Enquêtes répétées dans le temps

Guillaume Chauvet

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

06/10/2025

Introduction

Objectifs

L'objectif de ce cours est de présenter le cas des enquêtes répétées, et les principes d'estimation ponctuelle (à une date donnée) ou dans le temps (évolution entre deux dates).

Nous considérerons le cas :

- d'enquêtes où les unités d'échantillonnage sont les unités d'observation,
- d'enquêtes où les unités d'échantillonnage ne sont pas les unités d'observation, comme les enquêtes auprès des ménages.

Nous évoquerons les aspects liés au redressement des estimateurs, notamment pour tenir compte de la non-réponse.

Nous nous limiterons au cas de l'estimation d'un total : pas de difficultés spécifiques à l'estimation de paramètres complexes pour une enquête répétée dans le temps.

Terminologie

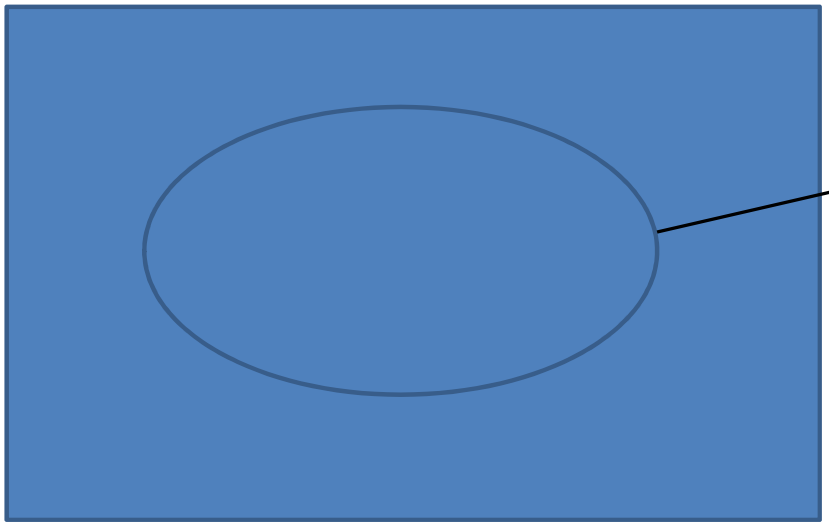
Nous pouvons nous intéresser à une population à une occasion, mais aussi à plusieurs occasions pour mesurer des changements et notamment les entrées et sorties d'unités de cette population (Juillard, 2016).

De façon générale, nous parlerons d'**enquête répétée dans le temps** dans le cas d'une collecte de données à plusieurs occasions, pas forcément sur les mêmes unités. Elles peuvent répondre à deux principaux objectifs :

1. estimer un paramètre à un temps donné à un niveau agrégé (**estimation transversale**),
2. calculer l'évolution d'un paramètre entre deux temps (**estimation longitudinale agrégée**), ou mesurer des évolutions individuelles entre deux temps (**estimation longitudinale individuelle**).

Principales notations

Population initiale



Population à la date t : $U(t)$

Estimation dans le temps

La situation se complique quand on souhaite réaliser une estimation à la date t+1, car la population $U(t)$ n'est pas stable dans le temps. La nouvelle population $U(t+1)$ contient de nouvelles unités, alors que d'autres ont disparu.

Il est donc important de bien spécifier le paramètre d'intérêt. Nous distinguerons en particulier :

- les estimations transversales ("cross-sectional estimation"), à une date donnée ;
- les estimations longitudinales ("longitudinal estimation"), où nous estimons l'évolution d'un paramètre entre deux dates.

Il est également important de bien spécifier le **champ** de l'étude, i.e. l'ensemble des unités pour lequel nous souhaitons produire une estimation.

Notation

Population transversale à la date t+1 : $U(t+1)$

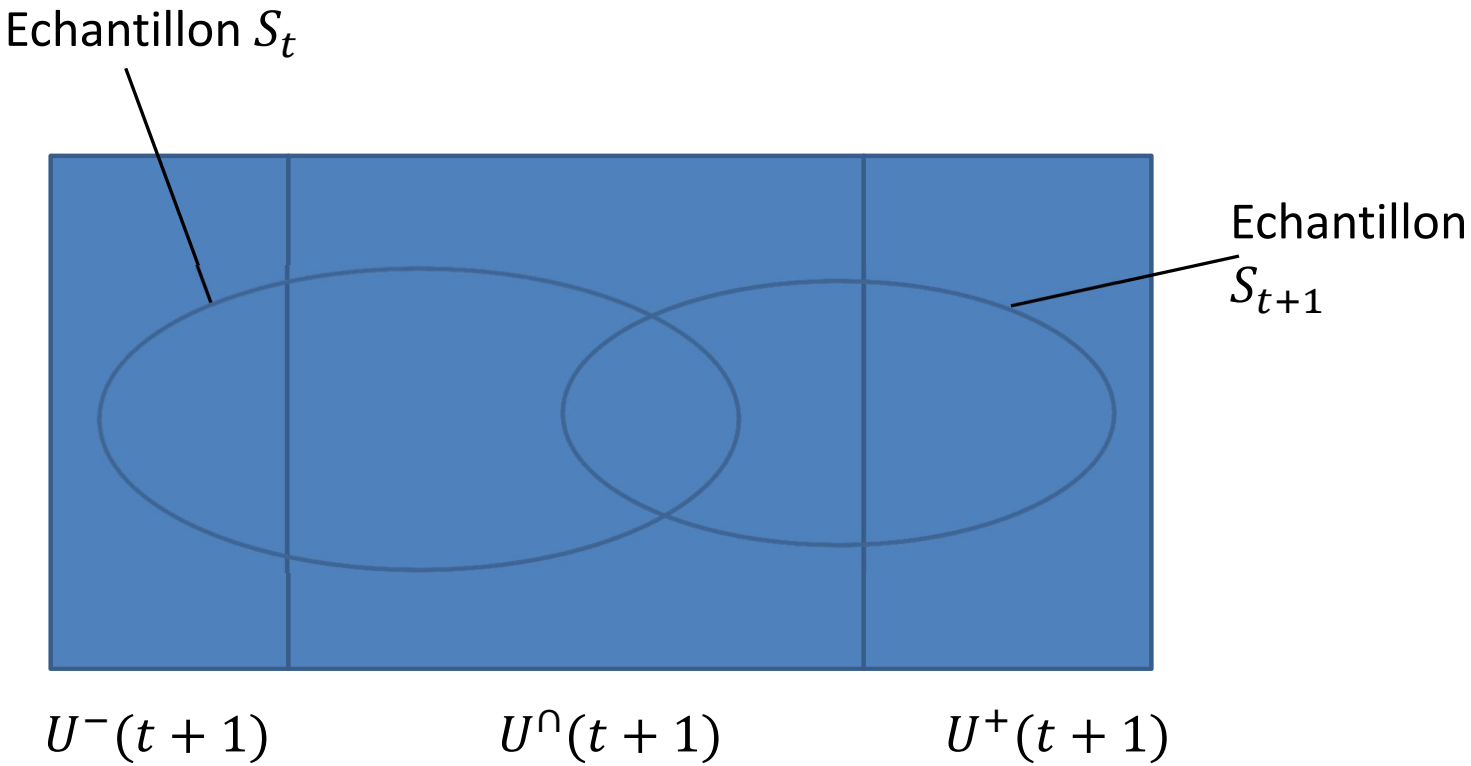
Population transversale à la date t : $U(t)$

$U^{-}(t+1)$: Décès entre les dates t et t+1

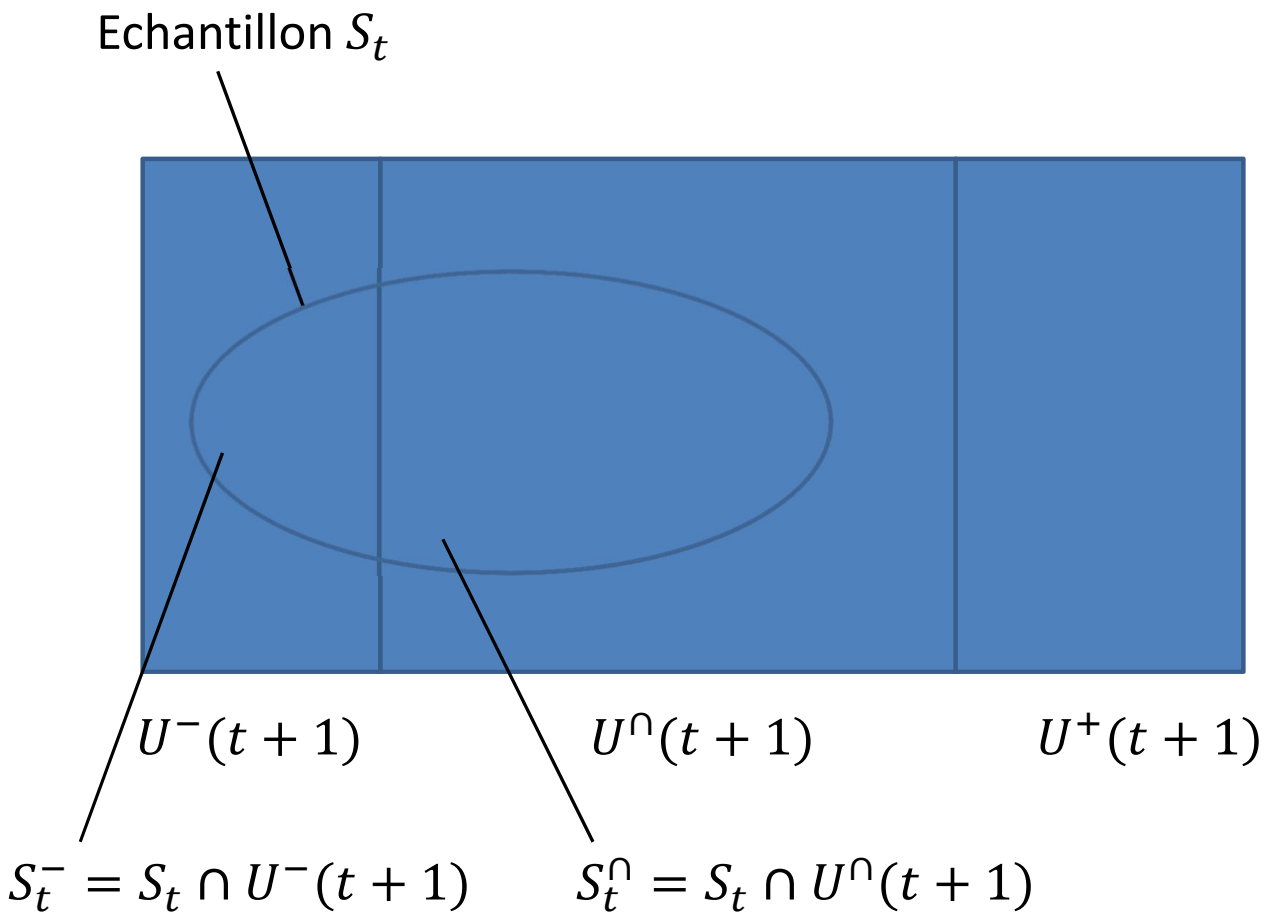
$U^{\cap}(t+1)$: Individus communs aux dates t et t+1

$U^{+}(t+1)$: Naissances entre les dates t et t+1

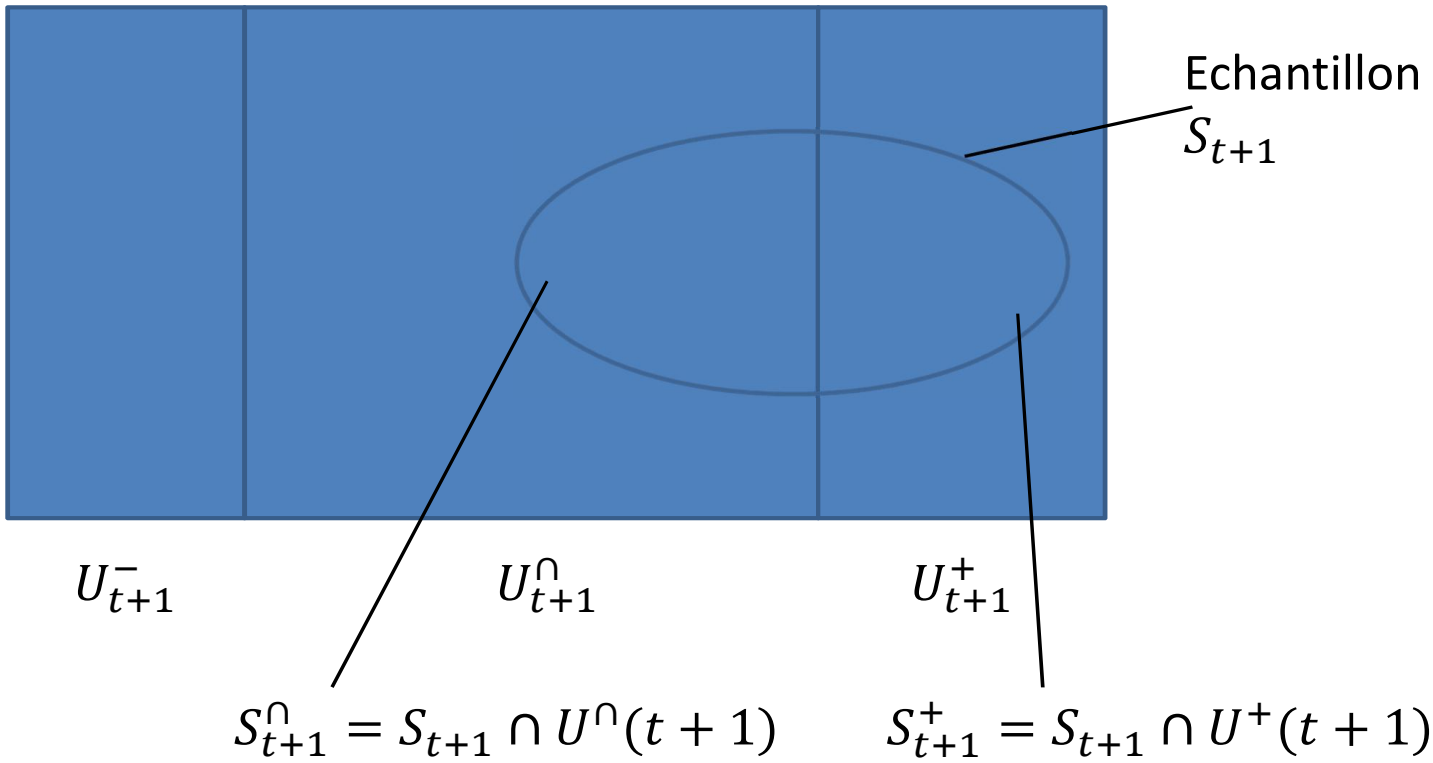
Notation



Notation


Echantillon S_t
$U^-(t+1)$
$U^\cap(t+1)$
$U^+(t+1)$
$S_t^- = S_t \cap U^-(t+1)$
$S_t^\cap = S_t \cap U^\cap(t+1)$

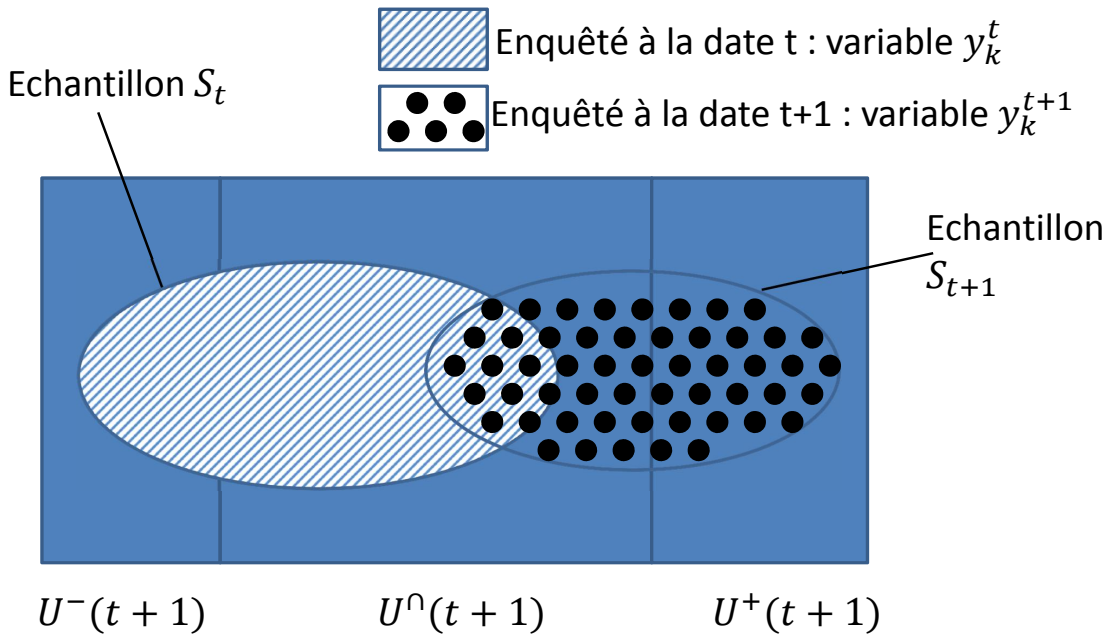
Notation


Echantillon
S_{t+1}
U^{-}_{t+1}
$U^{\cap}_{t+1}$
U^{+}_{t+1}
$S^{\cap}_{t+1} = S_{t+1} \cap U^{\cap}(t+1)$
$S^{+}_{t+1} = S_{t+1} \cap U^{+}(t+1)$

Types d'enquêtes répétées dans le temps

Enquêtes transversales répétées

Dans les **enquêtes transversales répétées** ("repeated cross-sectional survey"), nous sélectionnons à des dates différentes des échantillons **indépendants**.



Enquêtes transversales répétées (2)

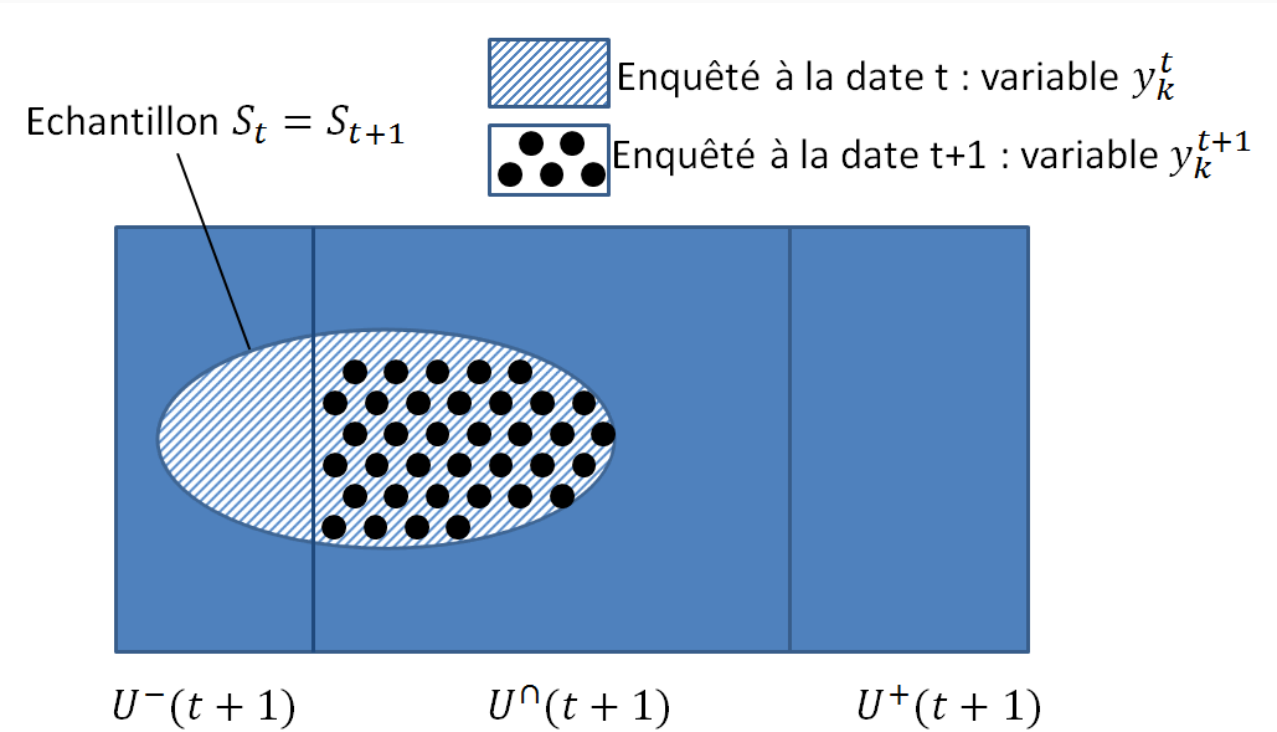
Dans les **enquêtes transversales répétées** ("repeated cross-sectional survey"), nous sélectionnons à des dates différentes des échantillons **indépendants**.

Si l'enquête est répétée régulièrement, on parle d'enquête périodique ("periodic survey") ou d'enquête continue ("continuing survey").
Ce sont des enquêtes adaptées à la production d'estimations transversales, et pas du tout à des mesures d'évolution individuelles (aucune garantie de sélectionner des individus communs à chaque temps).

Exemple : Enquête sur le patrimoine des ménages (version 2014-2015), qui a pour objectif de décrire les actifs financiers, immobiliers et professionnels des ménages. Sondage stratifié à deux degrés en métropole, à un degré dans les DOM. Passage à un panel rotatif (voir plus loin) depuis 2017.

Enquêtes par panel

Dans les **enquêtes par panel** ("panel surveys"), les mêmes mesures sont effectuées à différents temps sur le même échantillon. On parle encore d'enquête longitudinale.
C'est un plan de sondage parfaitement adapté pour une étude longitudinale individuelle.



Enquêtes par panel (2)

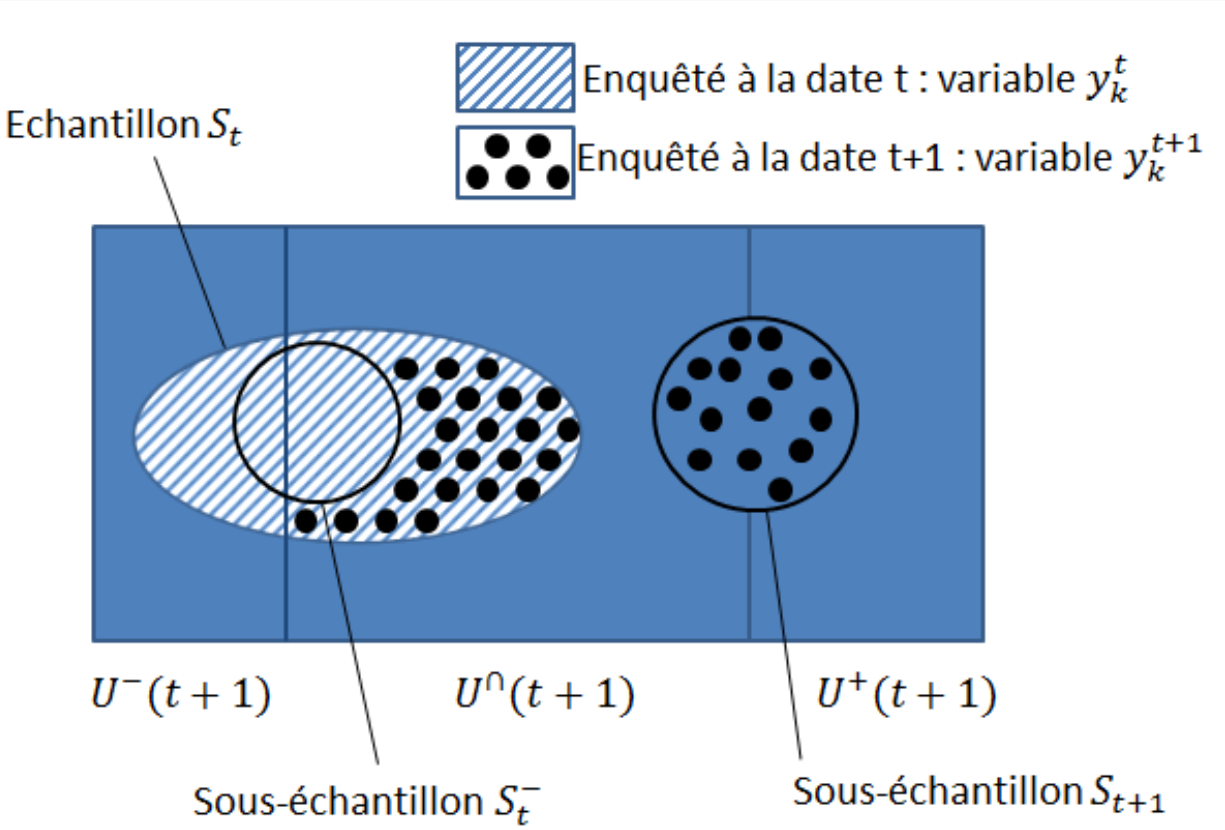
Dans les **enquêtes par panel** ("panel surveys"), les mêmes mesures sont effectuées à différents temps sur le même échantillon. On parle encore d'enquête longitudinale.
On parle plus spécifiquement d'une **cohorte** si les unités sont liées par un évènement commun (naissance, mariage).
C'est un plan de sondage parfaitement adapté pour une étude longitudinale individuelle.

Exemple : Etude Longitudinale Française depuis l'Enfance (ELFE), lancée en 2011. Consacrée au suivi des enfants de la naissance à l'âge adulte, elle aborde les multiples aspects de la vie de l'enfant sous l'angle des sciences sociales, de la santé et de la santé-environnement.

Plan de sondage produit, obtenu par croisement d'un échantillon de maternités et d'un échantillon de jours.

Enquêtes par panel rotatif

Dans les **enquêtes par panel rotatif** ("rotating panel surveys"), un sous-échantillon d'unités sort du panel à chaque temps, et un autre sous-échantillon entre pour le remplacer.



Enquêtes par panel rotatif (2)

Dans les **enquêtes par panel rotatif** ("rotating panel surveys"), un sous-échantillon d'unités sort du panel à chaque temps, et un autre sous-échantillon entre pour le remplacer.

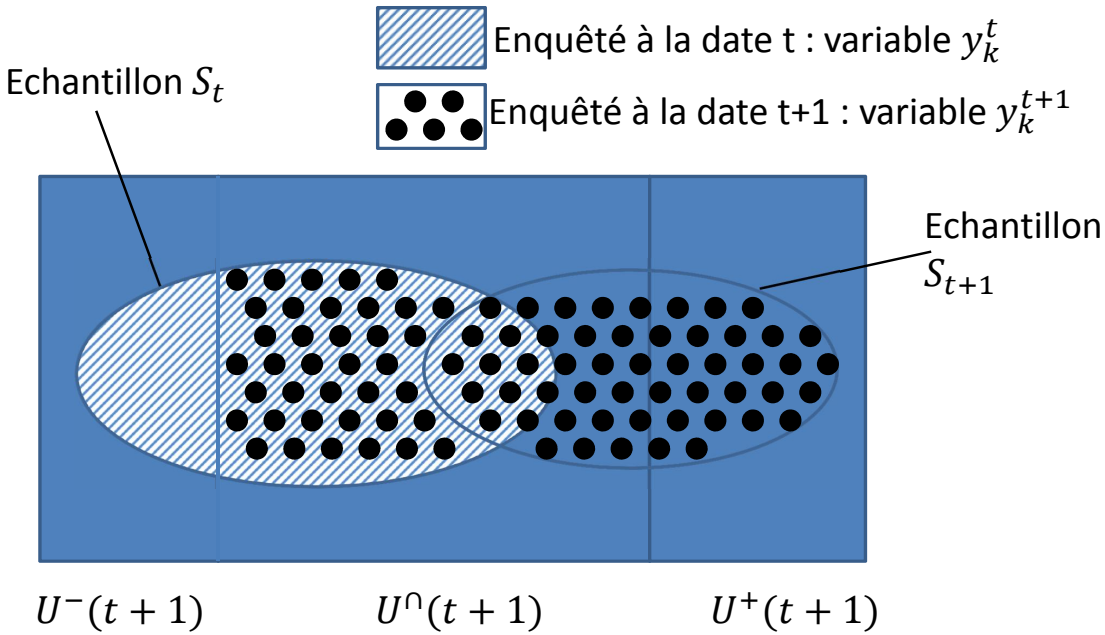
Ce type d'enquête est conçu pour répondre aux deux objectifs d'estimation.

Exemple : Enquête Emploi en Continu (EEC), réalisée par tirage de logements issus uniquement des fichiers de la taxe d'habitation.

Tirage de grappes de logements contigus selon un plan à deux degrés. Une grappe est enquêtée durant six trimestres consécutifs, puis remplacée dans l'échantillon par une autre grappe.

Enquêtes à échantillon partagé

Dans les **enquêtes à échantillon partagé** ("split panel surveys"), on utilise un panel que l'on complète à chaque date par un nouvel échantillon indépendant. Ce type d'enquête est conçu pour permettre des estimations transversales et longitudinales.



Enquêtes à échantillon partagé (2)

Dans les **enquêtes à échantillon partagé** ("split panel surveys"), on utilise un panel que l'on complète à chaque date par un nouvel échantillon indépendant.

Ce type d'enquête est également conçu pour répondre aux deux objectifs d'estimation.

Exemple : Enquête Panel Politique de la Ville (PPV) réalisée auprès de ménages de 2011 à 2014, qui s'intéresse aux différents aspects de la vie quotidienne des habitants des Zus.

Plan de sondage à deux degrés, avec tirage d'échantillons complémentaires de ménages de 2012 à 2014, notamment pour compenser de l'**attrition**.

Types d'enquêtes répétées dans le temps

	Estimation transversale	Estimation longitudinale agrégée	Estimation longitudinale individuelle
Enq. trans. répétées	☺	☹	☠
Enq. par panel	☠	☺	☺
Enq. par panel rotatif	☺	☺	☺
Enq. à échant partagé	☺	☺	☺

Remarques

Dans le cas d'une enquête répétée dans le temps, il existe généralement de multiples populations cible possibles, ce que l'on appelle encore le champ de l'étude. Il est nécessaire d'être précis sur l'objectif de l'estimation.

Les stratégies d'échantillonnage "optimales" diffèrent selon le type d'estimation, transversal ou longitudinal. Il est fréquent d'utiliser des stratégies mixtes permettant de faire les deux estimations (panel rotatif, échantillon partagé).

Le calcul d'une pondération peut rapidement devenir complexe pour les enquêtes longitudinales. On a recours à deux outils : la méthode de partage des poids (Lavallée, 2007) et la méthode de l'estimation composite.

Fil rouge : l'enquête PPV

L'enquête Panel Politique de la Ville (PPV) a été mise en place pour étudier les conditions de vie des habitants des quartiers de la politique de la ville. Les quatre vagues d'enquête (entre 2011 et 2014) visent à appréhender :

- la mobilité résidentielle dans les quartiers,
- la perception des politiques publiques,
- l'impact des politiques publiques sur les bénéficiaires.

L'échantillon initial est tiré selon un plan à 3 degrés (Couvert el al., 2016) :

- tirage d'un échantillon de quartiers de la politique de la ville, stratifié selon le degré d'avancement du programme de rénovation urbaine,
- dans les quartiers, tirage d'un échantillon de logements à l'aide d'une base de sondage constituée à partir des **EAR**,
- dans les logements, tirage d'une **unité de vie** ($\simeq$ un ménage) et de tous les individus de cette unité de vie.

Echantillonnage en population finie

Notations

Notation

Nous considérons à la date initiale t une population finie $U(t)$ d'unités statistiques, de taille N^t. Une variable d'intérêt y^t prend la valeur y_k^t pour chaque unité $k \in U(t)$.

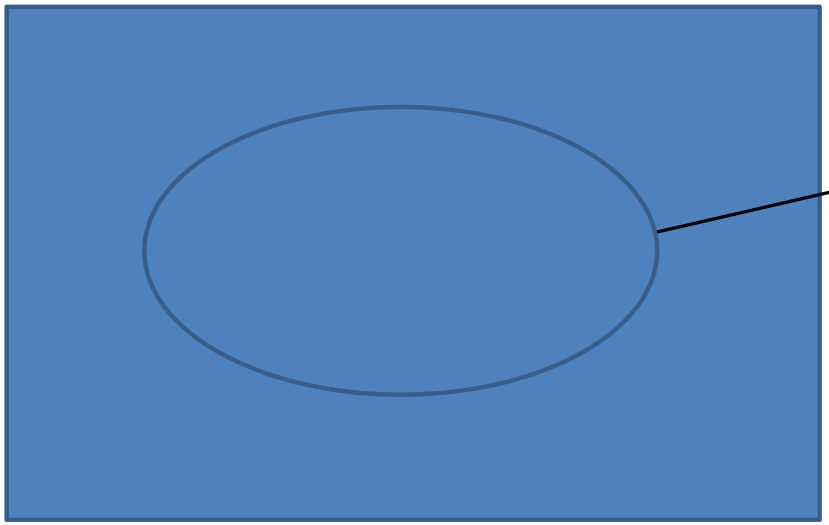
Nous sélectionnons un échantillon S_t, avec des probabilités d'inclusion d'ordre 1 $\pi_k^t > 0$, en supposant l'absence de biais de couverture.

Le total $Y(t) = \sum_{k \in U(t)} y_k^t$ peut être estimé sans biais par l'estimateur de Horvitz-Thompson

$$\hat{Y}_t(t) = \sum_{k \in S_t} \frac{y_k^t}{\pi_k^t} = \sum_{k \in S_t} d_k^t y_k^t,$$

en notant $d_k^t = 1/\pi_k^t$ le poids de sondage à la date t.

Notation



Population à la date t : $U(t)$

Variance de l'estimateur

La variance de l'estimateur $\hat{Y}_t(t)$ est donnée par

$$V\{\hat{Y}_t(t)\} = \sum_{k,l \in U(t)} \Delta_{kl}^t (d_k^t y_k^t)(d_l^t y_l^t),$$

$$\text{avec} \qquad \Delta_{kl}^t = \pi_{kl}^t - \pi_k^t \pi_l^t,$$

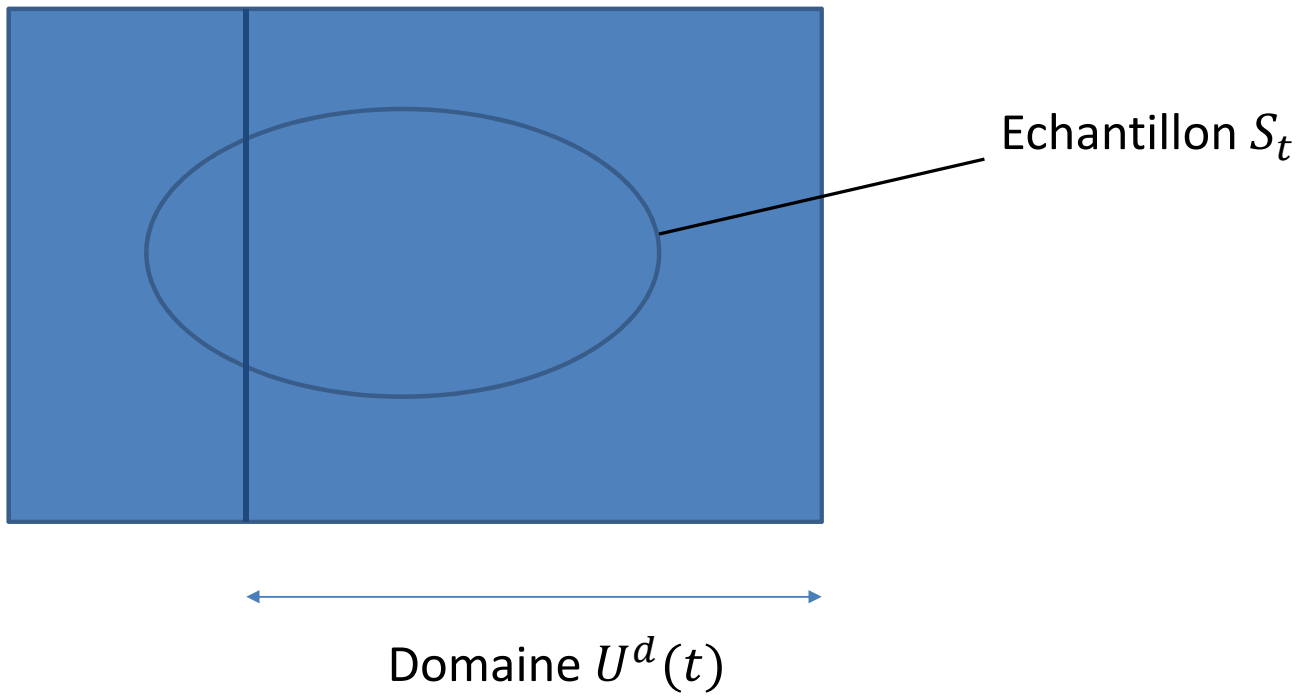
et π_{kl}^t la probabilité d'inclusion d'ordre 2 des unités k et l dans l'échantillon S_t.

Cette variance peut être estimée sans biais par

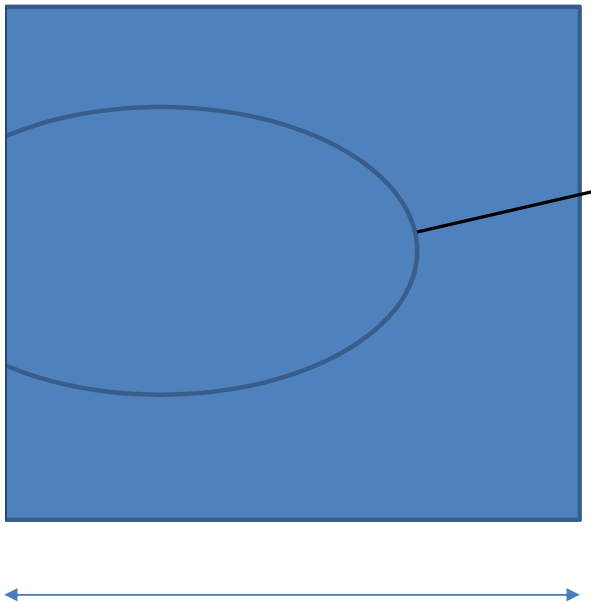
$$\hat{V}_t\{\hat{Y}_t(t)\} = \sum_{k,l \in S_t} \frac{\Delta_{kl}^t}{\pi_{kl}^t} (d_k^t y_k^t)(d_l^t y_l^t)$$

si tous les $\pi_{kl}^t > 0$.

Estimation sur domaine



Estimation sur domaine (2)



Estimation sur domaine (3)

Supposons que nous souhaitions estimer le total

$$Y^d(t) = \sum_{k \in U^d(t)} y_k^t. \quad (1)$$

de la variable y^t sur la sous-population (ou domaine) $U^d(t)$.

Nous pouvons l'estimer sans biais, en utilisant le sous-échantillon S_t^d qui tombe dans le domaine uniquement :

$$\hat{Y}_t^d(t) = \sum_{k \in S_t^d} d_k^t y_k^t,$$

$$V\{\hat{Y}_t^d(t)\} = \sum_{k,l \in U^d(t)} \Delta_{kl}^t (d_k^t y_k^t)(d_l^t y_l^t).$$

Réaliser une estimation sur domaine est particulièrement important pour une estimation à un temps $t > 1$, afin de bien isoler la sous-population cible.

Exemples de plans de sondage

Sondage aléatoire simple

Nous sélectionnons dans $U(t)$ un échantillon S_t par sondage aléatoire simple (SRS) de taille n^t. Nous pouvons estimer le total $Y(t)$ par

$$\hat{Y}_t(t) = N^t \bar{y}^t \quad \text{avec} \quad \bar{y}^t = \frac{1}{n^t} \sum_{k \in S_t} y_k^t.$$

Sa variance est donnée par

$$\begin{aligned} V\{\hat{Y}_t(t)\} &= (N^t)^2 \left(\frac{1}{n^t} - \frac{1}{N^t} \right) S_y^{2t}, \\ \text{avec } S_y^{2t} &= \frac{1}{N^t - 1} \sum_{k \in U(t)} (y_k^t - \mu_y^t)^2. \end{aligned}$$

Elle peut être estimée sans biais par

$$\begin{aligned} \hat{V}\{\hat{Y}_t(t)\} &= (N^t)^2 \left(\frac{1}{n^t} - \frac{1}{N^t} \right) s_y^{2t}, \\ \text{avec } s_y^{2t} &= \frac{1}{n^t - 1} \sum_{k \in S^t} (y_k^t - \bar{y}^t)^2. \end{aligned}$$

Exercice 1

Enquêtes entreprise

Population partitionnée en H strates $U_1^t, \ldots, U_H^t$ de tailles $N_1^t, \ldots, N_H^t$.

Nous sélectionnons dans chaque strate un échantillon S_{th} par sondage aléatoire simple (SRS) de taille n_h^t, et S_t est donné par la réunion de ces sous-échantillons.

$$\begin{aligned}
\hat{Y}_t(t) &= \sum_{h=1}^{H} N_h^t \bar{y}_h^t, \\
V\{\hat{Y}_t(t)\} &= \sum_{h=1}^{H} (N_h^t)^2 \left(\frac{1}{n_h^t} - \frac{1}{N_h^t} \right) S_{yh}^{2t}, \\
\hat{V}\{\hat{Y}_t(t)\} &= \sum_{h=1}^{H} (N_h^t)^2 \left(\frac{1}{n_h^t} - \frac{1}{N_h^t} \right) s_{yh}^{2t}.
\end{aligned}$$

Enquêtes ménage

Ces enquêtes visent à décrire les conditions de vie des ménages (enquête emploi, enquête logement, enquête patrimoine, ...). Elles sont souvent réalisées en utilisant un plan de sondage à **plusieurs degrés**.

Au temps t :

- un échantillon $S_{t,men}$ de ménages est sélectionné,
- dans chaque ménage $k \in S_{t,men}$, un échantillon $S_{tk,ind}$ d'individus est sélectionné.

L'échantillon final d'individus est donné par

$$S_{t,ind} = \bigcup_{k \in S_{t,men}} S_{tk,ind}.$$

Un individu est une unité statistique stable dans le temps, mais pas un ménage (séparation, mise en ménage, départ d'un individu, ...). C'est une des difficultés des enquêtes répétées dans le temps auprès des ménages.

Poids de sondage

Nous notons :

- $\pi_k^t = \mathbb{P}(k \in S_{t,men})$
  probabilité de sélectionner le ménage k dans $S_{t,men}$,
- $\pi_{l|k}^t = \mathbb{P}(l \in S_{t,ind} | k \in S_{t,men})$
  probabilité de sélectionner l'indiv. l si son ménage k est tiré.

D'après le plan de sondage utilisé, pour tout individu $l \in k$:

$$\pi_l^t \quad = \quad \pi_k^t \times \pi_{l|k}^t.$$

Chaque individu l de $S_{t,ind}$ a donc pour poids de sondage

$$d_l^t \quad = \quad \frac{1}{\pi_k^t} \times \frac{1}{\pi_{l|k}^t} \equiv d_k^t d_{l|k}^t \text{ pour tout } l \in k.$$

Traitement de la non-réponse

En pratique, nous observons un phénomène de non-réponse qui diminue la taille de l'échantillon effectivement observé. Nous nous concentrons ici sur le problème de **non-réponse totale**.

Cette non-réponse peut intervenir à la fois au niveau ménage et à la fois au niveau individuel.

Nous nous limitons ici au cas d'une enquête ménage avec interrogation de tous les individus du ménage. Les poids de départ sont donc :

$$\begin{aligned} d_k^t & \quad \text{pour le ménage } k, \\ d_l^t = d_k^t & \quad \text{pour l'individu } l \in \text{ ménage } k. \end{aligned}$$

Non-réponse au niveau ménage

En raison de la non-réponse des ménages, seul un sous-échantillon de répondants $S_{rt,men}$ est observé.

Chaque ménage $k \in S_{rt,men}$ possède le poids redressé de la non-réponse

$$d_{rk}^t = \underbrace{d_k^t}_{\text{poids de sondage}} \times \underbrace{\frac{1}{\hat{p}_k^t}}_{\text{poids de NR ménage}}.$$

Par exemple, dans le cas de l'enquête PPV, la correction de la non-réponse totale des ménages s'est faite par la méthode des Groupes Homogènes de Réponse (GHR) :

- variables explicatives identifiées par régression logistique : nb de pièces, HLM (oui/non), type d'habitation, année de construction.
- constitution des GHR par la méthode des scores (8 groupes).

Non-réponse au niveau individuel

En raison de la non-réponse des individus, seul un sous-échantillon de répondants $S_{rt,ind}$ est observé.

Chaque individu $l \in S_{rt}^{ind}$ possède le poids redressé de la non-réponse

$$d_{rl}^t = \underbrace{d_{rk}^t}_{\text{poids ménage}} \times \underbrace{\frac{1}{\hat{p}_l^t}}_{\text{poids de NR individuel}} \quad \text{avec } k \ni l. \tag{2}$$

Par exemple, dans le cas de l'enquête PPV, la correction de la non-réponse totale des individus s'est faite par la méthode des GHR :

- variables explicatives identifiées par régression logistique : âge, lieu de naissance, statut matrimonial.
- constitution des GHR par la méthode des scores (5 groupes).

Mise en commun de plusieurs échantillons

Exercice 2

Méthode d'estimation composite

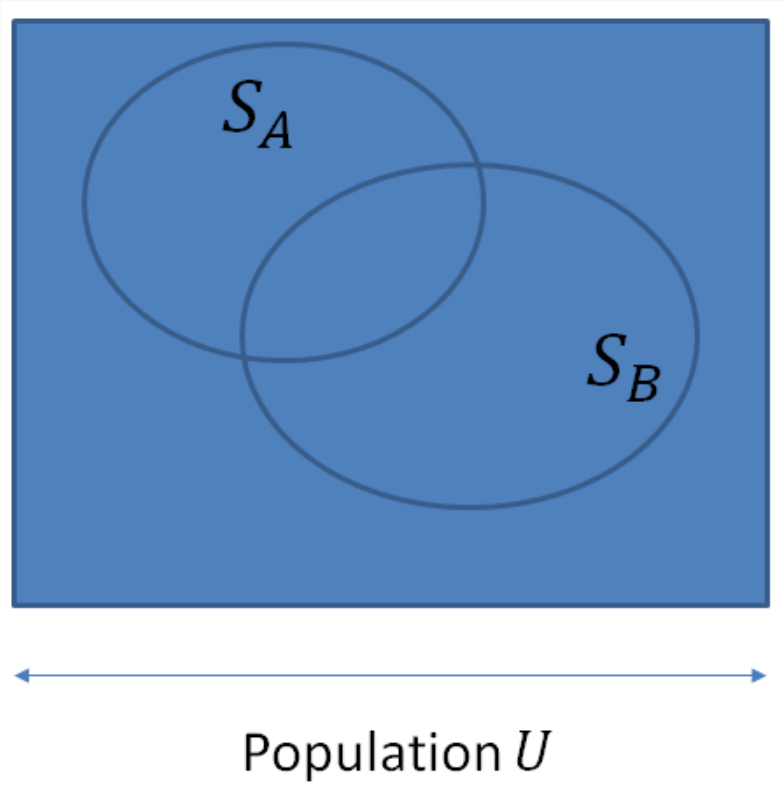
Principe

La méthode de l'estimation composite est très utile quand nous disposons de plusieurs échantillons pour couvrir une population, et que nous souhaitons combiner les estimations. Elle rentre dans la classe plus large des méthodes d'estimation sur bases de sondage multiples (Lohr, 2009).

Le principe est le suivant :

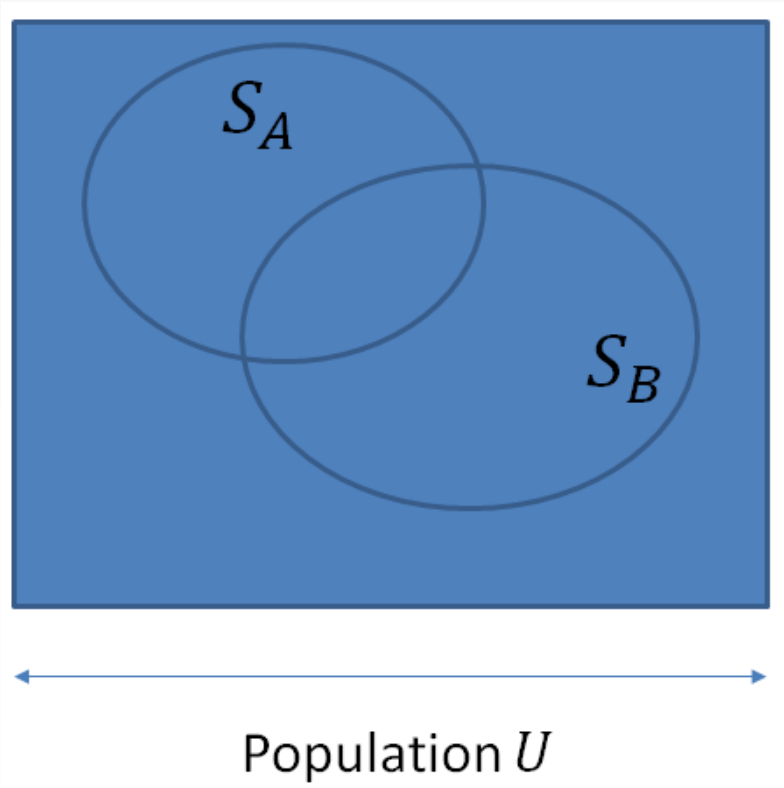
1. La population-cible U est partitionnée en sous-populations, en fonction du nombre d'échantillons permettant de couvrir chaque sous-population.
2. Dans chaque sous-population, nous obtenons une estimation non biaisée du total, en combinant les différents échantillons.
3. Nous ajoutons les différents estimateurs.

Exemple 1 : 2 échantillons pour une même population



- Soit $\hat{Y}_A = \sum_{k \in S_A} d_{Ak} y_k$ l'estimateur de HT calculé sur S_A.
  Soit $\hat{Y}_B = \sum_{k \in S_B} d_{Bk} y_k$ l'estimateur de HT calculé sur S_B.
- Chaque estimateur permet d'obtenir une estimation sans biais sur U.
- Nous pouvons combiner les deux à l'aide de l'estimateur composite $\hat{Y}(\theta) = \theta \hat{Y}_A + (1 - \theta) \hat{Y}_B$.

Exemple 1 : 2 échantillons pour une même population

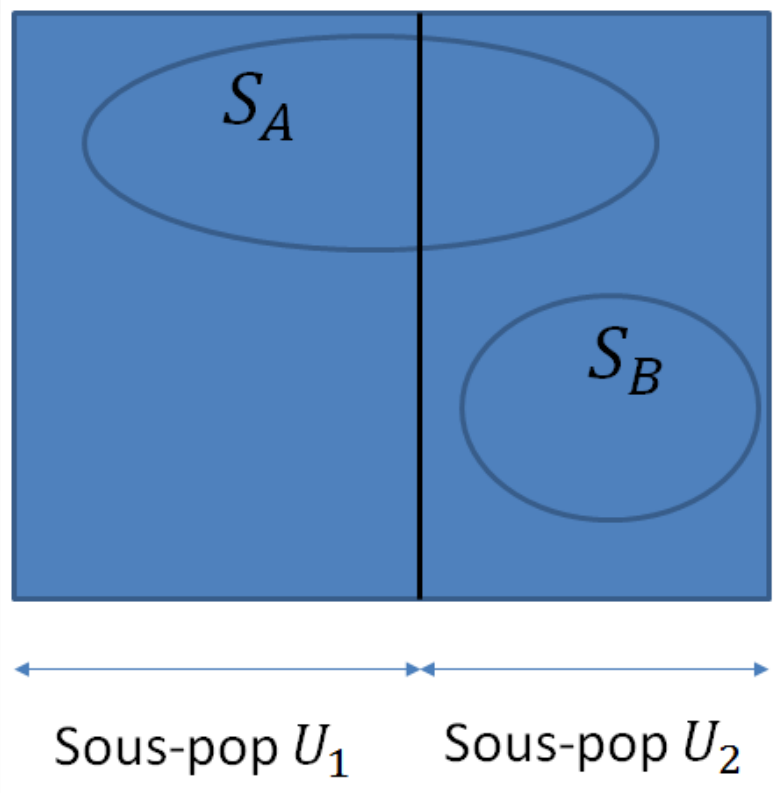


- Soit S l'échantillon réunion de S_A et de S_B. Nous obtenons sur S les poids

$$d_{\theta k} = \begin{cases} \theta d_{Ak} & \text{si } k \in S_A, \\ (1-\theta) d_{Bk} & \text{si } k \in S_B. \end{cases}$$

- Choix classiques pour le paramètre : $\theta = 0.5$ ou $\theta = \dfrac{n_A}{n_A + n_B}$.

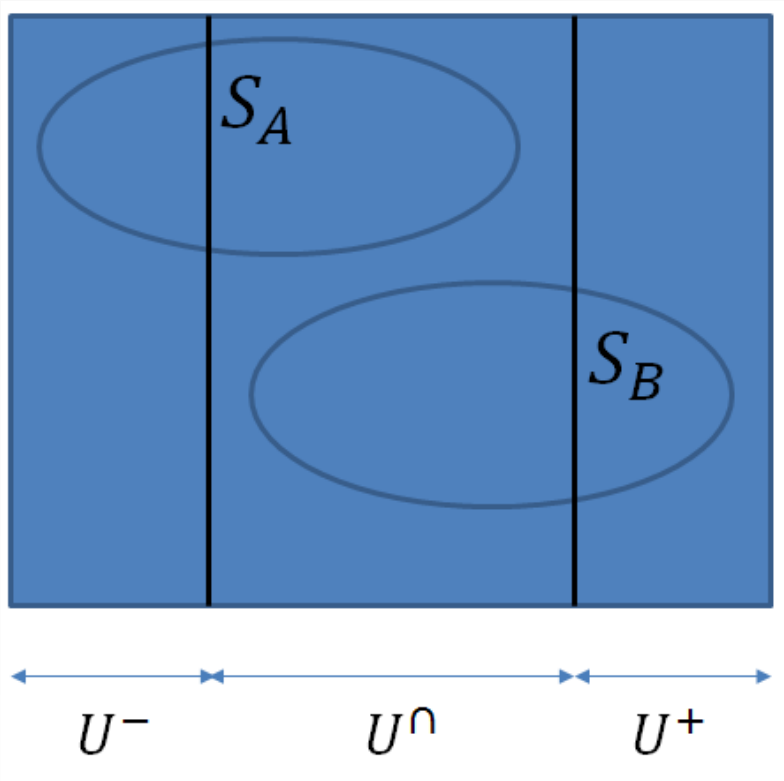
Exemple 2 : 2 échantillons dont l'un couvre une sous-population



- Soit $\hat{Y}_A = \sum_{k \in S_A} d_{Ak} y_k$ l'estimateur de HT calculé sur S_A.
  $\Rightarrow$ estimateur sans biais sur U.
- Soit $\hat{Y}_B = \sum_{k \in S_B} d_{Bk} y_k$ l'estimateur de HT calculé sur S_B.
  $\Rightarrow$ estimateur sans biais sur U_2.
- Soit $S = S_A \cup S_B$. Nous pouvons utiliser les poids

$$d_{\theta k} = \begin{cases} d_{Ak} & \text{si } k \in S_A \cap U_1, \\ \theta d_{Ak} & \text{si } k \in S_A \cap U_2, \\ (1-\theta) d_{Bk} & \text{si } k \in S_B. \end{cases}$$

Exemple 3 : 2 échantillons couvrant deux populations



- Soit $\hat{Y}_A = \sum_{k \in S_A} d_{Ak} y_k$ l'estimateur de HT calculé sur S_A.
  $\Rightarrow$ estimateur sans biais sur $U^- \cup U^\cap$.
- Soit $\hat{Y}_B = \sum_{k \in S_B} d_{Bk} y_k$ l'estimateur de HT calculé sur S_B.
  $\Rightarrow$ estimateur sans biais sur $U^\cap \cup U^+$.
- Soit $S = S_A \cup S_B$. Nous pouvons utiliser les poids

$$d_{\theta k} = \begin{cases} d_{Ak} & \text{si } k \in S_A \cap U^-, \\ \theta d_{Ak} & \text{si } k \in S_A \cap U^\cap, \\ (1-\theta) d_{Bk} & \text{si } k \in S_B \cap U^\cap, \\ d_{Bk} & \text{si } k \in S_B \cap U^+. \end{cases}$$

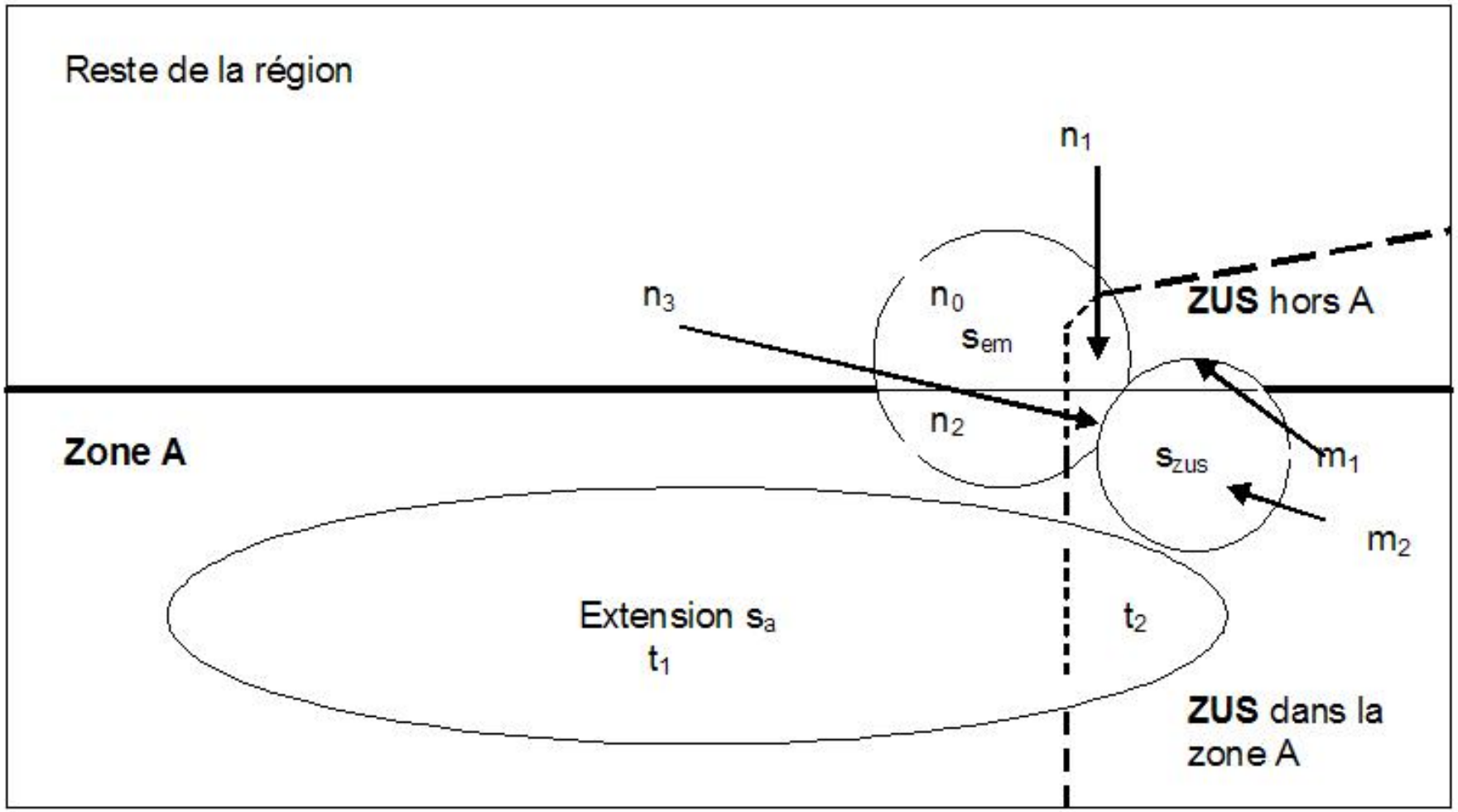
Cas de l'enquête Logement 2006

L'Enquête Logement 2006 est une enquête auprès des ménages, qui a donné lieu à une extension régionale et à plusieurs extensions locales au niveau de la région Bretagne notamment. Un complément d'échantillon a également été sélectionné dans des bases externes.

L'échantillon a été sélectionné en quatre temps :

- Sélection de l'échantillon national dans l'Echantillon Maître de 99 (RP99, BSLN),
- Sélection d'une extension régionale dans l'EMEX, pour les régions concernées,
- Sélection d'extensions d'échantillon au niveau local, pour les régions concernées,
- Sélection d'échantillons complémentaires dans des bases externes.

Schéma récapitulatif (Le Guennec, 2009)



Méthode de partage des poids

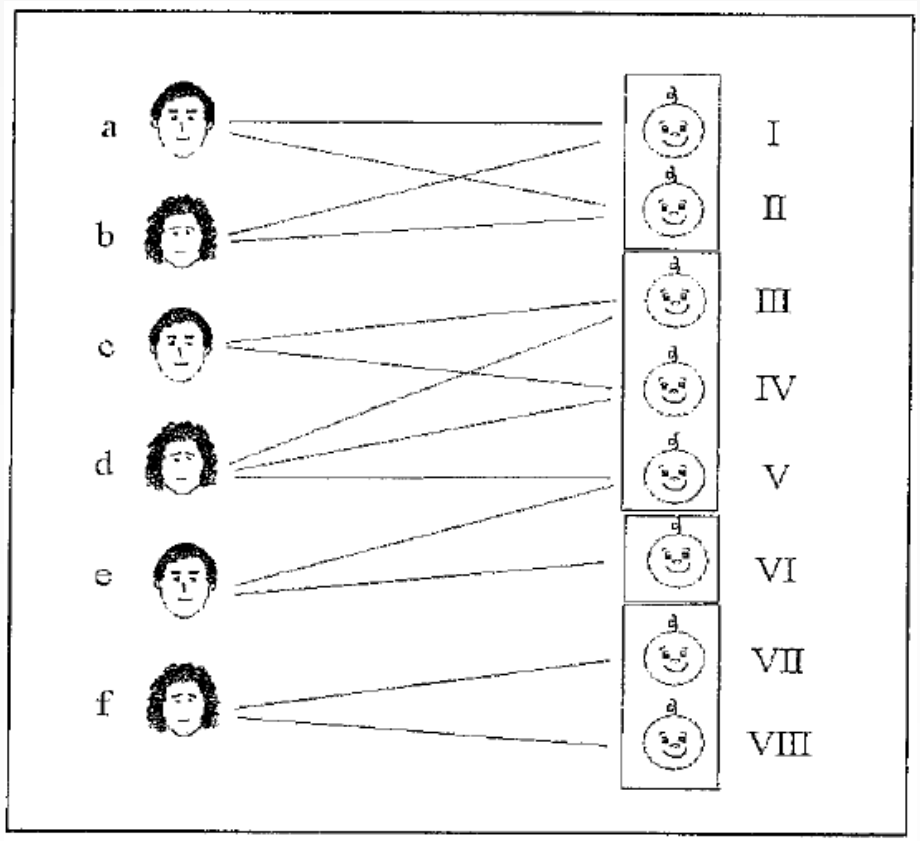
Motivation

Il arrive que les unités de la population-cible ne soient pas directement accessibles via une base de sondage.

- Enquêtes ménages : besoin d'estimations au niveau ménage et individuel. On procède par l'intermédiaire de la sélection d'un échantillon de logements.
- Enquête MORGOAT : enquête sur le tourisme en Bretagne conduite par l'Observatoire Régional du Tourisme de Bretagne (Deville et Maumy, 2006).

Il peut être plus pratique d'utiliser une population intermédiaire pour laquelle une base de sondage est disponible : on parle de méthodes **d'échantillonnage indirect**.

Exemple 1 (Lavallée, 2007, page 8)



Hypothèses

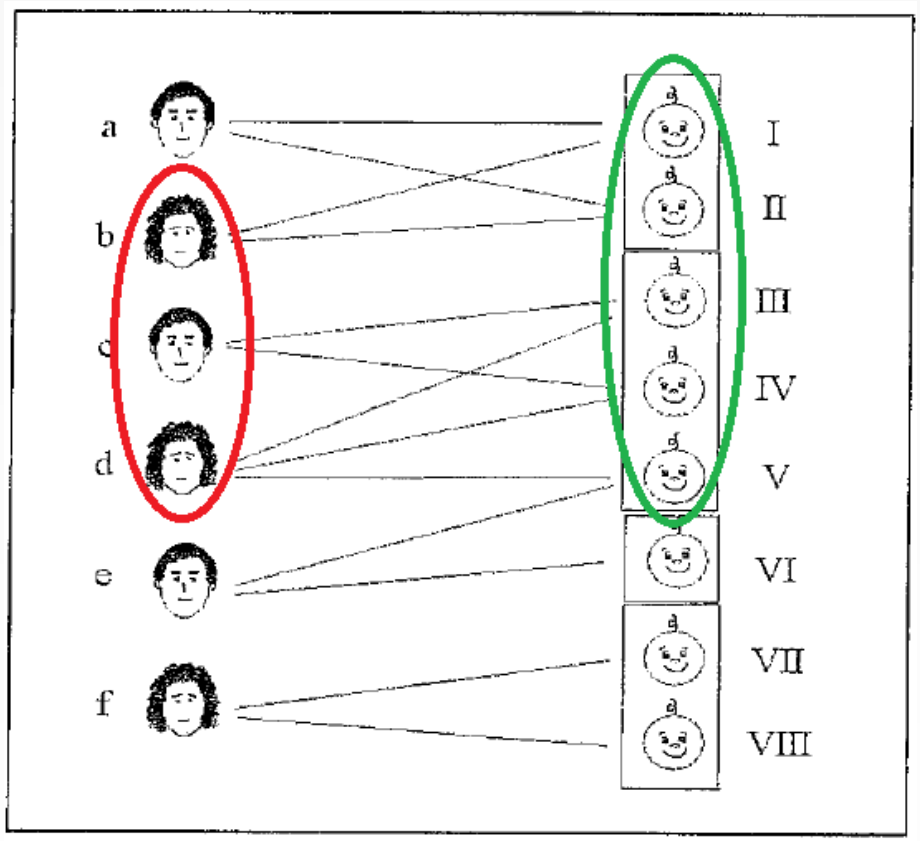
Nous nous intéressons à une population cible U pour laquelle une base de sondage n'est pas disponible.

Les unités de U sont **liées** aux unités d'une autre population U_A. **Nous supposons que chaque unité $k \in U$ possède au moins un lien avec une unité $i \in U_A$.**
Nous notons $L_{ik} = 1$ si les unités $i \in U_A$ et $k \in U$ sont liées, et $L_{ik} = 0$ sinon.

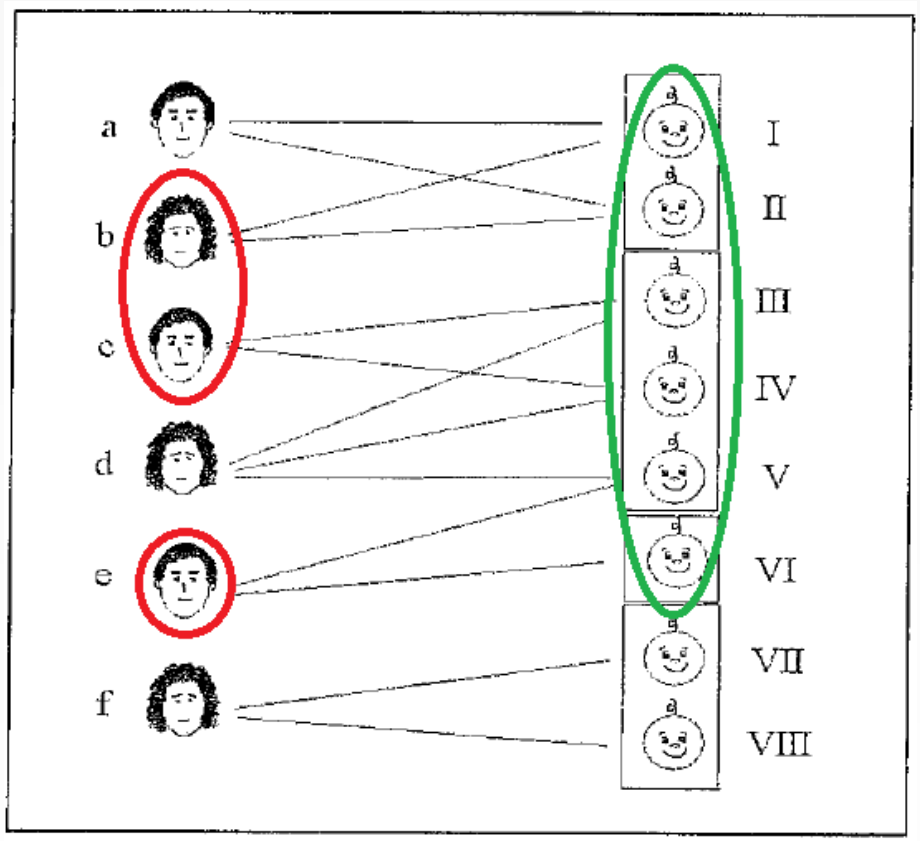
Une liste des unités statistiques est disponible pour U_A. Un échantillon S_A est sélectionné dans U_A, et nous enquêtons dans U toutes les unités liées à au moins une unité de S_A, i.e.

$$S = \{k \in U;\ \exists i \in S_A \text{ tel que } L_{ik} = 1\}.$$

Exemple 1 : un premier échantillon



Exemple 1 : un second échantillon



Dualité entre U_A et U

La méthode de partage des poids est basée sur un principe de dualité entre les deux populations.

Soit

$$N_{+k} = \sum_{i \in U_A} L_{ik}$$

le nombre total de liens entre l'unité $k \in U$ et U_A. Alors si $N_{+k} > 0$ pour chaque $k \in U$, le total $Y = \sum_{k \in U} y_k$ peut se réécrire :

$$\begin{aligned} Y &= \sum_{i \in U_A} z_i \quad \text{avec} \quad z_i = \sum_{k \in U} \frac{y_k L_{ik}}{N_{+k}} \\ &\equiv Z_A. \end{aligned}$$

Interprétation : chaque $y_k,\ k \in U$ est réparti à parts égales entre toutes les unités de U_A liées à k.

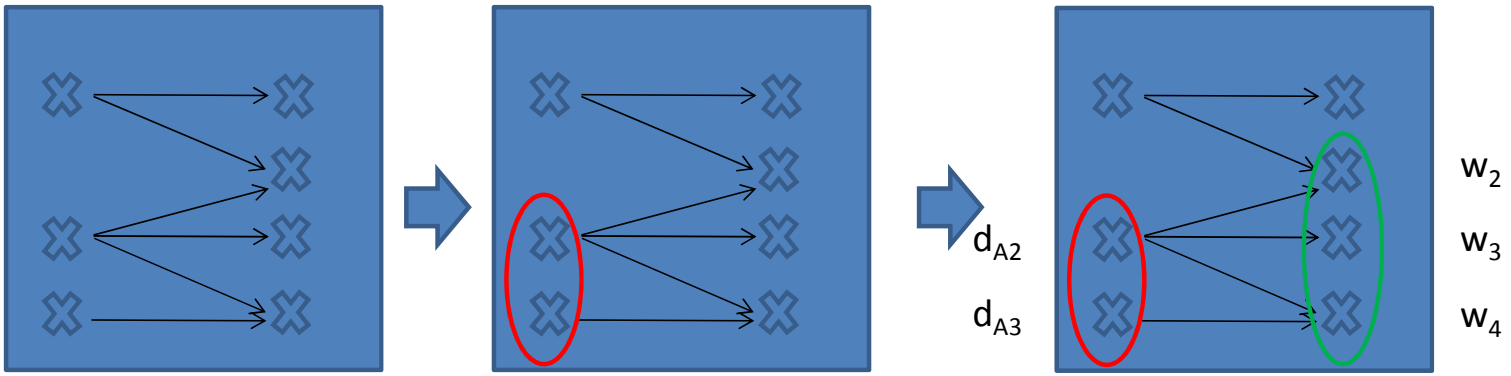
Estimateur par partage des poids

Comme Y peut être vu comme un total sur la population U_A, il est possible d'utiliser l'estimateur de Horvitz-Thompson sur S_A. Nous obtenons :

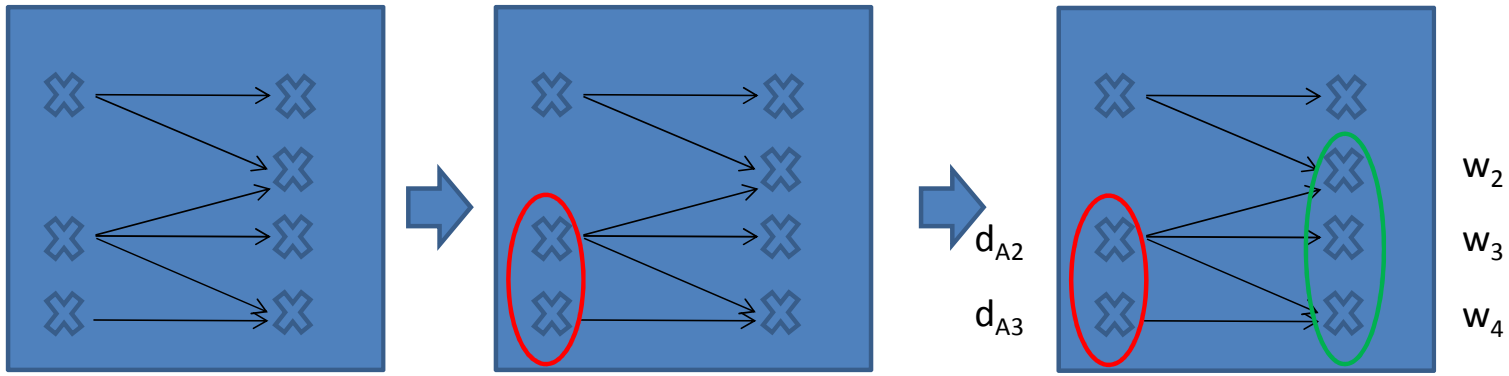
$$\begin{aligned}\hat{Y} &= \sum_{i\in S_A} d_{Ai} z_i = \hat{Z}_{A\pi} \\ &= \sum_{i\in S_A} d_{Ai} \sum_{k\in S} \frac{y_k L_{ik}}{N_{+k}} \\ &= \sum_{k\in S} w_k y_k \text{ avec } w_k = \frac{1}{N_{+k}} \sum_{i\in S_A} d_{Ai} L_{ik}.\end{aligned}$$

Le poids w_k de l'unité $k \in S$ est donné par la somme des poids des unités $i \in S_A$ liées à k, divisée par le nombre **total** de liens N_{+k}. **Cette information doit être collectée pendant l'enquête**.

Exemple 2



Exemple 2

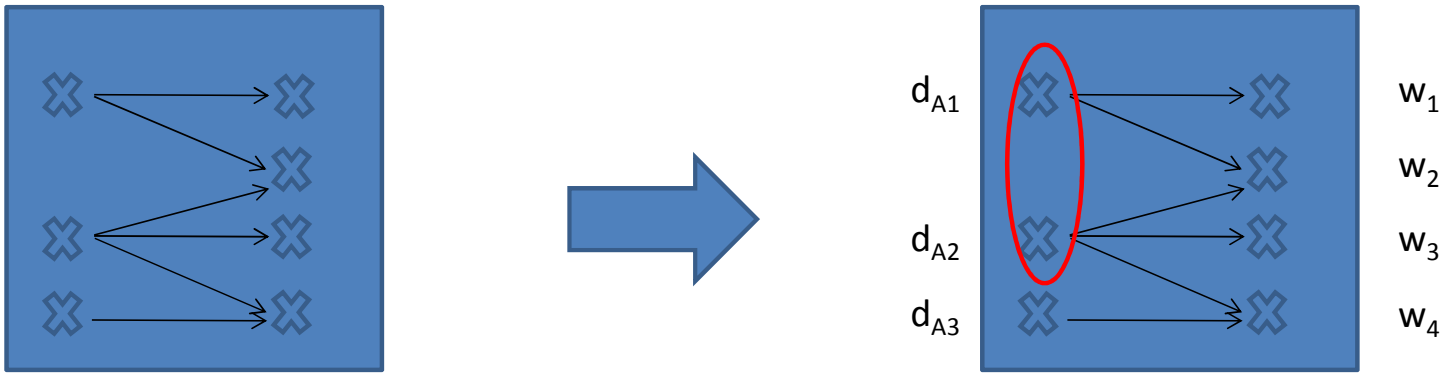


$$w_2 = \frac{d_{A2}}{2}$$

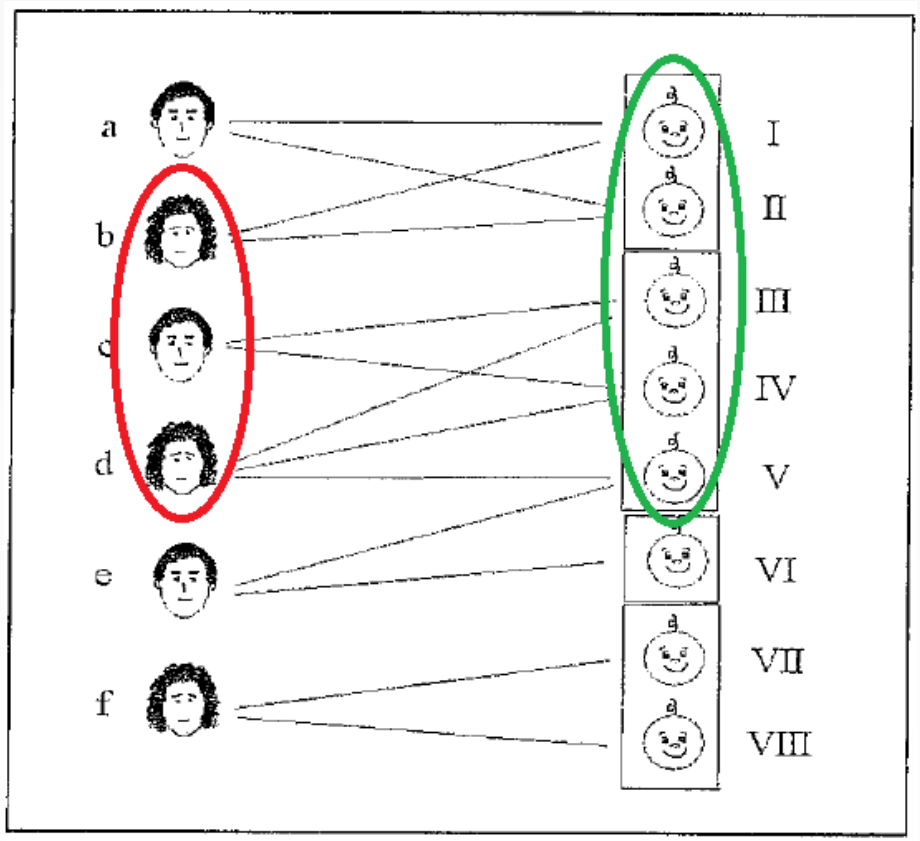
$$w_3 = \frac{d_{A2}}{1}$$

$$w_4 = \frac{d_{A2} + d_{A3}}{2}$$

Exemple 2 : à compléter



Exemple 1 : à compléter



Estimation de variance

La dualité entre U et U_A permet de produire facilement un estimateur de variance, en utilisant les propriétés du plan de sondage utilisé dans U_A.

Nous prenons

$$\begin{aligned} v\left(\hat{Y}\right) &= v_{HT}\left(\hat{Z}_{A\pi}\right) \\ &= \sum_{i,i' \in S_A} \frac{z_i}{\pi_{Ai}} \frac{z_{i'}}{\pi_{Ai'}} \frac{\Delta_{Aii'}}{\pi_{Aii'}}. \end{aligned}$$

Il est sans biais si tous les $\pi_{Aii'}$ sont strictement positifs.

Pour calculer cet estimateur de variance :

1. Calcul de la variable synthétique z_i permettant d'écrire $\hat{Y}$ comme un estimateur de Horvitz-Thompson dans U_A.
2. Utilisation de l'estimateur de variance correspondant au plan de sondage utilisé dans U_A.

Exercice 3

Estimateur direct de Horvitz-Thompson

En principe, il est possible de calculer les probabilités d'inclusion $\pi_k = Pr(k \in S)$ pour utiliser l'estimateur direct de Horvitz-Thompson sur S

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

En pratique, le calcul est difficile et souvent impossible. Pour chaque $k \in U$, soit G_k, de taille m, l'ensemble des unités de U_A qui sont liées à k. Nous avons

$$\begin{aligned} \pi_k &= Pr(\exists i \in S_A \quad L_{ik} = 1) \\ &= Pr\left(\cup_{i \in G_k} \{i \in S_A\}\right) \\ &= \sum_{a=1}^{m} (-1)^{a+1} \sum_{i_1 < \ldots < i_k \in G_k} Pr(\cap_{b=1}^{k} I_{Ai_b}). \end{aligned}$$

Méthode de partage des poids généralisée

La méthode peut être généralisée en utilisant des poids $\theta_{ik} > 0$ pour les liens. Nous pouvons réécrire

$$Y = \sum_{i \in U_A} z_{\theta i} \quad \text{où} \quad z_{\theta i} = \sum_{k \in U} \frac{y_k \theta_{ik} L_{ik}}{N^{\theta}_{+k}}$$

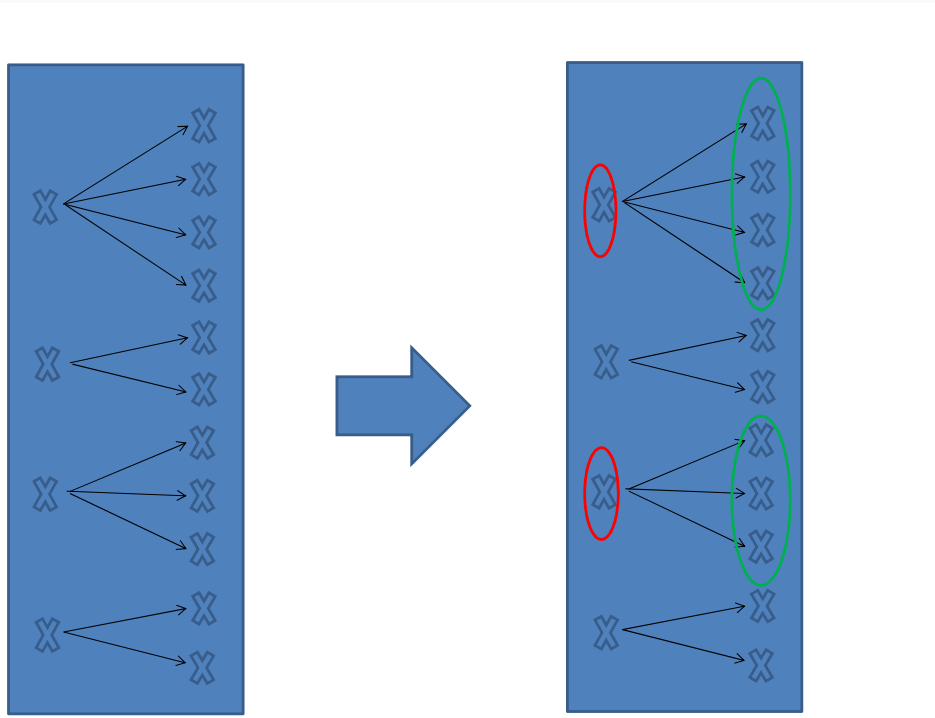
où $N^{\theta}_{+k} = \sum_{i \in U_A} \theta_{ik} L_{ik}$. Cela conduit à l'estimateur

$$\hat{Y}_{\theta} = \sum_{k \in S} w^{\theta}_k y_k \text{ où } w^{\theta}_k = \frac{1}{N^{\theta}_{+k}} \sum_{i \in S_A} d_{Ai} \theta_{ik} L_{ik}.$$

Les poids θ_{ik} peuvent être choisis de façon à réduire la variance (Deville et Lavallée, 2006).
Dans la pratique, des poids égaux $\theta_{ik} = 1$ sont souvent utilisés.

Exemple 3 : échantillonnage par grappes

Dans le cas particulier où chaque unité $k \in U$ est liée à une unité seulement $i \in U_A$, nous obtenons un échantillonnage par grappes.
C'est par exemple le cas si nous sélectionnons un échantillon de ménages (population U_A) dont tous les individus sont enquêtés (population U).



Exemple 3 : échantillonnage par grappes

Dans le cas d'un tirage par grappes, nous obtenons :

$$Y = \sum_{i \in U_A} z_i \quad \text{où} \quad z_i = \sum_{k \in U} \frac{y_k L_{ik}}{N_{+k}} = \sum_{k \in u_i} y_k.$$

La variable synthétique z_i correspondant au sous-total des valeurs y_k pour les unités $k \in u_i$.

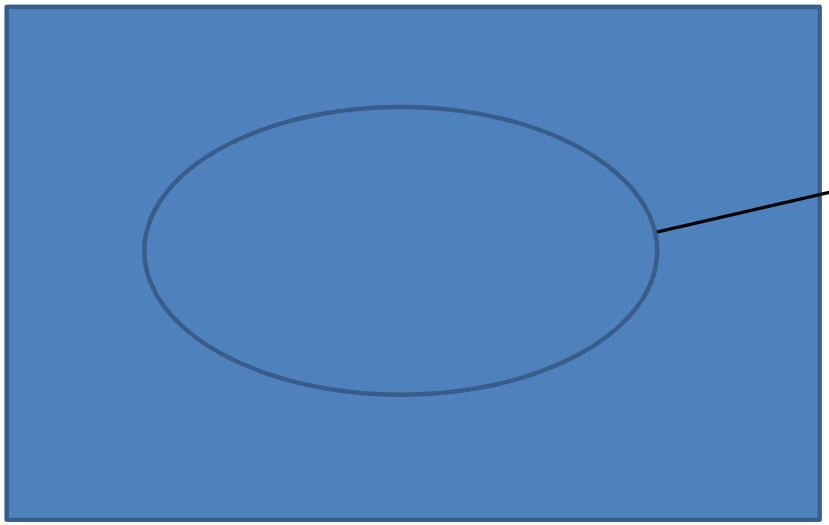
L'estimateur du total vaut

$$\hat{Y} = \sum_{i \in S_A} d_{Ai} z_i \quad \text{où} \quad z_i = \sum_{k \in u_i} y_k,$$

ce qui correspond à l'estimateur de Horvitz-Thompson pour un tirage par grappes.

Estimation dans le temps

Population initiale



Population à la date t : $U(t)$

Estimation initiale

Nous considérons à la date initiale t une population finie $U(t)$ d'unités statistiques, de taille N^t. Une variable d'intérêt y^t prend la valeur y_k^t pour chaque unité $k \in U(t)$.

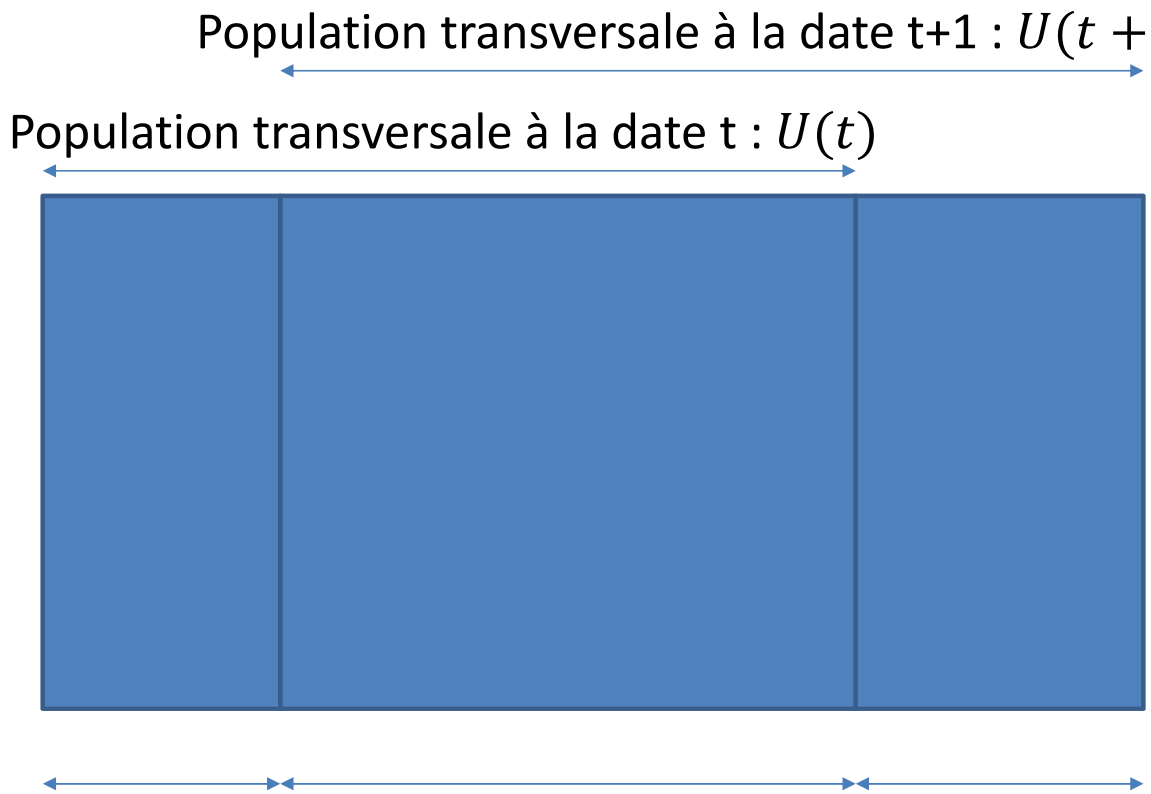
Nous sélectionnons un échantillon S_t, avec des probabilités d'inclusion d'ordre 1 $\pi_k^t > 0$ (pas de biais de couverture).

Le total $Y(t) = \sum_{k \in U(t)} y_k^t$ peut être estimé sans biais par l'estimateur de Horvitz-Thompson

$$\hat{Y}_t(t) = \sum_{k \in S_t} \frac{y_k^t}{\pi_k^t} = \sum_{k \in S_t} d_k^t y_k^t,$$

en notant $d_k^t = 1/\pi_k^t$ le poids de sondage à la date t.

Population au temps suivant

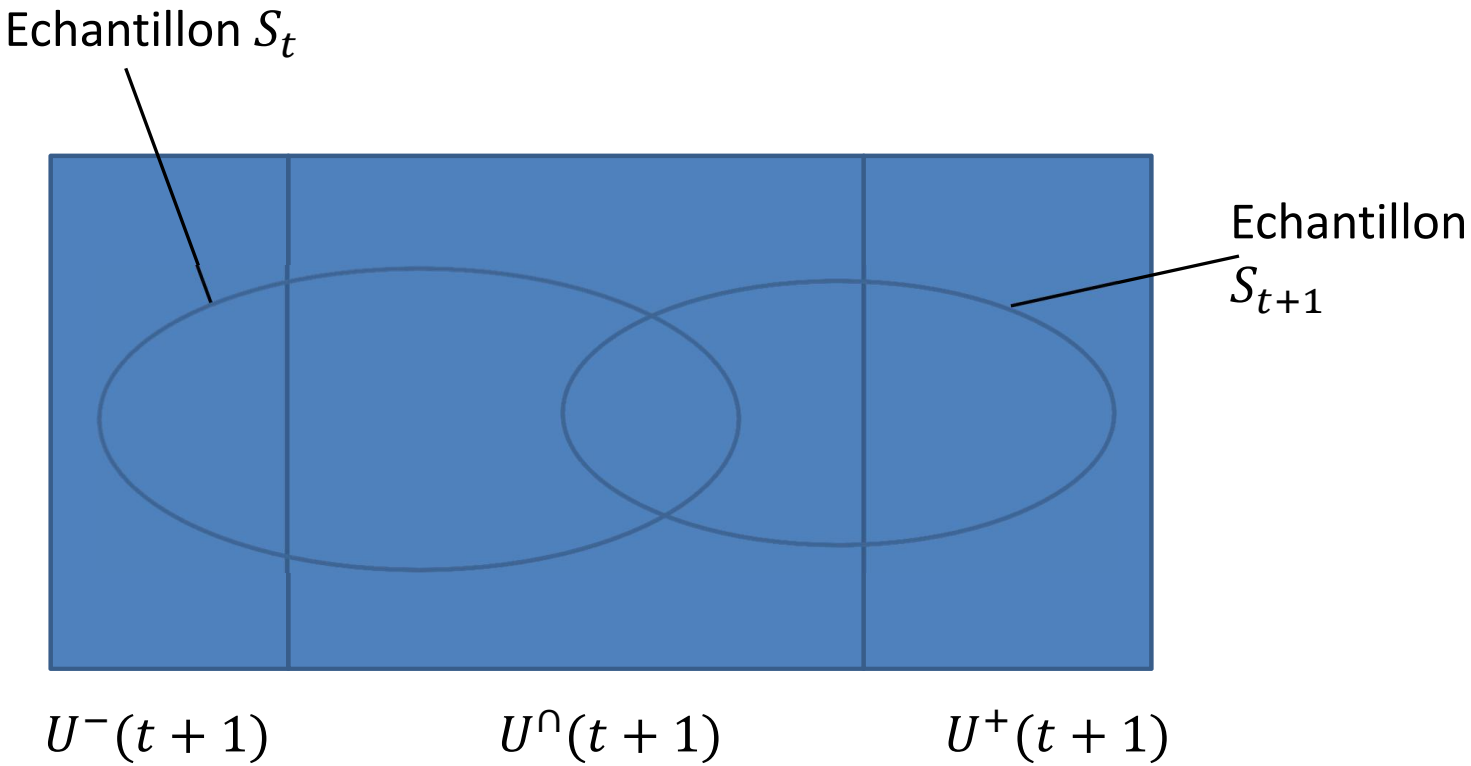


$U^{-}(t+1)$: Décès entre les dates t et t+1

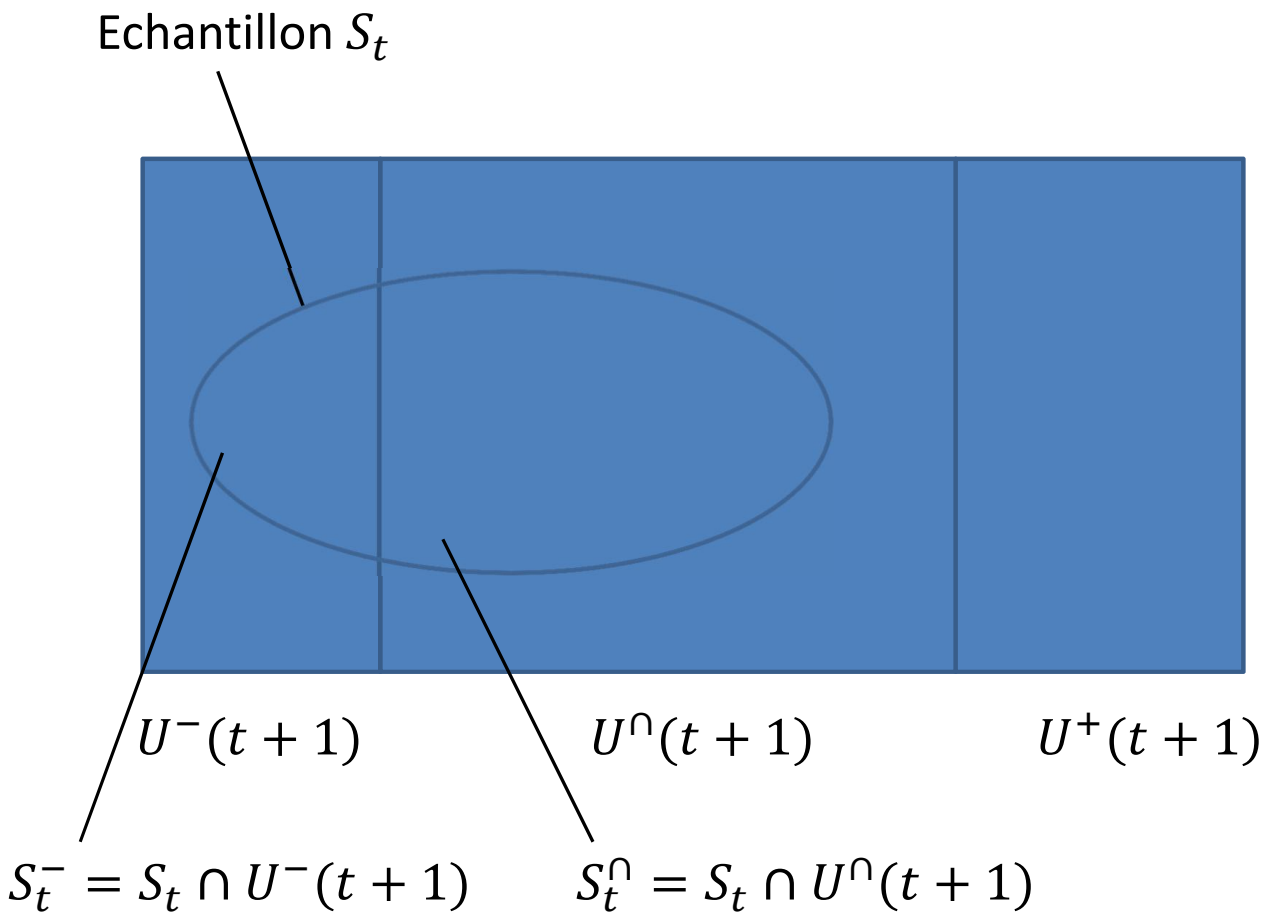
$U^{\cap}(t+1)$: Individus communs aux dates t et t+1

$U^{+}(t+1)$: Naissances entre les dates t et t+1

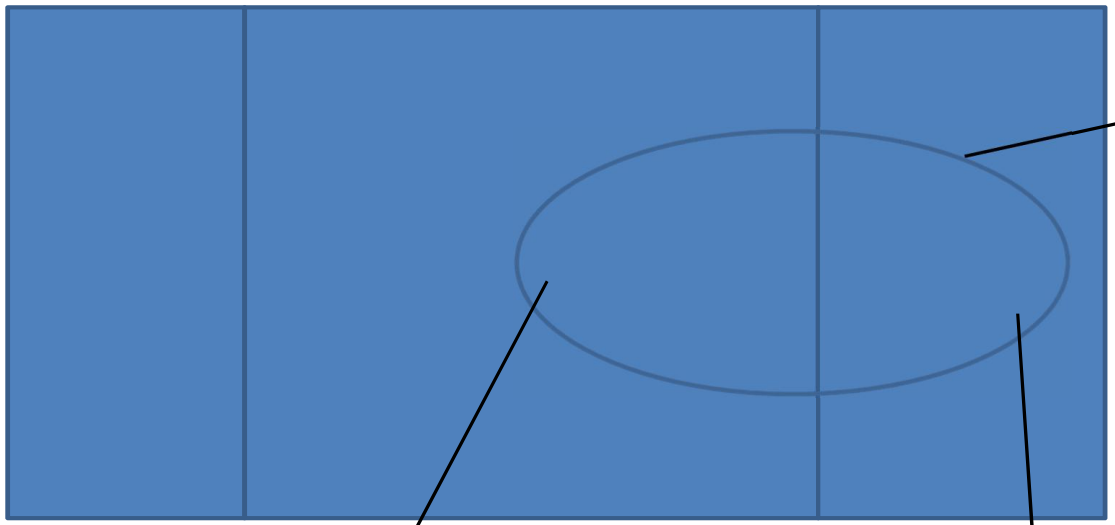
Population au temps suivant (2)



Population au temps suivant (3)



Population au temps suivant (4)



Estimation transversale

Estimation d'un paramètre (total, ratio, ...) à une date donnée. En règle générale, l'estimation va porter sur tous les individus présents à cette date.

Nous pouvons nous intéresser à l'estimation du total

$$Y(t) = \sum_{k \in U(t)} y_k^t \qquad \text{à la date t,}$$

$$Y(t+1) = \sum_{k \in U(t+1)} y_k^{t+1} \qquad \text{à la date t+1.}$$

Dans certains cas (e.g., enquêtes par panel), nous pouvons nous intéresser à la date t+1 à une estimation sur la population survivante $U^{\cap}(t+1)$:

$$Y^{\cap}(t+1) = \sum_{k \in U^{\cap}(t+1)} y_k^{t+1} \qquad \text{à la date t+1.}$$

Estimation longitudinale

Estimation de l'évolution d'un paramètre dans le temps. Nous nous limiterons ici au cas de l'évolution d'un total.

La population d'intérêt peut être définie :

- comme la réunion des populations transversales, que l'on cherche à comparer. Nous souhaitons alors estimer

$$\Delta = Y(t+1) - Y(t).$$

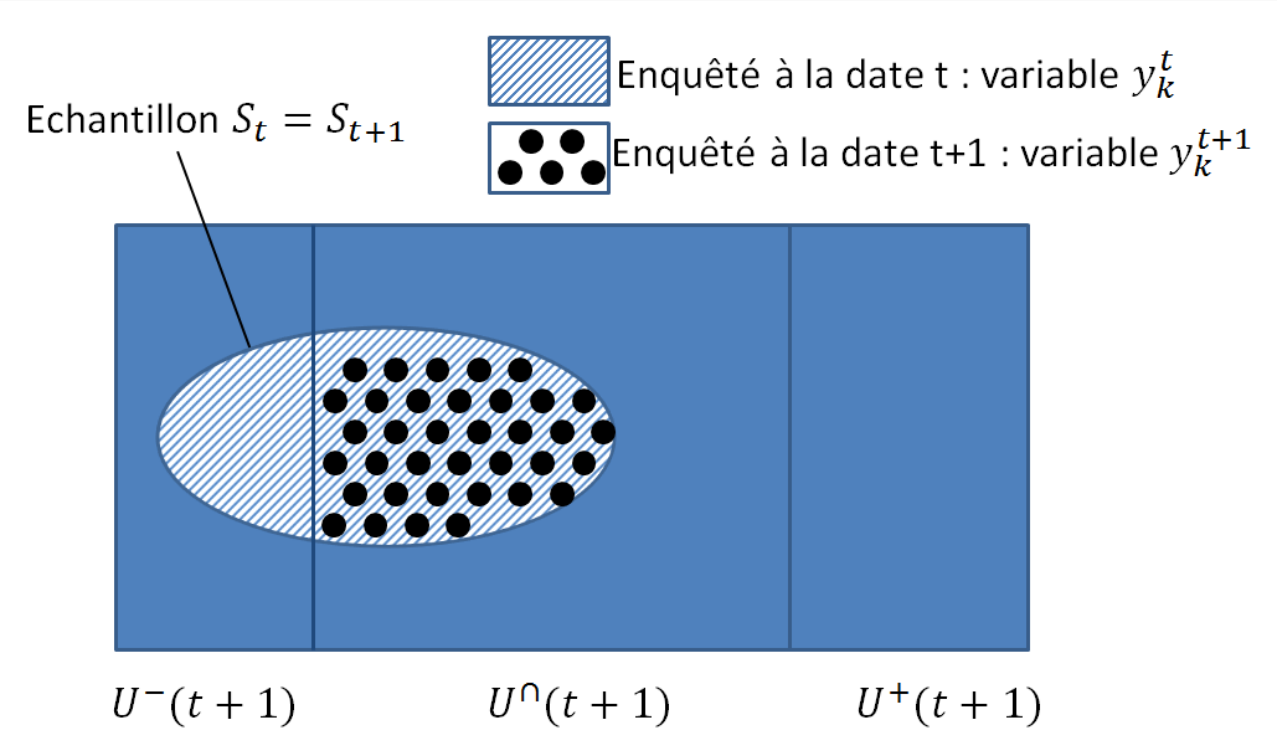
- comme l'intersection des populations transversales, i.e. les individus communs aux deux dates. Nous souhaitons alors estimer

$$\begin{aligned} \Delta^{\cap} &= \sum_{k \in U^{\cap}(t+1)} (y_k^{t+1} - y_k^t) \\ &= Y^{\cap}(t+1) - Y^{\cap}(t). \end{aligned}$$

Enquêtes par panel

Principe

Dans les **enquêtes par panel** ("panel surveys"), les mêmes mesures sont effectuées à différents temps sur le même échantillon.
C'est un plan de sondage parfaitement adapté pour une étude longitudinale individuelle.



Estimation transversale

Estimation transversale

Au temps t, nous avons

$$\hat{Y}_t(t) = \sum_{k \in S_t} d_k^t y_k^t \quad \text{avec} \quad V\{\hat{Y}_t(t)\} = \sum_{k,l \in U(t)} \Delta_{kl}^t (d_k^t y_k^t)(d_l^t y_l^t).$$

Au temps $t+1$:

- pas d'estimation transversale possible sur $U(t+1)$,
- estimation possible uniquement sur la population commune $U^{\cap}(t+1)$.

Nous avons

$$\begin{aligned} Y^{\cap}(t+1) &= \sum_{k \in U^{\cap}(t+1)} y_k^{t+1}, \\ \hat{Y}_t^{\cap}(t+1) &= \sum_{k \in S_t^{\cap}} d_k^t y_k^{t+1} \quad \text{avec} \quad S_t^{\cap} = S_t \cap U^{\cap}(t+1). \end{aligned}$$

Estimation transversale (2)

Dans le cas particulier d'un panel sélectionné selon un SRS, nous obtenons au temps $t+1$:

$$\hat{Y}_t^{\cap}(t+1) = \frac{N^t}{n^t} \sum_{k \in S_t^{\cap}} y_k^{t+1},$$

$$V\left\{\hat{Y}_t^{\cap}(t+1)\right\} \simeq (N^{\cap})^2 \left(\frac{1}{\bar{n}^{\cap}} - \frac{1}{N^{\cap}}\right) \left\{S_{\cap}^2 + \left(1 - \frac{N^{\cap}}{N^t}\right) (\mu_{\cap})^2\right\},$$

avec :

- $N^{\cap}$: taille de la pop. intersection $U^{\cap}(t+1)$,
- $\bar{n}^{\cap}$: taille moyenne de $S_t^{\cap}$,
- $\mu_{\cap} = \frac{1}{N^{\cap}} \sum_{k \in U^{\cap}(t+1)} y_k^{t+1}$: moyenne de y_k^{t+1} dans $U^{\cap}(t+1)$,
- $S_{\cap}^2 = \frac{1}{N^{\cap} - 1} \sum_{k \in U^{\cap}(t+1)} (y_k^{t+1} - \mu_{\cap})^2$: disp. de y_k^{t+1} dans $U^{\cap}(t+1)$.

Estimation transversale (3)

Dans le cas particulier d'un panel sélectionné selon un SRS, nous obtenons au temps $t+1$:

$$\hat{Y}_t^{\cap}(t+1) = \frac{N^t}{n^t} \sum_{k \in S_t^{\cap}} y_k^{t+1},$$

$$V\left\{\hat{Y}_t^{\cap}(t+1)\right\} \simeq (N^{\cap})^2 \left(\frac{1}{\bar{n}^{\cap}} - \frac{1}{N^{\cap}}\right) \left\{S_{\cap}^2 + \left(1 - \frac{N^{\cap}}{N^t}\right) (\mu_{\cap})^2\right\}.$$

La variance est celle que l'on obtiendrait avec un sondage aléatoire simple direct de taille $\bar{n}^{\cap}$ dans $U^{\cap}(t+1)$, avec un terme additionnel qui augmente avec le temps.

Cette augmentation est liée à la diminution de la taille de la population-cible $U^{\cap}(t+1)$ par rapport à la taille de la population d'origine $U(t)$.

Estimation longitudinale

Estimation longitudinale

L'estimation longitudinale ne peut porter que sur l'intersection $U^{\cap}(t+1)$ des populations transversales. Nous nous intéressons donc au paramètre

$$\Delta^{\cap} = \sum_{k \in U^{\cap}(t+1)} (y_k^{t+1} - y_k^t),$$

que nous estimons

$$\hat{\Delta}_t^{\cap} = \sum_{k \in S_t^{\cap}} d_k^t (y_k^{t+1} - y_k^t) \text{ où } S_t^{\cap} = S_t \cap U^{\cap}(t+1),$$

Dans le cas particulier d'un SRS au temps t, nous obtenons :

$$\hat{\Delta}_t^{\cap} = \frac{N^t}{n^t} \sum_{k \in S_t^{\cap}} (y_k^{t+1} - y_k^t).$$

Traitement de la non-réponse

En pratique, nous observons un phénomène d'**attrition** entre les temps t et $t+1$: seule une partie de l'échantillon d'origine S_t, noté S_{rt}, peut être effectivement suivi et enquêté.

Nous avons recours à une modélisation de la probabilité de réponse, notée $p_k^{t \to t+1}$, par exemple selon la méthode des GHR. Nous obtenons l'estimateur corrigé de l'attrition :

$$\hat{\Delta}_{rt}^{\cap} = \sum_{k \in S_{rt}^{\cap}} \frac{d_k^t}{\hat{p}_k^{t \to t+1}} (y_k^{t+1} - y_k^t).$$

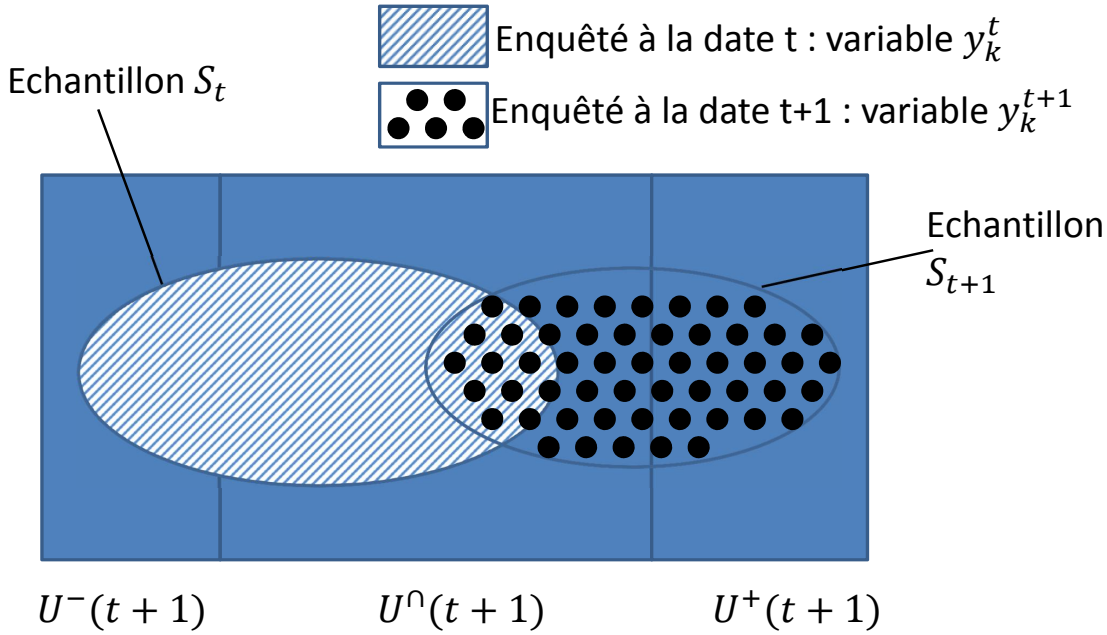
Il est généralement difficile d'obtenir de l'information auxiliaire sur la population intersection $U^{\cap}(t+1)$. Du coup, il n'est généralement pas possible de caler cet estimateur afin de diminuer sa variance.

Exercice 4

Enquêtes transversales répétées

Principe

Dans les **enquêtes transversales répétées** ("repeated cross-sectional survey"), nous sélectionnons à des dates différentes des échantillons indépendants dans une même population (modulo les naissances et les décès).



Estimation transversale

Estimation transversale

Les estimations transversales sont simplement basées sur les échantillons sélectionnés aux temps correspondants. Nous avons donc :

$$\begin{aligned} \hat{Y}_t(t) &= \sum_{k \in S_t} d_k^t y_k^t, \\ \hat{Y}_{t+1}(t+1) &= \sum_{k \in S_{t+1}} d_k^{t+1} y_k^{t+1}. \end{aligned}$$

Dans le cas particulier d'un SRS à chaque temps, nous obtenons :

$$\hat{Y}_t(t) = N^t \bar{y}^t \quad \text{avec} \quad V\{\hat{Y}_t(t)\} = (N^t)^2 \left(\frac{1}{n^t} - \frac{1}{N^t} \right) S_{yt}^2,$$

$$\hat{Y}_{t+1}(t+1) = N^{t+1} \bar{y}^{t+1} \quad \text{avec} \quad V\{\hat{Y}_{t+1}(t+1)\} = (N^{t+1})^2 \left(\frac{1}{n^{t+1}} - \frac{1}{N^{t+1}} \right) S_{y,t+1}^2.$$

Traitement de la non-réponse

En pratique, nous observons de la **non-réponse totale** à chaque temps : seule une partie de l'échantillon S_t (notée S_{rt}) et de l'échantillon S_{t+1} (notée $S_{r,t+1}$) peuvent être effectivement enquêtés.

Nous avons recours à une modélisation des probabilités de réponse aux temps t et $t+1$, notées p_k^t et p_k^{t+1}. Nous obtenons les estimateurs corrigés de la non-réponse totale :

$$\begin{aligned} \hat{Y}_{rt}(t) &= \sum_{k \in S_{rt}} \frac{d_k^t}{\hat{p}_k^t} y_k^t, \\ \hat{Y}_{r,t+1}(t+1) &= \sum_{k \in S_{r,t+1}} \frac{d_k^{t+1}}{\hat{p}_k^{t+1}} y_k^{t+1}. \end{aligned}$$

Ces estimateurs peuvent ensuite être calés sur de l'information auxiliaire connue sur la population-cible, afin de diminuer la variance.

Estimation longitudinale

Estimation longitudinale

Pour une estimation longitudinale sur la réunion des populations transversales, nous nous intéressonse au paramètre $\Delta = Y(t+1) - Y(t)$. Nous utilisons l'estimateur

$$\begin{aligned} \hat{\Delta} &= \hat{Y}_{t+1}(t+1) - \hat{Y}_t(t) \\ &= \sum_{k \in S_{t+1}} d_k^{t+1} y_k^{t+1} - \sum_{k \in S_t} d_k^t y_k^t, \\ \text{avec } V(\hat{\Delta}) &= V\{\hat{Y}_t(t)\} + V\{\hat{Y}_{t+1}(t+1)\}. \end{aligned}$$

Dans le cas particulier d'un SRS à chaque temps, nous obtenons :

$$\begin{aligned} \hat{\Delta} &= N^{t+1}\bar{y}^{t+1} - N^t\bar{y}^t, \\ \text{avec } V(\hat{\Delta}) &= (N^t)^2 \left(\frac{1}{n^t} - \frac{1}{N^t}\right) S_{yt}^2 \\ &+ (N^{t+1})^2 \left(\frac{1}{n^{t+1}} - \frac{1}{N^{t+1}}\right) S_{y,t+1}^2. \end{aligned}$$

Estimation longitudinale (2)

Pour une estimation longitudinale sur l'intersection des populations transversales, nous nous intéressonse au paramètre $\Delta^{\cap} = Y^{\cap}(t+1) - Y^{\cap}(t)$. Nous utilisons l'estimateur

$$\begin{aligned} \hat{\Delta}^{\cap} &= \hat{Y}^{\cap}_{t+1}(t+1) - \hat{Y}^{\cap}_{t}(t) \\ &= \sum_{k \in S^{\cap}_{t+1}} d_k^{t+1} y_k^{t+1} - \sum_{k \in S^{\cap}_{t}} d_k^{t} y_k^{t}, \\ \text{avec } V(\hat{\Delta}^{\cap}) &= V\{\hat{Y}^{\cap}_{t}(t)\} + V\{\hat{Y}^{\cap}_{t+1}(t+1)\}. \end{aligned}$$

Dans le cas particulier d'un SRS à chaque temps, nous obtenons :

$$\hat{\Delta}^{\cap} = \frac{N^{t+1}}{n^{t+1}} \sum_{k \in S^{\cap}_{t+1}} y_k^{t+1} - \frac{N^{t}}{n^{t}} \sum_{k \in S^{\cap}_{t}} y_k^{t}.$$

Là aussi, la variance de $\hat{\Delta}^{\cap}$ s'obtient en additionnant celles des deux composantes $\hat{Y}^{\cap}_{t+1}(t+1)$ et $\hat{Y}^{\cap}_{t}(t)$.

Comparaison panel - enquêtes transversales répétées

Comparaison

Dans le cas d'une estimation longitudinale sur la population intersection, nous obtenons avec des enquêtes transversales répétées l'estimateur

$$\hat{\Delta}^{\cap} = \sum_{k \in S^{\cap}_{t+1}} d_k^{t+1} y_k^{t+1} - \sum_{k \in S^{\cap}_{t}} d_k^{t} y_k^{t},$$

et avec un panel

$$\hat{\Delta}^{\cap} = \sum_{k \in S^{\cap}_{t}} d_k^{t} (y_k^{t+1} - y_k^{t}).$$

La stratégie de panel va conduire à une variance beaucoup plus faible si les variables y^t et y^{t+1} sont positivement corrélées, ce qui est souvent le cas pour des variables décalées dans le temps.

Etude par simulations

Pour comparer les deux stratégies, nous réalisons une étude par simulations. Nous générons :

- une pop. $U^-(t+1)$ de $20\ 000$ individus (décès $t \to t+1$),
- une pop. $U^\cap(t+1)$ de $80\ 000$ individus (indiv. communs),
- une pop. $U^+(t+1)$ de $20\ 000$ individus (naissances $t \to t+1$).

Nous générons deux variables auxiliaires x_1 et x_2 selon des lois gamma de paramètres 2 et 5, et deux variables d'intérêt :

$$\begin{aligned} y_k^1 &= 100 + x_{1k} + x_{2k} + \sigma u_{1k}, \quad (R^2 = 0.70) \\ y_k^2 &= \alpha y_k^1 + \sigma u_{2k}. \end{aligned}$$

Le coefficient α est choisi de façon à avoir un coefficient de corrélation linéaire entre y^1 et y^2 égal à 0.7, 0.8 ou 0.9.

Etude par simulations

Nous appliquons deux stratégies d'échantillonnage :

- Enquêtes transversales répétées :
  - échantillon S_t de taille n sélectionné dans $U(t)$, sur lequel y_k^t est observé,
  - échantillon S_{t+1} de taille n sélectionné dans $U(t+1)$, sur lequel y_k^{t+1} est observé.
- Panel : échantillon S_t de taille n sélectionné dans $U(t)$, sur lequel y_k^t et y_k^{t+1} sont observées.

Nous sélectionnons les échantillons selon un SRS avec $n = 500$, $1\ 000$ ou $2\ 000$. Nous nous intéressons à l'estimation de $\Delta^{\cap}$.

Nous comparons les deux stratégies d'échantillonnage en calculant

$$RE = \frac{Var_{pan}(\hat{\Delta}^{\cap})}{Var_{etr}(\hat{\Delta}^{\cap})}.$$

Résultats pour l'efficacité relative

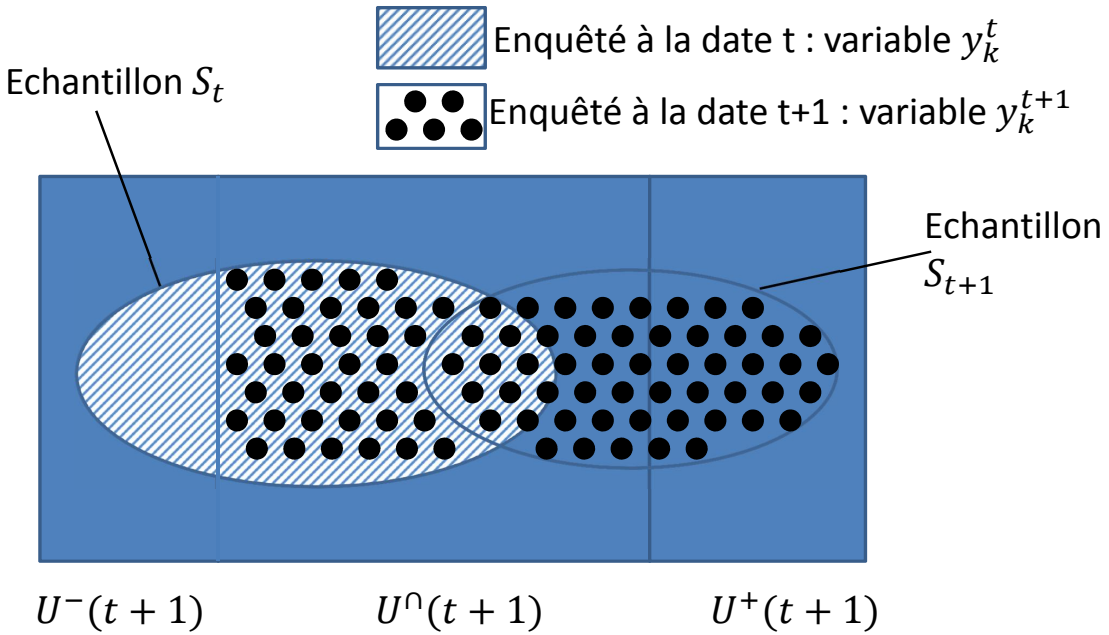
	Coef. de corrélation ρ^2		
	0.7	0.8	0.9
Taille d'échantillon n			
500	0.14	0.05	0.01
1 000	0.14	0.05	0.01
2 000	0.16	0.05	0.01

Nous constatons comme prévu que la stratégie de panel est beaucoup plus efficace pour une estimation longitudinale. Le gain est d'autant plus fort que la corrélation entre les variables y^t et y^{t+1} est forte.

Enquêtes à échantillon partagé

Principe

Dans les **enquêtes à échantillon partagé** ("split panel surveys"), nous utilisons un panel qui est complété à chaque date par un nouvel échantillon indépendant. Ce type d'enquête est conçu pour permettre des estimations transversales et longitudinales.



Estimation longitudinale

Estimation longitudinale

L'estimation longitudinale sur l'intersection $U^{\cap}(t+1)$ des populations transversales se fait comme pour une enquête par panel. Le paramètre

$$\Delta^{\cap} = \sum_{k \in U^{\cap}(t+1)} (y_k^{t+1} - y_k^t),$$

est estimé par

$$\hat{\Delta}_t^{\cap} = \sum_{k \in S_t^{\cap}} d_k^t (y_k^{t+1} - y_k^t) \text{ où } S_t^{\cap} = S_t \cap U^{\cap}(t+1).$$

Nous pouvons également nous intéresser à une estimation sur la réunion des population transversales (non considéré ici).

Traitement de la non-réponse

En pratique, nous observons un phénomène d'**attrition** entre les temps t et $t+1$: seule une partie de l'échantillon d'origine S_t, noté S_{rt}, peut être effectivement suivi et enquêté.

Nous avons recours à une modélisation de la probabilité de réponse, notée $p_k^{t \to t+1}$, par exemple selon la méthode des GHR. Nous obtenons l'estimateur corrigé de l'attrition :

$$\hat{\Delta}_{rt}^{\cap} = \sum_{k \in S_{rt}^{\cap}} \frac{d_k^t}{\hat{p}_k^{t \to t+1}} (y_k^{t+1} - y_k^t).$$

Il est généralement difficile d'obtenir de l'information auxiliaire sur la population intersection $U^{\cap}(t+1)$. Du coup, il n'est généralement pas possible de caler cet estimateur afin de diminuer sa variance.

Estimation transversale

Estimation

Pour une estimation transversale au temps t, nous avons

$$\hat{Y}_t(t) = \sum_{k \in S_t} d_k^t y_k^t \quad \text{avec} \quad V\{\hat{Y}_t(t)\} = \sum_{k,l \in U(t)} \Delta_{kl}^t (d_k^t y_k^t)(d_l^t y_l^t).$$

Pour une estimation transversale au temps $t+1$, les deux échantillons peuvent être utilisés :

- l'échantillon S_{t+1} permet de représenter toute la population $U(t+1)$,
- l'échantillon S_t permet de représenter la sous-population $U^{\cap}(t+1)$.

Nous considèrerons deux cas :

- celui où les unités d'échantillonnage sont les unités d'observation,
- celui où les unités d'échantillonnage (=ménages) ne sont pas les unités d'observation (=individus), qui est plus délicat pour le calcul des pondérations.

Cas où les unités d'échantillonnage sont les unités d'observation

Estimation au temps $t+1$

Afin de mettre en commun les échantillons au temps $t+1$ en évitant les problèmes de double compte, méthode d'estimation composite :

- Nous séparons la population entre $U^{+}(t+1)$ (représentée par un seul échantillon) et $U^{\cap}(t+1)$ (représentée par deux échantillons).
- Les individus de S_{t+1}^{+} conservent leur pondération.
- Nous réalisons une combinaison des pondérations pour les individus de $S_t^{\cap}$ et de $S_{t+1}^{\cap}$.

En résumé, pour les individus de l'échantillon réunion $S_t^{\cap} \cup S_{t+1}$:

$$d_{\theta k} = \begin{cases} \theta d_k^t & \text{si } k \in S_t^{\cap}, \\ (1-\theta) d_k^{t+1} & \text{si } k \in S_{t+1}^{\cap}, \\ d_k^{t+1} & \text{si } k \in S_{t+1}^{+}. \end{cases} \tag{3}$$

Choix du paramètre θ

Le choix $\theta = 0$ conduit à n'utiliser que l'échantillon S_{t+1}.

Le choix $\theta = 0.5$ est un choix assez courant en pratique, et conduit à donner aux échantillons S_t et S_{t+1} le même poids sur la sous-population $U_{t+1}^{\cap}$.

Il est également possible de chercher la valeur de α qui minimise la variance de l'estimateur composite (estimateur de Hartley).

Exercice 5

Cas des enquêtes ménage

Notations

On sélectionne initialement un échantillon S_t de ménages/individus :

- Un échantillon $S_{t,men}$ est sélectionné dans $U_{men}(t)$. Soit d_k^t le poids associé au ménage k.
- Un échantillon $S_{t,ind}$ d'individus est obtenu en interrogeant tous les individes des ménages de $S_{t,men}$. Nous avons :

$$d_l^t = d_k^t \quad \text{pour tout individu} \quad l \in \text{ ménage } k.$$

Les individus de $S_{t,ind}$ sont suivis au temps $t+1$: **individus longitudinaux**.

Nous supposons que le suivi s'effectue de la façon suivante : à la date $t+1$, tous les individus vivant dans un ménage contenant un individu longitudinal (on parle de **ménage longitudinal**) sont enquêtés.

Définitions

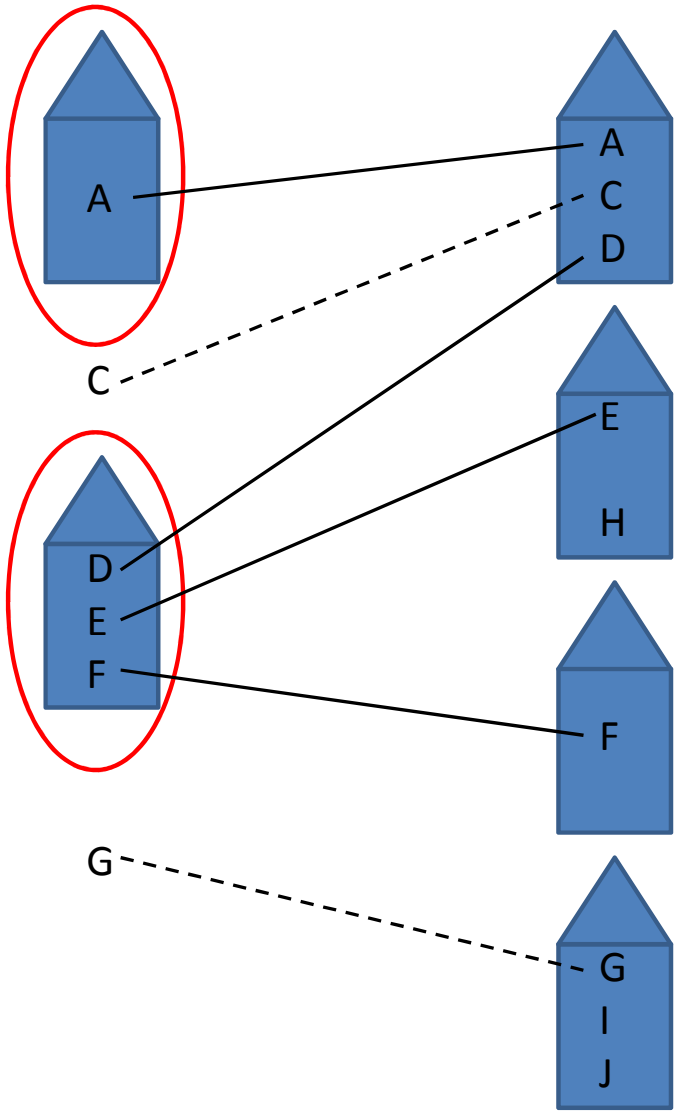
Individu initialement présent : individu qui faisait partie de la population au temps initial.

Individu initialement absent : individu qui ne faisait pas partie de la population au temps initial.
Immigrant : individu qui ne faisait pas partie de la population initiale, et la rejoint ensuite.
Nouveau-né : naissance.

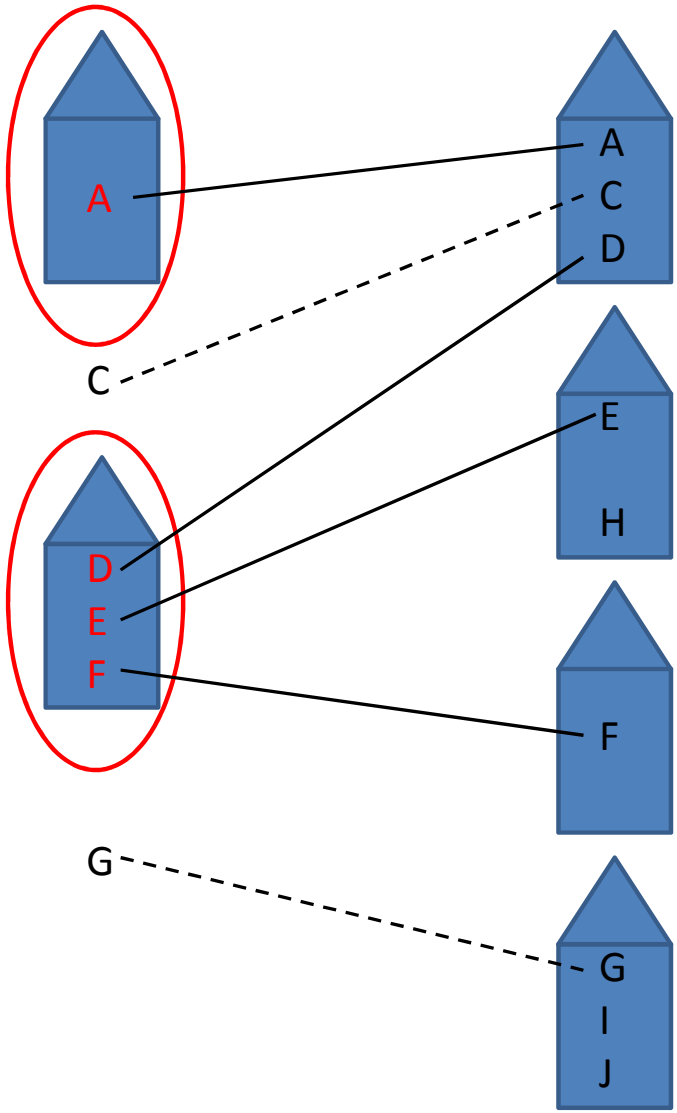
Individu longitudinal : individu sélectionné dans l'échantillon au temps initial.
Ménage longitudinal : ménage contenant au moins un individu longitudinal.
Cohabitant : individus (initialement absents ou initialement présents) qui rejoignent un ménage longitudinal.

Au temps $t+1$, nous enquêtons au titre de S_t l'ensemble des individus appartenant aux ménages longitudinaux.

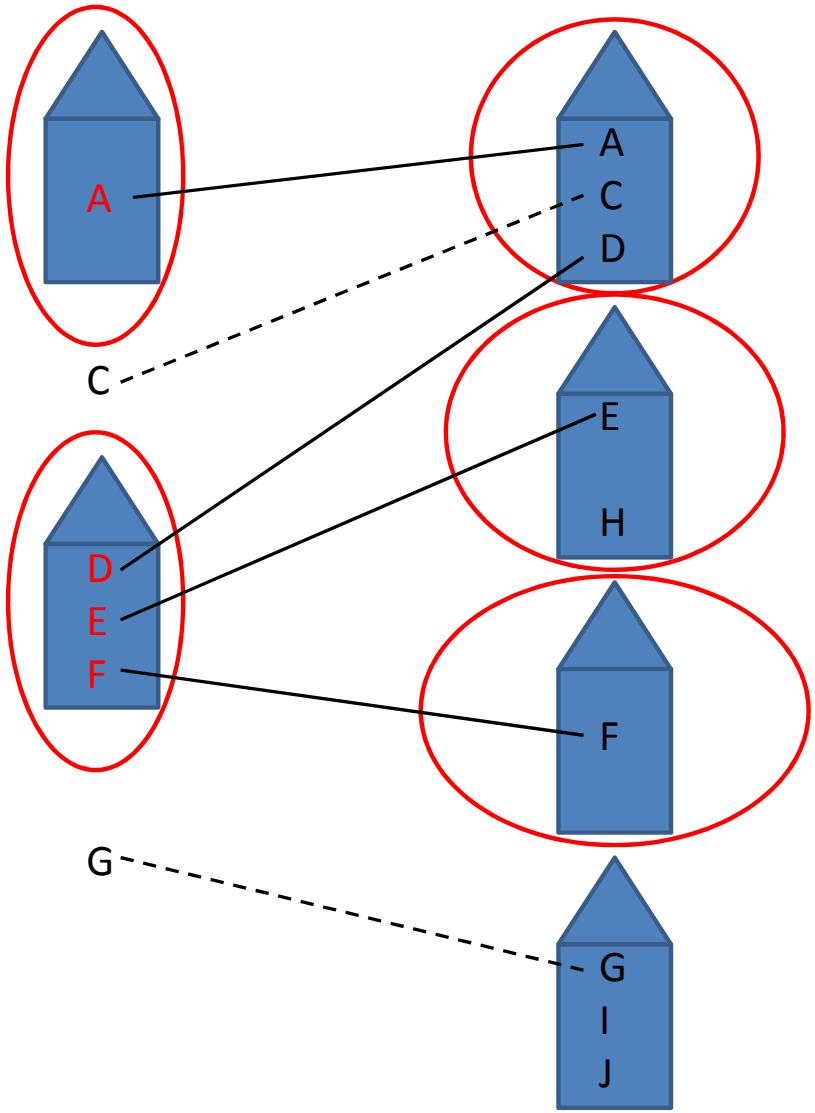
Exemple



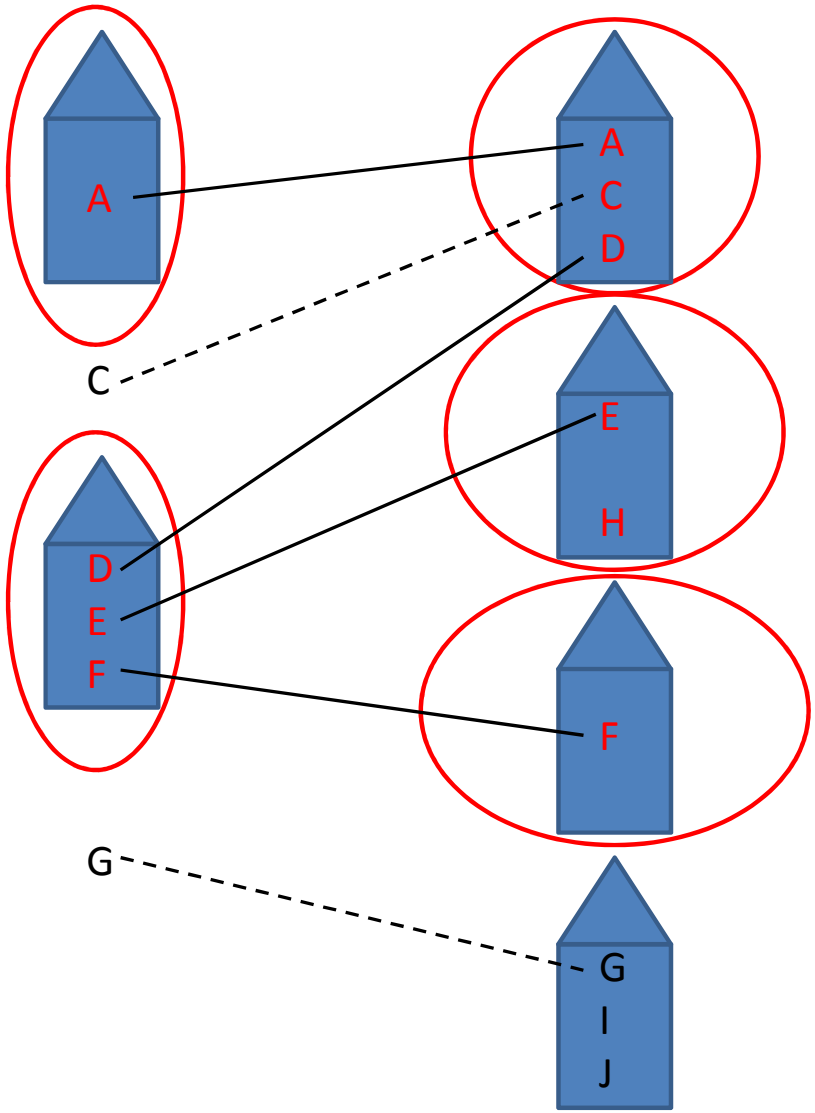
Exemple



Exemple



Exemple



Estimation transversale

L'estimation transversale est plus délicate. Notons $S_{men}^{t \to t+1}$ l'échantillon de ménages longitudinaux associé à S_t.
Afin de bien représenter la population transversale $U(t+1)$, il faut pouvoir le combiner à l'échantillon complémentaire de ménages $S_{t+1,men}$ sélectionné à la date $t+1$.

Nous procédons en plusieurs étapes :

1. Les poids des individus longitudinaux sont répartis sur les ménages de $S_{men}^{t \to t+1}$ à la date $t+1$, en utilisant la méthode de **partage des poids**.
2. L'échantillon $S_{men}^{t \to t+1}$ est combiné avec le nouvel échantillon $S_{t+1,men}$ en utilisant la technique de l'estimation composite.
3. Les individus reçoivent ensuite le poids de leur ménage.

Estimation transversale (2)

Un ménage longitudinal k va récupérer le poids

$$d_{t+1,k}^{t\to t+1} = \frac{\sum_{l\in S_{t,ind}^{\cap}} d_l^t L_{kl}}{N_{t+1,k}^{t\to t+1}},$$

en notant

- $L_{kl} = 1$ si l'individu l est dans le ménage k au temps $t+1$, et 0 sinon.
- $N_{t+1,k}^{t\to t+1} = \sum_{l\in U\cap(t+1)} L_{kl}$ le nombre d'individus du ménage k qui était dans le champ de l'enquête au temps t.

L'échantillon de ménages longitudinaux muni de ces poids permet de représenter les ménages présents à la date $t+1$, et qui contiennent au moins un individu qui appartenait à $U_{ind}(t)$ (individus initialement présents).

Estimation transversale (3)

Nous réunissons alors les deux échantillons de ménages en appliquant la technique de l'estimation composite :

- un ménage $k \in S_{men}^{t \to t+1}$ obtient le poids

$$w_k^{t+1} = \frac{d_{t+1,k}^{t \to t+1}}{2}.$$

- un ménage $k \in S_{t+1,men}$ obtient le poids

$$w_k^{t+1} = \begin{cases} \frac{d_{t+1,k}}{2} & \text{si } k \text{ contient un indiv. init. présent,} \\ d_{t+1,men} & \text{sinon.} \end{cases}$$

Comme tous les individus des ménages enquêtés sont également enquêtés, un individu l d'un ménage enquêté k récupère simplement le poids de son ménage w_k^{t+1}.

Exercice 6

Bibliographie

- Couvert, N., Dieusaert, P., et Henry, M. (2016), *Enquête Panel Politique de la Ville 4ème vague*, Dossier Comité du Label du Cnis.
- Deville, J-C., et Lavallée, P. (2006). Sondage indirect : les fondements de la méthode généralisée de partage des poids. Techniques d'enquête, 32, 185-196.
- Deville, J-C., et Maumy, M. (2006). Extensions de la méthode d'échantillonnage indirect et son application à l'enquête dans le tourisme : M.O.R.G.O.A.T. Techniques d'enquête, 32, 197-206.
- Juillard, H. (2016). Méthodes d'estimation et d'estimation de variance pour une enquête longitudinale. Thèse de l'université de Toulouse.
- Lavallée, P. (2007). Indirect sampling. Springer.
- Le Guennec, J. (2009). Les extensions régionales et locales de l'enquête Logement 2006. Journées de Méthodologie Statistique, Paris.
- Lohr, S. L. (2009). Multiple-frame surveys. Handbook of statistics, 29, pp. 71-88, Elsevier.
- Lynn, P. (2009). Méthodology of longitudinal surveys. Wiley.
- Menard, S. (2008). Handbook of longitudinal research. Elsevier.