Méthodes d'échantillonnage et d'estimation en population finie
Partie 2 : méthodes d'estimation

Guillaume Chauvet

École Nationale de la Statistique et de l'Analyse de l'Information

20/10/2025

Principaux objectifs du cours

Vu avec Laurent Costa : partie amont de l'enquête.

- Méthodes d'inférence dans le cas d'une population finie d'individus.
- Principales méthodes d'échantillonnage utilisées dans les enquêtes.

Nous nous intéressons ici à la partie aval de l'enquête :

- Méthodes de redressement qui permettent d'utiliser une information auxiliaire au moment de l'estimation.
- Estimation d'un paramètre complexe.

Base de sondage d'aéroports

A titre d'illustration, nous considérerons (encore) une base de sondage de $N = 12$ aéroports français, ayant accueilli entre 500 000 et 2 000 000 de passagers en 2019. Elle contient les variables :

- Nombre de passagers en 2019 (`Pass19`)
- Taille de l'Unité Urbaine en 2019 (`Pop19`)
  $\Rightarrow$ *variables auxiliaires*
- Nombre de passagers en 2020 (`Pass20`)
- Nombre de passagers en transit en 2020 (`Trans20`)
  $\Rightarrow$ *variables d'intérêt*

Base de données d'aéroports

	Pass19	Pop19	Pass20	Trans20
MONTPELLIER	1 900 000	620 000	*800 000*	*300*
BASTIA	1 600 000	100 000	*800 000*	*1 400*
AJACCIO	1 500 000	100 000	*900 000*	*1 300*
STRASBOURG	1 300 000	800 000	*500 000*	*1 300*
BREST	1 200 000	320 000	*500 000*	*1 800*
BIARRITZ	1 100 000	300 000	*400 000*	*200*
RENNES	900 000	730 000	*300 000*	*200*
FIGARI	700 000	20 000	*500 000*	*2 900*
PAU	600 000	240 000	*200 000*	*0*
TOULON	500 000	630 000	*200 000*	*0*
PERPIGNAN	500 000	320 000	*200 000*	*0*
TARBES	500 000	120 000	*100 000*	*0*
t_x	12 300 000	4 300 000		
μ_x	1 025 000	358 333		
S_x^2	$2.33\ 10^{11}$	$7.28\ 10^{10}$		
$cv_x = \sqrt{S_x^2}/\mu_x$	47%	76%		

Plan

Approche assistée par un modèle

Rappels sur le modèle linéaire

Supposons les valeurs de y dans la pop. U générées selon le modèle linéaire

$$y_k = \mathrm{x}_k^\top \boldsymbol{\beta} + \epsilon_k \text{ avec } \begin{cases} E_m(\epsilon_k) &= 0, \\ V_m(\epsilon_k) &= \sigma_k^2, \end{cases} \tag{1.1}$$

avec x_k un vecteur de q variables auxiliaires, et σ_k^2 un paramètre inconnu qui peut varier d'un individu à l'autre.

Dans le cas $\sigma_k^2 = \sigma^2$, nous retrouvons le modèle linéaire homoscédastique.

Rappels sur le modèle linéaire

Estimateurs des moindres carrés ordinaires

Nous notons

$$\underbrace{\mathbf{X}}_{(N,q)} = \begin{pmatrix} x_{11} & \cdots & x_{q1} \\ \vdots & & \vdots \\ x_{1N} & \cdots & x_{qN} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_N^\top \end{pmatrix} \quad \text{et} \quad \underbrace{\mathbf{Y}}_{(N,1)} = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}$$

Nous avons successivement

$$\mathbf{X}^\top\mathbf{X} = \begin{pmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_N \end{pmatrix} \begin{pmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_N^\top \end{pmatrix} = \sum_{k\in U} \mathbf{x}_k\mathbf{x}_k^\top,$$

$$\mathbf{X}^\top\mathbf{Y} = \begin{pmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_N \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix} = \sum_{k\in U} \mathbf{x}_k y_k,$$

$$\mathbf{B}_{MCO} = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\left(\mathbf{X}^\top\mathbf{Y}\right) = \left(\sum_{k\in U} \mathbf{x}_k\mathbf{x}_k^\top\right)^{-1} \sum_{k\in U} \mathbf{x}_k y_k.$$

Rappels sur le modèle linéaire

Estimateurs des moindres carrés généralisés

Pour un modèle hétéroscédastique, nous utiliserons plutôt l'*estimateur des moindres carrés généralisés*

$$\mathrm{B}_{MCG} = \left(\mathrm{X}^\top \boldsymbol{\Sigma}^{-1} \mathrm{X}\right)^{-1} (\mathrm{X}^\top \boldsymbol{\Sigma}^{-1} \mathrm{Y})$$

$$\text{avec } \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \sigma_N^2 \end{pmatrix} \text{ matrice de var-covar. du modèle.}$$

Nous obtenons successivement

$$\mathrm{X}^\top \boldsymbol{\Sigma}^{-1} \mathrm{X} = \sum_{k \in U} \frac{\mathrm{x}_k \mathrm{x}_k^\top}{\sigma_k^2} \quad \text{et} \quad \mathrm{X}^\top \boldsymbol{\Sigma}^{-1} \mathrm{Y} = \sum_{k \in U} \frac{\mathrm{x}_k y_k}{\sigma_k^2},$$

$$\text{puis } \mathrm{B}_{MCG} = \left(\sum_{k \in U} \frac{\mathrm{x}_k \mathrm{x}_k^\top}{\sigma_k^2} \right)^{-1} \left(\sum_{k \in U} \frac{\mathrm{x}_k y_k}{\sigma_k^2} \right).$$

Rappels sur le modèle linéaire

Exemple 0 : le modèle constant

Le *modèle constant* est le cas le plus simple. Il consiste à utiliser uniquement la constante ("intercept") dans le modèle avec une variance constante :

$$y_k = \beta_0 + \epsilon_k \text{ avec } \begin{cases} E_m(\epsilon_k) &= 0, \\ V_m(\epsilon_k) &= \sigma^2. \end{cases} \tag{1.2}$$

C'est un cas particulier du modèle linéaire (1.1) obtenu avec une seule variable auxiliaire $x_k = 1$, et $\sigma_k^2 = \sigma^2$ (homoscédasticité).

Nous avons

$$\begin{aligned} B_{MCG} &= \left(\sum_{k \in U} \frac{x_k x_k^\top}{\sigma_k^2} \right)^{-1} \left(\sum_{k \in U} \frac{x_k y_k}{\sigma_k^2} \right) \\ &= \frac{\sum_{k \in U} y_k}{\sum_{k \in U} 1} = \mu_y. \end{aligned}$$

Rappels sur le modèle linéaire

Résidus du modèle

La qualité de prédiction du modèle linéaire peut être résumée par les *résidus de régression*

$$E_k = y_k - \mathrm{x}_k^\top \mathrm{B}_{MCG}. \tag{1.3}$$

Plus les résidus sont faibles, plus la part de la variable d'intérêt expliquée par les variables auxiliaires x_k est importante. Nous utiliserons le critère

$$R^2 = 1 - \frac{\sum_{k\in U} E_k^2}{\sum_{k\in U}(y_k - \mu_y)^2} = 1 - \frac{\sum_{k\in U} E_k^2}{\sum_{k\in U} E_{0k}^2} \tag{1.4}$$

pour mesurer la qualité d'adéquation du modèle par rapport au modèle constant.

Rappels sur le modèle linéaire

Base de données d'aéroports

	Strate	Pass19	Pop19	Pass20	Trans20
MONTPELLIER	1	1 900 000	620 000	*800 000*	*300*
BASTIA	1	1 600 000	100 000	*800 000*	*1 400*
AJACCIO	1	1 500 000	100 000	*900 000*	*1 300*
STRASBOURG	1	1 300 000	800 000	*500 000*	*1 300*
BREST	1	1 200 000	320 000	*500 000*	*1 800*
BIARRITZ	1	1 100 000	300 000	*400 000*	*200*
RENNES	2	900 000	730 000	*300 000*	*200*
FIGARI	2	700 000	20 000	*500 000*	*2 900*
PAU	2	600 000	240 000	*200 000*	*0*
TOULON	2	500 000	630 000	*200 000*	*0*
PERPIGNAN	2	500 000	320 000	*200 000*	*0*
TARBES	2	500 000	120 000	*100 000*	*0*
t_x	$N_1 = 6$ $N_2 = 6$	12 300 000	4 300 000		

Rappels sur le modèle linéaire

Application à la base d'aéroports : modèle constant

Nous considérons les variables d'intérêt `Pass20` (nombre de passagers en 2020) et `Trans20` (nombre de passagers en transit en 2020).

Le modèle constant

$$y_k = \beta_0 + \epsilon_k \text{ avec } \begin{cases} E_m(\epsilon_k) &= 0, \\ V_m(\epsilon_k) &= \sigma^2. \end{cases}$$

conduit à prédire la variable y_k par sa valeur moyenne

$$\mu_y = \begin{cases} 450\ 000 & \text{pour } \texttt{Pass20}, \\ 783 & \text{pour } \texttt{Trans20}. \end{cases}$$

Rappels sur le modèle linéaire

Exemple 1 : modèle linéaire simple

Nous nous plaçons dans le cas où $q = 2$ et $x_k = (1, x_{1k})^\top$, et avec homoscédasticité. Le modèle s'écrit :

$$y_k = \beta_0 + \beta_1 x_{1k} + \epsilon_k \text{ avec } \begin{cases} E_m(\epsilon_k) = 0, \\ V_m(\epsilon_k) = \sigma^2. \end{cases} \quad (1.5)$$

Dans ce cas, l'estimateur des MCG coincide avec l'estimateur des MCO. Nous obtenons après calcul

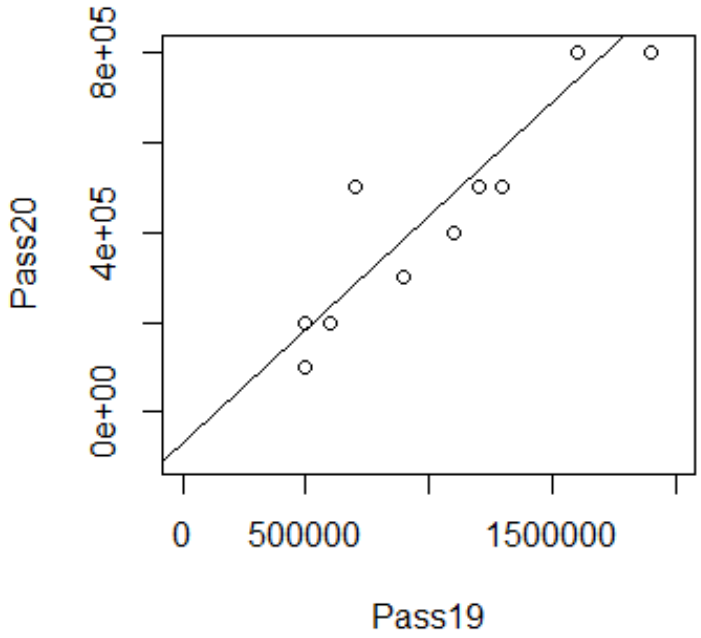
$$\begin{aligned} B_{1,MCG} &= \frac{\sum_{k\in U}(x_{1k} - \mu_{x1})(y_k - \mu_y)}{\sum_{k\in U}(x_{1k} - \mu_{x1})^2} = \frac{S_{xy}}{S_x^2}, \\ B_{0,MCG} &= \mu_y - B_{1,MCG} \times \mu_x, \\ E_k &= (y_k - \mu_y) - \frac{S_{xy}}{S_x^2}(x_{1k} - \mu_{x1}). \end{aligned}$$

Rappels sur le modèle linéaire

Application à la base d'aéroports : modèle linéaire simple $x = (1, Pass19)$

```
#Régression de Pass20 sur Pass19
> reg1 <- lm(Pass20 ~ Pass19)
> summary(reg1)
> plot(Pass19,Pass20,xlim=c(0,2000000),ylim=c(-100000,800000))
> abline(lm(Pass20~Pass19))
```

$$\begin{aligned} \text{Pass20}_k &= -68\,000 \\ &\quad +0.51\ \text{Pass19}_k + \epsilon_k, \\ R^2 &= 0.83 \end{aligned}$$

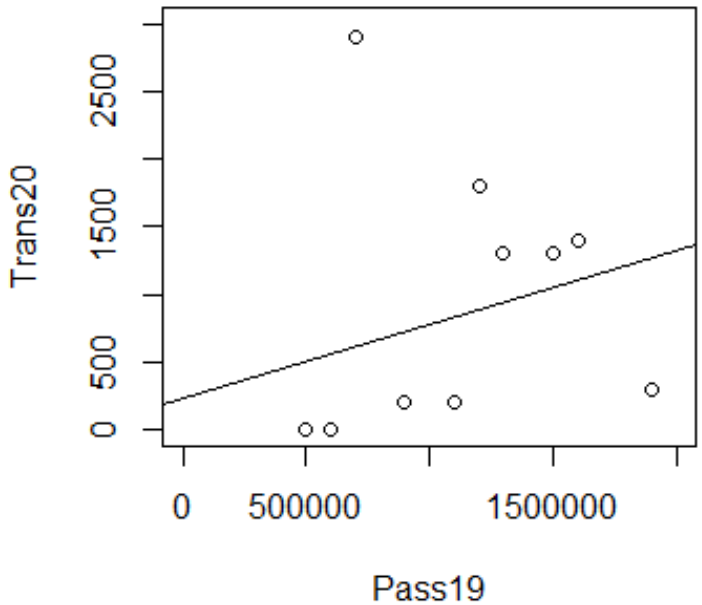


Rappels sur le modèle linéaire

Application à la base d'aéroports : modèle linéaire simple $x = (1, Pass19)$

```
#Régression de Trans20 sur Pass19
> reg2 <- lm(Trans20 ~ Pass19)
> summary(reg2)
> plot(Pass19,Trans20,xlim=c(0,2000000),ylim=c(0,3000))
> abline(lm(Trans20~Pass19))
```

$$\begin{aligned} \text{Trans20}_k &= 221 \\ &\quad +5.5\ 10^{-4}\ \text{Pass19}_k + \epsilon_k, \\ R^2 &= 0.08 \end{aligned}$$



Rappels sur le modèle linéaire

Exemple 2 : modèle ratio

Nous nous plaçons dans le cas où $q = 1$ et $x_k = x_{1k} > 0$, et avec hétéroscédasticité. Le *modèle ratio* s'écrit :

$$y_k = \beta_1 x_{1k} + \epsilon_k \text{ avec } \begin{cases} E_m(\epsilon_k) &= 0, \\ V_m(\epsilon_k) &= \sigma^2 x_{1k}. \end{cases} \tag{1.6}$$

L'estimateur des MCG se simplifie sous la forme :

$$\begin{aligned} B_{MCG} &= \left(\sum_{k \in U} \frac{x_k x_k^\top}{\sigma_k^2} \right)^{-1} \sum_{k \in U} \frac{x_k y_k}{\sigma_k^2} \\ &= \left(\sum_{k \in U} \frac{x_{1k}^2}{\sigma^2 x_{1k}} \right)^{-1} \sum_{k \in U} \frac{x_{1k} y_k}{\sigma^2 x_{1k}} = \frac{t_y}{t_{x1}}. \end{aligned}$$

Les résidus de régression sous ce modèle sont

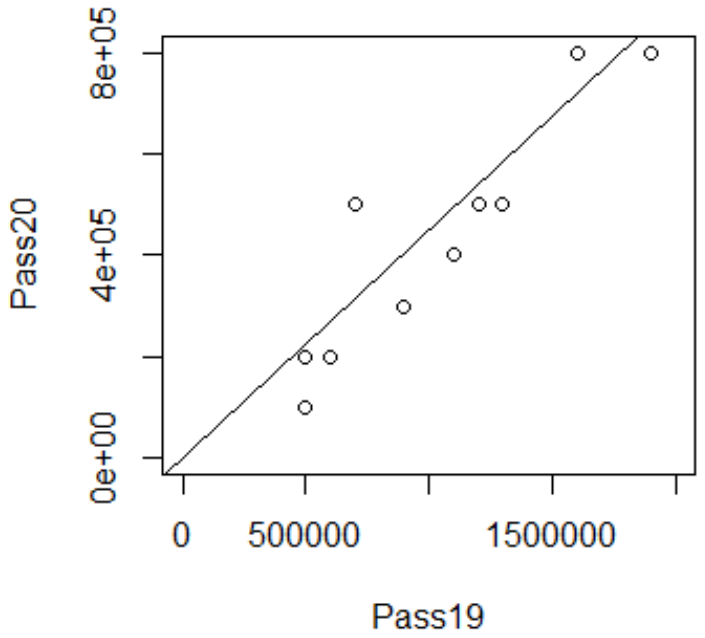
$$E_k = y_k - R\, x_{1k} \text{ avec } R = \frac{t_y}{t_{x1}}.$$

Rappels sur le modèle linéaire

Application à la base d'aéroports : modèle ratio x_1 = Pass19

```
#Régression de Pass20 sur Pass19 sans constante
#Poids Pass19^{-1}
> reg3 <- lm(Pass20~Pass19+0,weights=Pass19^{-1},)
> summary(reg3)
> plot(Pass19,Pass20,xlim=c(0,2000000),ylim=c(0,800000))
> abline(lm(Pass20~Pass19+0))
```

$$\begin{aligned} \text{Pass20}_k &= 0.44\ \text{Pass19}_k + \epsilon_k, \\ R^2 &= 0.82 \end{aligned}$$

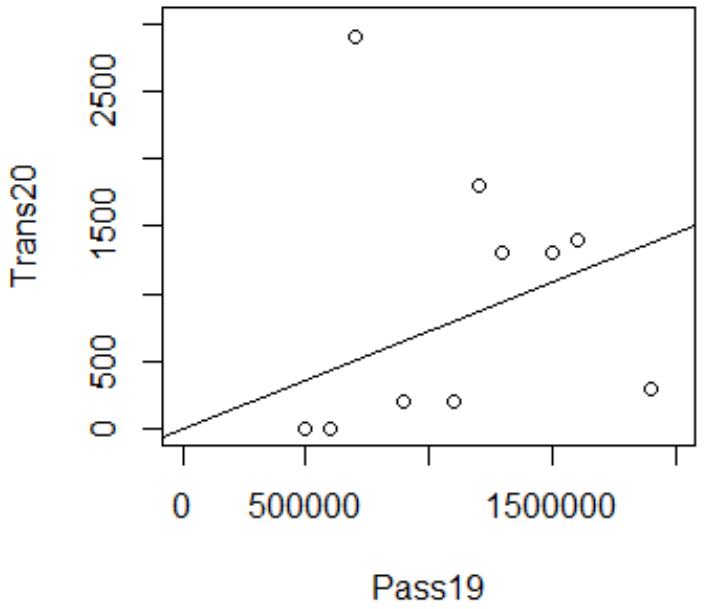


Rappels sur le modèle linéaire

Application à la base d'aéroports : modèle ratio x_1 = Pass19

```
#Régression de Trans20 sur Pass19 sans constante
#Poids Pass19^{-1}
> reg4 <- lm(Trans20~Pass19+0,weights=Pass19^{-1},)
> summary(reg4)
> plot(Pass19,Trans20,xlim=c(0,2000000),ylim=c(0,3000))
> abline(lm(Trans20~Pass19+0))
```

$$\begin{aligned} \text{Trans20}_k &= 0.00076\ \text{Pass19}_k + \epsilon_k, \\ R^2 &= 0.07 \end{aligned}$$



Rappels sur le modèle linéaire

Exemple 3 : modèle constant par strates

Population partitionnée en H strates $U_1, \ldots, U_H$.
Nous utilisons $\mathrm{x}_k = \{1(k \in U_1), \ldots, 1(k \in U_H)\}^\top$ avec homoscédasticité dans les strates. Le *modèle constant par strates* s'écrit :

$$y_k = \beta_h + \epsilon_k \text{ avec } \left\{ \begin{array}{ll} E_m(\epsilon_k) & = 0, \\ V_m(\epsilon_k) & = \sigma_h^2 \end{array} \right. \text{ pour } k \in U_h. \qquad (1.7)$$

L'estimateur des MCG se simplifie sous la forme :

$$\begin{aligned} B_{MCG} &= \left(\sum_{k \in U} \frac{\mathrm{x}_k \mathrm{x}_k^\top}{\sigma_k^2} \right)^{-1} \sum_{k \in U} \frac{\mathrm{x}_k y_k}{\sigma_k^2} \\ &= \begin{pmatrix} \frac{N_1}{\sigma_1^2} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \frac{N_H}{\sigma_H^2} \end{pmatrix}^{-1} \begin{pmatrix} \frac{t_{y1}}{\sigma_1^2} \\ \vdots \\ \frac{t_{yH}}{\sigma_H^2} \end{pmatrix} = \{\mu_{y1}, \ldots, \mu_{yH}\}^\top. \end{aligned}$$

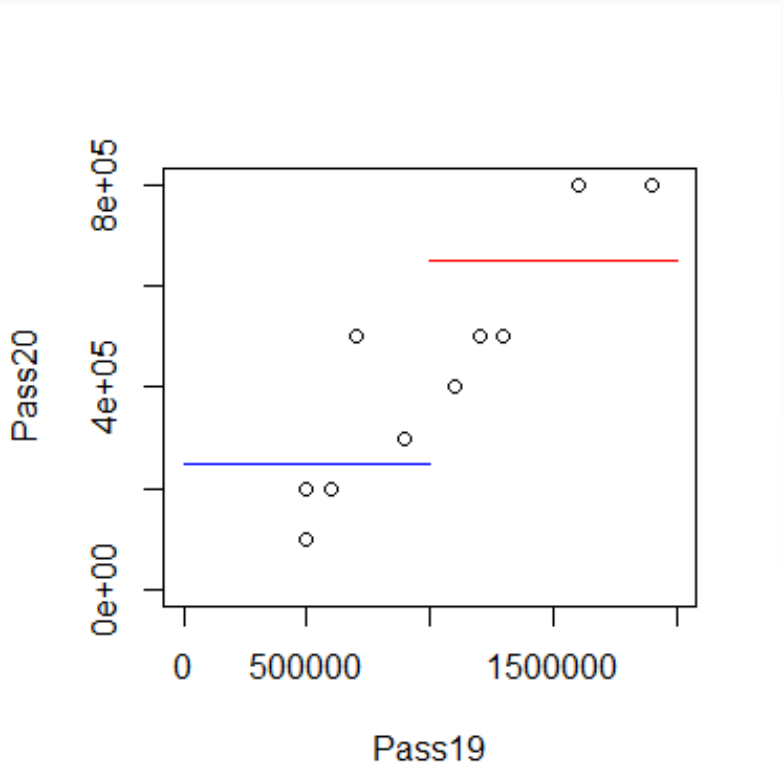
Les résidus sous ce modèle sont $E_k = y_k - \mu_{yh}$ pour $k \in U_h$.

Rappels sur le modèle linéaire

Modèle const. par strates $x = \{1(Pass19 \geq 1\ 000\ 000), 1(Pass19 < 1\ 000\ 000)\}$

```
#Modèle homogène par strates pour Pass20
> st1 <- c(1,1,1,1,1,1,0,0,0,0,0,0)
> st2 <- c(0,0,0,0,0,0,1,1,1,1,1,1)
> reg5 <- lm(Pass20~st1+st2+0)
> summary(reg5)
> plot(Pass19,Pass20,xlim=c(0,2000000),ylim=c(0,800000))
```

$$\begin{aligned} \text{Pass20}_k &= 650\ 000 \times 1(k \in U_1) \\ &+ 250\ 000 \times 1(k \in U_2) + \epsilon_k, \\ R^2 &= 0.61 \end{aligned}$$

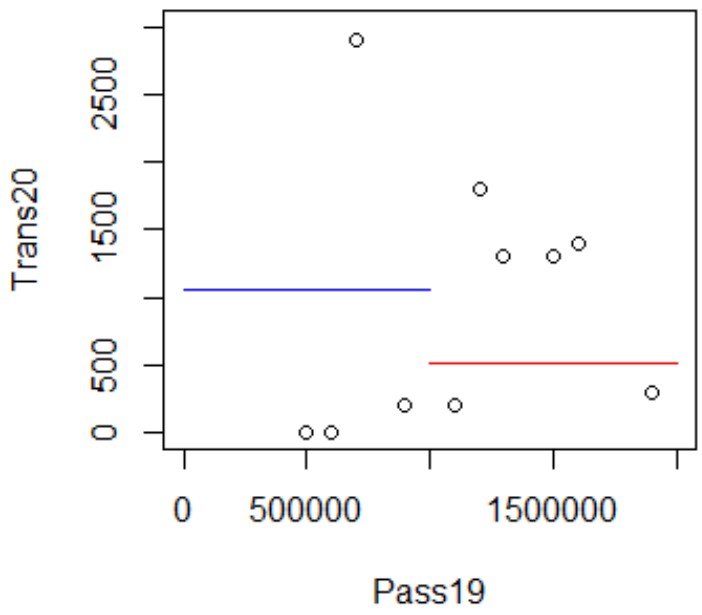


Rappels sur le modèle linéaire

Modèle const. par strates $x = \{1(Pass19 \geq 1\ 000\ 000), 1(Pass19 < 1\ 000\ 000)\}$

```
#Modèle homogène par strates pour Trans20
> reg6 <- lm(Trans20~st1+st2+0)
> summary(reg6)
> plot(Pass19,Trans20,xlim=c(0,2000000),ylim=c(0,3000))
```

$$\begin{aligned} \text{Trans20}_k &= 1\ 050 \times 1(k \in U_1) \\ &+ 517 \times 1(k \in U_2) + \epsilon_k, \\ R^2 &= 0.09 \end{aligned}$$



Modèle de travail

Principe

Lors du choix d'un plan de sondage, nous utilisons implicitement un *modèle de travail* de la forme (1.1) :

$$y_k = \mathrm{x}_k^\top \boldsymbol{\beta} + \epsilon_k \text{ avec } \begin{cases} E_m(\epsilon_k) &= 0, \\ V_m(\epsilon_k) &= \sigma_k^2, \end{cases}$$

avec un jeu de variables auxiliaires x_k spécifique. C'est une modélisation implicite de la variable d'intérêt y_k.

Le qualificatif "de travail" signifie que **le modèle n'a pas besoin de bien prédire la variable d'intérêt pour que l'estimateur de Horvitz-Thompson soit sans biais**. C'est de toute façon impossible dans une enquête où un même modèle ne peut pas être parfaitement adapté à toutes les variables collectées.

En revanche, la variance de l'estimateur de Horvitz-Thompson est réduite si le modèle est bien prédictif pour y_k (résidus de régression E_k faibles).

Modèle de travail

Exemple du sondage aléatoire simple

Sous un sondage aléatoire simple

$$V_p(\hat{t}_{y\pi}) \quad = \quad N^2 \frac{1-f}{n} S_y^2 \quad \text{où} \quad S_y^2 = \frac{1}{N-1} \sum_{k \in U} (y_k - \mu_y)^2.$$

La variance est donnée par la variable de résidus $E_k = y_k - \mu_y$. Elle est faible si les y_k sont peu dispersées autour de leur moyenne.

Le modèle de travail est donc le modèle constant (1.2) :

$$y_k \quad = \quad \beta_0 + \epsilon_k \text{ avec } \begin{cases} E_m(\epsilon_k) & = 0, \\ V_m(\epsilon_k) & = \sigma^2. \end{cases}$$

Modèle de travail

Exemple du sondage aléatoire simple stratifié

Sous un sondage aléatoire simple stratifié

$$V_p(\hat{t}_{y\pi}) = \sum_{h=1}^{H} (N_h)^2 \frac{1-f_h}{n_h} S_{yh}^2 \quad \text{où} \quad S_{yh}^2 = \frac{1}{N_h - 1} \sum_{k \in U_h} (y_k - \mu_{yh})^2.$$

La variance est donnée par la variable de résidus $E_k = y_k - \mu_{yh},\ k \in U_h$. Elle est faible si les y_k sont peu dispersées autour de leur moyenne au sein de chaque strate.

Le modèle de travail est le modèle constant par strates (1.7) :

$$y_k = \beta_h + \epsilon_k \text{ avec } \begin{cases} E_m(\epsilon_k) = 0, \\ V_m(\epsilon_k) = \sigma_h^2 \end{cases} \text{ pour } k \in U_h.$$

Modèle de travail

Information auxiliaire

Quand des variables auxiliaires x_k sont utilisées pour définir un plan de sondage, elles doivent être connues pour toutes les unités de la population. Par exemple, pour stratifier la population d'aéroports selon `Pass19`, cette variable doit être connue pour chaque aéroport.

Nous allons voir à l'aide de la méthode du calage comment utiliser au moment de l'estimation un q-vecteur x_k de variables auxiliaires dont seul le total sur la population $t_x = \sum_{k \in U} x_k$ est connu.
Objectif : passer des poids de sondage d_k à des poids calés w_k tels que

$$\sum_{k \in S} w_k x_k = \sum_{k \in U} x_k.$$

Autrement dit, les totaux des variables auxiliaires sont estimés sans erreur.

Approche assistée par un modèle

En résumé

Nous avons revu quelques exemples du modèle linéaire général :

- modèle constant : $y_k = \beta_0 + \epsilon_k$ avec $V_m(\epsilon_k) = \sigma^2$,
- modèle linéaire simple : $y_k = \beta_0 + \beta_1 x_{1k} + \epsilon_k$ avec $V_m(\epsilon_k) = \sigma^2$,
- modèle ratio : $y_k = \beta_1 x_{1k} + \epsilon_k$ avec $V_m(\epsilon_k) = \sigma^2 x_{1k}$,
- modèle const. par strates : $y_k = \beta_h + \epsilon_k$ avec $V_m(\epsilon_k) = \sigma_h^2$ si $k \in U_h$.

Nous avons vu que les propriétés d'un plan de sondage dépendent d'un modèle de travail, cas particulier du modèle linéaire général.

Le plan de sondage donne toujours des estimateurs de Horvitz-Thompson sans biais. La variance est faible si le modèle de travail est bien prédictif pour la variable d'intérêt y_k.

Estimateur par calage

Principe du calage

Nous supposons que l'échantillon S a été collecté. Nous supposons disponible un q-vecteur x_k de variables auxiliaires dont le total sur la population $t_{\mathrm{x}} = \sum_{k \in U} \mathrm{x}_k$ est connu.

Nous cherchons de nouveaux poids w_k qui

1. **restent proches**[1] des poids de départ d_k,
2. **vérifient les équations de calage**

$$\sum_{k \in S} w_k \mathrm{x}_k \quad = \quad t_{\mathrm{x}}. \tag{2.1}$$

Notre modèle de travail est le modèle linéaire général (1.1) :

$$y_k \quad = \quad \mathrm{x}_k^\top \boldsymbol{\beta} + \epsilon_k \text{ avec } \begin{cases} E_m(\epsilon_k) & = 0, \\ V_m(\epsilon_k) & = \sigma_k^2. \end{cases}$$

[1] au sens d'une fonction de distance dont nous donnons les propriétés un peu plus loin

Estimateur par calage

Sources pour les marges de calage

Les totaux des variables de calage peuvent être donnés par des registres :

- FIDELI[2] (Insee) : données descriptives des logements (adresse, caractéristiques, données fiscales) et des individus (informations socio-démographiques, données fiscales)
- SIRENE[3] : informations sur les entreprises et leurs établissements
- Autres registres : RNIPP[4], données d'état-civil, fichiers fiscaux de la DGFIP[5], fichiers de la sécurité sociale, DADS[6], ...

[2] Fichier démographique des logements et des individus
[3] Syst. nation. d'Identification et du Répertoire des Entrep. et de leurs Étab.
[4] Répertoire National d'Identification des Personnes Physiques
[5] Direction Générale des FInances Publiques
[6] Déclarations annuelles de Données Sociales

Estimateur par calage

Sources pour les marges de calage (2)

Les totaux des variables de calage peuvent être également estimés par des enquêtes jugées très fiables :

- estimation de population des annuelles de recensement (EAR), utilisées pour fournir des marges de calage pour les enquêtes ménages de l'Insee, les enquêtes Ined, la cohorte Constances[7], l'enquête PISA[8], ...
- estimations issues de l'enquête emploi en continu : enquêtes de la DARES[9], enquête Générations du Céreq[10], ...

[7] Cohorte des Consultants des centres d'examen de santé
[8] Programme international pour le suivi des acquis
[9] Service statistique du ministère du travail
[10] Centre d'études et de recherche sur les qualifications

Estimateur par calage

Objectifs du calage

Le calage vise :

- à garantir la cohérence entre les enquêtes
  - Caler les enquêtes auprès des ménages de l'Insee sur les mêmes variables (âge, sexe, région, taille du ménage, ...) permet de garantir que ces enquêtes fournissent des estimations cohérentes pour ces variables.
  - Contraintes imposées par Eurostat à certaines enquêtes européennes (LFS[11], SILC[12]) pour utiliser des variables de calage communes.
- à améliorer la précision des enquêtes
  - Quand une enquête est tirée selon un plan de sondage, le calage vise à réduire la variance des estimateurs.
  - Quand une enquête est obtenue en interrogeant des volontaires (access panels), le calage vise d'abord à réduire le biais des estimateurs.

[11] Labour Force Surveys

[12] Survey on Income and Life Conditions

Estimateur par calage

Mise en oeuvre (1)

Nous utilisons une fonction $G : \mathbb{R} \to \mathbb{R}^+$ appelée fonction de distance, et vérifiant les conditions suivantes :

1. G est convexe, à valeurs positives, dérivable sur son domaine de définition, avec $G(1) = 0$.
2. Soit $F(\cdot)$ la fonction inverse de $G'(\cdot)$. Nous avons $F(0) = F'(0) = 1$.

La quantité $G(w_k/d_k)$ mesure la distance associée à l'unité k. La condition 1 assure que $G(w_k/d_k)$ augmente quand le poids calé w_k s'éloigne du poids de sondage d_k.

Exemples :

$$\text{Méthode linéaire : } G(x) = \frac{1}{2}(x-1)^2 \text{ et } F(x) = 1 + x,$$

$$\text{Méthode raking ratio : } G(x) = x\ln(x) - x + 1 \text{ et } F(x) = e^x.$$

Estimateur par calage

Mise en oeuvre (2)

Nous résolvons le problème d'optimisation sous contraintes :

$$\min_{w_k} \sum_{k \in S} d_k \sigma_k^2 G\left(\frac{w_k}{d_k}\right) \quad \text{t.q.} \quad \sum_{k \in S} w_k x_k = t_x, \tag{2.2}$$

Nous utilisons le Lagrangien :

$$L(\boldsymbol{\lambda}) \quad = \quad \sum_{k \in S} d_k \sigma_k^2 G\left(\frac{w_k}{d_k}\right) - \boldsymbol{\lambda}^\top \left(\sum_{k \in S} w_k x_k - t_x\right) \tag{2.3}$$

avec $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_q)^\top$ un vecteur de multiplicateurs de Lagrange.

Estimateur par calage

Mise en oeuvre (3)

En calculant la dérivée partielle par rapport à w_k, nous obtenons

$$\frac{\partial L(\boldsymbol{\lambda})}{\partial w_k} = \sigma_k^2 G'\left(\frac{w_k}{d_k}\right) - \boldsymbol{\lambda}^\top \mathbf{x}_k = 0 \quad \Rightarrow \quad w_k = d_k F\left(\frac{\boldsymbol{\lambda}^\top \mathbf{x}_k}{\sigma_k^2}\right).$$

Les quantités w_k sont appelées *poids calés ou poids redressés*, et

$$\hat{t}_{yw} = \sum_{k \in S} w_k y_k \tag{2.4}$$

est appelé *estimateur calé* du total t_y. Les quantités

$$g_k \quad = \quad \frac{w_k}{d_k} = F\left(\frac{\lambda^\top \mathbf{x}_k}{\sigma_k^2}\right)$$

sont appelées les *g-poids*. Dans le package `sampling`, elles peuvent être calculées avec la fonction `calib`.

Estimateur par calage

Mise en oeuvre (4)

Le vecteur $\boldsymbol{\lambda}$ peut être déterminé en résolvant le système (non-linéaire) constitué par les équations de calage

$$\sum_{k \in S} d_k F\left(\frac{\boldsymbol{\lambda}^\top \mathrm{x}_k}{\sigma_k^2}\right) \mathrm{x}_k \quad = \quad t_{\mathrm{x}},$$

par exemple à l'aide de la méthode itérative de Newton-Raphson.

Il existe plusieurs fonctions de distance, programmées dans la fonction `calib`:

- méthode linéaire (`linear`),
- méthode raking ratio (`raking`),
- méthode logit (`logit`),
- méthode linéaire tronquée (`truncated`).

Estimateur par calage

Un exemple d'application

L'enquête emploi est un échantillon de logements rotatif (les logements entrants un trimestre donné sont enquêtés pendant six trimestres consécutifs). Environ 92 000 logements par trimestre, enquête en continu chaque semaine de l'année.

En France métropolitaine :

- tirage à deux degrés dans le répertoire FIDELI,
- calage sur des marges au niveau individuel (sexe, âge, région) et au niveau logement (nombre et type de logements, nombre de pièces, déciles de revenu du ménage, ...)

Dans les DOM :

- tirage stratifié (selon des zones géographiques), avec tirage systématique à probabilités égales dans les strates.
- calage sur des marges au niveau individuel (nombre, diplôme, lieu de naissance) et au niveau logement (zonage en aires urbaines, nombre et type de logements)

Propriétés de l'estimateur calé

Sous des conditions générales, l'estimateur de Horvitz-Thompson vérifie

$$N^{-1}\left(\hat{t}_{y\pi} - t_y\right) = O_p\left(n^{-\frac{1}{2}}\right).$$

Soit $\hat{t}_{yw}$ l'estimateur calé, obtenu avec une fonction de distance $G(\cdot)$ respectant les conditions énoncées en diapositive 33. Il vérifie

$$N^{-1}\left(\hat{t}_{yw} - t_y\right) = N^{-1}\left(\hat{t}_{E\pi} - t_E\right) + o_p\left(n^{-\frac{1}{2}}\right),$$

avec $E_k = y_k - \mathrm{x}_k^\top \mathrm{B}_{MCG}$ (résidus de régression) et $t_E = \sum_{k\in U} E_k$.

Le comportement de l'estimateur calé est asymptotiquement le même que celui de l'estimateur de HT du total des résidus. Nous en déduisons :

$$\begin{aligned} E_p(\hat{t}_{yw} - t_y) &\simeq 0, && \text{(estim. approx. sans biais),} \\ V_p(\hat{t}_{yw} - t_y) &\simeq V_p(\hat{t}_{E\pi}) && \text{(var. donnée par les résidus de régression).} \end{aligned}$$

Propriétés de l'estimateur calé

Variance et estimation de variance

La variance de l'estimateur calé $\hat{t}_{yw}$ est donc (approx.) donnée par

$$V_p(\hat{t}_{yw}) \simeq \sum_{k,l \in U} \Delta_{kl} \frac{E_k}{\pi_k} \frac{E_l}{\pi_l}. \quad (2.5)$$

Nous notons $e_k = y_k - \mathbf{x}_k^\top \hat{\mathbf{b}}_\pi$ les *résidus estimés de régression*, avec

$$\hat{\mathbf{b}}_\pi = \left(\sum_{k \in S} \frac{\mathbf{x}_k \mathbf{x}_k^\top}{\pi_k \ \sigma_k^2} \right)^{-1} \left(\sum_{k \in S} \frac{\mathbf{x}_k y_k}{\pi_k \ \sigma_k^2} \right)$$

La variance peut être estimée par

$$\hat{V}_{HT,1}(\hat{t}_{yw}) = \sum_{k,l \in S} \frac{\Delta_{kl}}{\pi_{kl}} \frac{e_k}{\pi_k} \frac{e_l}{\pi_l} \text{ ou } \hat{V}_{HT,2}(\hat{t}_{yw}) = \sum_{k,l \in S} \frac{\Delta_{kl}}{\pi_{kl}} \frac{g_k e_k}{\pi_k} \frac{g_l e_l}{\pi_l}. \quad (2.6)$$

Dans la fonction `calibev` de `sampling`, l'option `with=TRUE` donne le premier estimateur de variance, et l'option `with=FALSE` le second estimateur.

Mise en oeuvre pratique

Illustration : SRS de $n = 6$ aéroports

	Pass19	Pop19	Pass20	Trans20	d_k
AJACCIO	1 500 000	100 000	900 000	1 300	2
RENNES	900 000	730 000	300 000	200	2
FIGARI	700 000	20 000	500 000	2 900	2
TOULON	500 000	630 000	200 000	0	2
PERPIGNAN	500 000	320 000	200 000	0	2
TARBES	500 000	120 000	100 000	0	2
t_x	12 300 000	4 300 000			
$\hat{t}_\pi$	9 200 000	3 840 000	4 400 000	8 800	

Mise en oeuvre pratique

Illustration : SRS de $n = 6$ aéroports

```
#Echantillon SRS de 6 aéroports

#Donnees echantillonnees
> loc=c(0,0,1,0,0,0,1,1,0,1,1,1)
> ech=aeroports[loc==1,]

#Probabilites d'inclusion d'ordre 1 et 2
> nech=6
> Npop=12
> pi=rep(nech/Npop,Npop)
> pikl=UPmaxentropypi2(pi)

#Poids de sondage et pi_kl sur l'echantillon
> dech=rep(Npop/nech,nech)
> pikl_ech=pikl[ech==1,ech==1]
```

Elle correspond au choix $G(x) = \frac{1}{2}(x-1)^2$ et $F(x) = 1 + x$. Dans ce cas, les équations de calage admettent une solution explicite :

$$\boldsymbol{\lambda}_{lin} = \left(\sum_{k \in S} \frac{d_k \mathrm{x}_k \mathrm{x}_k^\top}{\sigma_k^2} \right)^{-1} (t_\mathrm{x} - \hat{t}_{\mathrm{x}\pi}).$$

Nous obtenons

$$\sum_{k \in s} w_k y_k = \hat{t}_{y\pi} + \hat{\mathrm{b}}_\pi^\top \left[t_\mathrm{x} - \hat{t}_{\mathrm{x}\pi} \right] \equiv \hat{t}_{y,greg}$$

$$\text{avec} \qquad \hat{\mathrm{b}}_\pi = \left(\sum_{k \in S} \frac{d_k \mathrm{x}_k \mathrm{x}_k^\top}{\sigma_k^2} \right)^{-1} \sum_{k \in S} \frac{d_k \mathrm{x}_k y_k}{\sigma_k^2}.$$

Il s'agit de l'*estimateur par la régression généralisée (GREG)*. Cette méthode de calage peut conduire à des poids finaux w_k négatifs.

Mise en oeuvre pratique

Application : méthode linéaire avec x = (1, *Pass*19)

```
#Variables de calage sur l'echantillon
> Xech=cbind(rep(1,nech),ech$Pass19)

#Totaux des variables de calage
> Xtot=c(12,12 300 000)

#g-poids
> gweight<-calib(Xech,dech,Xtot,
                 method="linear",description=FALSE)
> gweight
[1] 2.469828 1.267241 0.866379 0.465517 0.465517 0.465517

#Poids cales
> wech<-gweight*dech
> wech
[1] 4.939655 2.534483 1.732759 0.931035 0.931035 0.931035
```

Mise en oeuvre pratique

Fonctions de distance usuelles : méthode raking-ratio

La méthode raking ratio

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

Cette méthode permet d'assurer que les poids finaux w_k sont > 0.

```
#Calage par la méthode raking ratio
#g-poids
> gweight<-calib(Xech,dech,Xtot,
                 method="raking",description=FALSE)
> gweight
[1] 2.592926 1.019327 0.746715 0.547011 0.547011 0.547011

#Poids cales
> wech<-gweight*dech
> wech
[1] 5.185852 2.038654 1.493429 1.094021 1.094021 1.094021
```

Mise en oeuvre pratique

Fonctions de distance usuelles : méthode linéaire bornée

C'est une version tronquée de la méthode linéaire. Des bornes LO et UP sont spécifiées pour les rapports de poids pour assurer que pour tt $k \in S$:

$$LO \leq \frac{w_k}{d_k} \leq UP.$$

```
> gweight<-calib(Xech,dech,Xtot,method="truncated",
                 bounds=c(0.25,4),description=FALSE)
> gweight
[1] 2.469828 1.267241 0.866379 0.465517 0.465517 0.465517

> gweight<-calib(Xech,dech,Xtot,method="truncated",
                 bounds=c(0.50,2.42),description=FALSE)
> gweight
[1] 2.42 1.57 0.51 0.50 0.50 0.50

> gweight<-calib(Xech,dech,Xtot,method="truncated",
                 bounds=c(0.50,2),description=FALSE)
No convergence in 500 iterations with the given bounds.
The bounds for the g-weights are: -0.75 and 3.25
```

Mise en oeuvre pratique

Fonctions de distance usuelles : méthode logit

C'est une version tronquée de la méthode raking-ratio. Des bornes LO et UP sont également spécifiées pour les rapports de poids :

$$LO \leq \frac{w_k}{d_k} \leq UP.$$

```
> gweight<-calib(Xech,dech,Xtot,method="logit",
                 bounds=c(0.25,4),description=FALSE)
> gweight
[1] 2.577636 1.058856 0.744128 0.539795 0.539795 0.539795

> gweight<-calib(Xech,dech,Xtot,method="logit",
                 bounds=c(0.50,2.50),description=FALSE)
> gweight
[1] 2.496167 1.314498 0.640150 0.516394 0.516394 0.516394

> gweight<-calib(Xech,dech,Xtot,method="logit",
                 bounds=c(0.50,2.00),description=FALSE)
No convergence in 500 iterations with the given bounds.
The bounds for the g-weights are: 0.5 and 2
```

Mise en oeuvre pratique

Estimateur de variance d'un estimateur calé

En utilisant le premier estimateur de variance de l'équation (2.6), nous avons

$$\hat{V}_{HT,1}(\hat{t}_{yw}) = \sum_{k,l \in S} \frac{\Delta_{kl}}{\pi_{kl}} \frac{e_k}{\pi_k} \frac{e_l}{\pi_l}, \quad (2.7)$$

```
#Calage par la méthode linéaire
#Estimation du total de Pass19
>calibev(ech$Pass19,Xech,Xtot,pikl_ech,dech,gweight,with=TRUE)
$calest
[1] 12 300 011
$evar
[1] 8.724997e-13
#Estimation du total de Pass20
>calibev(ech$Pass20,Xech,Xtot,pikl_ech,dech,gweight,with=TRUE)
$calest
[1] 6 558 980
$evar
[1] 1.883e+11
```

Mise en oeuvre pratique

Estimateur de variance d'un estimateur calé (2)

En utilisant le second estimateur de variance de l'équation (2.6), nous avons

$$\hat{V}_{HT,2}(\hat{t}_{yw}) = \sum_{k,l \in S} \frac{\Delta_{kl}}{\pi_{kl}} \frac{g_k e_k}{\pi_k} \frac{g_l e_l}{\pi_l}. \quad (2.8)$$

```
#Calage par la méthode linéaire
#Estimation du total de Pass19
>calibev(ech$Pass19,Xech,Xtot,pikl_ech,dech,gweight,with=FALSE)
$calest
[1] 12 300 011
$evar
[1] 1.189e-12
#Estimation du total de Pass20
>calibev(ech$Pass20,Xech,Xtot,pikl_ech,dech,gweight,with=FALSE)
$calest
[1] 6 558 980
$evar
[1] 2.049e+11
```

Estimateur par calage

En résumé

Le calage consiste à obtenir de nouveaux poids, proches des poids de sondage d_k et vérifiant les équations de calage sur des totaux auxiliaires t_x.
Les marges de calage peuvent être fournies par des registres, ou estimées par de très grosses enquêtes.

L'estimateur par calage $\hat{t}_{yw} = \sum_{k \in S} w_k y_k$ est asymptotiquement sans biais, et sa variance n'est donnée que par les résidus de la régression de y_k sur les variables de calage x_k :

$$V_p(\hat{t}_{yw}) \simeq \sum_{k,l \in U} \Delta_{kl} \frac{E_k}{\pi_k} \frac{E_l}{\pi_l} \quad \text{avec} \quad E_k = y_k - x_k^\top B_{MCG},$$

$$\hat{V}_{HT,1}(\hat{t}_{yw}) = \sum_{k,l \in S} \frac{\Delta_{kl}}{\pi_{kl}} \frac{e_k}{\pi_k} \frac{e_l}{\pi_l} \quad \text{avec} \quad e_k = y_k - x_k^\top \hat{b}_\pi.$$

Exemples de méthodes de redressement

Exemples de méthodes de redressement

Estimateur par la régression généralisée (GREG)

Les exemples de cette section correspondent à l'utilisation de la méthode linéaire. Nous avons vu que dans ce cas, l'estimateur calé se réécrivait sous la forme de l'estimateur GREG :

$$\begin{aligned} \hat{t}_{y,greg} &= \hat{t}_{y\pi} + \hat{\mathrm{b}}_\pi^\top \left[t_{\mathrm{x}} - \hat{t}_{\mathrm{x}\pi} \right] \\ \text{avec} \qquad \hat{\mathrm{b}}_\pi &= \left(\sum_{k \in S} \frac{d_k \mathrm{x}_k \mathrm{x}_k^\top}{\sigma_k^2} \right)^{-1} \sum_{k \in S} \frac{d_k \mathrm{x}_k y_k}{\sigma_k^2}. \end{aligned}$$

Nous pouvons utiliser l'estimateur de variance

$$\hat{V}_{HT,1}(\hat{t}_{y,greg}) = \sum_{k,l \in S} \frac{\Delta_{kl}}{\pi_{kl}} \frac{e_k}{\pi_k} \frac{e_l}{\pi_l} \qquad \text{avec} \qquad e_k = y_k - \mathrm{x}_k^\top \hat{\mathrm{b}}_\pi.$$

Un autre estimateur de variance est donné dans l'équation (2.6).

Exemples de méthodes de redressement

Estimateur GREG utilisant $x_k = (1, \mathrm{Pass19}_k)$

	Pass19	Pop19	Pass20	Trans20	d_k	w_k
AJACCIO	1 500 000	100 000	900 000	1 300	2	4.94
RENNES	900 000	730 000	300 000	200	2	2.53
FIGARI	700 000	20 000	500 000	2 900	2	1.73
TOULON	500 000	630 000	200 000	0	2	0.93
PERPIGNAN	500 000	320 000	200 000	0	2	0.93
TARBES	500 000	120 000	100 000	0	2	0.93
t_x	12 300 000	4 300 000				
$\hat{t}_{\bullet\pi}$	9 200 000	3 840 000	4 400 000	8 800		
$\hat{V}_{HT}(\hat{t}_{\bullet\pi})$	$1.86\ 10^{12}$	$1.06\ 10^{12}$	$1.04\ 10^{12}$	$1.66\ 10^{7}$		
$\hat{t}_{\bullet w}$	12 300 000	3 375 000	6 538 000	11 950		
$\hat{V}_{HT}(\hat{t}_{\bullet w})$	0	$1.25\ 10^{12}$	$1.90\ 10^{11}$	$1.47\ 10^{7}$		

Exemples de méthodes de redressement

Estimateur GREG utilisant $x_k = (1, \text{Pass19}_k, \text{Pop19}_k)$

	Pass19	Pop19	Pass20	Trans20	d_k	w_k
AJACCIO	1 500 000	100 000	900 000	1 300	2	4.70
RENNES	900 000	730 000	300 000	200	2	3.47
FIGARI	700 000	20 000	500 000	2 900	2	1.06
TOULON	500 000	630 000	200 000	0	2	1.52
PERPIGNAN	500 000	320 000	200 000	0	2	0.84
TARBES	500 000	120 000	100 000	0	2	0.41
t_x	12 300 000	4 300 000				
$\hat{t}_{\bullet\pi}$	9 200 000	3 840 000	4 400 000	8 800		
$\hat{V}_{HT}(\hat{t}_{\bullet\pi})$	$1.86\ 10^{12}$	$1.06\ 10^{12}$	$1.04\ 10^{12}$	$1.66\ 10^{7}$		
$\hat{t}_{\bullet w}$	12 300 000	4 300 000	6 314 000	9 900		
$\hat{V}_{HT}(\hat{t}_{\bullet w})$	0	0	$1.03\ 10^{11}$	$1.10\ 10^{7}$		

Exemples de méthodes de redressement

Estimateur par le ratio

Nous supposons connu le total t_{x1} d'une seule variable auxiliaire (positive) x_{1k}. L'estimateur par le ratio est défini par

$$\hat{t}_{yR} = \hat{t}_{y\pi} \times \frac{t_{x_1}}{\hat{t}_{x_1\pi}} = \sum_{k \in S} w_k y_k$$

avec $w_k = d_k \times \frac{t_{x_1}}{\hat{t}_{x_1\pi}}$. Il est calé sur le total t_{x_1} : $\hat{t}_{x_1 R} = t_{x_1}$.

Cas particulier de l'estimateur GREG, obtenu sous le modèle ratio (1.6) :

$$y_k = \beta_1 x_{1k} + \epsilon_k \text{ avec } \begin{cases} E_m(\epsilon_k) &= 0, \\ V_m(\epsilon_k) &= \sigma^2 x_{1k}. \end{cases}$$

Exemples de méthodes de redressement

Estimateur par le ratio (2)

Le modèle ratio est un cas particulier du modèle linéaire général, obtenu avec $x_k = x_{1k}$ et $\sigma_k^2 = \sigma^2 \, x_{1k}$. Nous obtenons successivement :

$$\begin{aligned}
\hat{\mathrm{b}}_\pi &= \left(\sum_{k \in S} \frac{d_k \mathrm{x}_k \mathrm{x}_k^\top}{\sigma_k^2} \right)^{-1} \sum_{k \in S} \frac{d_k \mathrm{x}_k y_k}{\sigma_k^2} \\
&= \left(\sum_{k \in S} \frac{d_k x_{1k}^2}{x_{1k}} \right)^{-1} \sum_{k \in S} \frac{d_k x_{1k} y_k}{x_{1k}} = \frac{\hat{t}_{y\pi}}{\hat{t}_{x_1\pi}} = \hat{R}_\pi, \qquad (3.1) \\
\hat{t}_{y,greg} &= \hat{t}_{y\pi} + \hat{\mathrm{b}}_\pi^\top \left(t_{\mathrm{x}} - \hat{t}_{\mathrm{x}\pi} \right) \\
&= \hat{t}_{y\pi} + \hat{R}_\pi (t_{x_1} - \hat{t}_{x_1\pi}) = \hat{t}_{yR}. \qquad (3.2)
\end{aligned}$$

Exemples de méthodes de redressement

Estimateur par le ratio (3)

En utilisant les résultats obtenus pour l'estimateur GREG, l'estimateur par le ratio est approximativement non biaisé pour le total t_y. Sa variance est approximativement donnée par

$$V_p(\hat{t}_{yR}) \simeq V_p(\hat{t}_{E\pi})$$

avec $E_k = y_k - R\, x_{1k}$. La variance est donc réduite si les variables y_k et x_{1k} sont approximativement proportionnelles.

Nous pouvons utiliser l'estimateur de variance de Horvitz-Thompson :

$$\hat{V}_{HT,1}(\hat{t}_{yR}) = \sum_{k,l \in S} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} \frac{e_k}{\pi_k} \frac{e_l}{\pi_l}$$

avec $e_k = y_k - \hat{R}_\pi x_{1k}$.

Exemples de méthodes de redressement

Estimateur par le ratio (4)

Dans les fonctions calib et calibev, le paramètre q permet de spécifier que la structure du modèle de travail est hétéroscédastique.

```
Xech=ech$Pass19
Xtot=c(12 300 000)
#Variable Pass19
> gweight<-calib(Xech,dech,Xtot,q=Xech^{-1},method="linear")
> calibev(ech$Pass19,Xech,Xtot,pikl_ech,dech,gweight,
          q=Xech^{-1},with=TRUE)
$calest
[1] 12 300 000
$evar
[1] 0
#Variable Pass20
calibev(ech$Pass20,Xech,Xtot,pikl_ech,dech,gweight,
        q=Xech^{-1},with=TRUE)
$calest
[1] 5 882 609
$evar
[1] 2.402e+11
```

Exemples de méthodes de redressement

Estimateur par le ratio utilisant $x_k = (\mathrm{Pass19}_k)$

	Pass19	Pop19	Pass20	Trans20	d_k	w_k
AJACCIO	1 500 000	100 000	900 000	1 300	2	2.67
RENNES	900 000	730 000	300 000	200	2	2.67
FIGARI	700 000	20 000	500 000	2 900	2	2.67
TOULON	500 000	630 000	200 000	0	2	2.67
PERPIGNAN	500 000	320 000	200 000	0	2	2.67
TARBES	500 000	120 000	100 000	0	2	2.67
t_x	12 300 000	4 300 000				
$\hat{t}_{\bullet\pi}$	9 200 000	3 840 000	4 400 000	8 800		
$\hat{V}_{HT}(\hat{t}_{\bullet\pi})$	$1.86\ 10^{12}$	$1.06\ 10^{12}$	$1.04\ 10^{12}$	$1.66\ 10^{7}$		
$\hat{t}_{\bullet w}$	12 300 000	5 134 000	5 883 000	11 800		
$\hat{V}_{HT}(\hat{t}_{\bullet w})$	0	$1.62\ 10^{12}$	$2.40\ 10^{11}$	$1.47\ 10^{7}$		

Exemples de méthodes de redressement

Post-stratification

Supposons qu'après le tirage de l'échantillon, la population soit partitionnée en H groupes notés $U_1, \ldots, U_H$.
Les effectifs de ces *post-strates*, notés $N_1, \ldots, N_H$, sont supposés connus.
Soit S_h l'intersection de S et de U_h.

Ces effectifs peuvent être comparés avec leur estimateur de HT :

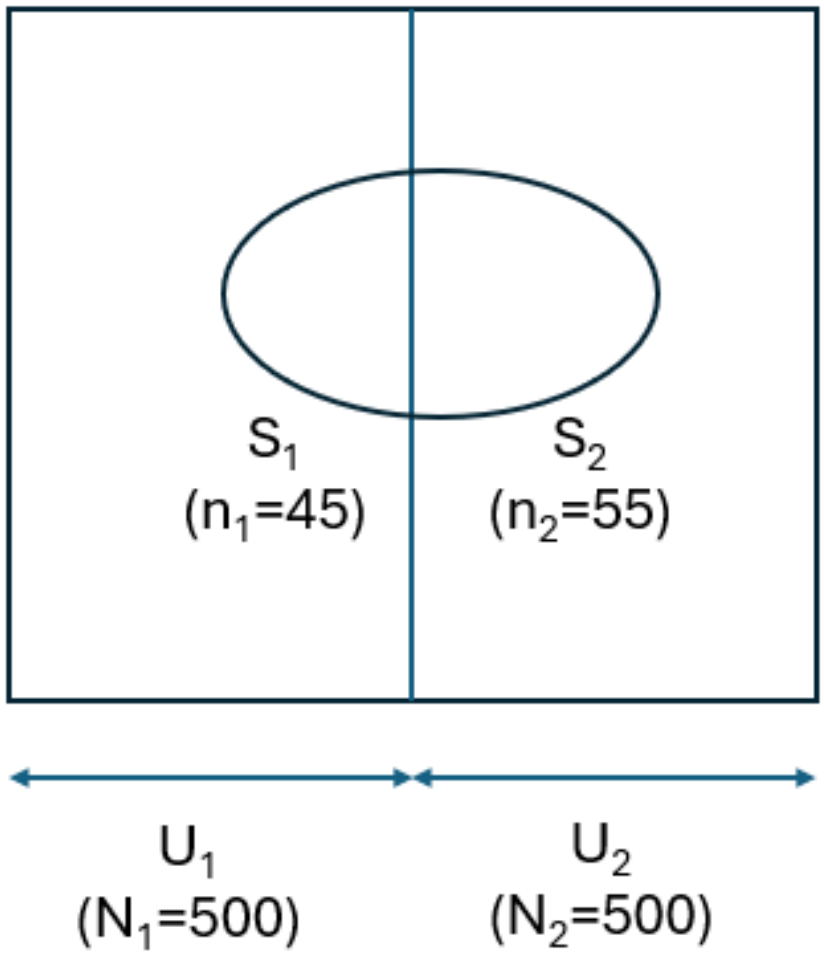
$$\begin{aligned} \hat{N}_h &= \sum_{k \in S_h} \frac{1}{\pi_k} = \sum_{k \in S} \frac{1(k \in U_h)}{\pi_k}, \\ E_p(\hat{N}_h) &= \sum_{k \in U} 1(k \in U_h) = N_h. \end{aligned}$$

L'estimateur post-stratifié est défini par

$$\hat{t}_{y,post} = \sum_{h=1}^{H} \frac{N_h}{\hat{N}_h} \hat{t}_{yh} \quad \text{avec} \quad \hat{t}_{yh} = \sum_{k \in S_h} \frac{y_k}{\pi_k}.$$

Exemples de méthodes de redressement

Post-stratification (2)



$$S \sim SRS(n = 100; U) \text{ et } d_k = 10,$$

$$N_1 = N_2 = 500,$$

$$\hat{N}_1 = \frac{N \times n_1}{n} = 450,$$

$$\hat{N}_2 = \frac{N \times n_2}{n} = 550,$$

$$\hat{t}_{y\pi} = 10 \sum_{k \in S} y_k,$$

$$\hat{t}_{y,post} \simeq 11.1 \sum_{k \in S_1} y_k + 9.1 \sum_{k \in S_2} y_k.$$

Exemples de méthodes de redressement

Post-stratification (3)

L'estimateur post-stratifié est motivé par le modèle homogène par strates

$$y_k = \beta_h + \epsilon_k \quad \text{et} \quad V_m(\epsilon_k) = \sigma_h^2 \text{ dans chaque strate } U_h.$$

C'est un cas particulier de l'estimateur par la régression généralisée, obtenu avec $\mathrm{x}_k = \{1(k \in U_1), \ldots, 1(k \in U_H)\}^T$ et $\sigma_k^2 = \sigma_h^2$ pour $k \in U_h$.
Nous avons :

$$\hat{\mathrm{b}}_\pi = \left\{ \sum_{k \in S} \frac{\mathrm{x}_k \mathrm{x}_k^\top}{\sigma_k^2 \, \pi_k} \right\}^{-1} \sum_{k \in S} \frac{\mathrm{x}_k y_k}{\sigma_k^2 \, \pi_k} \equiv \left\{ \frac{\hat{t}_{y1}}{\hat{N}_1}, \ldots, \frac{\hat{t}_{yH}}{\hat{N}_H} \right\}^T,$$

et

$$\begin{aligned} \hat{t}_{y,greg} &= \hat{t}_{y\pi} + \hat{\mathrm{b}}_\pi^\top \left(t_\mathrm{x} - \hat{t}_{\mathrm{x}\pi} \right) \\ &= \hat{t}_{y\pi} + \sum_{h=1}^{H} \frac{\hat{t}_{yh}}{\hat{N}_h} (N_h - \hat{N}_h) = \hat{t}_{y,post}. \end{aligned}$$

Exemples de méthodes de redressement

Post-stratification (4)

En utilisant les résultats obtenus pour l'estimateur GREG, l'estimateur post-stratifié est approximativement non biaisé pour le total t_y. Sa variance est approximativement donnée par

$$V_p(\hat{t}_{y,post}) \simeq V_p(\hat{t}_{E\pi})$$

avec $E_k = y_k - \mu_{yh}$ pour $k \in U_h$. La variance est donc réduite si la variable y est homogène à l'intérieur des post-strates.

En utilisant l'estimateur de variance de Horvitz-Thompson:

$$\hat{V}_{HT,1}(\hat{t}_{yR}) = \sum_{k,l \in S} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} \frac{e_k}{\pi_k} \frac{e_l}{\pi_l}$$

avec $e_k = y_k - \frac{\hat{t}_{yh}}{\hat{N}_h}$ pour $k \in S_h$.

Exemples de méthodes de redressement

Post-stratification (5)

Application sur la population d'aéroports :

```
> Xech=cbind(c(1,0,0,0,0,0),c(0,1,1,1,1,1))
> Xtot=c(6,6)
> gweight<-calib(Xech,dech,Xtot,method="linear")
> gweight
[1] 3.0 0.6 0.6 0.6 0.6 0.6

#Variable Trans20
> calibev(ech$Trans20,Xech,Xtot,pikl_ech,dech,gweight,
            with=TRUE)
$calest
[1] 11 520
$evar
[1] 15 667 200
```

Exemples de méthodes de redressement

Post-stratification (6)

	Pass19	Strate	Pass20	Trans20	d_k	w_k
AJACCIO	1 500 000	1	900 000	1 300	2	6
RENNES	900 000	2	300 000	200	2	1.2
FIGARI	700 000	2	500 000	2 900	2	1.2
TOULON	500 000	2	200 000	0	2	1.2
PERPIGNAN	500 000	2	200 000	0	2	1.2
TARBES	500 000	2	100 000	0	2	1.2
t_x	12 300 000	$N_1 = 6$				
$\hat{t}_{\bullet\pi}$	9 200 000	$\hat{N}_1 = 2$	4 400 000	8 800		
$\hat{V}_{HT}(\hat{t}_{\bullet\pi})$	$1.86\ 10^{12}$		$1.04\ 10^{12}$	$1.66\ 10^{7}$		
$\hat{t}_{\bullet w}$	12 720 000	$\hat{N}_{1,post} = 6$	6 960 000	11 500		
$\hat{V}_{HT}(\hat{t}_{\bullet w})$	$3.07\ 10^{11}$		$2.21\ 10^{11}$	$1.57\ 10^{7}$		

Exemples de méthodes de redressement

En résumé

L'estimateur par la régression généralisée s'écrit

$$\hat{t}_{y,greg} = \hat{t}_{y\pi} + \hat{\mathrm{b}}_{\pi}^{\top} \left\{ t_{\mathrm{x}} - \hat{t}_{\mathrm{x}\pi} \right\}.$$

Il conduit à une variance plus faible que l'estimateur de HT si les variables auxiliaires x_k sont fortement explicatives de y_k.

Deux cas particuliers sont :

- l'estimateur par le ratio $\hat{t}_{yR} = \hat{t}_{y\pi} \times \dfrac{t_{x_1}}{\hat{t}_{x_1\pi}}$, qui conduit à une variance faible si y_k est approximativement proportionnelle à x_{1k},
- l'estimateur post-stratifié $\hat{t}_{y,post} = \sum_{h=1}^{H} \dfrac{N_h}{\hat{N}_h} \hat{t}_{yh}$, qui conduit à une variance faible si les post-strates sont homogènes par rapport à y_k.

Estimation d'une fonction de totaux

Estimateur par substitution

Nous nous intéressons à un paramètre de la forme $\theta = f(t_y)$ avec $y_k = (y_{1k}, \ldots, y_{qk})^T$ un q-vecteur de variables d'intérêt, et $f : \mathbb{R}^q \to \mathbb{R}$.

Il est naturel d'estimer θ en remplaçant le total t_y inconnu par son estimateur de Horvitz-Thompson. Nous obtenons l'*estimateur par substitution* :

$$\hat{\theta}_\pi = f(\hat{t}_{y\pi}).$$

Si la fonction $f(\cdot)$ est différentiable au voisinage de t_y, nous avons :

$$N^{-1}\left(\hat{\theta}_\pi - \theta\right) = N^{-1}\left(\hat{t}_{u\pi} - t_u\right) + o_p(n^{-1/2}), \tag{4.1}$$

en notant $u_k = \{f'(t_y)\}^T \{y_k\}$ la *variable linéarisée* du paramètre θ

Exemple : variable linéarisée d'une ratio

Nous nous intéressons à $R = \frac{t_{y_1}}{t_{y_2}} = f(t_y)$, avec $y_k = (y_{1k}, y_{2k})^\top$ et

$$\begin{array}{rccc} f: & \mathbb{R}^2 & \longrightarrow & \mathbb{R} \\ & (u,v) & \mapsto & \dfrac{u}{v}. \end{array}$$

Nous l'estimons par substitution par

$$\hat{R}_\pi = f(\hat{t}_{y\pi}) = \frac{\hat{t}_{y_1\pi}}{\hat{t}_{y_2\pi}}.$$

Nous avons

$$f'(u,v) = \left(\frac{1}{v}, -\frac{u}{v^2}\right)^\top,$$

ce qui donne la variable linéarisée

$$u_k(R) = \frac{1}{t_{y_2}} y_{1k} - \frac{t_{y1}}{(t_{y_2})^2} y_{2k} = \frac{1}{t_{y2}}(y_{1k} - R y_{2k}).$$

Estimation de variance

L'approximation (4.1) nous donne

$$\begin{aligned} E_p(\hat{\theta}_\pi - \theta) &\simeq 0, && \text{(estimation approx. sans biais)} \\ V_p(\hat{\theta}_\pi - \theta) &\simeq V_p\left\{\hat{t}_{u\pi}\right\}. && \text{(variance donnée par la linéarisée).} \end{aligned}$$

La seconde ligne nous donne l'*approximation de variance par linéarisation*. Nous obtenons

$$\begin{aligned} V_p(\hat{\theta}_\pi) &\simeq \sum_{k,l\in U} \Delta_{kl} \frac{u_k}{\pi_k} \frac{u_l}{\pi_l}, \\ \hat{V}_{HT,1}(\hat{\theta}_\pi) &= \sum_{k,l\in S} \frac{\Delta_{kl}}{\pi_{kl}} \frac{\hat{u}_k}{\pi_k} \frac{\hat{u}_l}{\pi_l}, \end{aligned}$$

où la *variable linéarisée estimée* $\hat{u}_k = \left\{f'(\hat{t}_{y\pi})\right\}\{y_k\}$ s'obtient en remplaçant dans u_k les totaux inconnus par leurs estimateurs de HT.

Exemple : variable linéarisée d'une ratio

Nous nous intéressons à $R = \frac{t_{y_1}}{t_{y_2}}$. Nous avons vu que

$$u_k(R) = \frac{1}{t_{y2}}(y_{1k} - R y_{2k}).$$

En remplaçant dans $u_k(R)$ le total t_{y_1} par l'estimateur $\hat{t}_{y_1\pi}$ et le total t_{y_2} par l'estimateur $\hat{t}_{y_2\pi}$, nous obtenons

$$\hat{u}_k(R) = \frac{1}{\hat{t}_{y_2\pi}}(y_{1k} - \hat{R}_\pi y_{2k}).$$

Estimation de variance :

$$V_p(\hat{R}_\pi) \simeq \sum_{k,l\in U} \frac{u_k(R)}{\pi_k}\frac{u_l(R)}{\pi_l}\Delta_{kl},$$

$$\hat{V}_{HT,1}(\hat{R}_\pi) = \sum_{k,l\in S} \frac{\hat{u}_k(R)}{\pi_k}\frac{\hat{u}_l(R)}{\pi_l}\frac{\Delta_{kl}}{\pi_{kl}}.$$

Base de données d'aéroports

Estimation du taux de passagers en transit

Nous souhaitons estimer le taux de passagers en transit :

$$R = \frac{\text{Nb passagers en transit en 2020}}{\text{Nb passagers en 2020}} \equiv \frac{t_{y_1}}{t_{y_2}}.$$

Nous considérons le cas :

- d'un sondage aléatoire simple :

$$\hat{R}_\pi = \frac{\hat{t}_{y_1\pi}}{\hat{t}_{y_2\pi}} = \frac{\sum_{k\in S} y_{1k}}{\sum_{k\in S} y_{2k}},$$

- d'un sondage aléatoire simple stratifié :

$$\hat{R}_\pi = \frac{\hat{t}_{y_1\pi}}{\hat{t}_{y_2\pi}} = \frac{\sum_{h=1}^{H} N_h \bar{y}_{1h}}{\sum_{h=1}^{H} N_h \bar{y}_{2h}}.$$

Base de données d'aéroports

Estimation du taux de passagers en transit : SRS

	Pass19	Pop19	Pass20 (y_{2k})	Trans20 (y_{1k})	$\hat{u}_k$
AJACCIO	1 500 000	100 000	900 000	1 300	$-1.14\ 10^{-4}$
RENNES	900 000	730 000	300 000	200	$-9.09\ 10^{-5}$
FIGARI	700 000	20 000	500 000	2 900	$4.32\ 10^{-4}$
TOULON	500 000	630 000	200 000	0	$-9.09\ 10^{-5}$
PERPIGNAN	500 000	320 000	200 000	0	$-9.09\ 10^{-5}$
TARBES	500 000	120 000	100 000	0	$-4.55\ 10^{-5}$
			$\bar{y}_2 = 367\ 000$ $s^2_{y2} = 8.67\ 10^{10}$	$\bar{y}_1 = 733$ $s^2_{y1} = 1.38\ 10^{6}$	$\bar{\hat{u}} = 0$ $s^2_{\hat{u}} = 4.52\ 10^{-8}$

$$\hat{t}_{y_2\pi} = N\bar{y}_2 = 4\ 400\ 000 \qquad \hat{V}_{HT}\left(\hat{t}_{y_2\pi}\right) = N^2\frac{1-f}{n}s^2_{y2} = 1.04\ 10^{12}$$

$$\hat{t}_{y_1\pi} = N\bar{y}_1 = 8\ 800 \qquad \hat{V}_{HT}\left(\hat{t}_{y_1\pi}\right) = N^2\frac{1-f}{n}s^2_{y1} = 1.66\ 10^{7}$$

$$\hat{R} = \frac{\bar{y}_1}{\bar{y}_2} = 0.2\% \qquad \hat{V}_{HT}\left(\hat{R}\right) = N^2\frac{1-f}{n}s^2_{\hat{u}} = 5.43\ 10^{-7}.$$

Base de données d'aéroports

Estimation du taux de passagers en transit : SRS

```
#Recuperation du jeu de donnees
> aeroports <- read.csv("../aeroports.csv", header=TRUE)
#Echantillon selectionne par SRS
> ech=c(0,0,1,0,0,0,1,1,0,1,1,1)
> y2ech=aeroports[ech==1,4]
> y1ech=aeroports[ech==1,5]
# Probabilites d'inclusion
> n=6
> Npop=12
> pi=rep(n/Npop,Npop)
> pikl=UPmaxentropypi2(pi)
> pikl_ech=pikl[ech==1,ech==1]
> varest=vartaylor_ratio(y1ech,y2ech,pikl_ech)
> varest
$ratio
[1] 0.002
$estvar
[1] 5.429752e-07
```

Base de données d'aéroports

Estimation du taux de passagers en transit : STSRS

	Pass19	Pop19	Pass20 (y_{2k})	Trans20 (y_{1k})	$\hat{u}_k$
BASTIA	1 600 000	100 000	800 000	1 400	$-1.14\ 10^{-4}$
AJACCIO	1 500 000	100 000	900 000	1 300	$-9.09\ 10^{-5}$
STRASBOURG	1 300 000	800 000	500 000	1 300	$4.32\ 10^{-4}$
			$\bar{y}_{2,1} = 733\ 333$ $s^2_{y2,1} = 4.33\ 10^{10}$	$\bar{y}_{1,1} = 1\ 333$ $s^2_{y1,1} = 3.33\ 10^{3}$	$\bar{\hat{u}}_1 = 0$ $s^2_{\hat{u},1} = 2.50\ 10^{-7}$
RENNES	900 000	730 000	300 000	200	$-9.09\ 10^{-5}$
TOULON	500 000	630 000	200 000	0	$-9.09\ 10^{-5}$
PERPIGNAN	500 000	320 000	200 000	0	$-4.55\ 10^{-5}$
			$\bar{y}_{2,2} = 233\ 333$ $s^2_{y2,2} = 3.33\ 10^{9}$	$\bar{y}_{1,2} = 67$ $s^2_{y1,2} = 1.33\ 10^{4}$	$\bar{\hat{u}}_2 = 0$ $s^2_{\hat{u},2} = 5.00\ 10^{-9}$

$$\hat{t}_{y_2\pi} = \sum_{h=1}^{2} N_h \bar{y}_{2h} = 5\ 800\ 000 \qquad \hat{V}_{HT}\left(\hat{t}_{y_2\pi}\right) = \sum_{h=1}^{2} N_h^2 \frac{1-f_h}{n_h} s^2_{y_2 h} = 2.8\ 10^{11}$$

$$\hat{t}_{y_1\pi} = \sum_{h=1}^{2} N_h \bar{y}_{1h} = 8\ 400 \qquad \hat{V}_{HT}\left(\hat{t}_{y_1\pi}\right) = \sum_{h=1}^{2} N_h^2 \frac{1-f_h}{n_h} s^2_{y_1 h} = 10^5$$

$$\hat{R} = \frac{\sum_{h=1}^{2} N_h \bar{y}_{1h}}{\sum_{h=1}^{2} N_h \bar{y}_{2h}} = 0.14\% \qquad \hat{V}_{HT}\left(\hat{R}\right) = \sum_{h=1}^{2} N_h^2 \frac{1-f_h}{n_h} s^2_{\hat{u}h} = 1.53\ 10^{-6}.$$

En résumé

Un paramètre complexe qui s'écrit comme une fonction de totaux peut être estimé approximativement sans biais par substitution :

$$\theta = f(t_y) \quad \text{estimé par} \quad \hat{\theta}_\pi = f(\hat{t}_{y\pi}).$$

Nous obtenons un estimateur de variance en insérant la variable linéarisée estimée $\hat{u}_k = \left\{f'(\hat{t}_{y\pi})\right\}^T \{y_k\}$ dans l'estimateur de variance associé au plan de sondage :

$$\hat{V}_{HT,1}\left(\hat{\theta}_\pi\right) = \sum_{k,l \in S} \frac{\hat{u}_k}{\pi_k} \frac{\hat{u}_l}{\pi_l} \frac{\Delta_{kl}}{\pi_{kl}}.$$

En utilisant une approximation normale pour $\hat{\theta}_\pi$, nous obtenons l'intervalle de confiance

$$\left[\hat{\theta}_\pi \pm z_{1-\frac{\alpha}{2}} \sqrt{\hat{V}_{HT,1}\left(\hat{\theta}_\pi\right)}\right].$$

Références

- Ardilly, P. (2006). Les techniques de sondage. Editions Technip.
- Deville, J. C., and Särndal, C. E. (1992). Calibration estimators in survey sampling. Journal of the American statistical Association, 87(418), 376-382.
- Särndal, C. E., Swensson, B., and Wretman, J. (1992). Model assisted survey sampling. Springer Science and Business Media.
- Tillé, Y. (2006). Sampling algorithms (pp. 31-39). Springer New York.
- Tillé, Y. (2020). Sampling and estimation from finite populations. John Wiley and Sons.
- Tillé, Y. and Matei, A. (2023). Package "sampling", documentation de l'utilisateur.