Introduction aux méthodes d'échantillonnage avec application aux données de la mer et du littoral

Guillaume Chauvet, Elodie Plissonneau

Journées MAASC 2016

09/11/2016

ENSAI
École nationale de la statistique et de l'analyse de l'information

A quoi sert la théorie des Sondages ?

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

- quand on contrôle la façon dont l'échantillon est sélectionné (tirage probabiliste),
- quand on ne contrôle pas la façon dont l'échantillon 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 booking, skype, ...).
  - Non-réponse : diminue la taille de l'échantillon effectivement observé, et tend à sur-représenter des profils particuliers (risque de biais). Attention au fardeau de réponse.

ENSAI École nationale de la statistique et de l'analyse de l'information

A quoi sert la théorie des Sondages ?

On sélectionne un échantillon quand il est trop coûteux de travailler sur l'ensemble des données :

- Coût de collecte de l'information
  - Enquête auprès des ménages (e.g., enquête emploi)
  - Enquête auprès des entreprises (répertoire SIRUS)
  - Enquête d'opinion
  - Enquête épidémiologique
  - Enquête sur l'effort de pêche (ObsDeb, IFREMER)
- Coût de traitement de l'information
  - Flux de données (Big Data),
  - Exploitation d'une grosse base de données,
  - Recensement avant 1999 : exploitation au 1/4 ou au 1/20.

ENSAI École nationale de la statistique et de l'analyse de l'information

Exemple 1 : Enquêtes Annuelles de Recensement

Les Enquêtes Annuelles de Recensement (EAR) ont succédé en 2004 au Recensement traditionnel, approximativement décennal. Le plan de sondage utilisé (Godinot, 2005) distingue :

- les grandes communes (10 000 habitants ou plus au RP 1999). Au sein de chacune, sélection et enquête chaque année auprès de 8 % des logements.
- les petites communes (moins de 10 000 habitants au RP 1999). Au sein de chaque région, sélection et enquête chaque année auprès d'1/5 des petites communes.

Utilisation de l'**échantillonnage équilibré** (Deville et Tillé, 2004).

Les EAR servent de base de sondage pour les enquêtes ménages de l'Insee, et également pour d'autres enquêtes (Panel Politique de la Ville, étude MobiliSense, ...)

ENSAI École nationale de la statistique et de l'analyse de l'information

Exemple 2 : enquête auprès des ménages

Les enquêtes ménage de l'Insee visent à décrire les conditions de vie des ménages (emploi, logement, patrimoine, ...). Comme on ne dispose pas d'une **base de sondage**, i.e. d'un répertoire des ménages, une sélection directe de l'échantillon est impossible.

Les ménages enquêtés sont sélectionnés selon un principe de tirage à trois degrés : tirage d'un échantillon de zones appelé l'Echantillon-maître, puis de quartiers dans ces zones, puis de ménages dans ces quartiers.

Avantages : concentration des unités échantillonnées + base de sondage à constituer limitée.
Inconvénients : procédure complexe + estimation moins précise qu'avec un tirage direct.

ENSAI École nationale de la statistique et de l'analyse de l'information

Exemple 3 : enquêtes épidémiologiques et sociales

Utiliser plusieurs degrés de tirage est fréquemment utilisé dans les enquêtes, y compris hors Insee :

1. Enquête Panel Politique de la Ville (PPV) : sélection d'un échantillon de quartiers (UP), puis de logements (US), puis d'individus (UT) (Dieusaert et Henry, 2012).
2. Enquête épidémiologique : estimation de la contamination au plomb en tirant un échantillon d'hôpitaux (UP), puis d'enfants (US) dont les logements sont inspectés (Lucas, 2013).
3. Enquête PISA : sélection d'un échantillon de collèges (UP), puis d'un échantillon d'élèves de 15 ans (US).

ENSAI École nationale de la statistique et de l'analyse de l'information

Exemple 4 : enquête ObsDeb sur l'effort de pêche

Projet : améliorer la connaissance de la pêche professionnelle côtière en Méditerranée et DOM (navires de moins de douze mètres).

Plan de sondage : stratification des flottilles selon l'activité de l'année précédente et selon des strates géographiques. Tirage à deux degrés, avec échantillonnage de flottilles, puis de marées (=bateaux).

Estimation du nombre de sorties en mer et de l'effort de pêche par espèce.

Difficultés : contraintes techniques + forte dispersion des variables relevées.

ENSAI École nationale de la statistique et de l'analyse de l'information

Exemple 5 : Enquête VALPENA sur l'activité de pêche

Les enquêtes VALPENA visent à spatialiser l'activité de pêche maritime professionnelle afin de donner aux pêcheurs des arguments pour leur défense dans les problématiques de partage de l'espace marin.

Les enquêtes ont lieu chaque année, par les enquêteurs des comités de pêche, auprès des patrons pêcheurs, sur leur activité de l'année précédente.

Les 2 premières années d'enquête sont des années d'enquêtes exhaustives :

- Création d'une base navire fiable et complète.
- Deux limites : manque de moyens et surenquête des pêcheurs

La mise en place d'un plan de sondage avec tirage direct est donc possible :

- Stratification de la flotte suivant différents critères : géographique, activité et longueur,
- Nombre de navires à enquêter déterminé pour une certaine précision

Plan

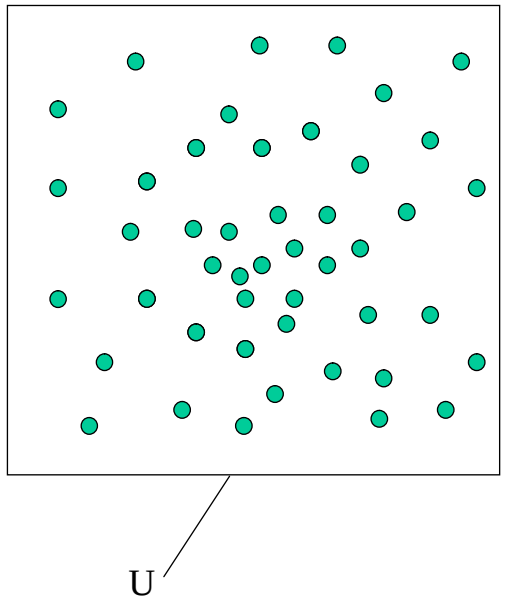
ENSAI École nationale de la statistique et de l'analyse de l'information

Echantillonnage en population finie

Notations

Notations

On se place dans le cadre d'une population finie U d'*individus statistiques* que l'on note $U = \{1, \ldots, k, \ldots, N\}$.
Sur chaque individu k de la population est défini une *variable d'intérêt* y_k.
On s'intéresse à l'estimation de grandeurs sur la population U.



Paramètre d'intérêt

On peut par exemple s'intéresser au total

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

d'une variable quantitative sur la population, ou encore à sa valeur 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, ...

ENSAI École nationale de la statistique et de l'analyse de l'information

Notations et Paramètre d'intérêt : cas de Valpena

La population U est la flotte de navires, notée $U = \{1, \ldots, N\}$, avec N le nombre de navires.

Sur chaque maille VALPENA l, et pour chaque navire k, on récolte l'information sur une variable d'intérêt notée y_k^l.

y_k^l peut être par exemple égal à

$$1 \text{ ou } 0 \quad \text{selon si le navire } k \text{ est présent ou non sur la maille,}$$
$$\text{le nombre de mois travaillé par le navire } k.$$

On peut s'intéresser au total de ces variables :

$$t_y^l = \text{Indicateur de Densité ou Indicateur d'Intensité}$$

ou à la moyenne

$$\mu_y^l = \text{ID en proportion ou II en proportion.}$$

ENSAI École nationale de la statistique et de l'analyse de l'information

Plan de sondage

Plan de sondage

On tire un échantillon S d'individus au moyen 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.

Pour un paramètre d'intérêt θ, on note $\hat{\theta}(S)$ son estimateur calculé sur la base des données relevées sur S. La valeur de cet estimateur pour les données effectivement recueillies est appelée l'estimation.

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

ENSAI École nationale de la statistique et de l'analyse de l'information

Exemple

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

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

L'échantillon S est sélectionné parmi les sous-ensembles suivants :

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

On a par exemple

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

ENSAI École nationale de la statistique et de l'analyse de l'information

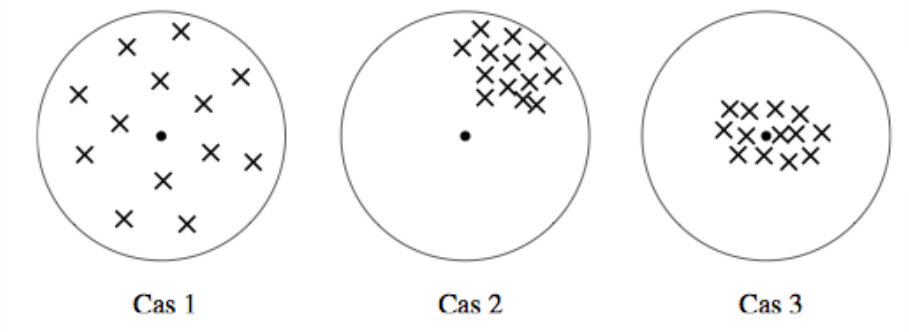
Mesures de précision

Deux mesures principales pour juger de la qualité d'un estimateur :

- le **biais** représente la valeur moyenne de l'écart entre l'estimateur et le paramètre d'intérêt,
- la **variance** représente la dispersion des estimations possibles.

Si on assimile notre problème d'estimation à un jeu de fléchettes où l'objectif est d'atteindre le centre :

- notre lancer est **non biaisé** si le centre de la cible est au coeur du nuage des fléchettes lancées,
- la **variance** du lancer correspond à l'écartement du nuage des fléchettes lancées.



Mesures de précision

En statistique, la valeur moyenne d'un estimateur est encore appelée l'espérance $E(\hat{\theta})$. Le biais est donc donné par

$$B(\hat{\theta}) = E(\hat{\theta} - \theta).$$

La variance est donnée par

$$V(\hat{\theta}) = E[\hat{\theta} - E(\hat{\theta})]^2.$$

L'Erreur Quadratique Moyenne est une mesure synthétique de l'incertitude, tenant compte à la fois du biais et de la variance :

$$EQM(\hat{\theta}) = E(\hat{\theta} - \theta)^2 = [B(\hat{\theta})]^2 + V(\hat{\theta}).$$

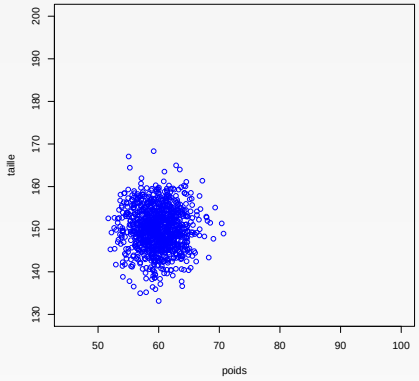
ENSAI École nationale de la statistique et de l'analyse de l'information

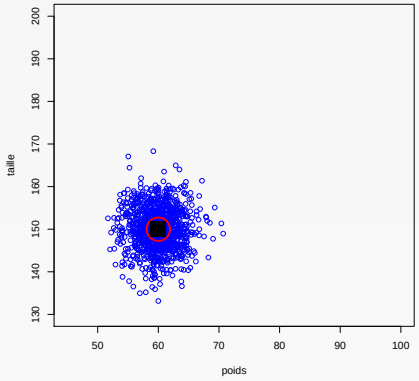
Quelques simulations

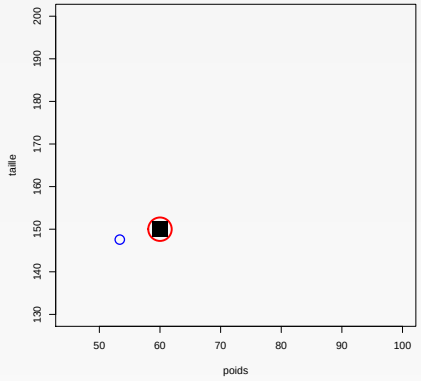
Pour illustrer la notion de biais et de variance, on considère l'exemple d'une population de $N = 1\,000$ individus âgés de 15 à 20 ans.

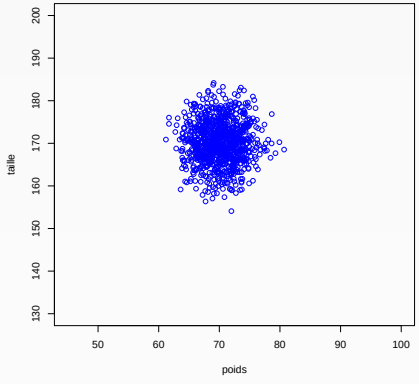
Dans cette population, un échantillon de taille $n = 50$ est sélectionné et enquêté. Pour chaque individu enquêté, on obtient son poids (en kg), sa taille (en cm) et son âge.

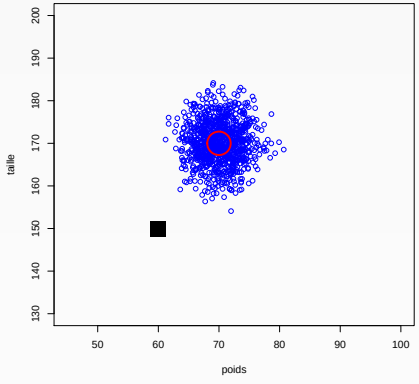
On s'intéresse à l'estimation du poids moyen et de la taille moyenne (carré noir). Chaque échantillon permet d'obtenir une estimation (points bleus) de ces paramètres. La moyenne des estimations est représentée par le point rouge.

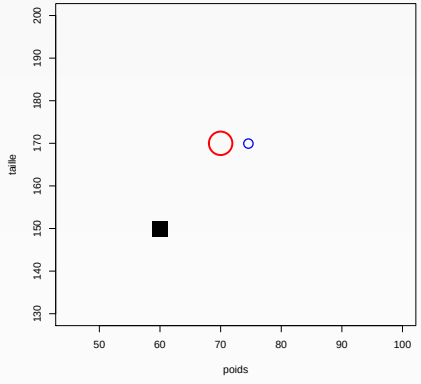

taille
poids
130
140
150
160
170
180
190
200
50
60
70
80
90
100

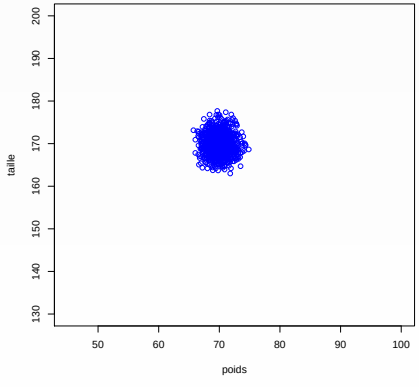

taille
poids
130
140
150
160
170
180
190
200
50
60
70
80
90
100

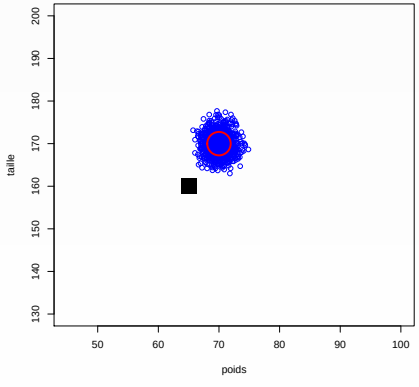

taille
poids
130
140
150
160
170
180
190
200
50
60
70
80
90
100

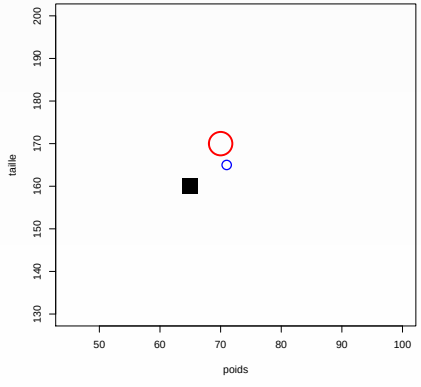

taille
poids
130
140
150
160
170
180
190
200
50
60
70
80
90
100


taille
poids
130
140
150
160
170
180
190
200
50
60
70
80
90
100


taille
poids
130
140
150
160
170
180
190
200
50
60
70
80
90
100


taille
poids
130
140
150
160
170
180
190
200
50
60
70
80
90
100


taille
poids
130
140
150
160
170
180
190
200
50
60
70
80
90
100


taille
poids
130
140
150
160
170
180
190
200
50
60
70
80
90
100

Estimation de Horvitz-Thompson

Probabilités d'inclusion

On note 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. Ces probabilités sont fixées avant le tirage à l'aide d'une **information auxiliaire**.

On note π_{kl} la probabilité que deux unités distinctes k et l soient sélectionnées conjointement dans l'échantillon. Ces probabilités doubles interviennent notamment dans la variance des estimateurs.

ENSAI École nationale de la statistique et de l'analyse de l'information

L'estimateur de Horvitz-Thompson

Si toutes les probabilités π_k sont > 0, le total t_y est estimé sans biais par l'estimateur de Horvitz-Thompson

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

C'est un estimateur pondéré, où les *poids de sondage* $d_k = 1/\pi_k$ ne dépendent pas de la variable d'intérêt.

Certaines probabilités d'inclusion peuvent être nulles :

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

ENSAI École nationale de la statistique et de l'analyse de l'information

Enquête "Sans-Domicile 2001" (De Peretti et al., 2006)

Sans-domicile : personne qui dort dans un lieu non prévu pour l'habitation ou prise en charge par un organisme fournissant un hébergement gratuit ou à faible participation.

Méthode d'échantillonnage indirect : sélection d'un échantillon de jours $\times$ services d'aide (hébergement, restauration).
Champ de l'enquête : sans-domicile ayant fréquenté, au moins une fois dans la semaine d'enquête, soit un service d'hébergement, soit une distribution de repas chauds.

Exclut les personnes :

- qui dorment dans la rue pour une période de temps courte et ne font pas appel à un centre ou à une distribution de repas,
- qui ne font pas (ou ne peuvent pas faire) appel au circuit d'assistance.

Autres mesures de précision

Intervalle de confiance

On suppose que $\hat{t}_{y\pi}$ estime sans biais t_y. Alors un intervalle de confiance pour t_y de niveau approximatif 95 % est donné par :

$$IC(t_y) = \left[\hat{t}_{y\pi} \pm 1.96\sqrt{\hat{V}(\hat{t}_{y\pi})}\right],$$

avec $\hat{V}(\hat{t}_{y\pi})$ un estimateur de variance.

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

Coefficient de variation

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

$$CV_p\left[\hat{t}_{y\pi}\right] = \frac{\sqrt{V_p\left(\hat{t}_{y\pi}\right)}}{t_y} \quad \text{estimé par} \quad \hat{CV}\left[\hat{t}_{y\pi}\right] = \frac{\sqrt{v\left(\hat{t}_{y\pi}\right)}}{\hat{t}_{y\pi}}.$$

Il s'agit d'une grandeur sans dimension, plus facile à comparer et à interpréter que la variance.

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

ENSAI École nationale de la statistique et de l'analyse de l'information

En résumé

On sélectionne un échantillon selon un plan de sondage $p(\cdot)$. La connaissance des probabilités d'inclusion permet de produire un estimateur sans biais

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

pour le total t_y.

Un intervalle de confiance de niveau approximatif 95 % est donné par

$$IC(t_y) = \left[\hat{t}_{y\pi} \pm 1.96 \sqrt{\hat{V}(\hat{t}_{y\pi})} \right],$$

avec $\hat{V}(\hat{t}_{y\pi})$ un estimateur de variance.

ENSAI École nationale de la statistique et de l'analyse de l'information

Méthodes d'échantillonnage

Sondage aléatoire simple sans remise

Sondage aléatoire simple sans remise

Il s'agit du plan de sondage qui donne la même probabilité à tous les échantillons de taille n d'être sélectionnés. Les probabilités d'inclusion sont égales à $\pi_k = n/N$.

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

$$\hat{t}_{y\pi} = N\bar{y} \quad \text{avec} \quad \bar{y} = \frac{1}{n}\sum_{k\in S} y_k.$$

La variance de $\hat{t}_{y\pi}$ est égale à

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

avec $f = n/N$ le **taux de sondage**.

ENSAI École nationale de la statistique et de l'analyse de l'information

Variance de la moyenne estimée

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

$$V_p[\bar{y}] = \frac{1-f}{n} S_y^2. \tag{1}$$

Remarques :

- Le facteur $(1-f)$ donne le gain de variance dû au tirage sans remise. On l'appelle *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.

ENSAI École nationale de la statistique et de l'analyse de l'information

Application : détermination de taille d'échantillon

On cherche une taille d'échantillon minimale permettant de respecter avec un niveau de confiance fixé (par exemple de 95 %) une contrainte de précision en termes d'*erreur relative* :

$$P \text{ connu à 8 \% près} \Leftrightarrow \left|\frac{\hat{P}-P}{P}\right| \leq 0.08.$$

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

$$\left|\frac{\hat{P}-P}{P}\right| \leq \gamma \quad \Leftrightarrow \quad n \geq \frac{1}{\frac{1}{N} + \frac{N-1}{N}\left[\frac{\gamma}{1.96}\right]^2 \frac{P}{1-P}}.$$

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

ENSAI École nationale de la statistique et de l'analyse de l'information

Algorithme de sélection pour un SRS (1)

Algorithme 1 Méthode de sélection draw by draw

1. Pour $k = 1, \ldots, n$, sélectionner une unité dans U à probabilités égales parmi les unités qui n'ont pas déjà été tirées.

Inconvénient : méthode lente, qui nécessite n lectures de fichier.

Algorithme de sélection pour un SRS (2)

Algorithme 2 Méthode du tri aléatoire

1. On attribue un nombre aléatoire $u_k \sim U[0,1]$ à chaque unité $k \in U$.
2. On trie la population selon les u_k croissants (ou décroissants).
3. L'échantillon est constitué des n premiers individus de la population triée.

Inconvénient : nécessite un tri du fichier.

ENSAI École nationale de la statistique et de l'analyse de l'information

Méthode du tri aléatoire : exemple

Individu	u_k
1	0.65
2	0.98
3	0.86
4	0.82
5	0.27
6	0.50
7	0.96
8	0.13

$\Rightarrow$

Individu	u_k
8	0.13
5	0.27
6	0.50
1	0.65
4	0.82
3	0.86
7	0.96
2	0.98

$\Rightarrow$

Individu	u_k	I_k
8	0.13	1
5	0.27	1
6	0.50	1
1	0.65	0
4	0.82	0
3	0.86	0
7	0.96	0
2	0.98	0

ENSAI École nationale de la statistique et de l'analyse de l'information

Le sondage aléatoire simple stratifié

Information auxiliaire

On parle d'*information auxiliaire* lorsqu'une information est connue sur l'ensemble de la population, sous forme détaillée ou synthétique.

Il est fréquent de disposer d'une information auxiliaire sur la population, qui va permettre de partitionner la population et d'obtenir un plan de sondage plus efficace que le SRS.

Exemples d'information auxiliaire :

- le sexe et l'âge, pour une enquête auprès d'individus physiques,
- la taille (nombre d'employés) pour les enquêtes-entreprise.

Information auxiliaire : cas de VALPENA

La base Navires de VALPENA (base de sondage) regorge d'informations auxiliaires qui peuvent être utilisées dans la stratification :

- Nom du navire
- Identifiant (immatriculation)
- Coordonnées
- Type de métiers exercés
- Lieu d'immatriculation
- Longueur
- Tonnage

La stratification retenue pour les PE de VALPENA est différente d'une région à l'autre :

- Port/Quartier maritime d'immatriculation
- 1er engin utilisé
- Longueur du navire

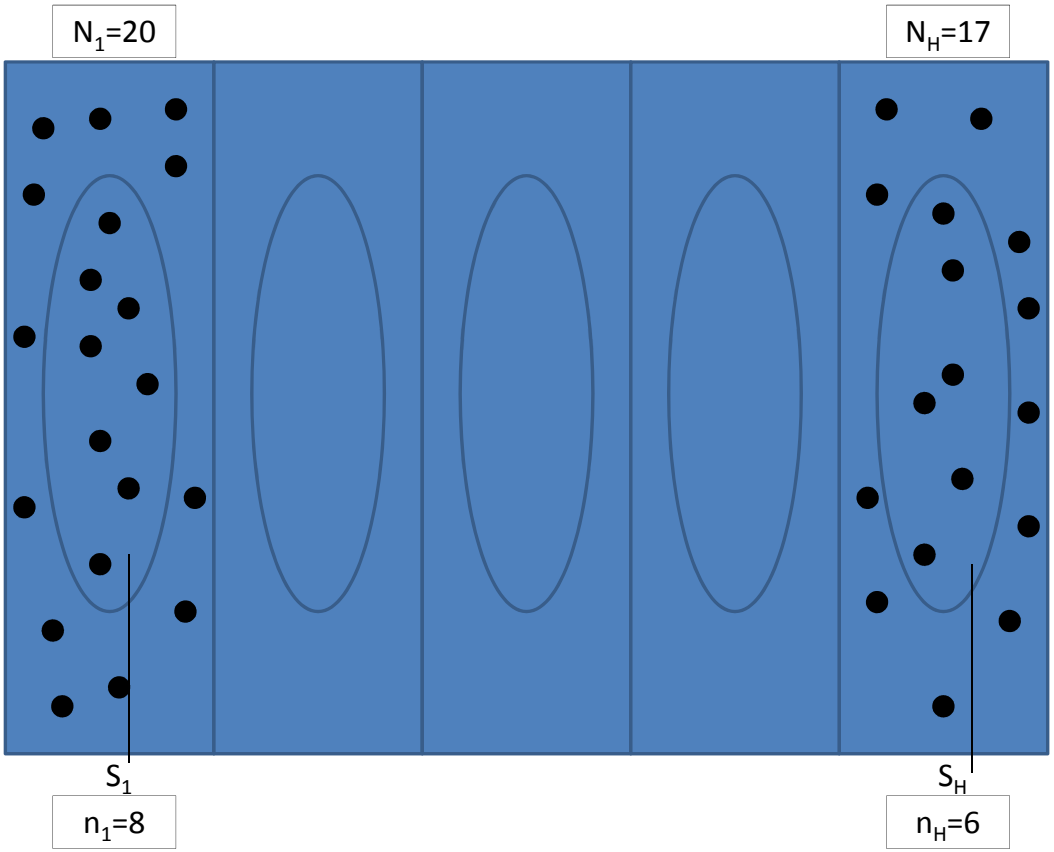
Motivations pour la stratification (Cochran, 1977)

- Précision maîtrisée pour des sous-populations,
  Ex : enquêtes-entreprise stratifiées par tranche de taille × type d'activité
- simplicité administrative (enquêtes conduites par différentes agences),
  Ex : enquête EU-SILC, conduite indépendamment dans chaque pays européen
- plans de sondage adaptés aux sous-populations,
  Ex : enquête de Recensement en population générale + auprès des SDF
- gain global de précision.

Principales questions :

1. Comment construire les strates ?
2. Quelle taille d'échantillon sélectionner dans chaque strate ?
3. Quel plan de sondage utiliser dans chaque strate ?

Notation et sondage stratifié


N_1=20
N_H=17
S_1
S_H
n_1=8
n_H=6

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.

ENSAI École nationale de la statistique et de l'analyse de l'information

Estimation d'un total

Le total t_y est estimé sans biais par

$$\hat{t}_{y\pi} = \sum_{h=1}^{H} N_h \bar{y}_h \quad \text{avec} \quad \bar{y}_h = \frac{1}{n_h} \sum_{k \in S_h} y_k.$$

La variance est donnée par

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

On l'estime par

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

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

ENSAI École nationale de la statistique et de l'analyse de l'information

Allocations pour le tirage stratifié

Allocation d'échantillon entre les strates

On suppose que la taille globale d'échantillon n est fixée, et que les strates ont été définies.

On doit choisir les tailles $n_1, \ldots, n_H$ des sous-échantillons à sélectionner dans chaque strate.

Nous revenons sur quelques allocations classiques pour le sondage aléatoire simple stratifié :

- Allocation Proportionnelle,
- Allocation Optimale.

ENSAI École nationale de la statistique et de l'analyse de l'information

Allocation Proportionnelle

Allocation Proportionnelle

Avec une allocation proportionnelle, le taux de sondage est le même dans chaque strate :

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

On peut le réécrire sous la forme

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

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

Equation de décomposition de la variance

La dispersion de la variable y dans la population U peut se décomposer sous la forme

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

$$\simeq \sum_{h=1}^{H} \frac{N_h}{N} S_{yh}^2 + \sum_{h=1}^{H} \frac{N_h}{N} (\mu_{yh} - \mu_y)^2$$

Le premier terme mesure la dispersion à l'intérieur des strates, alors que le second terme mesure la dispersion entre les strates.

ENSAI École nationale de la statistique et de l'analyse de l'information

Equation de décomposition de la variance

Notons que la dispersion globale S_y^2 est fixée. Le poids de chacune des deux composantes dépend de la variable de stratification choisie.

Exemple

k	1	2	3	4	5	6	7	8
y_k	1	1	1	1	5	5	5	5
x_{1k}	0	0	0	0	1	1	1	1
x_{2k}	0	0	1	1	1	1	0	0

Décomposition de la variance pour S_y^2 :

- si x_{1k} est la variable de stratification,
- si x_{2k} est la variable de stratification.

ENSAI École nationale de la statistique et de l'analyse de l'information

Retour vers l'allocation proportionnelle

La variance de l'estimateur stratifié avec allocation proportionnelle est approximativement donnée par

$$V_p\left[\hat{t}_{y\pi}\right] \simeq N^2 \frac{1-f}{n} S^2_{y,intra},$$

de sorte que :

- le SRS stratifié à allocation proportionnelle est (presque) toujours plus efficace que le SRS,
- la stratification devrait être choisie de façon à ce que la **dispersion à l'intérieur des strates** soit minimisée.

ENSAI École nationale de la statistique et de l'analyse de l'information

Allocation de Neyman

Principe

L'allocation de Neyman donne, pour une stratification donnée et une variable d'intérêt donnée, l'allocation d'échantillon pour laquelle la variance du π-estimateur est minimisée.

On obtient :

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

Notons en particulier que le calcul de cette allocation optimale nécessite la connaissance des dispersions dans les strates.

Principe

L'allocation de Neyman indique qu'il faut sélectionner un échantillon plus grand :

- dans les grandes strates,
- dans les strates présentant une forte dispersion.

L'allocation de Neyman peut conduire à des tailles d'échantillon supérieures aux tailles de strates, si ces dernières présentent une forte dispersion et/ou sont de grande taille.

Dans ce cas :

1. on effectue un recensement dans les strates concernées (on fixe $n_h = N_h$),
2. on recalcule l'allocation d'échantillon dans les autres strates.

Tirage à probabilités inégales

Algorithmes de tirage

Il existe en pratique des dizaines d'algorithmes de tirage permettant de respecter un jeu de probabilités d'inclusion fixé (voir Tillé, 2006). Nous présentons deux de ces algorithmes :

- le tirage systématique,
- la méthode du pivot.

Les différents algorithmes se distinguent par les probabilités d'inclusion d'ordre 2 obtenues, i.e. par la variance des estimateurs. Cependant, il n'existe pas d'algorithme uniformément préférable en termes de variance.

Le choix de la méthode à utiliser dépend de la connaissance que l'on a de la base de sondage mais aussi des contraintes pratiques sur l'échantillonnage.

Le tirage systématique

Principe

C'est une méthode simple et très rapide permettant de sélectionner un échantillon à probabilités inégales et de taille fixe.

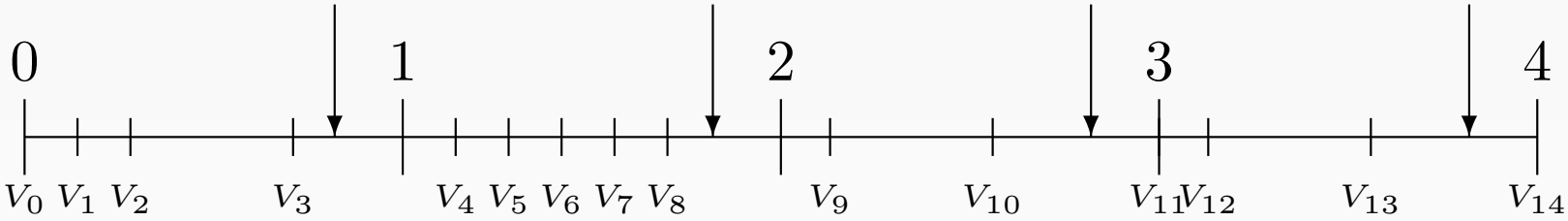
- On répartit les unités sur un segment de longueur n. Chaque unité k est représentée par un sous-segment de longueur π_k.
- On génère un nombre aléatoire u selon une loi uniforme $U[0, 1]$.
- On sélectionne l'unité k_0 dans le sous-segment contient u, puis l'unité k_1 dans le sous-segment contient $1 + u$, ...

ENSAI École nationale de la statistique et de l'analyse de l'information

Exemple

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

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



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

ENSAI École nationale de la statistique et de l'analyse de l'information

Applications du tirage systématique

Exemple 1 : sélection pour contrôle d'un sous-échantillon de questionnaires, arrivant à flux tendu.

Exemple 2 : enquête auprès des clients entrant dans un magasin.

Exemple 3 : tirage de logements dans un pâté de maison lors d'une enquête ménage.

La méthode du pivot

La méthode du pivot (Deville et Tillé, 1998)

Elle est basée sur des duels entre unités. A l'étape 1, les unités 1 et 2 s'affrontent :

- si $\pi_1 + \pi_2 \leq 1$, l'une des unités est éliminée et l'autre survit avec la somme des probabilités :

$$(\pi_1, \pi_2) = \begin{cases} (\pi_1 + \pi_2, 0) & \text{avec proba } \frac{\pi_1}{\pi_1+\pi_2}, \\ (0, \pi_1 + \pi_2) & \text{avec proba } \frac{\pi_2}{\pi_1+\pi_2}. \end{cases}$$

- si $\pi_1 + \pi_2 > 1$, l'une des unités est sélectionnée et l'autre survit avec la probabilité résiduelle :

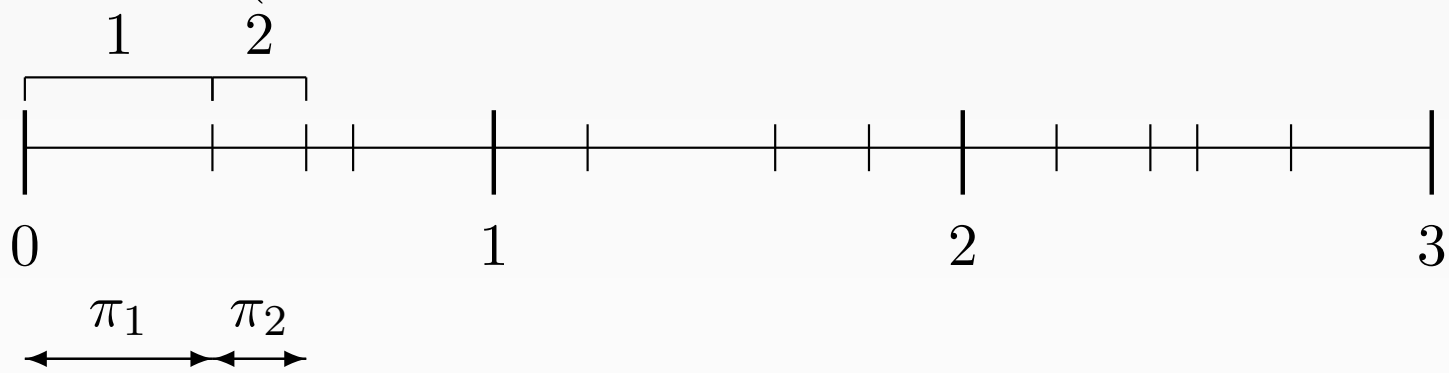
$$(\pi_1, \pi_2) = \begin{cases} (1, \pi_1 + \pi_2 - 1) & \text{avec proba } \frac{1-\pi_2}{2-\pi_1-\pi_2}, \\ (\pi_1 + \pi_2 - 1, 1) & \text{avec proba } \frac{1-\pi_1}{2-\pi_1-\pi_2}. \end{cases}$$

A l'étape t, le survivant affronte l'unité suivante $t+1$ selon le même principe. A l'étape $N-1$, un échantillon de taille n a été sélectionné, et les probabilités d'inclusion sont exactement respectées.

Exemple

Population U de taille $N = 11$, avec $n = 3$ et

$$\pi = (0.4 \;\; 0.2 \;\; 0.1 \;\; 0.5 \;\; 0.4 \;\; 0.2 \;\; 0.4 \;\; 0.2 \;\; 0.1 \;\; 0.2 \;\; 0.3)^\top .$$



Nous avons $(\pi_1, \pi_2) = (0.4, 0.2) = \begin{cases} (0.6, 0) & \text{avec proba } 0.4/0.6, \\ (0, 0.6) & \text{avec proba } 0.2/0.6 \end{cases}$

Si l'unité 2 survit, nous obtenons

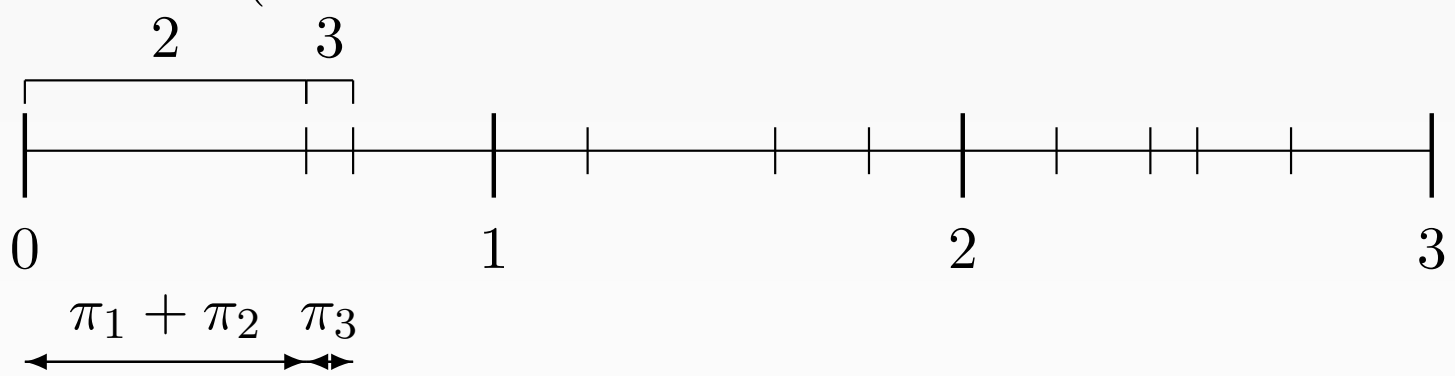
$$\pi^{(1)} = (0 \;\; 0.6 \;\; 0.1 \;\; 0.5 \;\; 0.4 \;\; 0.2 \;\; 0.4 \;\; 0.2 \;\; 0.1 \;\; 0.2 \;\; 0.3)^\top .$$

ENSAI École nationale de la statistique et de l'analyse de l'information

Exemple

Population U de taille $N = 11$, avec $n = 3$ et

$$\pi = (0.4 \;\; 0.2 \;\; 0.1 \;\; 0.5 \;\; 0.4 \;\; 0.2 \;\; 0.4 \;\; 0.2 \;\; 0.1 \;\; 0.2 \;\; 0.3)^\top .$$



Nous avons $(\pi_2^{(1)}, \pi_3^{(1)}) = (0.6, 0.1) = \begin{cases} (0.7, 0) & \text{avec proba } 0.6/0.7, \\ (0, 0.7) & \text{avec proba } 0.1/0.7 \end{cases}$.

Si l'unité 3 survit, nous obtenons

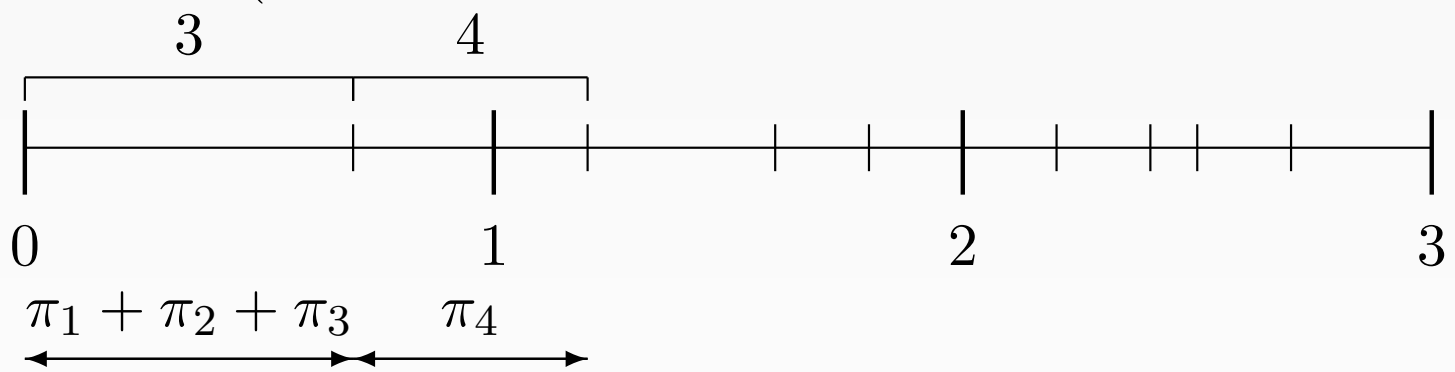
$$\pi^{(2)} = (0 \;\; 0 \;\; 0.7 \;\; 0.5 \;\; 0.4 \;\; 0.2 \;\; 0.4 \;\; 0.2 \;\; 0.1 \;\; 0.2 \;\; 0.3)^\top .$$

ENSAI École nationale de la statistique et de l'analyse de l'information

Exemple

Population U de taille $N = 11$, avec $n = 3$ et

$$\pi = (0.4 \;\; 0.2 \;\; 0.1 \;\; 0.5 \;\; 0.4 \;\; 0.2 \;\; 0.4 \;\; 0.2 \;\; 0.1 \;\; 0.2 \;\; 0.3)^\top .$$



Nous avons $(\pi_3^{(2)}, \pi_4^{(2)}) = (0.7, 0.5) = \begin{cases} (1, 0.2) & \text{avec proba } 0.5/(2 - 1.2), \\ (0.2, 1) & \text{avec proba } 0.3/(2 - 1.2) \end{cases}$

Si l'unité 3 survit, nous obtenons

$$\pi^{(3)} = (0 \;\; 0 \;\; 1 \;\; 0.2 \;\; 0.4 \;\; 0.2 \;\; 0.4 \;\; 0.2 \;\; 0.1 \;\; 0.2 \;\; 0.3)^\top , \ldots$$

ENSAI École nationale de la statistique et de l'analyse de l'information

Bibliographie

- Ardilly, P. (2006), *Les Techniques de Sondage*, Technip, Paris.
- Cochran, W.G (1977), *Sampling Techniques*, Wiley, New-York.
- De Peretti, P. et al (2006). *L'enquête sans-domicile 2001*. Insee Méthodes, 116, Paris.
- Särndal, C.-E., and Swensson, B., and Wretman, J.H. (1992), *Model Assisted Survey Sampling*, Springer-Verlag, New-York.
- Tillé, Y. (2006). *Sampling algorithms*, Springer, New-York.

ENSAI École nationale de la statistique et de l'analyse de l'information