R package **sampling**

8èmes Journées des Methodes Avancées pour l'Analyse de Sondages Complexes, Nantes, 2016

Alina Matei et Yves Tillé

Institut de Statistique
Université de Neuchâtel, Suisse

UNIVERSITÉ DE NEUCHÂTEL

Partie I

Vue d'ensemble

- Qu'est-ce que R ?
- Une petite introduction à la théorie des sondages.
- Le module sampling.

Qu'est-ce que R ?

- Le logiciel R est un logiciel gratuit (freeware) disponible sur le site `http://cran.r-project.org/`
- Il existe des versions pour Windows, Mac OS, Linux.
- Sont disponibles :
  - le programme de base.
  - des modules complémentaires (contributed packages).
- Nous avons créé un module complementaire nommé sampling.

Sampling

- Ce module répond à différents objectifs :
  - un logiciel libre pour le traitement d'une enquête,
  - des fonctions semblables à celles qu'offrent les logiciels payants.
- Ce projet a vu le jour en 2005 afin de servir d'outil pédagogique pour des cours avancés sur les méthodes d'échantillonnage, organisés par l'Office Fédéral de la Statistique Suisse sous l'égide d'Eurostat et de l'Association Européenne de Libre Echange (AELE).
- Version 2.7.
- Lien :

  `http://cran.r-project.org/web/packages/sampling/index.html`

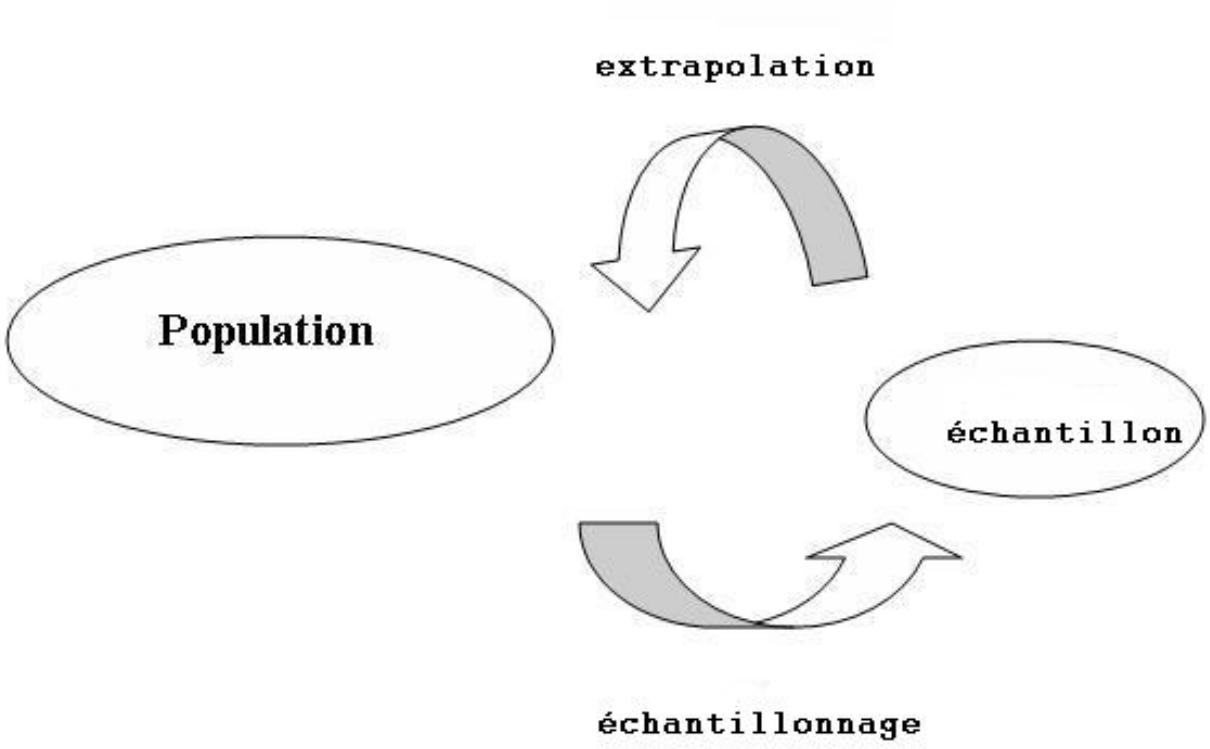
Introduction à la théorie des sondages

- La théorie des sondages occupe une place particulière en statistique (même si la pratique des sondages est très répandue).
- Bien que reposant sur les mêmes principes que la statistique mathématique classique, elle en diffère sensiblement dans son esprit en raison d'objectifs spécifiques.
- Elle se consacre essentiellement au problème de l'échantillonnage et de l'estimation.

Le formalisme

- On considère une population comprenant N individus parfaitement identifiés par un numéro d'ordre.
- Il suffit de ne retenir que ces numéros d'ordre et nous définissons ainsi la population $U = \{1, ..., k, ..., N\}$.
- Notons que les termes de population et individus sont purement conventionnels.
- Les parties de cette population sont appelées échantillons.

Echantillonnage et extrapolation



Stratégie de sondage : échantillonnage et extrapolation

On peut mettre en évidence la notion de stratégie de sondage en deux étapes :

- La première est l'échantillonnage aléatoire muni de propriétés probabilistes contrôlées : le plan de sondage $p(s)$ définit, pour chaque échantillon $s \subset U$, la probabilité qu'il soit sélectionné via un mécanisme aléatoire utilisé :

$$p(s) \geq 0, \text{ pour tout } s \subset U \text{ avec } \sum_{s \subset U} p(s) = 1.$$

- La seconde est la construction des estimations/extrapolations :
  - Elle consiste à définir pour l'échantillon aléatoire S une estimation. Par exemple pour le total d'une variable d'intérêt y

$$Y = \sum_{k \in U} y_k,$$

on veut définir un estimateur

$$\widehat{Y} = \sum_{k \in S} d_k y_k.$$

  - Les coefficients de pondération d_k jouent un role très important, car ils nous permettent d'extrapoler les résultats obtenus sur un échantillon vers ceux de la population.

Echantillonnage probabiliste

Les méthodes d'échantillonnage probabiliste les plus courantes sont :

- l'échantillonnage aléatoire simple
- l'échantillonnage à probabilités inégales (avec probabilité proportionnelle à la taille)
- l'échantillonnage stratifié
- l'échantillonnage en grappes
- l'échantillonnage à plusieurs degrés
- l'échantillonnage à plusieurs phases.

Probabilités d'inclusion

- La probabilité d'inclusion d'ordre un (π_k) est la probabilité d'une unité $k \in U$ d'appartenir à l'échantillon aléatoire S

$$\pi_k = \sum_{s \ni k} p(s), k \in U.$$

- La probabilité d'inclusion d'ordre deux $(\pi_{k\ell})$

$$\pi_{k\ell} = \sum_{s \ni k,\ell} p(s), k, \ell \in U.$$

Plans simples

Soit n la taille de l'échantillon.

- Plan simple avec remise : `srswr(n,N)`
- Plan simple sans remise : `srswor(n,N)`, `srswor1(n,N)`.

Plans simples : exemples

```
> library(sampling)
> srswor(4,8)
[1] 1 1 0 1 1 0 0 0
> srswor1(4,8)
[1] 0 1 0 1 0 1 0 1
> srswr(4,8)
[1] 0 0 1 3 0 0 0 0
# la fonction qui existe en R
> sample(8,4)
[1] 3 2 8 5
```

Plans à probabilités inégales

- Calcul des probabilités d'inclusion à partir d'une variable auxiliaire x connue sur toute la population : `pik=inclusionprobabilities(x,n)`
- Plan à probabilités inégales, de taille fixe, avec remise : `UPmultinomial(pik)`
- Plan à probabilités inégales, de taille aléatoire, sans remise : `UPpoisson(pik)`
- Plans à probabilités inégales, de taille fixe, sans remise : `UPbrewer(pik)`, `UPmaxentropy(pik)`, `UPmidzuno(pik)`, `UPpivotal(pik)`, `UPrandompivotal(pik)`, `UPminimalsupport(pik)`, `UPsampford(pik)`, `UPsystematic(pik)`, `UPrandomsystematic(pik)`, `UPtille(pik)`, où `pik` est un vecteur de probabilités d'inclusion dont la somme est entière, la taille de l'échantillon.

Plans à probabilités inégales : exemple

```
> # variable auxiliaire connue sur toute la population
> x=c(1,2,3,4,5,6,7,8,9)
> # taille de la population
> N=length(x)
> N
[1] 9
> # taille de l'échantillon
> n=4
> pik=inclusionprobabilities(x,n)
> pik
[1] 0.08888889 0.17777778 0.26666667 0.35555556 0.44444444
0.53333333 0.62222222 0.71111111 0.80000000
> sum(pik)
[1] 4
```

```
> # tirage d'un échantillon
> s=UPsystematic(pik)
> # s comme un vecteur de 0 et 1
> s
[1] 1 0 0 0 0 1 1 0 1
> # ou comme un vecteur d'étiquettes
> (1:N)[s==1]
[1] 1 6 7 9
```

Plans de sondages complexes

- plan stratifié : `strata()`, avec un tirage à probabilités égales ("srswor", "srswr") ou inégales ("poisson","systematic"),
- plan par grappes : `cluster()` avec un tirage à probabilités égales ("srswor", "srswr") ou inégales ("poisson","systematic"),
- plan à plusieurs degrés : `mstage()`, en utilisant un plan stratifié ou plan par grappes ou un plan simple, avec un tirage à probabilités égales ("srswor", "srswr") ou inégales ("poisson","systematic").

L'échantillonnage équilibré

- C'est un procédé d'échantillonnage aléatoire (dit 'équilibré') qui permet de respecter non seulement une taille fixée d'échantillon, mais encore la valeur du total de n'importe quel ensemble de variables auxiliaires $\mathbf{x}$ contenues dans la base de sondage :

$$\sum_{k \in S} \mathbf{x}_k \frac{1}{\pi_k} \approx \sum_{k \in U} \mathbf{x}_k.$$

- Cette technique permet d'augmenter considérablement la précision des estimations.
- La méthode du cube (Deville et Tillé, 2004) : `samplecube()`, et pour des plans complexes `balancedstratification()`, `balancedcluster()`, `balancedtwostage()`.

Echantillonnage équilibré : exemple

```
> # variables d'équilibrage
> X=cbind(c(1,1,1,1,1,1,1,1,1,1),
+ c(1.1,2.2,3.1,4.2,5.1,6.3,7.1,8.1,9.1,10),
+ c(2,3,4,6,1,2,4,5,6,4))
> X
      [,1] [,2] [,3]
 [1,]    1  1.1    2
 [2,]    1  2.2    3
 [3,]    1  3.1    4
 [4,]    1  4.2    6
 [5,]    1  5.1    1
 [6,]    1  6.3    2
 [7,]    1  7.1    4
 [8,]    1  8.1    5
 [9,]    1  9.1    6
[10,]    1 10.0    4
```

```
> # probabilités d'inclusion
> # taille de l'échantillon n=5
> pik=c(1/2,1/2,1/2,1/2,1/2,1/2,1/2,1/2,1/2,1/2)
> pik
 [1] 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5
 > # sélection d'un échantillon
> s=samplecube(X,pik,order=1,comment=TRUE)
.....
QUALITY OF BALANCING
  TOTALS HorvitzThompson_estimators Relative_deviation
1   10.0                      10.0           0.000000
2   56.3                      55.6          -1.243339
3   37.0                      38.0           2.702703
```

L'estimateur de Horvitz-Thompson

- Pour le total d'une variable d'intérêt y

$$Y = \sum_{k \in U} y_k,$$

l'estimateur de Horvitz-Thompson de Y est $(d_k = 1/\pi_k)$

$$\widehat{Y}_{HT} = \sum_{k \in S} d_k y_k = \sum_{k \in S} \frac{y_k}{\pi_k}.$$

- Fonctions : pour l'estimateur de HT du total Y `HTestimator(y,pik)` et son estimateur de la variance `varHT(y,pikl,method)`.

Exemple

```
> # tirage d'un échantillon
> s=UPsystematic(pik)
> # s comme un vecteur de 0 et 1
> s
[1] 1 0 0 0 0 1 1 0 1
> # ou comme un vecteur d'étiquettes
> (1:N)[s==1]
[1] 1 6 7 9
> # variable d'intérêt connue sur l'échantillon s
> y=c(2,4,3,2)
> # l'estimateur HT du total
> HTestimator(y,pik[s==1])
          [,1]
[1,] 37.32143
> # ou utiliser
> sum(y/pik[s==1])
[1] 37.32143
```

L'estimateur par calage

- Le calage assure l'amélioration de l'estimateur de Horvitz-Thompson : on maintient presque parfaitement son caractère sans biais et on diminue sa variance.
- On cherche à calculer des nouveaux poids w_k qui sont proches des poids initiaux $d_k = 1/\pi_k$, de telle manière que

$$\sum_{k \in S} w_k \mathbf{x}_k = \sum_{k \in U} \mathbf{x}_k.$$

- $\mathbf{x}_k$ est un vecteur de variables auxiliaires, car on peut utiliser plusieurs variables de calage.
- L'estimateur par calage (ou l'estimateur calé) de Y est $(w_k = g_k d_k = g_k/\pi_k)$

$$\widehat{Y}_{cal} = \sum_{k \in S} w_k y_k = \sum_{k \in S} \frac{g_k}{\pi_k} y_k.$$

Le calage

On peut utiliser plusieurs fonctions de calage F :

$$w_k = d_k F(q_k \boldsymbol{\lambda}' \mathbf{x}_k)$$

- linéaire $F(u) = 1 + u$,
- raking $F(u) = \exp(u)$,
- lineaire tronquée (avec bornes),
- logistique (avec bornes).
- Le calage est implementé à l'aide de la fonction `calib(Xs,d,total,q=rep(1,length(d)), method=c("linear","raking","truncated","logit"), bounds=c(low=0,upp=10),description=FALSE,max_iter=500)` qui retourne le vecteur des g-poids : $g_k = w_k/d_k = F(q_k \boldsymbol{\lambda}' \mathbf{x}_k)$. L'estimateur calé et sa variance estimée sont calculés à l'aide la fonction `calibev()`.

Le calage - exemple

```
> # on suppose que s a été tiré
> # variables de calage au niveau de s
> Xs=cbind(c(1,1,1,1,1,0,0,0,0,0), c(0,0,0,0,0,1,1,1,1,1),
+ c(1,2,3,4,5,6,7,8,9,10))
> Xs
      [,1] [,2] [,3]
 [1,]    1    0    1
 [2,]    1    0    2
 [3,]    1    0    3
 [4,]    1    0    4
 [5,]    1    0    5
 [6,]    0    1    6
 [7,]    0    1    7
 [8,]    0    1    8
 [9,]    0    1    9
[10,]    0    1   10
```

```
> # probabilités d'inclusion au niveau de s
> piks=rep(0.2,times=10)
> # totaux de la population pour X
> total=c(24,26,290)
> # les poids g en utilisant la méthode linéaire tronquée
> g=calib(Xs,d=1/piks,total,method="truncated",
bounds=c(0.75,1.2))
```

```
> # les poids g sont entre 0.75 et 1.2
> g
 [1] 0.75000 0.85125 0.95875 1.06625 1.17375 0.83875 0.94625
 1.05375 1.16125 1.20000
> # totaux de X au niveau de la pop.
> total
[1]  24  26 290
> # l'estimateur de Horvitz-Thompson de X
> colSums(Xs/piks)
[1]  25  25 275
> # l'estimateur calé de X
> colSums(Xs*g/piks)
[1]  24  26 290
```

```
> checkcalibration(Xs, d=1/piks, total, g)
$message
[1] "the calibration is done"

$result
[1] TRUE

$value
[1] 1e-06
> # variable d'intérêt connue sur l'échantillon s
> ys=c(3,4,5,6,1,2,4,5,2,1)
> # l'estimateur calé de Y
> sum(ys*g/piks)
[1] 161.3687
```

Autres fonctions

Pour :

- le calage généralisé,
- la poststratification,
- la correction de la non-réponse etc.

La liste complète des fonctions peut être obtenue en R en utilisant :

```
help(package=sampling)
```

Partie II

Probabilités d'inclusion dans les plans à probabilités inégales

Fonction : `inclusionprobabilities()`
Si l'on veut sélectionner exactement n unités d'observation avec des probabilités proportionnelles aux x_k, où x_k est une variable auxiliaire, corrélée avec la réponse y_k, on commence par calculer les quantités

$$\pi_k = \frac{nx_k}{\sum_{\ell \in U} x_\ell}, \text{ pour tout } k \in U.$$

- Les π_k calculés ainsi peuvent prendre des valeurs supérieures à 1. Pour éviter ce problème, on sélectionne d'office les unités ayant des π_k plus grands que 1.
- Ensuite, on recalcule les probabilités d'inclusion de la même manière sur les individus non sélectionnés d'office, en diminuant la taille de l'échantillon du nombre d'individus sélectionnés d'office.
- On répète ensuite cette opération jusqu'à ce que toutes les probabilités d'inclusion d'ordre un soient, ou égales à 1, ou strictement proportionnelles aux $x_k > 0$.

Plan simple sans remise

- Il existe plusieurs algorithmes de tirage.
- Par exemple, la **méthode du tri aléatoire** consiste à trier au hasard le fichier de données contenant la population. En pratique, on affecte un nombre aléatoire uniforme $[0, 1[$ à chaque individu de la population. On trie ensuite le fichier par ordre croissant (ou décroissant) des nombres aléatoires. Enfin, on choisit les n premiers (ou les n derniers) individus du fichier ainsi ordonné.
- Cette procédure de tirage est très aisée à mettre en œuvre. On doit cependant trier tout le fichier de données. Cette opération peut s'avérer très longue quand le fichier est grand.

Plan simple sans remise (prob. égales, taille fixe de s)

La méthode de sélection-rejet : `srswor1()`
N – la taille de la population; n – la taille de l'échantillon; $\pi_i = n/N$, pour toute $i \in U$.

$k = 0$;
$j = 0$;
Répéter tant que $j < n$:
  $u =$ nombre aléatoire uniforme $[0, 1[$;
  Si $u < \frac{n-j}{N-k}$ alors : sélectionner l'unité $k+1$; $j = j+1$;
  sinon passer l'unité $k+1$;
  $k = k+1$.

Plan de Poisson (prob. inégales, taille aléatoire de s)

Fonction : `UPpoisson()`

$$\left| \begin{array}{l} \text{Répéter pour } k = 1, \ldots, N \end{array} \right| \left| \begin{array}{l} u = \text{ variable aléatoire uniforme } [0, 1[; \\ \text{si } u < \pi_k \text{ sélectionner l'unité } k; \\ \text{sinon passer l'unité } k. \end{array} \right.$$

- La taille de l'échantillon est aléatoire, et il y a une probabilité non nulle de sélectionner un échantillon vide.
- Comme les unités sont sélectionnées indépendamment les unes des autres, on a $\pi_{k\ell} = \pi_k \pi_\ell$.

Plan systématique (prob. inégales, taille fixe de s)

Fonction : `UPsystematic()`
On veut sélectionner un échantillon de taille fixe n avec des probabilités d'inclusion π_k.

$u =$ une variable aléatoire uniforme [0,1[;
$a = -u$;

Répéter pour $k = 1, .., N$

- $b = a$;
- $a = a + \pi_k$;
- si $\lfloor a \rfloor \neq \lfloor b \rfloor$ sélectionner k.

La méthode est simple, mais elle a un défaut : beaucoup de probabilités d'inclusion d'ordre deux $\pi_{k\ell}$ sont nulles et la variance de l'estimateur de Horvitz-Thompson (HT) ne peut pas être estimée à l'aide de $\pi_{k\ell}$.

L'estimateur de Horvitz-Thompson et sa variance

Fonction : `HTestimator()`

Soit y la variable d'intérêt. On aimerait estimer le total de population

$$Y = \sum_{k \in U} y_k.$$

- l'estimateur de Horvitz-Thompson de Y $(\pi_k > 0, k \in U)$ est

$$\widehat{Y}_{HT} = \sum_{k \in S} y_k / \pi_k.$$

- la variance de $\widehat{Y}_{HT}$ $(\pi_k > 0, k \in U)$ est

$$Var(\widehat{Y}_{HT}) = \sum_{k \in U} \sum_{\ell \in U} \frac{y_k y_\ell}{\pi_k \pi_\ell} \Delta_{k\ell},$$

où

$$\Delta_{k\ell} = \begin{cases} \pi_{k\ell} - \pi_k \pi_\ell & \text{si } k \neq \ell \\ \pi_k (1 - \pi_k) & \text{si } k = \ell. \end{cases}$$

L'estimateur de la variance de $\widehat{Y}_{HT}$ pour un plan simple sans remise

$$\widehat{Var}(\widehat{Y}_{HT})_{srswor} = \frac{N(N-n)s_y^2}{n},$$

$$\text{où } s_y^2 = \frac{\sum_{k\in S}(y_k - \widehat{\bar{y}})^2}{n-1},$$

$$\widehat{\bar{y}} = \frac{1}{n}\sum_{k\in S} y_k.$$

L'estimateur de la variance de $\widehat{Y}_{HT}$ pour un plan de Poisson

Fonction : `varHT(.., method=1)`

Pour le plan de Poisson on a que $\pi_{k\ell} = \pi_k \pi_\ell$ et on déduit :

$$\widehat{Var}(\widehat{Y}_{HT})_{Poisson} = \sum_{k \in S} \frac{1 - \pi_k}{\pi_k^2} y_k^2.$$

L'estimateur de Sen-Yates-Grundy (SYG) de la variance de $\widehat{Y}_{HT}$

Fonction : `varHT(.., method=2)`

$$\widehat{Var}_{SYG}(\widehat{Y}_{HT}) = \sum_{k \in S} \sum_{\ell \in S} \left(\frac{y_k}{\pi_k} - \frac{y_\ell}{\pi_\ell} \right)^2 \frac{\pi_k \pi_\ell - \pi_{k\ell}}{\pi_{k\ell}}.$$

- Cet estimateur de variance n'est sans biais que si le plan est de taille fixe.
- Une condition suffisante pour qu'il soit positif est que $\pi_k \pi_\ell - \pi_{k\ell} \geq 0$, pour tout $k \in U, \ell \in U, k \neq \ell$.

Ces conditions sont dites conditions de Sen-Yates-Grundy.

L'estimateur de Deville de la variance de $\widehat{Y}_{HT}$ pour les plans de taille fixe à prob. inégales

Fonction : `varest()`

Deville, J.-C. (1993). *Estimation de la variance pour les enquêtes en deux phases*, INSEE.

$$\widehat{Var}_{Deville}(\widehat{Y}_{HT}) = \left(\frac{1}{1-\sum_{k\in S} a_k^2}\right) \sum_{k\in S}(1-\pi_k)\left(\frac{y_k}{\pi_k} - A\right)^2,$$

où $a_k = \frac{1-\pi_k}{\sum_{\ell\in S}(1-\pi_\ell)}$ et $A = \sum_{\ell\in S} \frac{a_\ell y_\ell}{\pi_\ell}$.

Plan équilibré

Un plan de sondage $p(s)$ est dit équilibré sur les caractères auxiliaires $\mathbf{x}_1, ..., \mathbf{x}_p$, si et seulement si il vérifie les équations d'équilibrage données par

$$\sum_{k \in S} \frac{x_{kj}}{\pi_k} = \sum_{k \in U} x_{kj},$$

pour tout $s \subset U$ tel que $p(s) > 0$, et pour tout $j = 1, ..., p$. En pratique on a :

$$\sum_{k \in S} \frac{x_{kj}}{\pi_k} \approx \sum_{k \in U} x_{kj}.$$

La **méthode du cube** (`samplecube()`) se décompose en deux phases :

1. la phase de vol : l'objectif est d'"arrondir" toutes les probabilités d'inclusion en vérifiant exactement les équations d'équilibrage ;
2. la phase d'atterrissage : consiste à gérer le mieux possible le fait que les équations d'équilibrage ne peuvent pas être exactement satisfaites.

Calage

Fonction : `calib()`

- On cherche à calculer des nouveaux poids w_k qui sont proches des poids initiaux $d_k = 1/\pi_k$, de telle manière que

$$\sum_{k \in S} w_k \mathbf{x}_k = \sum_{k \in U} \mathbf{x}_k,$$

où $w_k = d_k g_k = d_k F(q_k \boldsymbol{\lambda}' \mathbf{x}_k)$ et g_k sont les poids g.

- L'estimateur calé du total de la population est

$$\widehat{Y}_{cal} = \sum_{k \in S} w_k y_k = \sum_{k \in S} \frac{g_k}{\pi_k} y_k.$$

- Le calage est implémenté à l'aide de la fonction `calib()` qui retourne le vecteur des g-poids. L'estimateur calé et sa variance estimée sont calculés à l'aide de la fonction `calest()`.

Méthodes

- linéaire $F(u) = 1 + u$,
- raking $F(u) = \exp(u)$,
- linéaire tronquée (avec bornes $L < 1 < U$)

$$F(u) = \begin{cases} 1 + q_k u & \text{si} \quad (L-1)/q_k \leq u \leq (U-1)/q_k \\ U & \text{si} \quad u > (U-1)/q_k \\ L & \text{si} \quad u < (L-1)/q_k, \end{cases}$$

- logistique (avec bornes $L < 1 < U$),

$$F(u) = \frac{L(U-1) + U(1-L)\exp(Aq_k u)}{(U-1) + (1-L)exp(Aq_k u)} \in (L, U),$$

où $A = (U-L)/((1-L)(U-1))$.

L'estimateur de la variance d'un estimateur calé

- l'estimateur de Deville-Särndal (1992) : `calest()`
  Forme I :

$$\widehat{Var}_{DS}(\widehat{Y}_{cal}) = \sum_{k\in S}\sum_{\ell\in S}\frac{\pi_k\pi_\ell - \pi_{k\ell}}{\pi_{k\ell}}(w_k e_k)(w_\ell e_\ell),$$

où $e_k = y_k - \mathbf{x}_k'\widehat{\boldsymbol{B}}$ et $(\sum_{k\in S} w_k q_k \mathbf{x}_k\mathbf{x}_k')\widehat{\boldsymbol{B}} = \sum_{k\in S} w_k q_k \mathbf{x}_k y_k$.
Forme II :

$$\widehat{Var}_{DS}(\widehat{Y}_{cal}) = \sum_{k\in S}\sum_{\ell\in S}\frac{\pi_k\pi_\ell - \pi_{k\ell}}{\pi_{k\ell}}(d_k e_k)(d_\ell e_\ell),$$

- l'estimateur de Deville transformé : `varest()`

$$\widehat{Var}_{Deville}(\widehat{Y}_{cal}) = \left(\frac{1}{1-\sum_{k\in S} a_k^2}\right)\sum_{k\in S}(1-\pi_k)\left(\frac{e_k}{\pi_k} - A\right)^2,$$

où $a_k = \frac{1-\pi_k}{\sum_{\ell\in S}(1-\pi_\ell)}$ et $A = \sum_{\ell\in S}\frac{a_\ell e_\ell}{\pi_\ell}$.