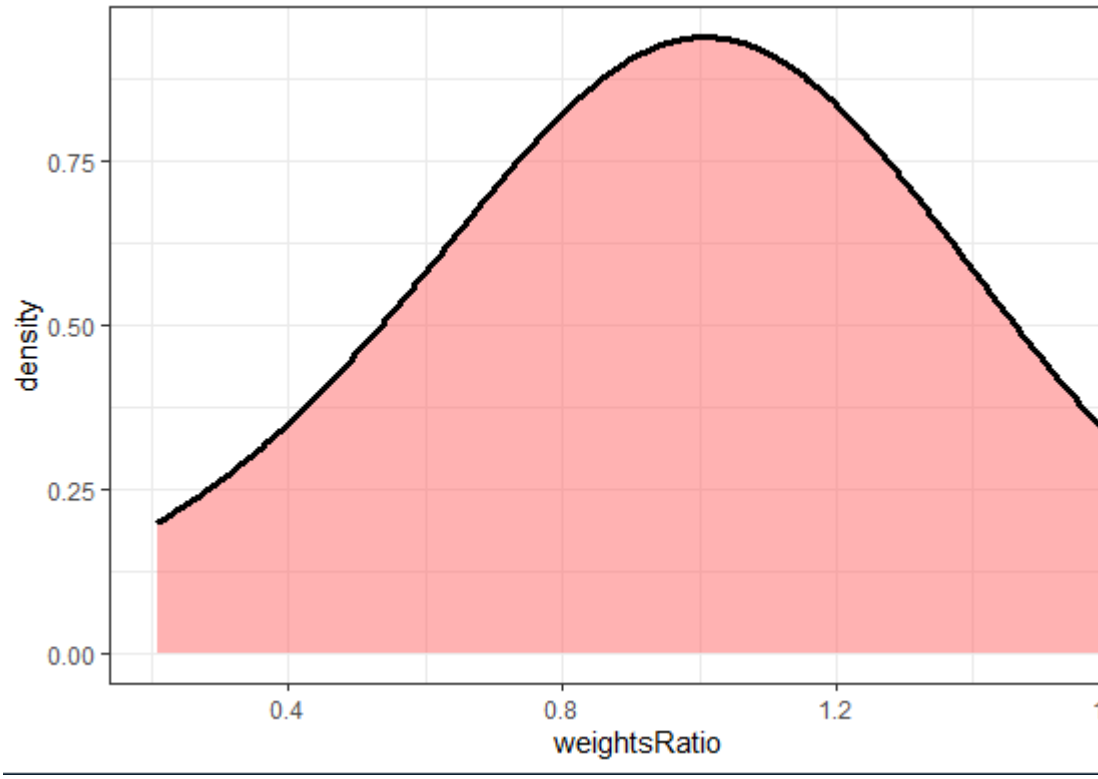
graphe de la densité des facteurs de calage

* Le calage pénalisé

```
 ##Le calage pénalisé
costs <- c(1,1,1,Inf)
wPenalise <- calibration(data=PS, marginMatrix=margins,
colWeights="poids", costs=costs,gap=1.4, popTotal=230)
```

```
################### Summary of before/after weight ratios ###################
Calibration method : linear
Mean : 0.9764
    0%     1%    10%    25%    50%    75%    90%    99%   100%
0.2097 0.2668 0.6314 0.7890 0.9884 1.1666 1.3210 1.5808 1.6096

################### 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             42.82  34.78
2              30.43             34.41  39.13
3              21.74             22.78  26.09

$sexe
  Before calibration After Calibration Margin
1              47.83             53.02  60.87
2              52.17             46.98  39.13

$service
  Before calibration After Calibration Margin
1              56.52             50.57  43.48
2              43.48             49.43  56.52

$salaire
Before calibration  After Calibration             Margin
            434000              470000             470000
```

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

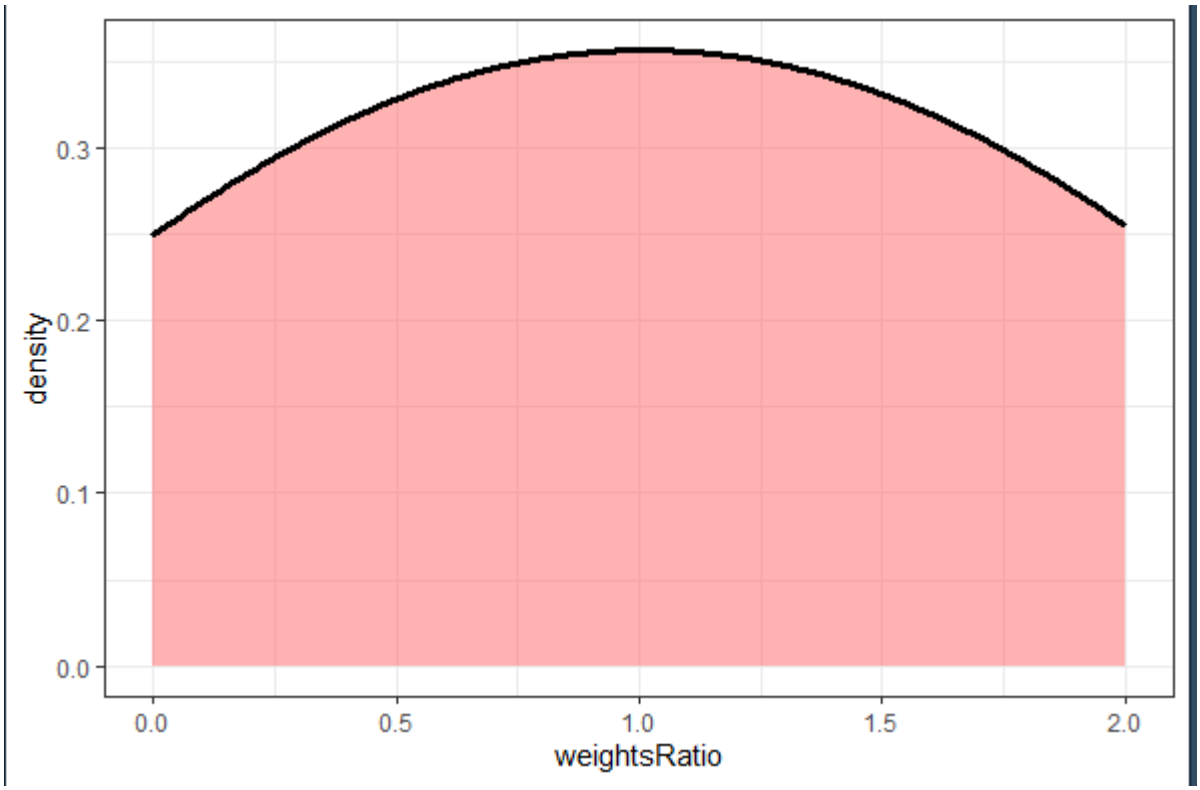
graphe de la densité des facteurs de calage

* Le calage sur bornes minimales

```
 ##Le calage sur bornes minimales
install.packages("Rglpk")
library(Rglpk)
wCalMin <- calibration(data=PS, marginMatrix=margins, colWeights="poids", method="min", popT
```

```
Solution found for calibration on minimal bounds:
L = 4.44089209850063e-16
U = 2

################### Summary of before/after weight ratios ###################
Calibration method : min
Mean : 1.0127
    0%     1%    10%    25%    50%    75%    90%    99%   100%
0.0000 0.0000 0.0000 0.5436 0.8162 1.5304 2.0000 2.0001 2.0001

################### 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
```

Les résulats de R (Rstiduo) sont donc les nombres U et L définis plus haut dans le calage sur bornes minimales.



graphe de la densité des facteurs de calage

3 Estimation d'un ratio

3.1 Rôle d'un ratio

Pour expliciter plus clairement le rôle d'un ratio prenom l'exemple suivant : La proportion d'électeurs qui, dans une élection présidentielle, choisissent un candidat particulier est le rapport :
Nombre de votants qui choisissent le candidat / Nombre de suffrages exprimés.Cette proportion doit être estimée comme un ratio car la taille de la population, c'est-à-dire le nombre d'électeurs qui votent n'est pas connue.

3.2 Partie Mathématique

Supposons une population U de ménages, y_k le revenu du ménage k et z_k le nombre de personnes composant le ménage. Le revenu moyen par tête dans cette population est le rapport des deux totaux de y_k et z_k sur une même population qui est inconnue.
Le ration est donné par la relation :

$$R = \frac{\sum_{k\in U} y_k}{\sum_{k\in U} z_k} = \frac{t_y}{t_z}$$

3.2.1 Cadre général de l'estimation d'un ratio

On dispose d'un plan de sondage de probabilités d'inclusion, k et kl. Un échantillon s est obtenu par ce plan et on observe y_k, z_k, $k \in s$. On estime le ratio R par le quotient des estimateurs de H-T(Horvitz-Thompson) des totaux :

$$\hat{R} = \frac{\hat{t}_{\prime y\pi}}{\hat{t}_{z\pi}}$$

C'est un estimateur non linéaire et on ne peut donc pas calculer exactement son espérance mathématique.
Nous en obtenons une expression approchée par une technique classique en sondages : la linéarisation.

Espérance mathématique et variance approchée de $\hat{R}$

Appelons f la fonction des totaux qui donne le ratio : $f(y, z) = y/z$ et écrivons le développement de Taylor à l'ordre 1 de f et au voisinage de $y_0 = t_y$ et $z_0 = t_z$. On obtient :

$$\hat{R} \approx \frac{t_y}{t_z} - \frac{R}{t_z}(\hat{t}_z - t_z) + \frac{1}{t_z}(\hat{t}_y - t_y)$$

ou

$$\hat{R} \approx R + \frac{1}{t_z}\sum_{k\in s} \frac{y_k - R_{zk}}{\pi_k} \quad (1)$$

Prenant l'espérance mathématique des deux expressions , on obtient : $E(\hat{R}) \approx R$.L'estimateur $\hat{R}$ est :

$$v_k = \frac{1}{t_z}(y_k - R_{zk})$$

est appelée linéarisée de $R = \frac{t_y}{t_z}$. On voit sur (1) que la variance linéarisée de $\hat{R}$, c'est-à-dire la variance du côté droit de (1), n'est autre que la variance de $\sum_{k\in s} \frac{v_k}{\pi_k}$,estimateur du total de la linéarisée.On peut donc appliquer les résultats obtenus pour l'estimation d'un total par les valeurs dilatées :

$$var(\hat{R}) \approx (\sum_{k\in s} \frac{v_k}{\pi_k}) = \sum\sum_{U} \Delta_{kl}\breve{v}_k\breve{v}_l$$

où $\breve{v}_k = v_k/\pi_k$. On ne connaît ni R ni t_z, on les remplace donc par $\widehat{R}$ et $\widehat{t_z}$ pour obtenir une estimation de la variance :

$$\widehat{var}(\widehat{R}) = \sum\sum_U \frac{\Delta_{kl}}{\pi_{kl}} \breve{\widehat{v}}_k \breve{\widehat{v}}_l$$

où $\widehat{v}_k = (y_k - \widehat{R}_{zk})/\widehat{t}_z$.

3.2.2 Estimation d'un ratio dans un plan SI

Par un plan SI(N,n) qui donne un échantillon s dans une population U, on obtient y_k, z_k, $k \in s$.
L'estimateur du ratio est :

$$\widehat{R} = \frac{\widehat{t}_y}{\widehat{t}_z} = \frac{\bar{y}_s}{\bar{z}_s}$$

On applique ensuite les formules spécifiques au plan SI pour l'estimation de la variance du total $\sum_U \widehat{v}_k$ de la linéarisée. On obtient ainsi :

$$\widehat{var}(\widehat{R}) = N^2(\frac{1}{n} - \frac{1}{N})S_{\widehat{vs}} = \frac{1}{\bar{z}_s^2}(\frac{1}{n} - \frac{1}{N})S^2_{y-\widehat{R}_{z,s}}$$

Avec

$$S^2_{\widehat{vs}} = \frac{1}{n-1}\sum_s (\widehat{v}_k - \bar{\widehat{v}})^2$$

3.3 Pratique de Ratio sur R

```
library(foreign)
library(survey)
Data <-
  read.dta(
    "https://stats.idre.ucla.edu/stat/books/sop/tab7pt1.dta" ,
    convert.factors = FALSE
  )
head(Data)
summary(Data[, c("pharmexp", "totmedex")])
```

Nous abtenons les six premières lignes et le résumé de notre jeu de données extraire directement sur le site de STATA.

	area	pharmexp	totmedex	totcnt	wt1
1	1	100000	300000	8	1.142857
2	2	50000	200000	8	1.142857
3	3	75000	300000	8	1.142857
4	4	200000	600000	8	1.142857
5	5	150000	450000	8	1.142857
6	6	175000	520000	8	1.142857

```
    pharmexp          totmedex
 Min.    : 50000   Min.    :200000
 1st Qu.: 87500   1st Qu.:300000
 Median :150000   Median :450000
 Mean   :128571   Mean   :402857
 3rd Qu.:162500   3rd Qu.:485000
 Max.    :200000   Max.    :600000
```

```
mydesign <-
  svydesign (  id = ~1 ,
           data = Data ,
          weights = ~wt1 ,
             fpc = ~totcnt
            )
svyratio (~pharmexp, ~totmedex, mydesign)
```

On obtient le ratio et l'erreur standard de ces deux variables :"pharmexp" et "totmedex".

```
Ratio estimator: svyratio.survey.design2(~pharmexp, ~totmedex, mydesign)
Ratios=
          totmedex
pharmexp 0.3191489
SEs=
             totmedex
pharmexp 0.004006652
```

4 Estimateur de regression

Supposons que la population finie U est elle-même obtenue par des tirages indépendants dans une population infinie et définissons le modèle de superpopulation :

$$E_\xi(y_k) = \sum_{j=1}^{J} \beta_j x_{jk} = x'_k\beta \qquad k \in U$$

L'indice fait référence à la loi sur la population parente infinie. Le problème qui nous intéresse est toujours d'estimer le total de y sur U. Quel que soit le

plan de sondage mis en œuvre, on observe y et les x pour chaque élément de l'échantillon et on connaît les valeurs des x pour tout U.
On s'inspire de l'estimateur par différence. Il faut donc estimer le coefficient A apparaissant dans cette méthode. La méthode des moindres carrés pondérés suggère un tel estimateur. Ceci observé, la construction se déroule naturellement.
Si l'échantillon au sens de la superpopulation, c'est-à-dire U tout entier, était disponible, on estimerait β par :

$$B = T^{-}1t$$

$$T = \sum_{U} \frac{x_k x'_k}{\sigma_k^2}$$

et

$$t = \sum_{U} \frac{x_k y_k}{\sigma_k^2}$$

dans ces expressions, T est l'habituel : $(XWX)^{-}1$ de la méthode des moindres carrés pondérés, t est $X'W_y$ où W est la matrice diagonale des poids $1/\sigma_k^2$
. Mais T et t sont des matrices de totaux sur U, donc non calculables. On les estime à partir de s.

$$\widehat{B} = \widehat{T}^{-}1\widehat{t}$$

$$\widehat{T} = \sum_{s} \frac{x_k x'_k}{\sigma_k^2 \pi_k}$$

et

$$\widehat{t} = \sum_{s} \frac{x_k y_k}{\sigma_k^2 \pi_k}$$

∗L'estimateur du total par régression est alors défini en les mêmes termes que l'estimateur par différence.

$$\widehat{t}_{yr} := \widehat{t}_{y\pi} + (t_{XU} - \widehat{t}_{X\pi})'\widehat{B} = \widehat{t}_{y\pi} + \sum_{j=1}^{J} (t_{xj} - \widehat{t}_{xj,\pi})\widehat{B}_j$$

Une autre écriture :

$$\widehat{t}_{yr} = \sum_{s} g_{ks} \breve{y}_k$$

où $g_{ks} = (t_{XU} - \widehat{t}_{X\pi})'\widehat{t}^{-}1\frac{x_k}{\sigma_k^2}$

L'estimateur par régression est donc un estimateur pondéré mais les poids dépendent de l'échantillon, au contraire de ce qui se passe avec l'estimateur par les valeurs dilatées.

4.1 Partie pratique sur R

```
a <- matrix( c(130 , 935, 255 , 1170,
```

```
              510 , 1920, 340 , 1500,
              450 , 1900), nrow= 5, byrow=T, ncol=2)
x <- a[,1]
y <- a[,2]
colnames(a) <- c("Surf","Prix")

X <- matrix(c(rep(1,5),x) ,nrow= 5, ncol=2, byrow=F)
w <- diag(1/x)
n <- 5
N <- 1000
T <- (N/n)*t(X) %*%w%*% X
t <- (N/n)*t(X) %*%w%*% y
B <- solve(T,t)
txu <- 315000
ty.ht <- N *mean(y)
ty.reg <- ty.ht + (N - N)*B[1,1] + (txu - N * mean(x))* B[2,1]
ty.reg
```

```
> w
             [,1]        [,2]        [,3]        [,4]        [,5]
[1,] 0.007692308 0.000000000 0.000000000 0.000000000 0.000000000
[2,] 0.000000000 0.003921569 0.000000000 0.000000000 0.000000000
[3,] 0.000000000 0.000000000 0.001960784 0.000000000 0.000000000
[4,] 0.000000000 0.000000000 0.000000000 0.002941176 0.000000000
[5,] 0.000000000 0.000000000 0.000000000 0.000000000 0.002222222
```

```
> B
           [,1]
[1,] 550.230594
[2,]   2.773796
```

```
> ty.ht
[1] 1485000
```

Et l'estimateur de regression

```
> ty.reg
[1] 1423976
```

5 Bibliographie

Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association, 87(418) :376–382*

Andersson, P. G. and Thorburn, D. (2005). An optimal calibration distance

leading to the optimal regression estimator. *Survey Methodology, 31(1) :95–99* Kim, J. K. and Park, M. (2010). Calibration estimation in survey sampling. *Inter- national Statistical Review, 78(1) :21–39.* Krapavickaite, D. and Plikusas, A. (2005). Estimation of ratio in finite population. *Informatica, 16 :347–364.* Andersson, P. G. (2006). A conditional perspective of weighted variance estimation of the optimal regression estimator. *Journal of Statistical Planning and Inference,136(1) :221–234.*

`coloraer les vchoses`