Méthodes de sondages
Partie 1 : échantillonnage en population finie

Guillaume Chauvet

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

30/09/2025

Problématique

Utiliser un échantillon d'unités pour tirer des conclusions sur un ensemble de données plus grand :

- quand nous contrôlons la façon dont l'échantillon est sélectionné (tirage probabiliste) :
  - Sondage aléatoire simple,
  - Sondage stratifié,
  - Sondage à plusieurs degrés.
- quand nous ne contrôlons pas la façon dont l'échant. est sélectionné :
  - Méthodes de tirage empiriques (échantillonnage par quotas) pour connaître une opinion politique, les habitudes en termes de médias.
  - Echantillons de volontaires (enquête de satisfaction).
  - Non-réponse : diminue la taille de l'échantillon effectivement observé, et tend à sur-représenter des profils particuliers (risque de biais).

Principaux objectifs du cours

- Présenter les méthodes d'inférence pour une population finie d'individus.
- Donner les principales méthodes d'échantillonnage dans les enquêtes.
- Décrire les méthodes de redressement qui permettent d'utiliser une information auxiliaire au moment de l'estimation.
- Décrire les méthodes de correction de la non-réponse dans les enquêtes.

Nous utiliserons :

- le package R `sampling` pour l'échantillonnage,
- le package R `gustave` pour l'estimation de variance (créé et maintenu par Martin Chevalier et Khaled Larbi, Insee).

```
#Appel des packages
> library(sampling)
> help(package="sampling")
> library(gustave)
> help(package="gustave")
```

Package sampling

Le package `sampling` contient des fonctions permettant :

- de sélectionner des échantillons, à probabilités égales ou inégales, en stratifiant la population,
- de réaliser des estimations de totaux,
- de calculer des estimateurs par *calage*,
- de réaliser des estimations de variance,
- de corriger de la non-réponse totale.

Base de sondage d'aéroports

A titre d'illustration, nous considérerons en fil rouge une *base de sondage*[1] 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`) et taille de l'*Unité Urbaine*[2] en 2019 (`Pop19`) : ce seront nos *variables auxiliaires*, supposées connues sur la population entière.
- Nombre de passagers en 2020 (`Pass20`) et Nombre de passagers en transit en 2020 (`Trans20`) : ce seront nos *variables d'intérêt*, supposées observées sur un échantillon seulement.

[1] Liste des individus dont nous disposons et dans laquelle nous échantillonnons pour réaliser une estimation dans la population d'intérêt.

[2] Commune ou ensemble de communes présentant une zone de bâti continu (pas de coupure de plus de 200 mètres entre deux constructions) et qui compte au moins 2 000 habitants.

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*
Total t_x	12 300 000	4 300 000		
Moyenne μ_x	1 025 000	358 333		
Dispersion S_x^2	$2.33\ 10^{11}$	$7.28\ 10^{10}$		
$cv_x = \sqrt{S_x^2}/\mu_x$	47%	76%		

Base de sondage de municipalités

Nous utiliserons également la base de sondage `belgianmunicipalities` disponibles avec `sampling`.

Elle fournit des informations sur les 589 communes de Belgique au 01/07/2004, ainsi que des informations financières datées de 2001.

```
#Récupération de deux bases de données du package
> data("belgianmunicipalities")
> attach(belgianmunicipalities)
```

Variables de "belgianmunicipalities"

Commune	Municipality name
INS	INS Code INS
Province	Province number
Arrondiss	Administrative division number
Men04	Number of men on July 1, 2004
Women04	Number of women on July 1, 2004
Tot04	Total population on July 1, 2004
Men03	Number of men on July 1, 2003
Women03	Number of women on July 1, 2003
Tot03	Total population on July 1, 2003
Diffmen	Men04 minus Men03
Diffwom	Women04 minus Women03
DiffTOT	Tot04 minus Tot03
TaxableIncome	Total taxable income in euros in 2001
Totaltaxation	Total taxation in euros in 2001
Averageincome	Average of the income-tax return in euros in 2001
Medianincome	median of the income-tax return in euros in 2001.

Plan de la première partie

1. Echantillonnage en population finie
   - Notations
   - Plan de sondage
   - Estimation de Horvitz-Thompson
   - Calcul de précision
2. Méthodes d'échantillonnage
   - Sondage aléatoire simple
   - Sondage aléatoire simple stratifié
   - Tirage à probabilités inégales
   - Echantillonnage à plusieurs degrés

Echantillonnage en population finie

Notations

Notations

Nous nous plaçons dans le cadre d'une population finie U d'*individus* ou *unités statistiques*, supposés identifiables par un label. Nous noterons

$$U = \{1, \ldots, k, \ldots, N\}.$$

où N désigne la taille de la population U.

Nous nous intéressons à une *variable d'intérêt* y (souvent vectorielle), qui prend la valeur y_k sur l'individu k de U.
La variable y est vue ici comme non aléatoire : la population U étant fixée, **la valeur prise par y sur chaque individu est parfaitement définie et déterministe**.

Nous souhaitons estimer des paramètres de la population U. Le *champ de l'enquête* désigne les caractéristiques des unités de la population d'intérêt. Il est très important de les définir aussi précisément que possible, et en particulier les catégories qui en sont exclues.

Types d'unités

Nous appellerons *unité d'échantillonnage* une unité élémentaire susceptible d'être tirée lorsque nous procédons à l'échantillonnage, et par l'intermédiaire de laquelle l'information est collectée.

L'*unité d'observation* est l'unité de base sur laquelle l'information est collectée. L'ensemble de ces unités constitue la *population d'intérêt*.

Par exemple :

- une enquête auprès des ménages peut procéder en échantillonnant des logements (unité d'échantillonnage), et en enquêtant tout ou partie des ménages (unité d'observation) situés dans ces logements;
- l'inventaire forestier national procède en échantillonnant des points sur le territoire (unité d'échantillonnage), puis en mesurant des arbres (unité d'observation) situés à proximité de ces points.

Paramètres d'intérêt

Nous nous intéresserons principalement à l'estimation d'un total

$$t_y = \sum_{k \in U} y_k$$

d'une variable quantitative sur la population, ou encore à celle de sa moyenne

$$\mu_y = \frac{1}{N} \sum_{k \in U} y_k.$$

Exemple :
Chiffre d'affaires total des entreprises d'un secteur d'activité, pourcentage d'étudiants fumeurs, ...

Nous étudierons plus loin des paramètres plus complexes comme un ratio

$$R = \frac{\sum_{k \in U} y_k}{\sum_{k \in U} x_k} = \frac{t_y}{t_x}.$$

Paramètres d'intérêt (2)

Un cas particulier important est celui de l'estimation sur une sous-population U_d (appelée *domaine*) de :

$$t_{yd} = \sum_{k \in U_d} y_k : \text{sous-total sur le domaine,}$$

$$\mu_{yd} = \frac{1}{N_d} \sum_{k \in U_d} y_k : \text{moyenne sur le domaine,}$$

avec N_d la taille du domaine U_d.

Il peut s'agir d'un domaine au sens géographique (habitants d'une région), socio-démographique (individus de moins de 20 ans), ou encore temporel (individus présents à une date donnée).

Paramètres d'intérêt (3)

Dans certains cas, il sera utile de voir une moyenne comme un cas particulier de ratio.

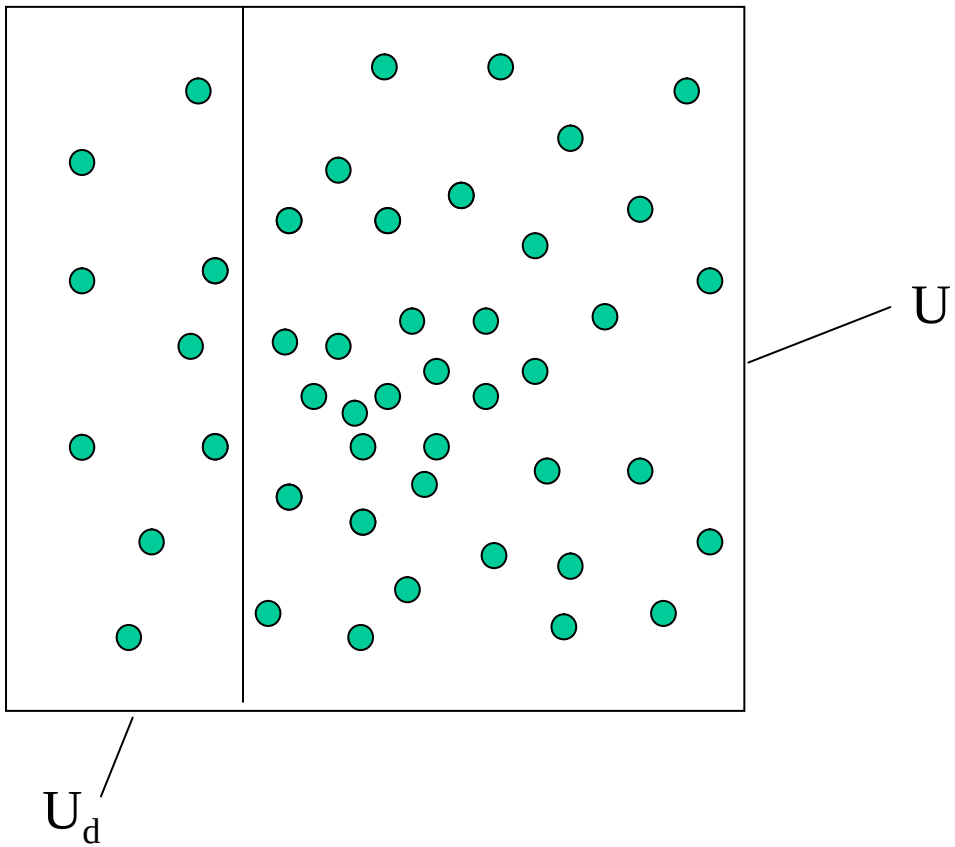
La moyenne sur la population peut se réécrire

$$\mu_y = \frac{t_y}{t_x} \quad \text{avec} \quad x_k = 1.$$

La moyenne sur un domaine peut se réécrire

$$\mu_{yd} = \frac{t_z}{t_x} \quad \text{avec} \quad \left\{ \begin{array}{ll} z_k & = y_k \, 1\left(k \in U_d\right), \\ x_k & = 1\left(k \in U_d\right). \end{array} \right.$$

Une population et un domaine d'intérêt



Plan de sondage

Plan de sondage

La sélection de l'échantillon aléatoire S se fait à l'aide d'un *plan de sondage* p sur U, c'est à dire à l'aide d'une loi de probabilité sur les parties de U :

$$\forall s \subset U \quad p(s) \geq 0 \text{ et } \sum_{s \subset U} p(s) = 1. \tag{1.1}$$

Nous notons S l'échantillon aléatoire, et nous distinguerons :

- le *paramètre d'intérêt* $\theta(y_k \ , \ k \in U) \equiv \theta$: quantité déterministe,
- l'*estimateur* $\hat{\theta}(y_k \ , \ k \in S) \equiv \hat{\theta}(S) \equiv \hat{\theta}$: variable aléatoire,
- l'*estimation* $\hat{\theta}(y_k \ , \ k \in s) \equiv \hat{\theta}(s)$: réalisation de l'estimateur, pour une partie donnée $s \subset U$.

Nous appellerons *algorithme d'échantillonnage* une méthode pratique permettant de sélectionner un échantillon selon le plan de sondage choisi.

Exemple

Soit la population $U = \{1, 2, 3, 4\}$, et $p(\cdot)$ le plan de sondage défini par :

$$\begin{array}{lllllllll} p(\{1,2\}) & = & 0.2 & p(\{1,4\}) & = & 0.1 & p(\{3,4\}) & = & 0.3 \\ p(\{1,2,3\}) & = & 0.3 & p(\{2,3,4\}) & = & 0.1 & & & \end{array}$$

La variable aléatoire S prend ses valeurs dans

$$\{\{1,2\}, \{1,4\}, \{3,4\}, \{1,2,3\}, \{2,3,4\}\}.$$

Nous avons par exemple

$$\mathbb{P}(S = \{1,2\}) = p(\{1,2\}) = 0.2$$

A la différence des lois de probabilités classiques (normale, exponentielle, binomiale, ...) l'aléatoire ne porte pas sur la variable mais sur le sous-ensemble d'individus observés.

Comparaison avec une variable aléatoire réelle

Soit X une variable aléatoire distribuée selon une loi de Poisson $\mathcal{P}(\lambda)$. La variable aléatoire X prend ses valeurs dans

$$\mathbb{N} = \{0, 1, 2, \ldots\}.$$

Nous avons pour $k \in \mathbb{N}$:

$$\mathbb{P}(X = k) = \exp^{-\lambda} \times \frac{\lambda^k}{k!}.$$

L'espérance de X correspond à la valeur moyenne de ses valeurs possibles, pondérées par leurs probabilités :

$$\begin{aligned} E[X] &= \sum_{k \in \mathbb{N}} k \times \mathbb{P}(X = k) \\ &= \lambda. \end{aligned}$$

Mesures de précision

L'espérance d'un estimateur $\hat{\theta}(S)$ se définit de façon analogue :

$$E_p\left[\hat{\theta}(S)\right] = \sum_{s\subset U} \hat{\theta}(s) \times \mathbb{P}(S=s) = \sum_{s\subset U} p(s)\ \hat{\theta}(s). \qquad (1.2)$$

Le biais d'un estimateur $\hat{\theta}(S)$ correspond à son erreur moyenne :

$$B_p\left[\hat{\theta}(S)\right] = E_p\left[\hat{\theta}(S)-\theta\right] = \sum_{s\subset U} p(s)\left[\hat{\theta}(s)-\theta\right]. \qquad (1.3)$$

Nous nous intéresserons aussi à la Variance et à l'Erreur Quadratique Moyenne (EQM) :

$$V_p\left[\hat{\theta}(S)\right] = \sum_{s\subset U} p(s)\left\{\hat{\theta}(s) - E_p[\hat{\theta}(S)]\right\}^2, \qquad (1.4)$$

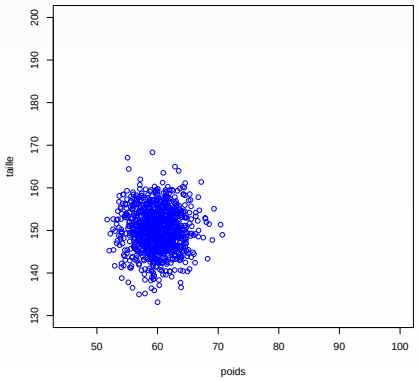
$$EQM_p\left[\hat{\theta}(S)\right] = E_p\left[\left\{\hat{\theta}(S)-\theta\right\}^2\right] = B_p\left[\hat{\theta}(S)\right]^2 + V_p\left[\hat{\theta}(S)\right].$$

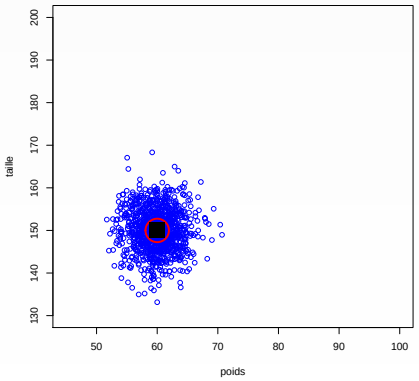
Quelques simulations (cas 1)

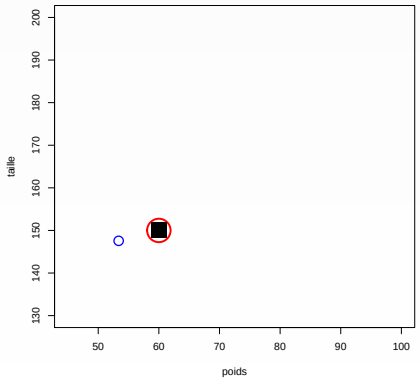
Pour illustrer la notion de biais et de variance, nous considérons l'exemple d'une population de $N = 1\ 000$ individus âgés de 15 à 20 ans.

Dans cette population, un échantillon de taille $n = 50$ est sélectionné et enquêté. Pour chaque individu enquêté, nous obtenons son poids (en kg) et sa taille (en cm).

Nous nous intéressons à l'estimation du poids moyen et de la taille moyenne (carré noir). Chaque échantillon permet d'obtenir une estimation (points bleus) de ces paramètres. La moyenne des estimations est représentée par le point rouge.





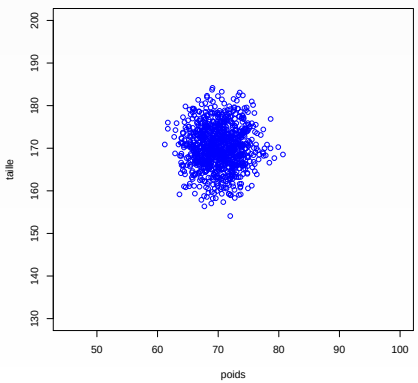


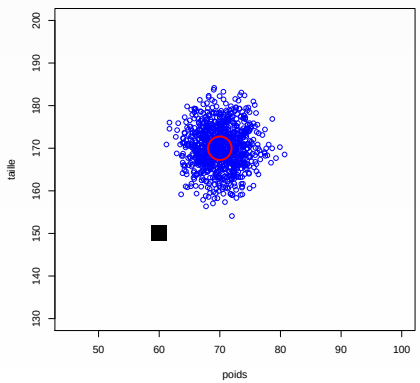
Quelques simulations (cas 2)

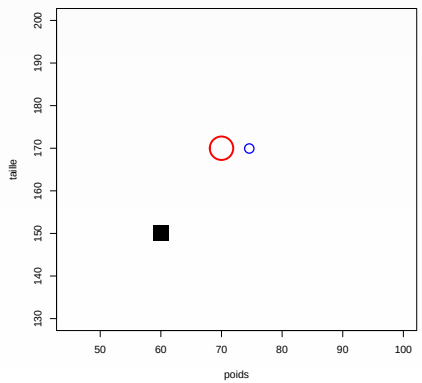
Pour illustrer la notion de biais et de variance, nous considérons l'exemple d'une population de $N = 1\ 000$ individus âgés de 15 à 20 ans.

Dans cette population, un échantillon de taille $n = 50$ est sélectionné et enquêté. Pour chaque individu enquêté, nous obtenons son poids (en kg) et sa taille (en cm).

Nous nous intéressons à l'estimation du poids moyen et de la taille moyenne (carré noir). Chaque échantillon permet d'obtenir une estimation (points bleus) de ces paramètres. La moyenne des estimations est représentée par le point rouge.





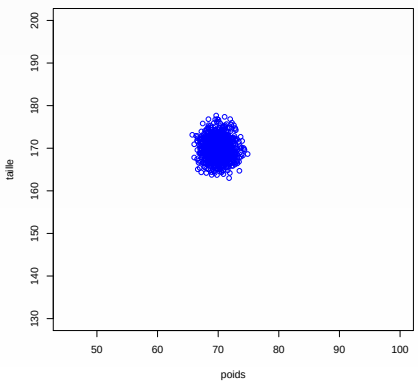


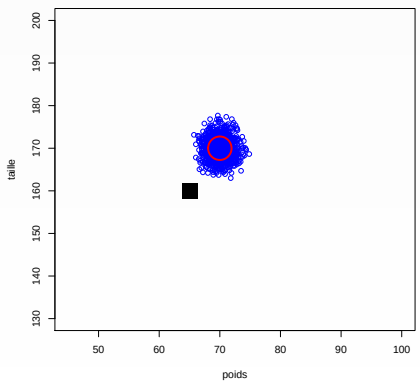
Quelques simulations (cas 3)

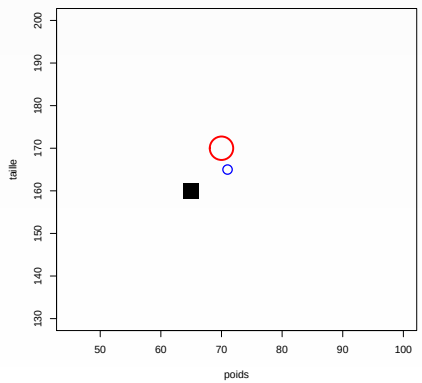
Pour illustrer la notion de biais et de variance, nous considérons l'exemple d'une population de $N = 1\ 000$ individus âgés de 15 à 20 ans.

Dans cette population, un échantillon de taille $n = 50$ est sélectionné et enquêté. Pour chaque individu enquêté, nous obtenons son poids (en kg) et sa taille (en cm).

Nous nous intéressons à l'estimation du poids moyen et de la taille moyenne (carré noir). Chaque échantillon permet d'obtenir une estimation (points bleus) de ces paramètres. La moyenne des estimations est représentée par le point rouge.







Probabilités d'inclusion

Nous notons π_k la *probabilité d'inclusion* de l'unité k, c'est à dire la probabilité que l'unité k soit retenue dans l'échantillon :

$$\pi_k \quad = \quad \mathbb{P}(k \in S) = \sum_{s/k \in s} p(s). \tag{1.5}$$

En pratique, les probabilités d'inclusion π_k sont fixées avant le tirage à l'aide d'une information auxiliaire. Nous utilisons ensuite un plan de sondage qui respecte ces probabilités d'inclusion.

Nous notons π_{kl} la probabilité que deux unités distinctes k et l soient sélectionnées conjointement dans l'échantillon :

$$\pi_{kl} \quad = \quad \mathbb{P}(k, l \in S) = \sum_{s/k,l \in s} p(s). \tag{1.6}$$

Ces probabilités doubles interviennent notamment dans la variance des estimateurs. Il est souvent difficile de les calculer exactement.

Variables indicatrices

L'utilisation de la variable $I_k = 1(k \in S)$, indiquant l'appartenance à l'échantillon de l'unité k, permet souvent de simplifier les calculs.

Pour deux unités k et l distinctes, nous avons les propriétés suivantes :

$$E_p(I_k) = \pi_k, \tag{1.7}$$
$$V_p(I_k) = \pi_k(1-\pi_k), \tag{1.8}$$
$$Cov_p(I_k, I_l) = \pi_{kl} - \pi_k\pi_l \equiv \Delta_{kl}, \tag{1.9}$$
$$E_p[n(S)] = \sum_{k\in U} \pi_k. \tag{1.10}$$

Nous notons $\Delta = [\Delta_{kl}]_{k,l\in U}$ la matrice de variance-covariance du plan de sondage $p(\cdot)$.

En résumé

Un plan de sondage est une loi de probabilité sur les parties de U. L'alea porte sur le sous-ensemble S d'individus observés.

Les notions d'espérance et de variance d'un estimateur $\hat{\theta}(S)$ s'adaptent de façon naturelle :

$$\begin{aligned} B_p\left[\hat{\theta}(S)\right] &= \sum_{s\subset U} p(s)\left[\hat{\theta}(s)-\theta\right], \\ V_p\left[\hat{\theta}(S)\right] &= \sum_{s\subset U} p(s)\left\{\hat{\theta}(s)-E_p[\hat{\theta}(S)]\right\}^2. \end{aligned}$$

Nous appelons probabilités d'inclusion d'ordre 1 et 2 :

$$\begin{aligned} \pi_k &= \mathbb{P}(k\in S), \\ \pi_{kl} &= \mathbb{P}(k,l\in S). \end{aligned}$$

Exercice

Soit la population $U = \{1, 2, 3, 4\}$, et $p(\cdot)$ le plan de sondage défini par :

$$\begin{array}{llllll} p(\{1,2\}) = 0.2 & p(\{1,4\}) = 0.1 & p(\{3,4\}) = 0.3 \\ p(\{1,2,3\}) = 0.3 & p(\{2,3,4\}) = 0.1 \end{array}$$

Calculer les probabilités d'inclusion d'ordre 1. $[\pi_1 = 0.6]$

Montrer que la taille moyenne d'échantillon obtenue est égale à 2.4.

Donner les probabilités d'inclusion d'ordre 2 :

- des unités 1 et 2, $[\pi_{12} = 0.5]$
- des unités 1 et 4,
- des unités 2 et 4.

Estimation de Horvitz-Thompson

Estimateur de Horvitz-Thompson

Nous nous intéressons à l'estimation du total

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

L'estimateur de Horvitz-Thompson est défini par

$$\hat{t}_{y\pi} = \sum_{k \in S} \frac{y_k}{\pi_k} = \sum_{k \in U;\ \pi_k > 0} \frac{y_k}{\pi_k} I_k. \tag{1.11}$$

C'est un estimateur pondéré, utilisable pour n'importe quelle variable d'intérêt.

Principe : un individu k de l'échantillon représente $d_k = 1/\pi_k$ individus de la population.

Estimateur de Horvitz-Thompson (2)

Proposition 1

$$E_p\left[\hat{t}_{y\pi}\right] = t_y - \sum_{\substack{k\in U \\ \pi_k=0}} y_k. \tag{1.12}$$

L'estimateur de Horvitz-Thompson est donc non biaisé pour le total t_y si tous les π_k sont > 0, ce que nous supposerons dans la suite du cours.

Certaines probabilités d'inclusion peuvent être nulles, notamment :

- en cas de défaut de couverture de la base de sondage (liste des individus pas à jour, ou individus impossibles à joindre),
- quand nous choisissons délibérément de laisser de côté une partie de la population (cut-off sampling, parfois utilisé dans les enquêtes-entreprise).

Dans ce cas, il faut parfois redéfinir le champ de l'enquête.

Variance

La variance de l'estimateur de Horvitz-Thompson est donnée par

$$V_p(\hat{t}_{y\pi}) = \sum_{k,l \in U} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} \Delta_{kl} \quad \text{où} \quad \Delta_{kl} = \pi_{kl} - \pi_k \pi_l. \tag{1.13}$$

Cette variance peut être estimée sans biais par

$$\hat{V}_{HT}(\hat{t}_{y\pi}) = \sum_{k,l \in S} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} \frac{\Delta_{kl}}{\pi_{kl}} \tag{1.14}$$

si tous les π_{kl} sont strictement positifs. Il s'agit de l'*estimateur de variance de Horvitz-Thompson*.

Principe : un couple (k, l) d'individus de l'échantillon représente $1/\pi_{kl}$ couples de la population.

Variance pour un plan de taille fixe

Formule de Sen-Yates-Grundy

Un plan de sondage est dite *de taille fixe* égale à n si seuls les échantillons de taille n ont une probabilité non nulle d'être tirés.

Pour un plan de taille fixe, la variance de $\hat{t}_{y\pi}$ peut se réécrire

$$V_p(\hat{t}_{y\pi}) = -\frac{1}{2} \sum_{k \neq l \in U} \left[\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right]^2 \Delta_{kl}. \tag{1.15}$$

D'après cette formule, la variance est nulle si $\pi_k \propto y_k$. Ce choix n'est pas possible en pratique (les variables d'intérêt sont inconnues avant l'enquête).

Il est possible de s'en rapprocher en choisissant les probabilités d'inclusion proportionnellement à une mesure de taille des unités. On parle de *tirage à probabilités proportionnelles à la taille*.

Exemple d'échantillonnage et d'estimation en R

```
#Calcul de probabilités d'inclusion prop. à la taille
> n=50
> pi_50=inclusionprobabilities(averageincome,n)
> summary(pi_50)
[1]  Min.  1st Qu. Median   Mean 3rd Qu.   Max.
[1]0.05693 0.07675 0.08375 0.08489 0.09113 0.14076

#Tirage selon un plan de taille fixe a entropie maximale
> ech=UPmaxentropy(pi_50)

#Estimation de HT du total de TaxableIncome
> y=TaxableIncome
> est_ht=HTestimator(y[ech==1],pi_50[ech==1])
```

Exemple d'échantillonnage et d'estimation en R

```
#Calcul des probabilités d'inclusion d'ordre 2 pour
#le tirage a entropie maximale
> pikl_rej_50=UPmaxentropypi2(pi_50)

#Estimation de variance de HT
> vest_ht=varHT(y[ech==1],pikl_rej_50[ech==1,ech==1],1)
#Estimation de variance de SYG
> vest_yg=varHT(y[ech==1],pikl_rej_50[ech==1,ech==1],2)
> options("scipen"=-100,digits="4")
> est_ht
[1,] 1.033e+11
> vest_ht
[1] 9.941e+19
> vest_yg
[1] 9.927e+19
```

Exercice

1) Montrer que

$$V_p\left[n(S)\right] = \sum_{k,l \in U} \Delta_{kl}.$$

2) Montrer que pour un plan de sondage de taille fixe, nous avons :

$$\sum_{l \in U} \Delta_{kl} \quad = \quad 0 \text{ pour tout } k \in U.$$

3) En déduire que pour un plan de sondage de taille fixe, $V_p\left[n(S)\right] = 0$. Obtenez directement ce résultat à partir de l'équation (1.15).

Calcul de précision

Intervalle de confiance

Nous supposons que $\hat{t}_{y\pi}$ estime sans biais t_y. Alors un intervalle de confiance pour t_y de niveau approximatif $1-\alpha$ est donné par :

$$IC_{1-\alpha}(t_y) = \left[\hat{t}_{y\pi} \pm z_{1-\frac{\alpha}{2}} \sqrt{V_p(\hat{t}_{y\pi})}\right], \quad (1.16)$$

avec $z_{1-\frac{\alpha}{2}}$ le quantile d'ordre $1-\frac{\alpha}{2}$ d'une loi normale centrée réduite $\mathcal{N}(0,1)$.

Rappel :

- $\alpha = 0.05 \Rightarrow z_{0.975} = 1.96$
- $\alpha = 0.10 \Rightarrow z_{0.95} = 1.64$

Interprétation (pour $\alpha = 0.05$) : le vrai total t_y est contenu dans l'intervalle de confiance pour (approximativement) 95% des échantillons.

Intervalle de confiance

Comme la vraie variance $V_p(\hat{t}_{y\pi})$ est généralement inconnue, nous la remplaçons par un estimateur noté $\hat{V}(\hat{t}_{y\pi})$.

Nous obtenons l'intervalle de confiance estimé :

$$\widehat{IC}_{1-\alpha}(t_y) = \left[\hat{t}_{y\pi} \pm z_{1-\frac{\alpha}{2}} \sqrt{\hat{V}(\hat{t}_{y\pi})}\right]. \quad (1.17)$$

L'intervalle de confiance est (approximativement) valide :

- si l'estimateur centré réduit $\dfrac{\hat{t}_{y\pi} - t_y}{\sqrt{V_p(\hat{t}_{y\pi})}}$ suit asymptotiquement une loi normale $\mathcal{N}(0,1)$,
- si l'estimateur de variance $\hat{V}(\hat{t}_{y\pi})$ est faiblement consistant pour $V_p(\hat{t}_{y\pi})$.

Coefficient de variation

La précision de l'estimation du total peut également être donnée sous la forme du *coefficient de variation*

$$CV_p(\hat{t}_{y\pi}) = \frac{\sqrt{V_p\left(\hat{t}_{y\pi}\right)}}{t_y} \text{ estimé par } \hat{CV}(\hat{t}_{y\pi}) = \frac{\sqrt{\hat{V}(\hat{t}_{y\pi})}}{\hat{t}_{y\pi}}. \qquad (1.18)$$

Il s'agit d'une grandeur sans dimension, plus facile à comparer et à interpréter que la variance. Avec un niveau de confiance de 0.95, l'intervalle de confiance du total est donné par

$$\begin{aligned} \widehat{IC}_{0.95}(t_y) &= \left[\hat{t}_{y\pi} \pm 1.96 \sqrt{\hat{V}(\hat{t}_{y\pi})}\right] \\ &= \hat{t}_{y\pi}\left[1 \pm 1.96\ \hat{CV}(\hat{t}_{y\pi})\right]. \end{aligned}$$

Interprétation : un CV de $x\%$ correspond à un total connu à plus ou moins 2 $x\%$, avec un niveau de confiance de 0.95.

En résumé

La connaissance des probabilités d'inclusion d'ordre 1 permet de calculer l'estimateur de Horvitz-Thompson du total

$$\hat{t}_{y\pi} = \sum_{k \in S} \frac{y_k}{\pi_k}.$$

Pour un plan de sondage quelconque, sa variance est estimée sans biais par

$$\hat{V}_{HT}(\hat{t}_{y\pi}) = \sum_{k,l \in S} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} \frac{\Delta_{kl}}{\pi_{kl}}$$

si tous les π_{kl} sont strictement positifs.

En utilisant une approximation normale pour $\hat{t}_{y\pi}$, nous obtenons l'intervalle de confiance :

$$\widehat{IC}_{1-\alpha}(t_y) = \left[\hat{t}_{y\pi} \pm z_{1-\frac{\alpha}{2}} \sqrt{\hat{V}_{HT}(\hat{t}_{y\pi})} \right].$$

Exercice

Nous reprenons l'exercice de la diapositive 29. Nous supposons que l'échantillon $\{1, 2\}$ est sélectionné, et que les valeurs observées sont $y_1 = 6$ et $y_2 = 3.6$.

1) L'estimateur de Horvitz-Thompson est-il sans biais pour le pds considéré?
2) Donner la valeur de cet estimateur pour l'échant. sélectionné. $[\hat{t}_{y\pi} = 16]$

Nous souhaitons maintenant calculer une mesure de précision associée à $\hat{t}_{y\pi}$.

3) L'estimateur de variance de Horvitz-Thompson est-il sans biais pour le pds considéré?
4) Donner la valeur de cet estimateur pour l'échantillon sélectionné.
5) Donner l'estimation du coefficient de variation. $[\hat{CV}(\hat{t}_{y\pi}) = 59\%]$

Méthodes d'échantillonnage

Sondage aléatoire simple

Sondage aléatoire simple sans remise

Le sondage aléatoire simple sans remise (SRS) de taille n est le plan de sondage qui donne la même probabilité à tous les échantillons de taille n d'être sélectionnés. Nous obtenons :

$$p(s) = \begin{cases} 1/C_N^n & \text{si } n(s) = n, \\ 0 & \text{sinon.} \end{cases} \tag{2.1}$$

Un algorithme de tirage pour le SRS procède de la façon suivante : nous tirons une unité parmi N à probabilités égales, puis une unité à probabilités égales parmi les $N-1$ unités restantes, et ainsi de suite jusqu'à obtenir les n unités (méthode *draw by draw*).

En pratique, il existe d'autres algorithmes de tirage plus efficaces ne nécessitant qu'une seule lecture de fichier (Tillé, 2011, Section 4.4).

Echantillonnage aléatoire simple en R

Le package `sampling` permet de faire un sondage aléatoire simple en utilisant un algorithme de sélection unité par unité (fonction `srswor`), ou l'algorithme de sélection-rejet, plus rapide (fonction `srswor1`).

```
> n=100
> Npop=589
> ech_srs=srswor(n,Npop)
> ech2_srs=srswor1(n,Npop)
```

Estimateur de Horvitz-Thompson

Proposition 2

Soient k et l deux unités distinctes quelconques. Alors :

$$\pi_k = \frac{n}{N}, \qquad \pi_{kl} = \frac{n(n-1)}{N(N-1)}. \tag{2.2}$$

L'estimateur de Horvitz-Thompson du total peut donc se réécrire sous la forme

$$\hat{t}_{y\pi} = \frac{N}{n} \sum_{k \in S} y_k = N\, \bar{y}, \tag{2.3}$$

avec $\bar{y}$ la moyenne simple sur l'échantillon S sélectionné.

Variance de l'estimateur de Horvitz-Thompson

La variance de l'estimateur de Horvitz-Thompson s'obtient à partir de la formule de Sen-Yates-Grundy :

$$V_p(\hat{t}_{y\pi}) = N^2 \frac{1-f}{n} S_y^2 \quad \text{avec} \quad S_y^2 = \frac{1}{N-1} \sum_{k \in U} (y_k - \mu_y)^2. \quad (2.4)$$

Elle est estimée sans biais par

$$\hat{V}_{HT}(\hat{t}_{y\pi}) = N^2 \frac{1-f}{n} s_y^2 \quad \text{avec} \quad s_y^2 = \frac{1}{n-1} \sum_{k \in S} (y_k - \bar{y})^2. \quad (2.5)$$

La quantité $f = n/N$ est appelée le *taux de sondage*.

Estimation pour un SRS

```
#Tirage d'un échantillon aléatoire simple
>n <- 100
>Npop <- 589
>ech=srswor1(n,Npop)
#Estimation
>pi <- rep(n/Npop,Npop)
>y=TaxableIncome
>est_ht=HTestimator(y[ech==1],pi_50[ech==1])
>est_ht
[1,] 2.517e+11
#Estimation de variance pour un SRS (PACKAGE SAMPLING)
>vest_srs=varest(y[ech==1],,pi[ech==1],)
>vest_srs
[1] 1.16e+20
#Estimation de variance pour un SRS (PACKAGE GUSTAVE)
>vest_srs_gus=var_srs(y[ech==1],pi[ech==1])
>vest_srs_gus
[1] 1.16e+20
```

Estimation de la moyenne μ_y

Par linéarité, la moyenne μ_y peut être estimée sans biais par

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

La variance de cet estimateur est donnée par

$$V_p(\bar{y}) = \frac{1-f}{n} S_y^2. \tag{2.6}$$

Remarques :

- Le facteur $(1-f)$ donne le gain de variance dû au tirage sans remise. Il est appelé *correction de population finie*. Ce gain peut être très important (cas des enquêtes-entreprise).
- Si le taux de sondage est faible, la variance ne dépend que de la taille d'échantillon n.

Cas d'une proportion

Dans le cas particulier où le paramètre d'intérêt est une proportion notée P, la variable d'intérêt y est une variable indicatrice dont on cherche à estimer la moyenne.

Exemple : proportion d'étudiants portant des lunettes dans la promotion,

$$y_k = \begin{cases} 1 & \text{si l'étudiant } k \text{ porte des lunettes,} \\ 0 & \text{sinon.} \end{cases}$$

En particulier, le paramètre peut s'écrire sous la forme

$$P = \frac{1}{N} \sum_{k \in U} y_k,$$

et être estimé par

$$\hat{P} = \frac{1}{n} \sum_{k \in S} y_k.$$

Proposition 3

Dans le cas d'une variable indicatrice (0/1) y, nous avons :

$$S_y^2 = \frac{N}{N-1} P(1-P), \tag{2.7}$$

$$s_y^2 = \frac{n}{n-1} \hat{P}(1-\hat{P}). \tag{2.8}$$

La variance de l'estimateur de la moyenne $\hat{P}$ peut alors se réécrire

$$V_p(\hat{P}) = \frac{1-f}{n} \frac{N}{N-1} P(1-P), \tag{2.9}$$

et être estimée sans biais par

$$\hat{V}_{HT}(\hat{P}) = \frac{1-f}{n-1} \hat{P}(1-\hat{P}). \tag{2.10}$$

Application : détermination de taille d'échantillon

Nous cherchons la taille d'échantillon minimale permettant de respecter avec un niveau de confiance fixé (par exemple de 95 %) une contrainte de précision en termes :

1. soit d'*erreur absolue* :
   P connu à plus ou moins 0.02 $\Leftrightarrow |\hat{P} - P| \leq 0.02$.
2. soit d'*erreur relative* :
   P connu à 8 % près $\Leftrightarrow |\frac{\hat{P}-P}{P}| \leq 0.08$.

Application : détermination de taille d'échantillon

Erreur absolue

Avec un niveau de confiance de 95 % la contrainte de précision peut se réécrire :

$$
\begin{aligned}
|\hat{P} - P| \leq \beta \quad &\Leftrightarrow \quad 1.96\sqrt{V_p(\hat{P})} \leq \beta \\
&\Leftrightarrow \quad 1.96\sqrt{\left[\frac{1}{n} - \frac{1}{N}\right]\frac{N}{N-1}P(1-P)} \leq \beta \\
&\Leftrightarrow \quad n \geq \frac{1}{\frac{1}{N} + \frac{N-1}{N}\left[\frac{\beta}{1.96}\right]^2 \frac{1}{P(1-P)}}. \qquad (2.11)
\end{aligned}
$$

Il est toujours possible de se placer dans le pire des cas en prenant $P = 0.5$, mais il est préférable de disposer d'un a priori (même vague) sur le paramètre P.

Application : détermination de taille d'échantillon

Erreur relative

Avec un niveau de confiance de 95 % la contrainte de précision peut se réécrire :

$$\left|\frac{\hat{P}-P}{P}\right| \leq \gamma \quad \Leftrightarrow \quad 1.96\ CV_p(\hat{P}) \leq \gamma$$

$$\Leftrightarrow \quad 1.96\sqrt{\left[\frac{1}{n}-\frac{1}{N}\right]\frac{N}{N-1}\frac{1-P}{P}} \leq \gamma$$

$$\Leftrightarrow \quad n \geq \frac{1}{\frac{1}{N}+\frac{N-1}{N}\left[\frac{\gamma}{1.96}\right]^2\frac{P}{1-P}}. \qquad (2.12)$$

Calculer cette borne nécessite de disposer d'un a priori sur le paramètre P, ou au moins d'un majorant pour ce paramètre.

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*

Un échantillon aléatoire simple

	Pass19	Pop19	Pass20	Trans20
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
FIGARI	700 000	20 000	500 000	2 900
PAU	600 000	240 000	200 000	0
TARBES	500 000	120 000	100 000	0
		$\bar{y}$	500 000	1 150
		$\hat{IC}_{0.95}$	[321 000, 679 000]	[540, 1 760]
		$\hat{CV}$	18%	27%
		μ_y	*450 000*	*783*

Un autre échantillon aléatoire simple

	Pass19	Pop19	Pass20	Trans20
AJACCIO	1 500 000	100 000	900 000	1 300
RENNES	900 000	730 000	300 000	200
FIGARI	700 000	20 000	500 000	2 900
TOULON	500 000	630 000	200 000	0
PERPIGNAN	500 000	320 000	200 000	0
TARBES	500 000	120 000	100 000	0
		$\bar{y}$	367 000	733
		$I\hat{C}_{0.95}$	[200 000, 533 000]	[70, 1 400]
		$\hat{CV}$	23%	46%
		μ_y	*450 000*	*783*

En résumé

Formule générale	Formule pour un plan SRS
$\hat{t}_{y\pi} = \sum_{k \in S} \frac{y_k}{\pi_k}$	$\hat{t}_{y\pi} = N\bar{y}$
$V_p(\hat{t}_{y\pi}) = \sum_{k,l \in U} \Delta_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}$	$V_{SRS}(\hat{t}_{y\pi}) = N^2 \frac{1-f}{n} S_y^2$
$\hat{V}_{HT}(\hat{t}_{y\pi}) = \sum_{k,l \in S} \frac{\Delta_{kl}}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}$	$\hat{V}_{HT}(\hat{t}_{y\pi}) = N^2 \frac{1-f}{n} s_y^2$

Estimation d'une proportion avec un plan SRS

$$\hat{P} = \frac{1}{n} \sum_{k \in S} y_k$$

$$V_{SRS}(\hat{P}) = \frac{1-f}{n} \frac{N}{N-1} P(1-P)$$

$$\hat{V}_{HT}(\hat{P}) = \frac{1-f}{n-1} \hat{P}(1-\hat{P})$$

Exercice

Parmi les 350 étudiants de l'Ensai, nous voulons estimer la proportion P d'étudiants qui portent des lunettes.
1) Quelle taille d'échantillon faut-il sélectionner pour que cette proportion soit estimée à 10% près, avec un niveau de confiance de 0.95 :

1. en utilisant l'information suivante : 50% des personnes de la population française portent des lunettes ; $[n = 184]$
2. en utilisant maintenant l'information suivante : 20% des $15 - 25$ ans portent des lunettes. $[n = 286]$

Nous sélectionnons finalement un échantillon de $n = 70$ étudiants, parmi lesquels 20 portent des lunettes.
2) Donner une estimation de P, et un intervalle de confiance à 95 %. $\left[\widehat{IC} = [0.19, 0.38]\right]$

Sondage aléatoire simple stratifié

Base de données d'aéroports

Nous considérons à nouveau la population de $N = 12$ aéroports, pour lesquels nous voulons estimer le nombre moyen de passagers en 2020 et le nombre moyen de passagers en transit en 2020. Nous sélectionnons pour cela un échantillon de $n = 6$ aéroports.

Nous coupons la population en deux groupes :

- 6 aéroports (sous-pop. U_1) ayant transporté plus de 1 000 000 de passagers en 2019,
- 6 aéroports (sous-pop. U_2) ayant transporté moins de 1 000 000 de passagers en 2019.

Base de données d'aéroports

Stratégies d'échantillonnage envisagées :

- SRS(6) dans la population entière,
- SRS(3) dans U_1 et SRS(3) dans U_2,
- SRS(4) dans U_1 et SRS(2) dans U_2.

		Pass20	Trans20
SRS(6)	V_p	$5.98\ 10^9$	$7.42\ 10^4$
	CV_p	17.2%	34.8%
SRS(3)+SRS(3)	V_p	$2.58\ 10^9$	$7.45\ 10^4$
	CV_p	11.3%	34.9%
SRS(4)+SRS(2)	V_p	$2.48\ 10^9$	$1.23\ 10^5$
	CV_p	11.1%	44.8%

Information auxiliaire

Une *information auxiliaire* désigne une information connue sur l'ensemble de la population :

- soit sous forme détaillée (e.g., sexe et âge, pour une pop. d'individus),
- soit sous forme synthétique (e.g., nombre total d'individus par sexe et par tranche d'âge).

Pour utiliser une information auxiliaire à l'étape du plan de sondage, elle doit être connue de façon détaillée. Elle peut par exemple permettre de partitionner la population en groupes, pour obtenir un plan de sondage plus efficace que le SRS.

Le gain de précision obtenu dépend du lien entre la variable auxiliaire et la variable d'intérêt.

Motivations pour la stratification (Cochran, 1977)

- Précision maîtrisée pour des sous-populations,
  Ex : enquêtes-entreprise stratifiées par tranche de taille $\times$ type d'activité
- Simplicité administrative (enquêtes conduites par différentes agences),
  Ex : enquête EU-SILC, conduite indépendamment dans chaque pays européen. Tirage direct dans un registre (Danemark, Suède, Luxembourg, Pays-Bas, Malte), ou tirage à 2 degrés (Espagne, France, Italie, Irlande, Pologne, République Tchèque, Slovénie, Lettonie).
- Plans de sondage adaptés aux sous-populations,
  Ex : Enquêtes de Recensement réalisée séparément pour les individus vivant en logement ordinaire (répertoire Fidéli), les individus vivant en communauté, et les personnes sans-domicile.
- Gain global de précision.

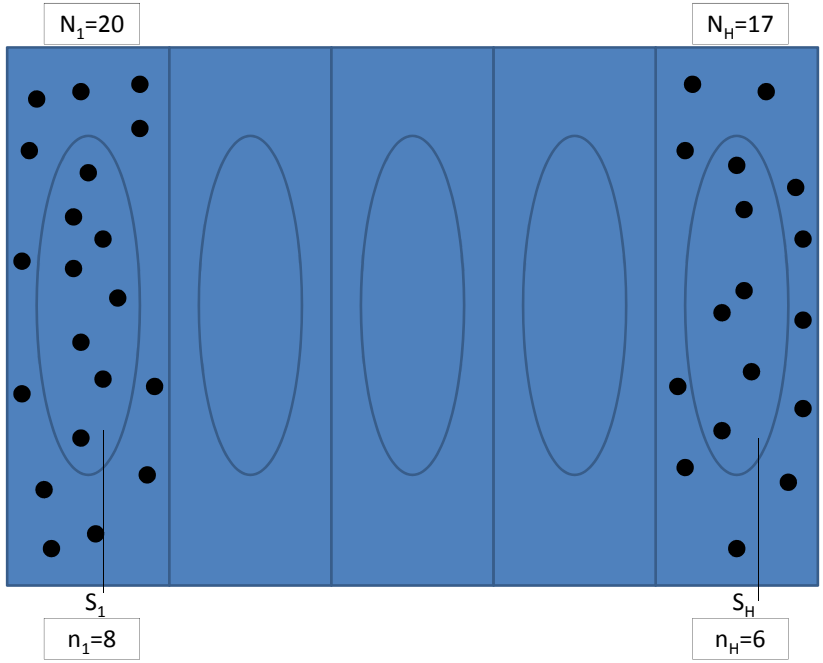
Principales questions :

1. Comment construire les strates?
2. Quelle taille d'échantillon sélectionner dans chaque strate?

Définition

La population U est dite *stratifiée* quand les unités sont partitionnées en H sous-populations $U_1, \ldots, U_H$ appelées *strates*.

Le plan de sondage est stratifié quand des **échantillons indépendants** sont sélectionnés dans chaque strate. Nous considérerons le cas particulier du *sondage aléatoire simple stratifié (STSRS)* où un SRS est appliqué dans chaque strate.



Exemple : enquêtes entreprises

Les échantillons pour les enquêtes auprès des entreprises sont souvent tirés selon des plans de sondages aléatoires simples stratifiés. La stratification est obtenue en croisant :

- un critère d'activité (nomenclature d'activités française NAF),
- un critère de taille (tranches d'effectifs salariés et/ou tranches de chiffres d'affaires).

Par exemple (voir Demoly et al., 2014), l'enquête sur les technologies de l'information et de la communication (TIC) a été tirée en stratifiant selon :

- le secteur d'activité,
- la tranche d'effectif de l'entreprise (10-19, 20-49, 50-249, 250-499, 500 et +),
- le chiffre d'affaires,

avec un seuil d'exhaustivité pour les plus grandes tranches d'effectif et les plus gros chiffres d'affaires.

Notations et estimation

Nous notons N_h la taille de la strate U_h. Le total t_y et la moyenne μ_y se décomposent sous la forme

$$t_y = \sum_{h=1}^{H} t_{yh} \quad \text{avec} \quad t_{yh} = \sum_{k \in U_h} y_k \text{ le sous-total sur } U_h, \tag{2.13}$$

$$\mu_y = \sum_{h=1}^{H} \frac{N_h}{N} \mu_{yh} \quad \text{avec} \quad \mu_{yh} = \frac{t_{yh}}{N_h} \text{ la moyenne sur } U_h. \tag{2.14}$$

Sous un STSRS, les estimateurs de Horvitz-Thompson s'obtiennent en utilisant les estimateurs sans biais dans les strates :

$$\hat{t}_{y\pi} = \sum_{h=1}^{H} N_h \bar{y}_h \quad \text{et} \quad \hat{\mu}_{y\pi} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h, \tag{2.15}$$

avec $\bar{y}_h = \frac{1}{n_h} \sum_{k \in S_h} y_k$ la moyenne dans le sous-échant. S_h tiré dans U_h.

Probabilités d'inclusion

Proposition 4

Dans le cas d'un STSRS :

- *Pour tout* $h = 1, \ldots, H$ *et pour tout* $k \in U_h$ *:*

$$\pi_k = \frac{n_h}{N_h}. \tag{2.16}$$

- *Pour tout* $h = 1, \ldots, H$ *et pour tous* $k \neq l \in U_h$ *:*

$$\pi_{kl} = \frac{n_h(n_h - 1)}{N_h(N_h - 1)}. \tag{2.17}$$

- *Pour tous* $h \neq h' = 1, \ldots, H$*, pour tout* $k \in U_h$ *et* $l \in U_{h'}$ *:*

$$\pi_{kl} = \frac{n_h n_{h'}}{N_h N_{h'}}. \tag{2.18}$$

Estimation d'un total

Par indépendance des tirages, la variance de $\hat{t}_{y\pi}$ s'obtient par sommation:

$$V_p(\hat{t}_{y\pi}) = \sum_{h=1}^{H} N_h^2 \frac{1-f_h}{n_h} S_{yh}^2 \text{ avec } S_{yh}^2 = \frac{1}{N_h - 1} \sum_{k \in U_h} (y_k - \mu_{yh})^2. \quad (2.19)$$

Elle est estimée par

$$\hat{V}_{HT}(\hat{t}_{y\pi}) = \sum_{h=1}^{H} N_h^2 \frac{1-f_h}{n_h} s_{yh}^2 \text{ avec } s_{yh}^2 = \frac{1}{n_h - 1} \sum_{k \in S_h} (y_k - \bar{y}_h)^2, \quad (2.20)$$

en notant $f_h = n_h/N_h$ le taux de sondage dans la strate U_h.

Estimation d'une moyenne

La variance de $\hat{\mu}_{y\pi}$ s'obtient de façon analogue :

$$V_p(\hat{\mu}_{y\pi}) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \frac{1-f_h}{n_h} S_{yh}^2. \quad (2.21)$$

Elle est estimée sans biais par

$$\hat{V}_{HT}(\hat{\mu}_{y\pi}) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \frac{1-f_h}{n_h} s_{yh}^2. \quad (2.22)$$

Allocation d'échantillon

Avant de réaliser le tirage d'échantillon et l'estimation, nous devons déterminer l'allocation d'échantillon, i.e. la façon dont la taille globale d'échantillon n est répartie entre les strates.

L'*allocation proportionnelle* consiste à allouer l'échantillon dans les strates, proportionnellement à leur importance.

L'*allocation optimale* (ou *allocation de Neyman*) consiste à allouer l'échantillon de façon à minimiser la variance de $\hat{t}_{y\pi}$, pour une variable d'intérêt particulière y_k.

Allocation proportionnelle

L'*allocation proportionnelle* consiste à allouer l'échantillon dans les strates, proportionnellement à leur importance :

$$n_h = n\frac{N_h}{N}. \tag{2.23}$$

Autrement dit, plus la strate est grande, plus l'échantillon sélectionné dedans est grand.

Cette allocation conduit à une fraction de sondage constante par strate

$$f_h = \frac{n_h}{N_h} = \frac{n}{N} = f,$$

ce qui n'est pas le cas avec les autres allocations.

Allocation proportionnelle

Dans le cas d'une allocation proportionnelle, la variance de $\hat{t}_{y\pi}$ se réécrit

$$V_p\left[\hat{t}_{y\pi}\right] = N^2 \frac{1-f}{n} \sum_{h=1}^{H} \frac{N_h}{N} S_{yh}^2 \simeq N^2 \frac{1-f}{n} S_{y,intra}^2. \quad (2.24)$$

Rappelons que selon l'équation de décomposition de la variance :

$$S_y^2 = \underbrace{\sum_{h=1}^{H} \frac{N_h - 1}{N-1} S_{yh}^2}_{S_{y,intra}^2} + \underbrace{\sum_{h=1}^{H} \frac{N_h}{N-1} (\mu_{yh} - \mu_y)^2}_{S_{y,inter}^2}. \quad (2.25)$$

La stratification avec allocation proportionnelle permet de gommer la variabilité entre les strates $S_{y,inter}^2$.
La stratification devrait être choisie de façon à ce que la **dispersion à l'intérieur des strates** soit minimisée.

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*

Base de données d'aéroports

Stratégies d'échantillonnage envisagées :

- SRS(6) dans la population entière,
- SRS(3) dans U_1 et SRS(3) dans U_2 (allocation proportionnelle).

		Pass20	Trans20
SRS(6)	V_p	$5.98\ 10^9$	$7.42\ 10^4$
	CV_p	17.2%	34.8%
SRS(3)+SRS(3)	V_p	$2.58\ 10^9$	$7.45\ 10^4$
	CV_p	11.3%	34.9%
	$S^2_{y,intra}$	$2.82\ 10^{10}$	$8.13\ 10^5$
	$S^2_{y,inter}$	$4.36\ 10^{10}$	$7.76\ 10^4$
	$\frac{S^2_{y,inter}}{S^2_y}$	61%	9%

Echantillonnage stratifié dans R

```
> aeroports$U <- 1+(aeroports$Pass19<=1000000)
> ech=strata(data=aeroports,stratanames="U",
                size=c(3,3),method = "srswor")
> ech
   U ID_unit Prob Stratum
2  1       2  0.5       1
3  1       3  0.5       1
5  1       5  0.5       1
7  2       7  0.5       2
8  2       8  0.5       2
11 2      11  0.5       2
```

Allocation optimale

L'*allocation optimale* (ou *allocation de Neyman*) consiste à allouer l'échantillon de façon à minimiser la variance de $\hat{t}_{y\pi}$, pour une variable d'intérêt particulière y_k :

$$\min_{n_h} \sum_{h=1}^{H} \left[\frac{1}{n_h} - \frac{1}{N_h}\right] N_h^2 S_{yh}^2 \quad \text{t.q.} \quad \sum_{h=1}^{H} n_h = n. \tag{2.26}$$

En utilisant une technique de Lagrangien, nous obtenons :

$$n_h = n \frac{N_h \sqrt{S_{yh}^2}}{\sum_{j=1}^{H} N_j \sqrt{S_{yj}^2}}. \tag{2.27}$$

Le calcul de cette allocation optimale nécessite la connaissance des dispersions S_{yh}^2 dans les strates.

Allocation optimale

L'allocation de Neyman indique qu'il faut sélectionner un échantillon plus grand dans les grandes strates et/ou dans les strates présentant une forte dispersion.

L'allocation n'est optimale que pour la variable d'intérêt considérée.

La formule de l'allocation de Neyman peut conduire à $n_h > N_h$. Dans ce cas, la strate est recensée ($n_h = N_h$: strate dite exhaustive), et l'allocation est recalculée dans les autres strates :

$$n_j = (n - N_h) \frac{N_j \sqrt{S_{yj}^2}}{\sum_{i \neq h=1}^{H} N_i \sqrt{S_{yi}^2}} \text{ pour } j \neq h = 1, \ldots, H. \quad (2.28)$$

Base de données d'aéroports

Stratégies d'échantillonnage envisagées :

- SRS(6) dans la population entière,
- SRS(3) dans U_1 et SRS(3) dans U_2 (allocation proportionnelle),
- SRS(4) dans U_1 et SRS(2) dans U_2 (allocation Pass19-optimale).

	Var. d'optimisation			
	Pass19	Pop19	*Pass20*	*Trans20*
	Avant arrondi			
U_1	3.89	3.01	*3.60*	*2.14*
U_2	2.11	2.99	*2.40*	*3.86*
	Après arrondi			
U_1	4	3	*4*	*2*
U_2	2	3	*2*	*4*

Base de données d'aéroports

Stratégies d'échantillonnage envisagées :

- SRS(6) dans la population entière,
- SRS(3) dans U_1 et SRS(3) dans U_2 (allocation proportionnelle),
- SRS(4) dans U_1 et SRS(2) dans U_2 (allocation Pass19-optimale).

		Pass20	Trans20
SRS(6)	V_p	$5.98\ 10^9$	$7.42\ 10^4$
	CV_p	17.2%	34.8%
SRS(3)+SRS(3)	V_p	$2.58\ 10^9$	$7.45\ 10^4$
	CV_p	11.3%	34.9%
SRS(4)+SRS(2)	V_p	$2.48\ 10^9$	$1.23\ 10^5$
	CV_p	11.1%	44.8%

En résumé

Formule générale	Formule pour un plan STSRS
$\hat{t}_{y\pi} = \sum_{k \in S} \frac{y_k}{\pi_k}$	$\hat{t}_{y\pi} = \sum_{h=1}^{H} N_h \bar{y}_h$
$V_p(\hat{t}_{y\pi}) = \sum_{k,l \in U} \Delta_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}$	$V_p(\hat{t}_{y\pi}) = \sum_{h=1}^{H} N_h^2 \frac{1 - f_h}{n_h} S_{yh}^2$
$\hat{V}_{HT}(\hat{t}_{y\pi}) = \sum_{k,l \in S} \frac{\Delta_{kl}}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}$	$\hat{V}_{HT}(\hat{t}_{y\pi}) = \sum_{h=1}^{H} N_h^2 \frac{1 - f_h}{n_h} s_{yh}^2.$

Allocations d'échantillon	
Proportionnelle	$n_h = n \frac{N_h}{N}$
Optimale	$n_h = n \frac{N_h \sqrt{S_{yh}^2}}{\sum_{j=1}^{H} N_j \sqrt{S_{yj}^2}}$

Exercice

Dans la base d'aéroports, nous considérons la stratification suivante :

- $V_0 = \{\text{MONTPELLIER}, \text{FIGARI}\}$,
- $V_1 = U_1 \setminus \{\text{MONTPELLIER}\}$ et $V_2 = U_2 \setminus \{\text{FIGARI}\}$.

1) Donner l'allocation d'échantillon pour un tirage de taille $n = 6$, avec V_0 strate exhaustive, et allocation proportionnelle dans V_1 et V_2.
[$n_0 = 2, n_1 = 2, n_2 = 2$]

2) Donner les probabilités d'inclusion de chaque aéroport.

3) En supposant que `BASTIA` et `BREST` sont tirées dans V_1, et que `RENNES` et `PAU` sont tirées dans V_2, donner une estimation du total de `Pass20` et `Trans20`.
[5 800 000 et 11 700]

Tirage à probabilités inégales

Introduction

Nous avons vu précedemment que la stratification permettait de réduire la variance des estimateurs. Si les strates sont homogènes relativement à la variable d'intérêt (dispersion intra faible), le sondage aléatoire simple stratifié est une stratégie efficace d'échantillonnage.

En pratique, il peut subsister une forte hétérogénéité dans les strates. Dans ce cas, nous pouvons rechercher une stratégie d'échantillonnage plus efficace en individualisant les probabilités de sélection π_k de chacun des individus.

Nous devons ensuite choisir un *algorithme de tirage*, i.e. une méthode pratique de sélection respectant les probabilités d'inclusion choisies.
Nous étudierons deux de ces algorithmes : le *tirage de Poisson* et le *tirage systématique*.

Fonction de base sample

```
> sample(x,[n],size,replace = FALSE, prob = NULL)
```

- `x` : vecteur dans lequel sélectionner, ou entier positif.
- `size` : taille d'échantillon (entier positif).
- `replace`: échantillonnage sans remise (FALSE) ou avec remise (TRUE). **L'option FALSE donne une méthode biaisée d'échantillonnage à probabilités inégales**.
- `prob` : vecteur de probabilités, les probabilités d'inclusion sont proportionnelles à `prob` (NULL pour un tirage à probabilités égales).

Fonction de base sample

```
> sample(1:10,6,replace = FALSE)
```

Sélection d'un échantillon de 6 unités parmi les 10 premiers entiers selon un SRS : Ok.

```
> prob <- c(1,1,1,1,1,2,2,2,2,2)
> sample(1:10,6,replace = TRUE,prob)
```

Sélection d'un échantillon de 6 unités parmi les 10 premiers entiers. Tirage avec remise à probabilités inégales : Ok.

```
> prob <- c(1,1,1,1,1,2,2,2,2,2)
> sample(1:10,6,replace = FALSE,prob)
```

Sélection d'un échantillon de 6 unités parmi les 10 premiers entiers. Tirage sans remise à probabilités inégales : méthode de tirage fausse.

Probabilités proportionnelles à la taille

La taille moyenne d'échantillon sélectionné est donnée par

$$E_p[n(S)] = \sum_{k \in U} \pi_k.$$

En notant n la taille d'échantillon souhaitée, les *probabilités d'inclusion proportionnelles à une mesure de taille* (pps) $x_k \geq 0$ sont données par :

$$\pi_k = n \frac{x_k}{\sum_{l \in U} x_l}. \tag{2.29}$$

Si certaines unités sont particulièrement grosses (au sens de la variable x_k), cette formule peut donner des probabilités d'inclusion > 1. Dans ce cas, les unités correspondantes sont sélectionnées d'office, et les probabilités d'inclusion des autres unités sont recalculées.

Recalcul des probabilités d'inclusion

```
#Calcul de probas d'inclusion proportionnelles à la taille
> n=50
> pi_50=inclusionprobabilities(averageincome,n)
> summary(pi_50)
[1]  Min.  1st Qu. Median   Mean 3rd Qu.   Max.
[1]0.05693 0.07675 0.08375 0.08489 0.09113 0.14076

> n=400
> pi_400=inclusionprobabilities(averageincome,n)
> summary(pi_400)
[1]  Min.  1st Qu. Median   Mean 3rd Qu.   Max.
[1]0.4556 0.6142  0.6702 0.6791 0.7293 1.0000
```

Base de données d'aéroports

Probabilités proportionnelles à Pass19

	Pass19	Pop19	π_k
MONTPELLIER	1 900 000	620 000	0.93
BASTIA	1 600 000	100 000	0.78
AJACCIO	1 500 000	100 000	0.73
STRASBOURG	1 300 000	800 000	0.63
BREST	1 200 000	320 000	0.59
BIARRITZ	1 100 000	300 000	0.54
RENNES	900 000	730 000	0.44
FIGARI	700 000	20 000	0.34
PAU	600 000	240 000	0.29
TOULON	500 000	630 000	0.24
PERPIGNAN	500 000	320 000	0.24
TARBES	500 000	120 000	0.24

Base de données d'aéroports

Probabilités proportionnelles à Pop19

	Pass19	Pop19	π_k	
			Essai 1	Essai 2
MONTPELLIER	1 900 000	620 000	0.87	0.90
BASTIA	1 600 000	100 000	0.14	0.14
AJACCIO	1 500 000	100 000	0.14	0.14
STRASBOURG	1 300 000	800 000	1.12	1.00
BREST	1 200 000	320 000	0.45	0.46
BIARRITZ	1 100 000	300 000	0.42	0.43
RENNES	900 000	730 000	1.02	1.00
FIGARI	700 000	20 000	0.03	0.03
PAU	600 000	240 000	0.33	0.35
TOULON	500 000	630 000	0.88	0.91
PERPIGNAN	500 000	320 000	0.45	0.46
TARBES	500 000	120 000	0.17	0.17

Le tirage de Poisson

Tirage de Poisson

C'est un principe de piles ou faces indépendants, avec une pièce et un lancer différents pour chaque unité.

- Etape 1 : génération de $u_1 \sim U[0,1]$. Si $u_1 \leq \pi_1$, l'unité 1 est retenue dans l'échantillon.
- Etape 2 : génération de $u_2 \sim U[0,1]$ indépendamment de u_1. Si $u_2 \leq \pi_2$, l'unité 2 est retenue dans l'échantillon.
- ...
- Etape N : génération de $u_N \sim U[0,1]$ indépendamment de $u_1, \ldots, u_{N-1}$. Si $u_N \leq \pi_N$, l'unité N est retenue dans l'échantillon.

En utilisant les propriétés d'une loi $U[0,1]$ et l'indépendance des tirages :

$$\begin{aligned} \mathbb{P}(k \in S) &= \mathbb{P}(u_k \leq \pi_k) = F_U(\pi_k) = \pi_k, \\ \pi_{kl} &= \pi_k \, \pi_l \text{ si } k \neq l. \end{aligned}$$

Estimateur de Horvitz-Thompson

La variance s'obtient à partir de l'expression générale de HT :

$$V_{pois}(\hat{t}_{y\pi}) = \sum_{k \in U} \left(\frac{y_k}{\pi_k} \right)^2 \pi_k (1 - \pi_k). \quad (2.30)$$

Elle est estimée sans biais par

$$\hat{V}_{HT}(\hat{t}_{y\pi}) = \sum_{k \in S} \left(\frac{y_k}{\pi_k} \right)^2 (1 - \pi_k). \quad (2.31)$$

En particulier, cela implique que la taille d'échantillon est aléatoire :

$$V_{pois}[n(S)] = \sum_{k \in U} \pi_k (1 - \pi_k).$$

Base de données d'aéroports

Tirage de Poisson à probabilités proportionnelles à Pass19

	Pass19	Pop19	π_k	u_k	I_k
MONTPELLIER	1 900 000	620 000	0.93	0.46	
BASTIA	1 600 000	100 000	0.78	0.75	
AJACCIO	1 500 000	100 000	0.73	0.58	
STRASBOURG	1 300 000	800 000	0.64	0.71	
BREST	1 200 000	320 000	0.59	0.79	
BIARRITZ	1 100 000	300 000	0.54	0.14	
RENNES	900 000	730 000	0.44	0.93	
FIGARI	700 000	20 000	0.34	0.36	
PAU	600 000	240 000	0.29	0.40	
TOULON	500 000	630 000	0.24	0.51	
PERPIGNAN	500 000	320 000	0.24	0.41	
TARBES	500 000	120 000	0.24	0.40	

Utilisation

Le tirage de Poisson présente une grande variance d'échantillonnage. Il est utilisé pour certaines enquêtes auprès des entreprises, car il permet de simplifier la *coordination* du tirage de plusieurs échantillons.

On parle de coordination :

- négative quand on tire plusieurs échantillons afin qu'ils soient aussi disjoints que possible,
- positive quand on tire plusieurs échantillons afin qu'ils se recouvrent autant que possible.

Le tirage de Poisson est également utilisé dans un contexte de *non-réponse*, pour modéliser le mécanisme de réponse dans l'échantillon S complet (voir la Section 4).

Mise en oeuvre sous R

```
#Probabilités d'inclusion proportionnelles à la taille
> n=50
> pi_50=inclusionprobabilities(averageincome,n)

#Tirage de Poisson et estimation du total de TaxableIncome
> ech_poi=UPpoisson(pi_50)
> y=TaxableIncome
> HTestimator(y[ech_poi==1],pi_50[ech_poi==1])
[1,] 1.220165e+11

#Estimation de variance de HT
> pikl_poi_50=pi_50 %*% t(pi_50) +diag(pi_50-pi_50*pi_50)
> varHT(y[ech_poi==1],pikl_poi_50[ech_poi==1,ech_poi==1],1)
[1] 6.1382e+20

#Estimation de variance (package GUSTAVE)
> y_mat <- matrix(y, ncol = 1)
> var_pois(y_mat[ech_poi==1, , drop =
    FALSE],pi_50[ech_poi==1])
[1] 6.1382e+20
```

En résumé

Formule générale	Formule pour un plan de Poisson
$\hat{t}_{y\pi} = \sum_{k \in S} \frac{y_k}{\pi_k}$	-
$V_p(\hat{t}_{y\pi}) = \sum_{k,l \in U} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} \Delta_{kl}$	$V_p(\hat{t}_{y\pi}) = \sum_{k \in U} \left(\frac{y_k}{\pi_k} \right)^2 \pi_k (1 - \pi_k)$
$\hat{V}_{HT}(\hat{t}_{y\pi}) = \sum_{k,l \in S} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} \frac{\Delta_{kl}}{\pi_{kl}}$	$\hat{V}_{HT}(\hat{t}_{y\pi}) = \sum_{k \in S} \left(\frac{y_k}{\pi_k} \right)^2 (1 - \pi_k).$

Le tirage systématique

Principe

C'est une méthode simple et très rapide permettant de sélectionner un échantillon à probabilités inégales et de taille fixe.
C'est la méthode la plus utilisée en pratique, même pour un tirage à probabilités égales.

Principe :

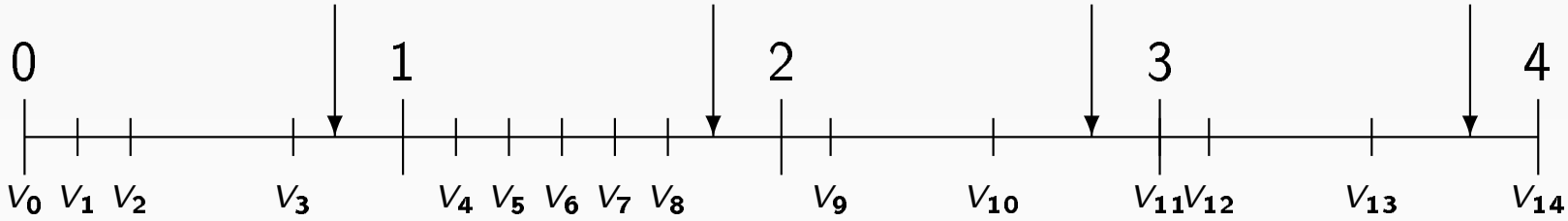
- Les unités de la population sont représentées sur un segment de longueur n. Chaque unité k est représentée par un segment de longueur π_k.
- Nous générons un nombre aléatoire $u \sim U[0,1]$, puis les nombres $u_i = u + (i-1),\ i = 1, \ldots, n-1$.
- Une unité est sélectionnée si un de ces nombres aléatoire tombe dans son segment.

```
#Probabilités d'inclusion proportionnelles à la taille
> n=50
> pi_50=inclusionprobabilities(averageincome,n)
#Tirage systématique
> ech_sys=UPsystematic(pi_50)
```

Exemple

Population U de taille $N = 14$ avec $n = 4$:

- $\pi_1 = \pi_2 = \pi_5 = \pi_6 = \pi_7 = \pi_8 = \pi_{12} = 1/7$,
- $\pi_3 = \pi_4 = \pi_9 = \pi_{10} = \pi_{11} = \pi_{13} = \pi_{14} = 3/7$.



$$
\begin{aligned}
u = 0.82 \in [V_3, V_4] &\Rightarrow \text{l'unité 4 est sélectionnée,} \\
1 + u = 1.82 \in [V_8, V_9] &\Rightarrow \text{l'unité 9 est sélectionnée,} \\
2 + u = 2.82 \in [V_{10}, V_{11}] &\Rightarrow \text{l'unité 11 est sélectionnée,} \\
3 + u = 3.82 \in [V_{13}, V_{14}] &\Rightarrow \text{l'unité 14 est sélectionnée.}
\end{aligned}
$$

Probabilités d'inclusion

Les probabilités d'inclusion π_k sont exactement respectées. Les probabilités d'inclusion d'ordre deux sont calculables (Tillé, 2011, p. 126), mais beaucoup d'entre elles sont nulles. Par conséquent, il n'existe pas d'estimateur sans biais de variance pour l'estimateur HT.

```
#Probabilités d'inclusion d'ordre 2
> pikl_sys=UPsystematicpi2(pi_50)
> pikl_sys[1:6,1:6]
        [,1]    [,2]    [,3]    [,4]    [,5]   [,6]
[1,] 0.1144 0.00000 0.00000 0.00000 0.00000 0.0000
[2,] 0.0000 0.07468 0.00000 0.00000 0.00000 0.0000
[3,] 0.0000 0.00000 0.09965 0.00000 0.00000 0.0000
[4,] 0.0000 0.00000 0.00000 0.07412 0.00000 0.0000
[5,] 0.0000 0.00000 0.00000 0.00000 0.09011 0.0000
[6,] 0.0000 0.00000 0.00000 0.00000 0.00000 0.1034
```

Base de données d'aéroports

Tirage systématique à probabilités proportionnelles à Pass19

					I_k	
	Pass19	π_k	V_{k-1}	V_k	$u = 0.11$	$u = 0.88$
MONTPELLIER	1 900 000	0.93	0	0.93		
BASTIA	1 600 000	0.78	0.93	1.71		
AJACCIO	1 500 000	0.73	1.71	2.44		
STRASBOURG	1 300 000	0.64	2.44	3.08		
BREST	1 200 000	0.59	3.08	3.67		
BIARRITZ	1 100 000	0.54	3.67	4.21		
RENNES	900 000	0.44	4.21	4.65		
FIGARI	700 000	0.34	4.65	4.99		
PAU	600 000	0.29	4.99	5.28		
TOULON	500 000	0.24	5.28	5.52		
PERPIGNAN	500 000	0.24	5.52	5.76		
TARBES	500 000	0.24	5.76	6.00		

Utilisation

Le tirage systématique permet de bénéficier d'un effet de stratification. Si la population est préalablement triée selon une variable auxiliaire x_k, les unités de l'échantillon vont être tirées dans toute la distribution de x_k.

Si la variable d'intérêt y_k est liée avec la variable de tri, cet effet de stratification peut conduire à une réduction de la variance par rapport à d'autres algorithmes de tirage utilisant les mêmes probabilités d'inclusion.

Le tirage systématique est également souvent utilisé avec des probabilités égales de tirage, en remplacement du SRS.

Cas de probabilités d'inclusion égales

Dans le cas de probabilités d'inclusion égales, la méthode est généralement plus efficace que le SRS si la population est triée avant le tirage selon une variable auxiliaire x_k corrélée avec la variable d'intérêt.

```
#Corrélation entre Tot04 et TaxableIncome
> y=TaxableIncome
> cor(Tot04,y)
[1] 0.988

#Tri de la population selon la variable Tot04
> permutation <- order(Tot04)
> Tot04_rank <- Tot04[permutation]
> y_rank <- y[permutation]

#Paramètres de l'échantillonnage (probabilités égales)
> n <- 50
> Npop <- 589
> pi0_50 <- rep(n/Npop,Npop)
```

Cas de probabilités d'inclusion égales

Comparaison entre SRS et tirage systématique

```
#Probabilités d'inclusion d'ordre 2 pour un SRS
> pikl_srs <- UPsampfordpi2(pi0_50)
#Variance exacte sous un SRS
> var_srs <-    t(y_rank/pi0_50)
            %*%(pikl_srs-pi0_50%*%t(pi0_50))
            %*%(y_rank/pi0_50)
#Probabilités d'inclusion d'ordre 2 pour le SYS
> pikl_sys <- UPsystematicpi2(pi0_50)
#Variance exacte sous un SYS
> var_sys <-    t(y_rank/pi0_50)
             %*%(pikl_sys-pi0_50%*%t(pi0_50))
             %*%(y_rank/pi0_50)

> options("scipen"=-100,digits="3")
> var_srs
[1,] 6.56e+20
> var_sys
[1,] 3.08e+20
```

Exercice

1) Pour l'échantillon sélectionné selon un plan de Poisson (cf diapositive 96), donner :

- une estimation du total de `Pass20` et `Trans20`, [3.86 10^6 et 4 269]
- une estimation du CV associé. [25% et 30%]

2) Pour chacun des deux échantillons sélectionnés selon un tirage systématique (cf diapositive 104), donner une estimation du total de `Pass20` et de `Trans20`.
[5.40 10^6 et 7 319]
[5.50 10^6 et 13 034]

Echantillonnage à plusieurs degrés

Motivation

Les différentes méthodes vues précedemment supposent que l'on peut constituer une base de sondage, i.e. une liste des unités de la population U. Souvent, une telle base de sondage n'est pas disponible et il n'est donc pas possible de tirer directement les individus.

Nous utilisons alors des méthodes de tirage indirect comme l'échantillonnage à plusieurs degrés, où l'échantillon est sélectionné en plusieurs temps.

Ce type d'échantillonnage a également l'avantage d'être moins cher pour des enquêtes en face à face, car il permet de concentrer géographiquement les unités enquêtées.

Ex. 1 : enquêtes auprès des ménages de l'Insee

Base de sondage constituée à partir de sources fiscales. Le répertoire statistique des individus et des logements (RESIL) devrait remplacer prochainement l'ancienne base FIDELI. Tirage à 2 ou 3 degrés :

1. Tirage de zones (obtenues par agrégation ou découpage de communes) pour obtenir l'**échantillon-maître** de l'Insee $\Rightarrow$ Unités Primaires (UP)
2. Dans ces UP, tirage d'un échantillon de ménages $\Rightarrow$ Unités Secondaires (US)
3. Dans ces ménages, l'enquête est réalisée auprès de tous les individus du ménage ou d'un représentant tiré aléatoirement (**individu Kish**).

Le tirage à plusieurs degrés pourrait être remis en cause par la disponibilité d'un répertoire, et par le fait que beaucoup d'enquêtes de l'Insee sont maintenant réalisées en bimode séquentiel (internet, puis téléphone).

Faivre, S. (2017). *Echantillonnage des enquêtes auprès des ménages* dans la source fiscale.

Ardilly, P. (2024). *Etude efficacité-coût de l'échantillon-maître Insee*. 13ème colloque francophone sur les sondages, Luxembourg, 5-8 novembre 2024.

Ex. 2 : enquêtes épidémiologiques (Santé Publique France)

L'Étude de Santé sur l'Environnement, la Biosurveillance, l'Activité physique et la Nutrition (ESTEBAN) réalisée entre 2014 et 2016 visait à mesurer le niveau d'imprégnation de la population métropolitaine à différents polluants, et à établir des mesures de référence.

Le plan de sondage est similaire à celui des enquêtes ménage de l'Insee : L'enquête était réalisée selon un plan de sondage à 3 degrés :

1. Tirage d'un échantillon de zones au premier degré.
2. Tirage de deux échantillons de ménages au second degré (un parmi les ménages contenant un enfant âge de 6 à 17 ans, un contenant un adulte âgé de 18 à 74 ans).
3. Tirage d'un individu Kish par ménage.

La principale différence est que chaque échantillon de ménages est sélectionné par génération aléatoire de numéros de téléphone.

Ex. 3 : enquêtes PISA

Le Programme International pour le Suivi des Acquis des élèves (PISA) mesure la capacité des jeunes de 15 ans à utiliser leurs compétences en compréhension de l'écrit, en mathématiques et en sciences.

L'enquête est réalisée selon un plan à 2 ou 3 degrés :

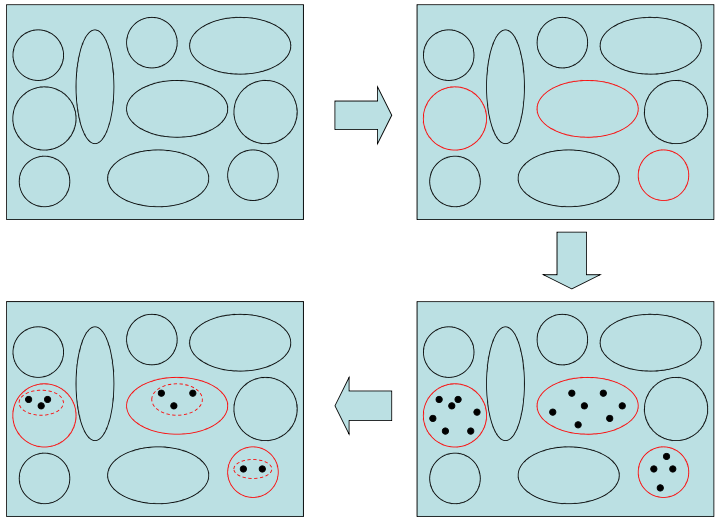
- Etablissement d'une base de sondage des écoles contenant des étudiants de 15 ans (équivalent classes de 5ème ou supérieures). Tirage stratifié[3] d'écoles, à probabilités proportionnelles au nombre d'élèves, par tirage systématique dans les strates (stratification implicite).
- Tirage d'un échantillon d'étudiants éligibles au second degré, à probabilités égales.

OCDE (2022). *PISA 2022: Technical Report*.

[3] En France, selon la région, le type d'école et leur taille

Principe de l'échantillonnage à deux degrés

La population U d'individus est partitionnée en N_I grosses unités appelées **Unités Primaires** (UP). Les unités de U sont appelées les **Unités Secondaires** (US).
Un échantillon d'UP est tiré au premier degré, puis un échantillon d'US dans chaque UP.



Motivation

L'échantillonnage multi-degrés est avant tout utilisé pour des considérations pratiques :

- **Réduction des coûts d'enquête**. Si les unités de la population sont très dispersées géographiquement, un tirage direct (e.g., selon un SRS) conduirait à un échantillon également fortement dispersé. L'utilisation de plusieurs degrés de tirage permet de concentrer les unités échantillonnées.
- **Constitution de la base de sondage**. On ne doit disposer d'une liste des unités de la population (US) que pour les UP sélectionnées.

Le problème des coûts de déplacement se pose pour une enquête en face à face, mais pas pour une enquête par téléphone ou par internet. La méthode d'échantillonnage dépend donc du **mode de collecte** utilisé.

Principe

Nous sélectionnons l'échantillon S en deux temps. Au premier degré, un échantillon S_I d'UP est tiré. Nous notons :

$$\pi_{Ii} = \mathbb{P}(u_i \in S_I)$$

la probabilité de sélectionner l'UP u_i dans S_I.

Au second degré, nous tirons un échantillon S_i d'US dans chaque UP u_i sélectionnée. Nous notons

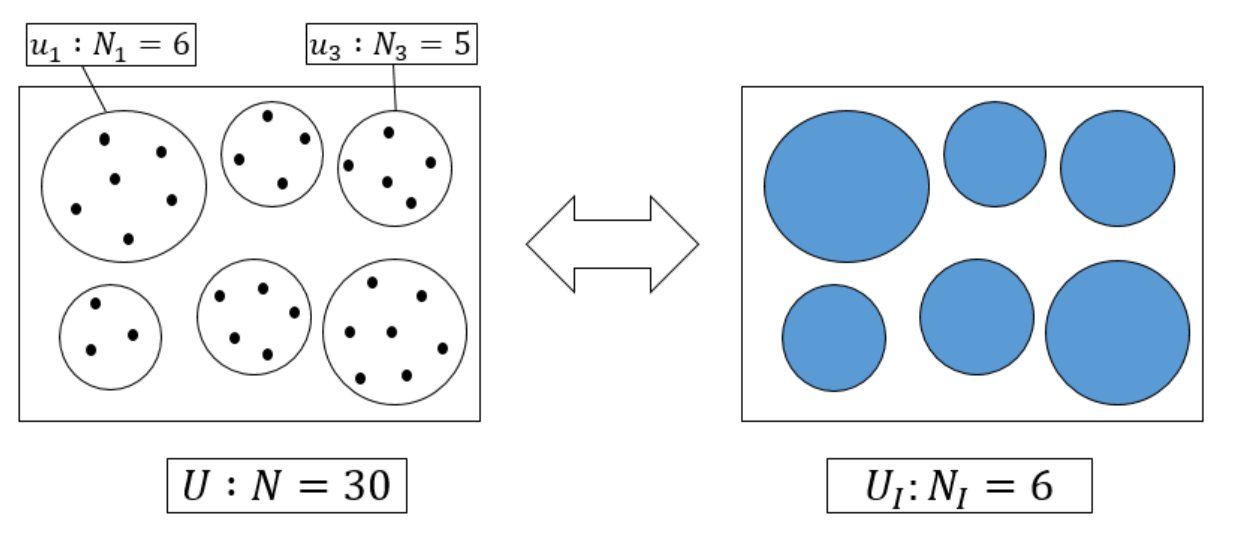
$$\pi_{k|i} = \mathbb{P}(k \in S_i | u_i \in S_I)$$

la probabilité de sélectionner une US k si son UP a été tirée au 1er degré.

L'échantillon global est $S = \bigcup_{u_i \in S_I} S_i$, et les probabilités d'inclusion valent

$$\pi_k = \pi_{Ii} \times \pi_{k|i} \quad \text{pour tout} \quad k \in u_i.$$

Exemple : SRS au premier degré



Nous sélectionnons :

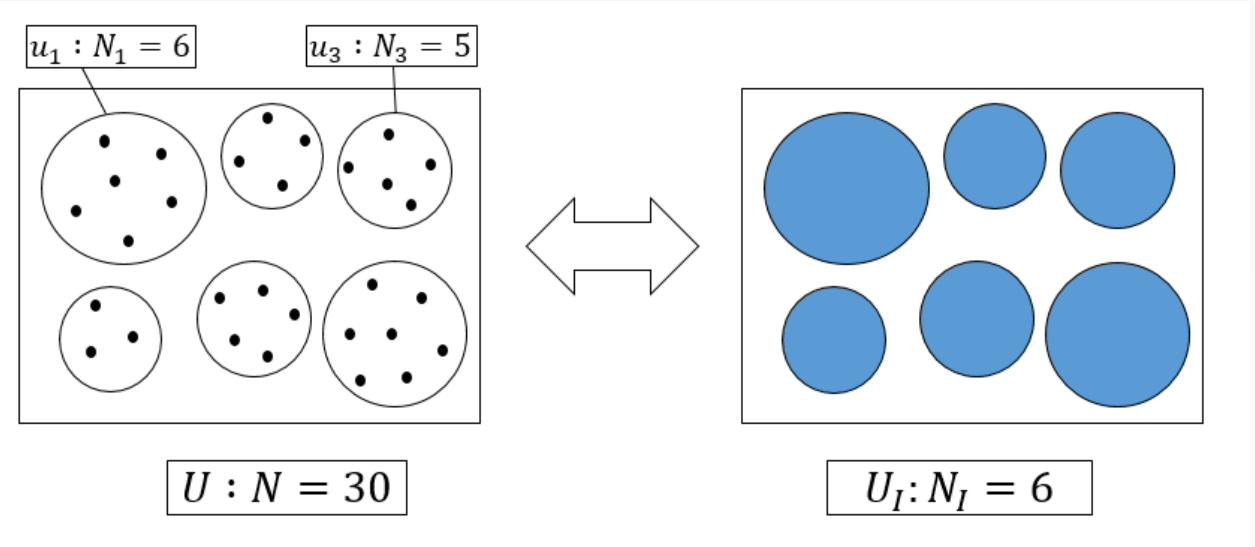
- un échantillon S_I de $n_I = 3$ UP selon un SRS,
- un échantillon S_i de $n_i \equiv n_0 = 2$ US selon un SRS dans chaque UP.

Nous avons par exemple :

$$\text{pour tout } k \in u_1 \quad \pi_k =$$

$$\text{pour tout } k \in u_3 \quad \pi_k =$$

Exemple : PPS au premier degré



Nous sélectionnons :

- un échantillon S_I de $n_I = 3$ UP à probabilités prop. à la taille,
- un échantillon S_i de $n_i \equiv n_0 = 2$ US selon un SRS dans chaque UP.

Nous avons par exemple :

$$\text{pour tout } k \in u_1 \quad \pi_k =$$

$$\text{pour tout } k \in u_3 \quad \pi_k =$$

Estimation

Pour une variable d'intérêt y, nous pouvons estimer son total t_y par

$$\hat{t}_{y\pi} = \sum_{k \in S} \frac{y_k}{\pi_k} = \sum_{u_i \in S_I} \frac{\hat{Y}_i}{\pi_{Ii}} \quad \text{avec} \quad \hat{Y}_i = \sum_{k \in S_i} \frac{y_k}{\pi_{k|i}}.$$

Sa variance peut s'obtenir en conditionnant sur l'échantillon S_I :

$$\begin{aligned} V_p(\hat{t}_{y\pi}) &= V_p\ E_p(\hat{t}_{y\pi}|S_I) + E_p\ V_p(\hat{t}_{y\pi}|S_I) \\ &= \underbrace{V_p\left(\sum_{u_i \in S_I} \frac{Y_i}{\pi_{Ii}}\right)}_{V_{UP}} + \underbrace{E_p\ V_p(\hat{t}_{y\pi}|S_I).}_{V_{US}} \end{aligned} \tag{2.32}$$

Le premier terme de variance est généralement prépondérant :

- Il dépend de la variabilité inter UP, d'autant plus grande que les UP sont grandes (effet taille) et/ou homogènes en intra (effet de grappe).
- Sous des hypothèses raisonnables, il est possible de montrer que

$$V_{UP} = O\left(\frac{N^2}{n_I}\right) \quad \text{et} \quad V_{US} = O\left(\frac{N^2}{n_I\ n_0}\right).$$

Sondage aléatoire simple à chaque degré

L'estimateur de Horvitz-Thompson se réécrit

$$\hat{t}_{y\pi} = \frac{N_I}{n_I} \sum_{u_i \in S_I} N_i \bar{y}_i \text{ avec } \bar{y}_i = \frac{1}{n_i} \sum_{k \in S_i} y_k. \tag{2.33}$$

Sa variance est donnée par :

$$V_p(\hat{t}_{y\pi}) = \underbrace{N_I^2 \left(1 - \frac{n_I}{N_I}\right) \frac{S_{YI}^2}{n_I}}_{V_{UP}} + \underbrace{\frac{N_I}{n_I} \sum_{u_i \in U_I} N_i^2 \left(1 - \frac{n_i}{N_i}\right) \frac{S_{yi}^2}{n_i}}_{V_{US}}, \tag{2.34}$$

avec

$$S_{YI}^2 = \frac{1}{N_I - 1} \sum_{u_i \in U_I} \left(Y_i - \frac{Y}{N_I}\right)^2 \quad \rightarrow \quad \text{Variance des } Y_i \text{ sur les UP}$$

$$S_{yi}^2 = \frac{1}{N_i - 1} \sum_{k \in u_i} \left(y_k - \frac{Y_i}{N_i}\right)^2 \quad \rightarrow \quad \text{Variance de } y \text{ dans l'UP } u_i$$

Plan à deux degrés auto-pondéré

Un plan à deux degrés couramment utilisé en pratique consiste :

- à sélectionner un échantillon d'UP avec des probabilités proportionnelles au nombre d'US,
- à tirer un échantillon de n_0 US dans chaque UP sélectionnée.

Cela conduit aux probabilités d'inclusion

$$\pi_k = n_I \frac{N_i}{N} \times \frac{n_0}{N_i} = \frac{n_I n_0}{N} = \frac{n}{N},$$

d'où le nom de plan auto-pondéré.

Dans une enquête auprès des ménages, ce plan de sondage permet :

- de donner de grosses probabilités de tirage aux plus grandes communes, ce qui réduit la variance,
- de tirer un même nombre de ménages dans chaque commune, ce qui équilibre la charge de travail des enquêteurs.

Exemple : enquêtes MICS

L'Enquête par grappes à indicateurs multiples (MICS) est un programme international d'enquête sur les ménages élaboré et appuyé par l'UNICEF. MICS est conçu pour recueillir des estimations sur les indicateurs clés qui sont utilisés pour évaluer la situation des enfants et des femmes.

C'est une source importante de données sur la protection de l'enfance, l'éducation de la petite enfance, et sur la santé et la nutrition des enfants.

Depuis le lancement des MICS dans les années 1990, plus de 300 enquêtes ont été réalisées dans plus de 100 pays. La dernière vague d'enquêtes a eu lieu en 2021, et la prochaine est prévue en 2025.

Lien : https://mics.unicef.org/

Exemple : enquêtes MICS

Extrait du rapport d'enquête de l'enquête MICS 2021 au Nigéria.

The sample for the MICS 2021 was designed to provide estimates for a large number of indicators on the situation of children and women at the national, rural/urban levels, for 36 states and the Federal Capital Territory (FCT), Abuja, as well as the six geo-political zones of Nigeria.

States were identified as the main sampling strata and the sample of households was selected in two stages. Within each stratum, at the first sampling stage a specified number of census enumeration areas (EAs) were selected systematically with probability proportional to size.

After a household listing was carried out within the selected EAs, a systematic sample of 20 households was drawn in each sample EA. The total target sample size for the main Nigeria MICS was 1,850 clusters and 37,000 households.

Estimation de variance

Il est possible de reprendre la décomposition de variance (2.32) pour calculer un estimateur de variance, comprenant un terme pour chaque degré. C'est l'approche utilisée dans le package `Gustave`.

Beaucoup d'enquêtes utilisent plutôt l'estimateur de variance simplifié

$$\hat{V}_{wr}(\hat{t}_{y\pi}) = \frac{n_I}{n_I - 1} \sum_{u_i \in S_I} \left(\frac{\hat{Y}_i}{\pi_{Ii}} - \frac{\hat{t}_{y\pi}}{n_I} \right)^2 . \tag{2.35}$$

Sous l'hypothèse d'un tirage avec remise au premier degré, il est sans biais pour l'ensemble de la variance, quel que soit le nombre de degrés de tirage. Il est également très simple à calculer car (en dehors des identifiants des UP), il ne nécessite de connaître que les poids de sondage.
Il est généralement conservatif dans les enquêtes réelles.