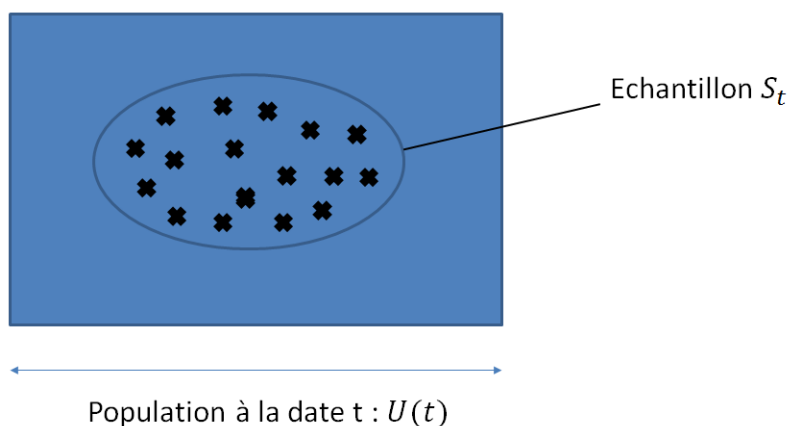


Enquêtes répétées dans le temps

Exercices

Exercice 1

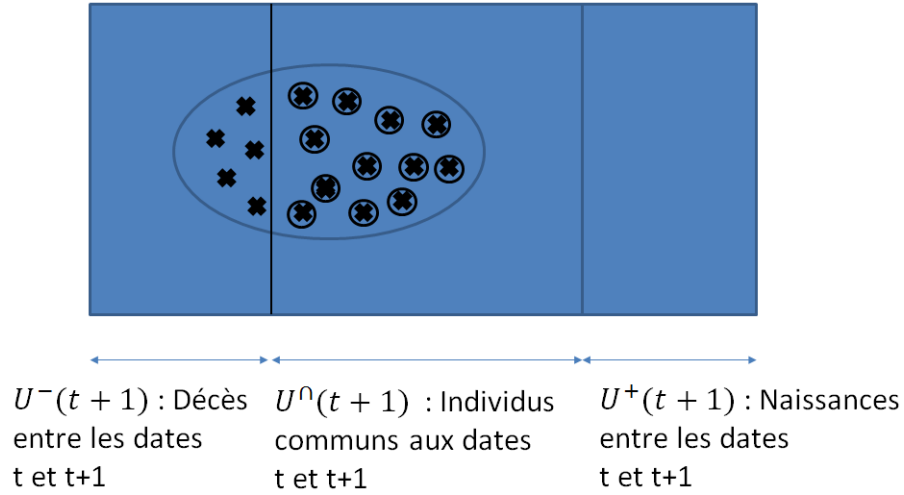
Dans la population des 1 000 habitants d'une commune, nous sélectionnons une année t un échantillon de 200 personnes par sondage aléatoire simple. Parmi ces 200 personnes interrogées, 120 souhaitent la construction d'une piscine.



- 1) Donner une estimation du nombre de personnes intéressées par la construction de la piscine l'année t .
- 2) Donner l'intervalle de confiance à 95 % associé. Il est rappelé que pour une variable y_k^t binaire, la dispersion calculée sur l'échantillon vaut :

$$s_y^{2t} = \frac{n^t}{n^t - 1} \bar{y}^t (1 - \bar{y}^t).$$

L'année $t + 1$, 30 personnes de l'échantillon ne résident plus dans la commune (déménagement ou décès). Parmi les 170 personnes restantes de l'échantillon, 110 souhaitaient l'année t la construction de la piscine et 100 souhaitent toujours l'année $t + 1$ la construction de la piscine.



Il est proposé d'étudier l'évolution du nombre de personnes étudiées par la construction de la piscine, avec deux indicateurs possibles :

- la différence entre le nombre total de personnes intéressées à la date $t + 1$, et le nombre total de personnes intéressées à la date t :

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

- la différence entre le nombre total de personnes intéressées à la date $t + 1$ et le nombre total de personnes intéressées à la date t , parmi les personnes présentes aux deux dates :

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

3) Lequel de ces deux paramètres choisissez-vous d'estimer?

4) Montrer que le paramètre $\Delta^\cap$ peut se réécrire sous la forme

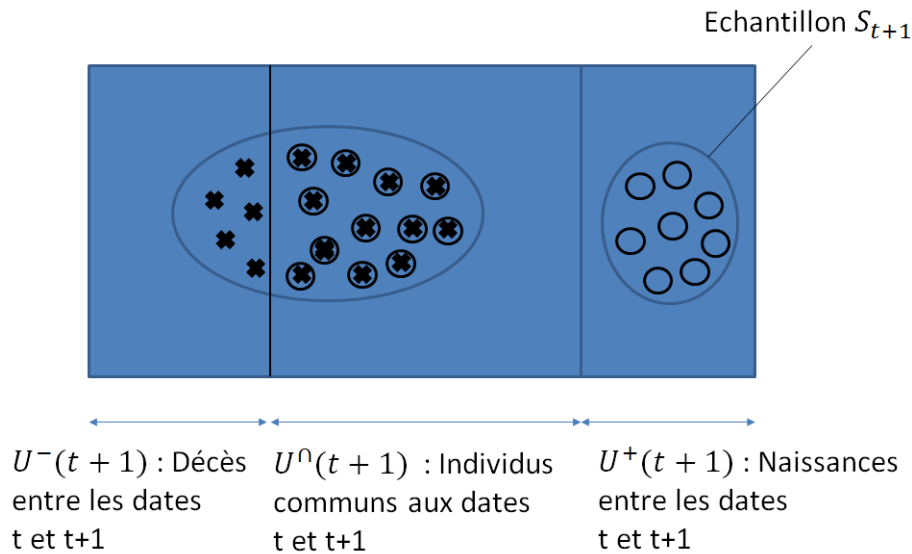
$$\Delta^\cap = \sum_{k \in U(t)} z_k^{t+1},$$

pour une variable z_k^{t+1} à préciser, et en indiquant ses valeurs possibles.

5) Donner une estimation du paramètre, et l'intervalle de confiance à 95 % associé. Nous avons sur l'échantillon :

$$s_z^{2,t+1} = 0.148$$

Parmi les 100 nouveaux habitants de la commune, nous sélectionnons un échantillon de 50 individus par sondage aléatoire simple. Parmi ces derniers, 15 souhaitent la construction d'une piscine.



6) Donner une estimation du nombre de personnes intéressées par la construction de la piscine l'année $t+1$.

7) En utilisant les questions 1 et 6, donner une estimation du paramètre Δ .

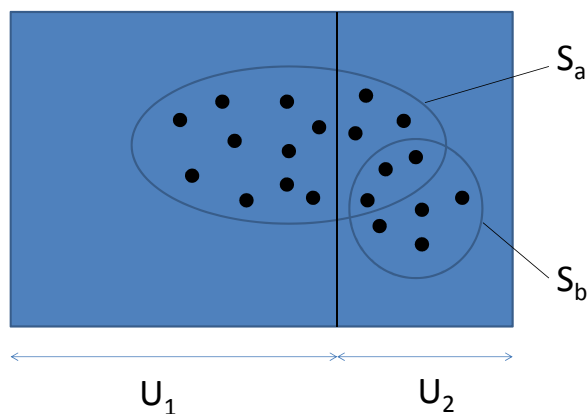
8) Pourquoi la sélection de ce nouvel échantillon semble-t-elle difficile en pratique? Comment pourrait-on sélectionner ce nouvel échantillon?

Exercice 2

Nous souhaitons estimer le total Y d'une variable d'intérêt y_k définie sur une population U , partitionnée en deux sous-populations U_1 et U_2 .

Nous sélectionnons un échantillon S_a dans U , avec des probabilités d'inclusion $\pi_{ak} > 0$ pour tout $k \in U$. Nous notons d_{ak} les poids d'extrapolation associés.

Nous sélectionnons un échantillon S_b dans U_2 , avec des probabilités d'inclusion $\pi_{bk} > 0$ pour tout $k \in U_2$. Nous notons d_{bk} les poids d'extrapolation associés.



- 1) Montrer que $\hat{Y}_a = \sum_{k \in S_a} d_{ak} y_k$ est sans biais pour Y .
- 2) Proposer un estimateur sans biais du sous-total $Y_2 = \sum_{k \in U_2} y_k$ utilisant l'échantillon S_b uniquement, et un estimateur utilisant l'échantillon S_a uniquement.

Nous considérons tout d'abord le système de pondération suivant :

$$\begin{aligned} \text{Dans } U_1 \quad d_k &= d_{ak} \quad \text{pour } k \in S_a. \\ \text{Dans } U_2 \quad d_k &= \frac{d_{ak}}{2} \quad \text{pour } k \in S_a, \\ &= \frac{d_{bk}}{2} \quad \text{pour } k \in S_b. \end{aligned}$$

- 3) Montrer que l'estimateur du total que nous obtenons peut s'écrire sous la forme :

$$\hat{Y} = \hat{Y}_a^1 + \frac{1}{2} \hat{Y}_a^2 + \frac{1}{2} \hat{Y}_b^2,$$

avec des notations à préciser.

4) En déduire que cette pondération conduit à une estimation sans biais du total Y .

5) Nous supposons dans cette question seulement :

- que la sous-population U_2 contient 10 % des unités,
- que S_a est tiré par sondage aléatoire simple de taille $n_a = 100$,
- que S_b est tiré par sondage aléatoire simple de taille $n_b = 50$.

Quel problème se pose sur U_2 avec la méthode d'estimation retenue?

Nous considérons maintenant le système de pondération suivant, avec θ une constante fixée :

$$\begin{array}{lll} \text{Dans } U_1 & d_k = d_{ak} & \text{pour } k \in S_a. \\ \text{Dans } U_2 & d_k = \theta d_{ak} & \text{pour } k \in S_a, \\ & d_k = (1 - \theta) d_{bk} & \text{pour } k \in S_b. \end{array}$$

6) Montrer que l'estimateur du total que nous obtenons peut s'écrire sous la forme :

$$\hat{Y} = \hat{Y}_a^1 + \theta \hat{Y}_a^2 + (1 - \theta) \hat{Y}_b^2.$$

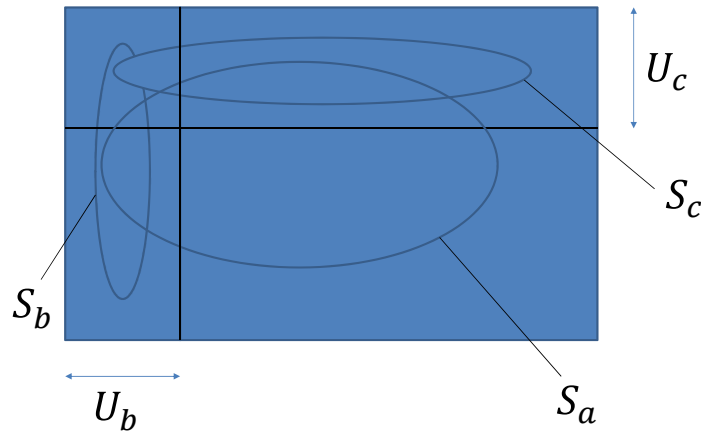
7) En déduire que cette pondération conduit à une estimation sans biais du total Y .

8) Comment peut-on choisir le paramètre θ en pratique?

Nous considérons une population U , dans laquelle :

- nous tirons un échantillon S_a dans U , avec des probabilités d'inclusion $\pi_{ak} > 0$,
- nous tirons un échantillon S_b dans une sous-population U_b , avec des probabilités d'inclusion $\pi_{bk} > 0$,
- nous tirons un échantillon S_c dans une sous-population U_c , avec des probabilités d'inclusion $\pi_{ck} > 0$.

La situation est résumée dans la figure ci-dessous.



9) Proposer un système de pondération inspiré de celui donné à la question 3, permettant d'utiliser les 3 échantillons et d'obtenir un estimateur sans biais du total Y .

Exercice 3

Un jour donné, nous sélectionnons un échantillon de $n = 3$ sites touristiques de Rennes et toutes les personnes visitant les sites sont enquêtées. Soit U_A la population des sites touristiques, et U la population des visiteurs.

Une personne peut visiter plusieurs sites pendant la journée.

- le premier site sélectionné possède un poids de sondage d_1 , et contient les visiteurs $V1$, $V2$ et $V3$,
- le second site sélectionné possède un poids de sondage d_2 , et contient les visiteurs $V1$, $V4$ et $V5$,
- le troisième site sélectionné possède un poids de sondage d_3 , et contient les visiteurs $V2$, $V4$, $V6$ et $V7$.

La table suivante donne, pour tous les visiteurs sélectionnés dans l'échantillon, le nombre de sites visités pendant la journée et une indicatrice y_k égale à 1 si le visiteur est étranger et à 0 sinon:

	V1	V2	V3	V4	V5	V6	V7
Nb de sites	4	5	3	4	2	3	3
Etranger (0/1)	1	1	0	0	0	1	0
Poids w_k							

- 1) Calculer les poids w_k obtenus par la méthode de partage des poids.
- 2) En supposant que $d_1 = d_2 = d_3 = 4$, donner une estimation du nombre total de visiteurs étrangers.
- 3) Montrer que l'estimateur obtenu peut se réécrire sous la forme

$$\hat{Y} \equiv \hat{Z}_A = \frac{N_A}{n_A} \sum_{i \in S_A} z_i,$$

en précisant la formule donnant la variable d'intérêt z_i , et les valeurs de cette variable sur l'échantillon S_A .

- 4) En supposant que S_A est sélectionné par sondage aléatoire simple sans remise, donner un estimateur de variance pour $\hat{Y}$ et calculer sa valeur.

- 5) Donner une estimation du pourcentage de visiteurs étrangers.

Exercice 4

Dans la population des 1 000 habitants d'une commune présents à la date t , nous tirons un échantillon S_t par sondage aléatoire simple de taille 200. Parmi les individus enquêtés, 80 sont propriétaires de leur logement.

1) Donner une estimation du nombre total de propriétaires à la date t .

Les individus échantillonnés sont suivis dans le temps, et enquêtés à nouveau l'année suivante $t + 1$. Parmi les 200 individus de l'échantillon :

- 20 sont décédés ou ont quitté la commune.
- 100 ont plus de 60 ans. Parmi eux, 80 acceptent de répondre l'année suivante. Parmi ces 80 personnes, 38 étaient propriétaires l'année t et 40 le sont l'année $t + 1$
- 80 ont moins de 60 ans. Parmi eux, 40 acceptent de répondre l'année suivante. Parmi ces 40 personnes, 10 étaient propriétaires l'année t et 15 le sont l'année $t + 1$.

Nous souhaitons estimer l'évolution du nombre d'individus propriétaires.

2) Quel est le champ de cette estimation? Combien d'individus de l'échantillon sont encore dans le champ de l'enquête?

Nous supposons que les probabilités de réponse sont homogènes au sein des deux classes d'âge (moins de 60 ans/plus de 60 ans).

3) Donner une estimation de la taille des GHR dans la population, en utilisant les poids de sondage.

4) Donner une estimation des probabilités de réponse.

5) En utilisant ces probabilités estimées, donner les poids corrigés de la non-réponse. Montrer que ces poids sont calés sur les tailles estimées des deux GHR (cf question 3).

6) Donner une estimation de l'évolution du nombre d'individus propriétaires.

Exercice 5

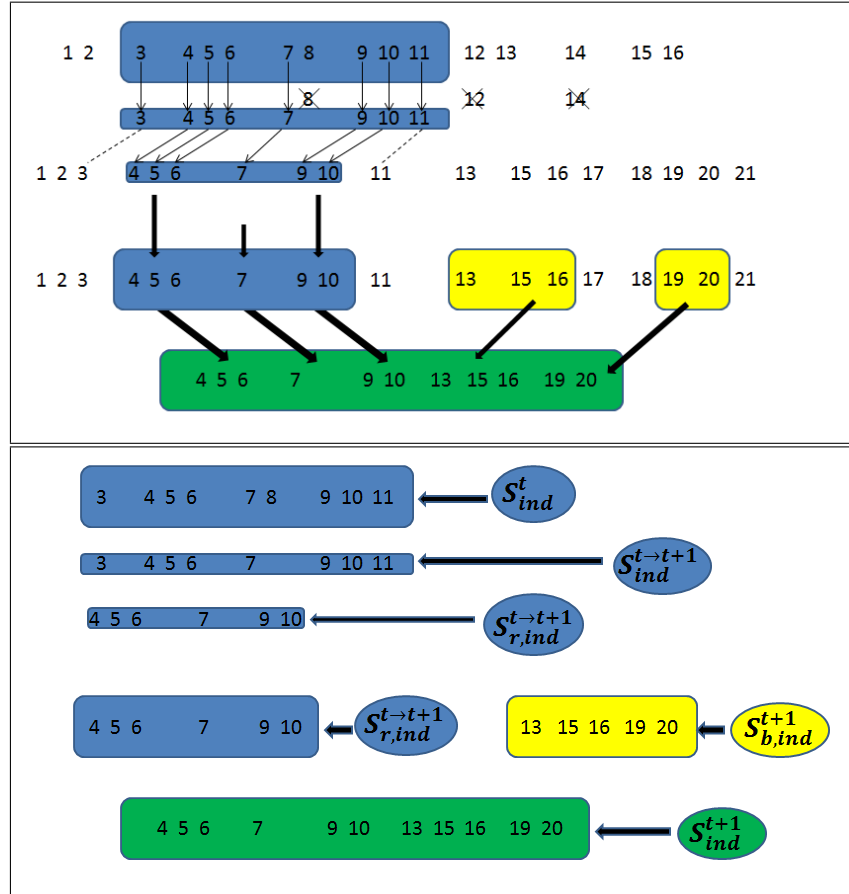


Figure 1: Enquête à échantillon partagé : suivi d'individus dans le temps

Une population $U_{ind}(t)$ contient 16 individus, numérotés de 1 à 16. Nous sélectionnons parmi ceux-ci un échantillon S_{ind}^t , selon un sondage aléatoire simple. Nous obtenons

$$S_{ind}^t = \{3, 4, 5, 6, 7, 8, 9, 10, 11\}.$$

Ces individus sont suivis dans le temps l'année $t + 1$, mais :

- 1 individu de l'échantillon (8) ne fait plus partie du champ de l'enquête
 $\Rightarrow$ sous-échantillon $S_{ind}^{t \rightarrow t+1}$,

- 2 individus de l'échantillon (3 et 11) ne répondent pas ou sont perdus de vue
 $\Rightarrow$ sous-échantillon $S_{r,ind}^{t \rightarrow t+1}$.

Parmi l'ensemble des individus présents l'année $t + 1$ (population $U_{ind}(t + 1)$), nous sélectionnons un échantillon $S_{b,ind}^{t+1}$ selon un sondage aléatoire simple. Nous obtenons

$$S_{b,ind}^{t+1} = \{13, 15, 16, 19, 20\}.$$

Nous notons S_{ind}^{t+1} la réunion de $S_{r,ind}^{t \rightarrow t+1}$ et de $S_{b,ind}^{t+1}$. Les notations sont résumées dans la Figure 1.

Population au temps t

- 1) Donner les pondérations initiales des individus de S_{ind}^t .
- 2) Quelle population représente t-on avec cet échantillon muni de ces poids.

Population au temps $t + 1$: échantillon initial

- 3) En vous aidant de la Figure 1, donner les individus constituant les sous-populations $U_{ind}^-(t + 1)$ (décès), $U_{ind}^\cap(t + 1)$ (individus communs aux deux dates), et $U_{ind}^+(t + 1)$ (naissances).
- 4) Quelle population les individus de $S_{ind}^{t \rightarrow t+1}$ permet-elle de représenter? Donner les poids associés aux individus.

Nous supposons que le mécanisme de réponse conduisant à $S_{r,ind}^{t \rightarrow t+1}$ est uniforme ou MCAR.

- 5) Quelle population les individus de $S_{r,ind}^{t \rightarrow t+1}$ permet-elle de représenter?
- 6) Donner les poids associés aux individus de $S_{r,ind}^{t \rightarrow t+1}$.

Population au temps $t + 1$: échantillon complémentaire

- 7) Quelle population les individus de $S_{b,ind}^{t+1}$ permet-elle de représenter? Donner les poids associés aux individus.
- 8) En utilisant la technique de l'estimation composite, proposer une façon de calculer les poids des individus de S_{ind}^{t+1} .

Exercice 6

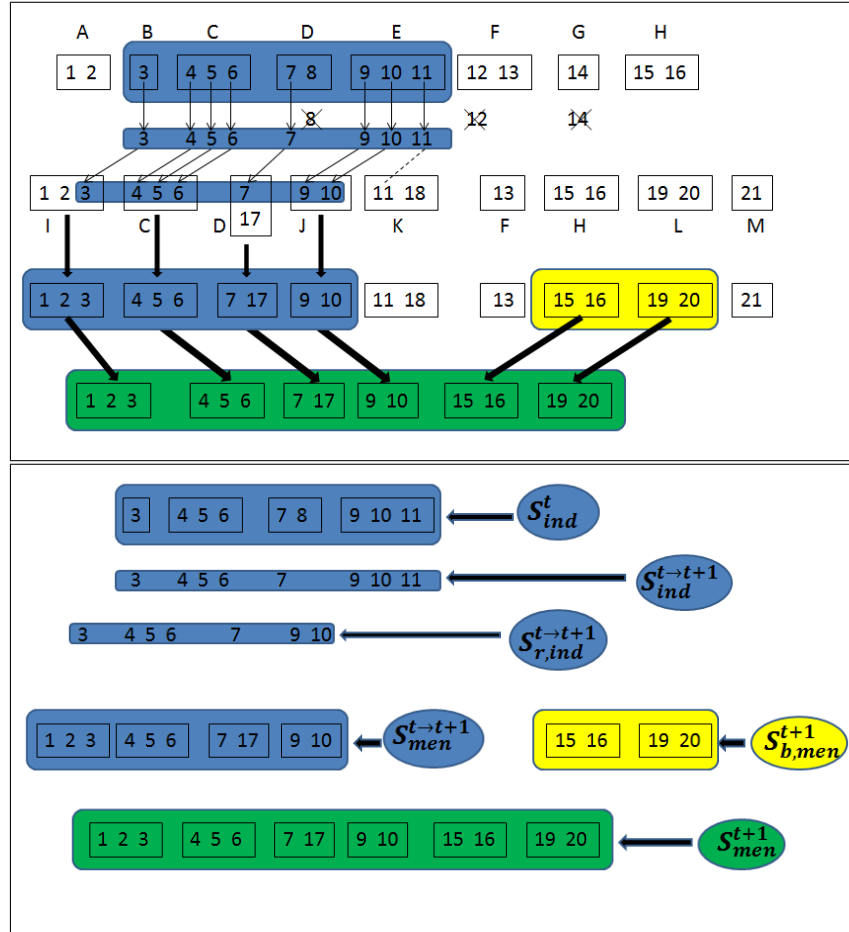


Figure 2: Enquête à échantillon partagé : suivi d'individus dans le temps

La population $U_{men}(t)$ contient 8 ménages, numérotés de A à H. Nous sélectionnons un échantillon S_{men}^t , selon un sondage aléatoire simple, et tous les individus de ces ménages sont enquêtés. Cela conduit aux échantillons

$$\begin{aligned} S_{men}^t &= \{B, C, D, E\}, \\ S_{ind}^t &= \{3, 4, 5, 6, 7, 8, 9, 10, 11\}. \end{aligned}$$

Les individus correspondant sont suivis dans le temps l'année $t + 1$, mais :

- 1 individu de l'échantillon (8) ne fait plus partie du champ de l'enquête
 $\Rightarrow$ sous-échantillon $S_{ind}^{t \rightarrow t+1}$,
- 1 individu de l'échantillon (11) ne répond pas ou est perdu de vue
 $\Rightarrow$ sous-échantillon $S_{r,ind}^{t \rightarrow t+1}$.

Nous enquêtons les ménages contenant au moins un individu de $S_{r,ind}^{t \rightarrow t+1}$, et tous les individus appartenant à ces ménages.

Parmi l'ensemble des ménages présents l'année $t+1$ (population $U_{men}(t+1)$), nous sélectionnons un échantillon $S_{b,men}^{t+1}$ selon un sondage aléatoire simple.

Nous obtenons $S_{b,men}^{t+1} = \{H, L\}$.

Nous notons S_{men}^{t+1} la réunion des ménages enquêtés au temps $t+1$. Tous les individus de ces ménages sont enquêtés au temps $t+1$.

Les notations sont résumées dans la Figure 2.

Population au temps t

- 1) Donner les pondérations initiales des ménages de S_{men}^t .
- 2) Quelle population représente t-on avec cet échantillon muni de ces poids?
- 3) Donner les pondérations initiales des individus de S_{ind}^t .
- 4) Quelle population représente t-on avec cet échantillon muni de ces poids?

Population au temps $t+1$: échantillon initial

- 5) En vous aidant de la Figure 2, donner les individus constituant les sous-population $U_{ind}^-(t+1)$ (décès), $U_{ind}^\cap(t+1)$ (individus communs aux deux dates), et $U_{ind}^+(t+1)$ (naissances).
- 6) Donner les individus longitudinaux, les individus initialement présents, les individus initialement absents, les ménages longitudinaux, les cohabitants.
- 7) Quelle population les individus de $S_{ind}^{t \rightarrow t+1}$ permet-elle de représenter? Donner les poids associés aux individus.

Nous supposons que le mécanisme de réponse conduisant à $S_{r,ind}^{t \rightarrow t+1}$ est uniforme ou MCAR.

- 8) Quelle population les individus de $S_{r,ind}^{t \rightarrow t+1}$ permet-elle de représenter?

- 9) Donner les poids associés aux individus de $S_{r,ind}^{t \rightarrow t+1}$.
- 10) Quel échantillon de ménages $S_{men}^{t \rightarrow t+1}$ est associé à $S_{r,ind}^{t \rightarrow t+1}$?
- 11) En utilisant la méthode de partage des poids, donner les poids des ménages de $S_{men}^{t \rightarrow t+1}$.
- 12) Quelle population les ménages de $S_{men}^{t \rightarrow t+1}$ munis de ces poids permet-elle de représenter?

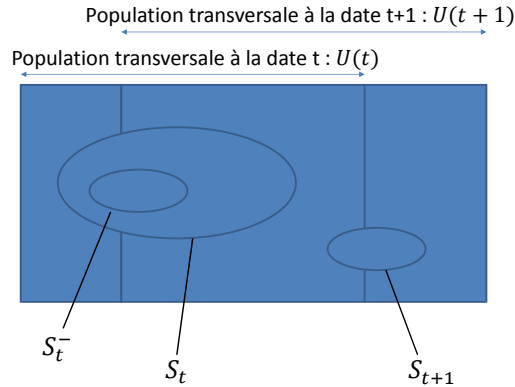
Population au temps $t + 1$: échantillon complémentaire

- 13) Quelle population les ménages de $S_{b,men}^{t+1}$ permet-elle de représenter? Donner les poids associés aux ménages.
- 14) En utilisant la technique de l'estimation composite, proposer une façon de calculer les poids des ménages de S_{men}^{t+1} .
- 15) Donner les pondérations des individus associés aux ménages de S_{men}^{t+1} . Quelle population ces individus représentent-ils?

Exercice 7

Dans une population d'individus $U(t)$ de taille $N^t = 1\,000$, on sélectionne un échantillon S_t par SRS de taille $n_t = 200$. A la date $t + 1$, certaines unités disparaissent (sous-population $U^-(t + 1)$, de taille 200) et d'autres apparaissent (sous-population $U^+(t + 1)$, de taille 400). On renouvelle l'enquête suivant le principe des échantillons rotatifs :

- on sélectionne dans S_t un sous-échantillon S_t^- de taille $n_{t+1} = 50$ par SRS. Les individus de S_t^- ne participent pas à l'enquête à la date $t + 1$. Les autres individus sont réenquêtés.
- on sélectionne dans U_{t+1} un nouvel échantillon S_{t+1} de taille n_{t+1} par SRS. Ces individus sont enquêtés à la date $t + 1$.



- 1) Donner la pondération et l'échantillon utilisé pour une estimation transversale à la date t .
- 2) Donner la pondération et l'échantillon utilisé pour une estimation longitudinale à la date $t + 1$ sur la population intersection.
- 3) Donner une pondération et l'échantillon utilisé pour une estimation transversale à la date $t + 1$:

- en donnant la même importance aux deux échantillons pour l'estimation dans $U^\cap(t + 1)$,

- en pondérant les deux échantillons disponibles dans $U^\cap(t+1)$ selon leur taille. On supposera que 120 unités de $S_t \setminus S_t^-$ sont dans $U^\cap(t+1)$, et que 30 unités de S_{t+1} sont dans $U^\cap(t+1)$.