Université Montpellier 2

Master MIND

Sondages et Enquêtes

Mathieu Ribatet

Table des matières

Liste des symboles

$1_{\{x\in A\}}$	Variable indicatrice
$\hat{t}_{y,\pi}, \hat{\mu}_{y,\pi}$	π–estimateur du total t_y et de la moyenne μ_y
$\mathbb{E}(X)$	Espérance de la variable aléatoire X
$\mathcal{U}$	Population
$\mathcal{U}_h$	Strate de la population
$\mathcal{U}_i$	Grappe de la population
$\mathscr{L}$	Lagrangien du problème d'optimisation sous contraintes
$\mathscr{S}, \tilde{\mathscr{S}}$	Ensemble des échantillons non ordonnés sans remise et ordonnés avec remise
$\pi_k, \pi_{k\ell}$	Probabilités d'inclusion d'ordre un et deux
σ_y^2	Variance du caractère y sur la population
$\mathrm{Var}(X)$	Variance de la variable aléatoire X
k	Unité de la population
N	Taille de la population
N_h	Taille de la strate $\mathcal{U}_h$
N_i	Taille de la grappe $\mathcal{U}_i$
n_S, n	Taille de l'échantillon S
$p(\cdot)$	Plan de sondage
S	Échantillon
S_y^2	Variance corrigée du caractère y sur la population
t_y, μ_y	Total et moyenne du caractère y sur la population
$t_{y,h}, \mu_{y,h}$	Total et moyenne du caractère y sur la strate/grappe $\mathcal{U}_h$
y	Caractère défini sur la population

Remarques préliminaires

Ce polycopié se veut une introduction à la théorie des sondages. Je dois avouer que les résultats présentés ici sont outrageusement pompés du livre de Yves Tillé *Théorie des Sondages : Échantillonnage et estimation en populations finies*. L'étudiant motivé pourra donc s'orienter vers ce livre s'il veut approfondir ses connaissances sur la thématique — ou tout simplement voir une rédaction bien meilleure que la mienne !

En lisant se document pour la première fois (j'espère que ce ne sera pas quelques jours seulement avant l'examen), vous constaterez qu'il y a des "trous" dans le texte. Ces trous masquent les démonstrations ou encore les solutions aux exercices et seront complétés lors des séances de cours. Comme quoi ça vaut la peine de venir en cours non ?

Le cours comprends énormément de formules. Si je peux me permettre un conseil, n'essayez pas d'apprendre par coeur toutes ces formules mais seulement les plus importantes. En effet, les formules "secondaires" se retrouvent généralement très rapidement par de simples calculs.

Enfin si vous trouvez des coquilles (ce qui est plus que fort probable !) dans ce support de cours, j'apprécierai que vous me les fassiez connaître.

Bonne lecture et travail donc...

Mathieu Ribatet

Chapitre 1

Formalisation mathématique d'un sondage

Ce chapitre pose les bases de la théorie des sondage en introduisant le vocabulaire, la notion d'aléatoire spécificique aux sondages et les estimateurs principaux.

1.1 Population, Caractère et Fonction d'intérêt

En sondage on s'intéresse à une **population** (ou **univers**) finie $\mathcal{U}$ constituée de N **unités** (ou **individus**) notées $u_1, \ldots, u_N$. On supposera que ces unités sont **identifiables** si chacune d'entre elle peut se voir attribuer un numéro d'identification **unique**. Ainsi par abus de notations, on écrira indifféremment

$$\mathcal{U} = \{u_1, \ldots, u_N\}, \qquad \mathcal{U} = \{1, \ldots, N\}.$$

Remarque. La définition de la population $\mathcal{U}$ est souvent problématique. Par exemple, pour l'étude des *habitants de plus de 18 ans d'un pays* la population n'est pas parfaitement identifiée si l'on ne suppose pas une date de référence pour l'âge et si l'on ne précise pas certains critères comme : France métropolitaine, Résidants ou Nationalité, ...

L'objectif d'un sondage ne porte pas sur les unités elles mêmes mais plutôt sur un **caractère** y qui est mesuré sur chaque unité de $\mathcal{U}$. Ainsi la valeur prise par le caractère y sur la $k$ème unité est notée y_k.

Remarque. Les valeurs prises par le caractère **ne sont pas aléatoires**. C'est d'ailleurs pour cela que l'on parle de *caractère* plutôt que de *variable* ; cette dernière ayant une connotation aléatoire.

Dans un monde idéal, on aimerait donc connaître le **vecteur paramètre** $\mathbf{y}_N = (y_1, \ldots, y_N)$; mais il est clair que ceci relève de l'impossible. Comment connaître ces N valeurs à partir de n observations ($n \ll N$) ? Souvent on visera seulement (et c'est déjà bien suffisant) un résumé du vecteur paramètre $\mathbf{y}_N$ comme par exemple la moyenne, une proportion, ... Plus formellement, on souhaite estimer une fonction θ de $\mathbf{y}_N$

$$\theta = \theta(y_k \colon k \in \mathcal{U}).$$

Exemple 1.1.1. La fonction d'intérêt θ peut être
- un total : $t_y = \sum_{k\in\mathcal{U}} y_k$;
- une moyenne : $\mu_y = N^{-1} \sum_{k\in\mathcal{U}} y_k$;
- un ratio : $R = t_y/t_x$ où x est un deuxième caractère d'intérêt.

1.2 Échantillon

Dans ce cours nous allons croiser essentiellement deux types d'échantillons : avec remise et ordonné et sans remise ni ordre.

Exemple 1.2.1. Soit une population $\mathcal{U} = \{1, 2\}$. L'ensemble des échantillons ordonnés avec remise est

$$\tilde{\mathscr{S}} = \{(1), (2), (1, 1), (1, 2), (2, 1), (2, 2), (1, 1, 1), (1, 1, 2), \ldots\}.$$

En particulier puisqu'il y a remise, la taille de l'échantillon peut être supérieure à la taille de la population !

Exemple 1.2.2. Soit une population $\mathcal{U} = \{1, 2, 3\}$. L'ensemble des échantillons non ordonnés et sans remise est

$$\mathscr{S} = \{\{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}.$$

Il est commode de se représenter un échantillon non ordonné et sans remise comme un **sous ensemble non vide** de $\mathcal{U}$. En effet un ensemble est par définition non ordonné et sans répétition. Ainsi l'ensemble des échantillons non ordonnés et sans remise est l'ensemble des parties non vides de $\mathcal{U}$, i.e.,

$$\mathscr{S} = \{s \colon s \subset \mathcal{U}\} \setminus \emptyset.$$

Par conséquent la taille de l'échantillon est au plus égale à la taille de la population et $|\mathscr{S}| = 2^N - 1$.

Remarque. Clairement il est possible de passer de $\tilde{\mathscr{S}}$ à $\mathscr{S}$ en supprimant l'information sur l'ordre et la multiplicité à l'aide d'une **fonction de réduction** $r \colon \tilde{\mathscr{S}} \mapsto \mathscr{S}$.

Exemple 1.2.3. Pour $\mathcal{U} = \{1, 2, 3\}$, on a

$$r\{(1, 1, 2)\} = r\{(1, 2)\} = r\{(2, 1)\} = r\{(1, 2, 2)\} = \{1, 2\}.$$

1.3 Plan de sondage

Définition 1.3.1. Un plan de sondage non ordonné et sans remise p est une **loi de probabilité** sur $\mathscr{S}$, i.e,

$$p(s) \geq 0, \qquad s \in \mathscr{S},$$

et

$$\sum_{s \in \mathscr{S}} p(s) = 1.$$

De même on définit un plan de sondage ordonné avec remise $\tilde{p}$ comme une loi de probabilité sur $\tilde{\mathscr{S}}$.

Clairement la fonction de réduction r permet de définir un plan de sondage sur $\mathscr{S}$ à l'aide d'un plan de sondage sur $\tilde{\mathscr{S}}$, i.e.,

$$p(s) = \sum_{\tilde{s} \in \tilde{\mathscr{S}}} \tilde{p}(\tilde{s}) 1_{\{r(\tilde{s}) = s\}}, \qquad s \in \mathscr{S}.$$

Exemple 1.3.1. Pour $\mathcal{U} = \{1, 2, 3\}$, on considère le plan de sondage consistant à sélectionner 2 unités avec remise et probabilités égales.

— Le plan de sondage sur $\tilde{\mathscr{S}}$ est alors

$$\begin{aligned}
&\tilde{p}\{(1,1)\} = 1/9, && \tilde{p}\{(1,2)\} = 1/9, && \tilde{p}\{(1,3)\} = 1/9,\\
&\tilde{p}\{(2,1)\} = 1/9, && \tilde{p}\{(2,2)\} = 1/9, && \tilde{p}\{(2,3)\} = 1/9,\\
&\tilde{p}\{(3,1)\} = 1/9, && \tilde{p}\{(3,2)\} = 1/9, && \tilde{p}\{(3,3)\} = 1/9,
\end{aligned}$$

— et celui sur $\mathscr{S}$ est

$$\begin{aligned}
&p(\{1\}) = 1/9, && p(\{1,2\}) = 2/9, && p(\{1,3\}) = 2/9\\
&p(\{2\}) = 1/9, && p(\{2,3\}) = 2/9, && p(\{3\}) = 1/9.
\end{aligned}$$

Notez que la taille de l'échantillon pour le plan de sondage sur $\mathscr{S}$ est aléatoire.

Puisqu'un plan de sondage n'est rien d'autre qu'une loi de probabilité, nous pouvons définir des échantillons aléatoires S et $\tilde{S}$, i.e., des variables aléatoires à valeurs dans $\mathscr{S}$ et $\tilde{\mathscr{S}}$ respectivement. Les lois de S et $\tilde{S}$ sont donc données par

$$\Pr(S = s) = p(s), \quad s \in \mathscr{S}, \qquad \text{et} \qquad \Pr(\tilde{S} = \tilde{s}) = \tilde{p}(\tilde{s}), \quad \tilde{s} \in \tilde{\mathscr{S}}.$$

Remarque. Comme nous l'avons vu dans l'exemple précédent, la taille de l'échantillon notée n_S peut être aléatoire. Lorsque $\mathrm{Var}(n_S) = 0$, l'échantillon est dit de **taille fixe**.

1.4 Probabilités d'inclusion

Soit un échantillon aléatoire S, la variable aléatoire $1_{\{k\in S\}}$, $k \in \mathcal{U}$, nous sera très utile. Notons que c'est bien une variable aléatoire puisque S est aléatoire.

Définition 1.4.1. La probabilité d'inclusion de la $k$ème unité, notée π_k, correspond à la probabilité que cette $k$ème unité appartienne à l'échantillon, i.e.,

$$\pi_k = \Pr(k \in S) = \sum_{s\in\mathscr{S}} p(s) 1_{\{k\in s\}} = \sum_{s\ni k} p(s), \qquad k \in \mathcal{U}.$$

Notons également que par définition, $\pi_k = \mathbb{E}(1_{\{k\in S\}})$.

De même nous pouvons définir des probabilités d'inclusion d'ordre supérieur.

Définition 1.4.2. La probabilité d'inclusion d'ordre 2 est la probabilité que deux unités distinctes appartiennent simultanément à un échantillon, i.e.,

$$\pi_{k\ell} = \Pr(k \in S, \ell \in S) = \sum_{s\ni k,\ell} p(s), \qquad k, \ell \in \mathcal{U}, \quad k \neq \ell.$$

Notons que comme précédemment, $\pi_{k\ell} = \mathbb{E}(1_{\{k\in S\}} 1_{\{\ell\in S\}})$.

On a

$$\mathrm{Var}(1_{\{k\in S\}}) = \mathbb{E}(1^2_{\{k\in S\}}) - \mathbb{E}(1_{\{k\in S\}})^2 = \pi_k(1-\pi_k),$$

et

$$\mathrm{Cov}(1_{\{k\in S\}}, 1_{\{\ell\in S\}}) = \mathbb{E}(1_{\{k\in S\}} 1_{\{\ell\in S\}}) - \mathbb{E}(1_{\{k\in S\}})\mathbb{E}(1_{\{\ell\in S\}}) = \pi_{k\ell} - \pi_k\pi_\ell,$$

avec $k, l \in \mathcal{U}$, $k \neq \ell$.

Dans la suite on notera

$$\Delta_{k\ell} = \begin{cases} \mathrm{Cov}(1_{\{k\in S\}}, 1_{\{\ell\in S\}}), & k \neq \ell \\ \mathrm{Var}(1_{\{k\in S\}}), & k = \ell. \end{cases}$$

1.5 Plans simples et de taille fixe

La théorie des sondages revient souvent à caractériser certaines propriétés de plan de sondage donnés. Ici nous nous intéressons aux plans dit **simples** et les plans de **taille fixe**.

Définition 1.5.1. Un plan est dit **simple** si tous les échantillons de même taille ont la même probabilité d'être sélectionnés.

Définition 1.5.2. Un plan est dit de **taille fixe** si $\mathrm{Var}(|S|) = \mathrm{Var}(n_S) = 0$, où $|A|$ représente la cardinalité d'un ensemble A. On notera alors $n = n_S$ la taille de l'échantillon.

Les plans de taille fixe ont des probabilités d'inclusion bien spécifiques.

Théorème 1.5.1. *Si un plan est de taille fixe n, alors*

$$\begin{aligned}
&\sum_{k\in\mathcal{U}} \pi_k = n,\\
&\sum_{\substack{k\in\mathcal{U}\\ k\neq\ell}} \pi_{k\ell} = (n-1)\pi_\ell, \qquad \ell \in \mathcal{U},\\
&\sum_{k\in\mathcal{U}} \Delta_{k\ell} = 0, \qquad \ell \in \mathcal{U}.
\end{aligned}$$

Démonstration.

□

Définition 1.5.3. Un plan sans remise est dit **simple** si tous les échantillons de même taille ont la même probabilité d'être sélectionnés, i.e.,

$$p(s_1) = p(s_2), \qquad s_1, s_2 \in \mathscr{S},\ |s_1| = |s_2|.$$

Remarque. Clairement on a $\binom{N}{n}$ échantillons (non ordonnés) de taille n dans $\mathcal{U}$. Ainsi si le plan est simple **et** de taille fixe on a pour tout $s \in \mathscr{S}$

$$p(s) = \begin{cases} \binom{N}{n}^{-1}, & |s| = n \\ 0, & \text{sinon,} \end{cases}$$

avec $\binom{N}{n} = N!/\{n!(N-n)!\}$.

En revanche, si le plan n'est pas de taille fixe on a

$$p(s) = \binom{N}{n}^{-1} \Pr(|S| = n),$$

où $|s| = n$.

1.6 Le π–estimateur

C'est sans aucun doute l'estimateur qu'il faut à tout prix connaître lorsque l'on s'intéresse aux sondages.

1.6.1 Estimation d'un total et d'une moyenne

Horvitz et Thompson (1952) ont introduit un estimateur linéaire sans biais d'un total t_y pour tout plan de sondage

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

Cet estimateur est appelé le **π–estimateur**, l'estimateur d'**Horvitz–Thompson** ou encore l'estimateur **des valeurs dilatées**.

Théorème 1.6.1. *Si $\pi_k > 0$ pour tout $k \in \mathcal{U}$, alors $\hat{t}_{y,\pi}$ estime t_y sans biais.*

Démonstration.

□

Remarque. Si certaines probabilités d'inclusion sont nulles alors l'estimateur est biaisé puisque

$$\mathbb{E}(\hat{t}_{y,\pi}) = \mathbb{E}\left(\sum_{k \in S} \frac{y_k}{\pi_k}\right) = \mathbb{E}\left(\sum_{\substack{k \in \mathcal{U} \\ \pi_k > 0}} \frac{y_k}{\pi_k} 1_{\{k \in S\}}\right) = \sum_{\substack{k \in \mathcal{U} \\ \pi_k > 0}} \frac{y_k}{\pi_k} \pi_k = t_y - \sum_{\substack{k \in \mathcal{U} \\ \pi_k = 0}} y_k.$$

Notons que, lors de la deuxième égalité, la restriction aux unités telles que $\pi_k > 0$ sous le signe de sommation est justifiée par le fait qu'une unité dont la probabilité d'inclusion d'ordre un est nulle n'appartiendra jamais à l'échantillon aléatoire S.

Nous avons introduit le π–estimateur pour estimer le total t_y mais nous pouvons également l'utiliser pour estimer la moyenne μ_y par

$$\hat{\mu}_{y,\pi} = \frac{1}{N} \sum_{k \in S} \frac{y_k}{\pi_k}.$$

Notons toutefois que pour utiliser cet estimateur il faut que la taille de la population N soit connue—ce n'est malheureusement pas toujours le cas...

Cela dit puisque $N = \sum_{k \in \mathcal{U}} 1$, on peut estimer N par Horvitz–Thompson, i.e.,

$$\hat{N}_\pi = \sum_{k \in S} \frac{1}{\pi_k}.$$

1.6.2 Variance du π–estimateur

Il est également possible de connaître la variance du π–estimateur.

Théorème 1.6.2. *Soit $\hat{t}_{y,\pi}$ le π–estimateur d'un total t_y. Si $\pi_k > 0$ pour tout $k \in \mathcal{U}$, alors*

$$\mathrm{Var}(\hat{t}_{y,\pi}) = \sum_{k,\ell \in \mathcal{U}} \frac{y_k y_\ell}{\pi_k \pi_\ell} \Delta_{k\ell}.$$

Démonstration.

□

1.6.3 Variance pour les plans de taille fixe

Dans le cas de plans de **taille fixe**, on peut réécrire la variance du π–estimateur sont une forme différente.

Théorème 1.6.3. *Soit $\hat{t}_{y,\pi}$ le π–estimateur d'un total t_y. Si le plan est de taille fixe et que $\pi_k > 0$ pour tout $k \in \mathcal{U}$, alors*

$$\mathrm{Var}(\hat{t}_{y,\pi}) = -\frac{1}{2} \sum_{\substack{k,\ell \in \mathcal{U} \\ k \neq \ell}} \left(\frac{y_k}{\pi_k} - \frac{y_\ell}{\pi_\ell} \right)^2 \Delta_{k\ell}.$$

Démonstration.

□

1.6.4 Estimation de la variance du π–estimateur

L'idée de base du π–estimateur peut être naturellement étendue au contexte des fonctions de deux variables $f(\cdot,\cdot)$.

Théorème 1.6.4. *Soit $f(\cdot,\cdot)$ une fonction de deux variables quelconque. Si $\pi_{k\ell} > 0$, pour tout $k, \ell \in \mathcal{U}$ $k \neq \ell$, alors*

$$\sum_{\substack{k,\ell\in\mathcal{U}\\ k\neq\ell}} \frac{g(y_k, y_\ell)}{\pi_{k\ell}} 1_{\{k\in S, \ell\in S\}}$$

est un estimateur sans biais de

$$\sum_{\substack{k,\ell\in\mathcal{U}\\ k\neq\ell}} \frac{g(y_k, y_\ell)}{\pi_{k\ell}}.$$

Démonstration.

□

On peut donc se servir du théorème précédent afin de construire un estimateur sans biais de $\mathrm{Var}(\hat{t}_{y,\pi})$. On a donc à partir de l'expression donnée en Section 1.6.2 l'estimateur

$$\begin{aligned}\widehat{\mathrm{Var}}(\hat{t}_{y,\pi}) &= \sum_{k\in\mathcal{U}} \frac{y_k^2 \pi_k^{-2} \Delta_{kk}}{\pi_k} 1_{\{k\in S\}} + \sum_{\substack{k,\ell\in\mathcal{U}\\ k\neq\ell}} \frac{y_k y_\ell \pi_k^{-1} \pi_\ell^{-1} \Delta_{k\ell}}{\pi_{k\ell}} 1_{\{k\in S, \ell\in S\}} \\ &= \sum_{k\in\mathcal{U}} \frac{y_k^2 (1-\pi_k)}{\pi_k^2} 1_{\{k\in S\}} + \sum_{\substack{k,\ell\in\mathcal{U}\\ k\neq\ell}} \frac{y_k y_\ell}{\pi_k \pi_\ell \pi_{k\ell}} \Delta_{k\ell} 1_{\{k\in S, \ell\in S\}}.\end{aligned}$$

Si le plan est à **taille fixe** alors nous pouvons utiliser l'expression donnée lors de la Section 1.6.3 ; ce qui nous conduit à l'estimateur

$$\widehat{\mathrm{Var}}(\hat{t}_{y,\pi}) = -\frac{1}{2}\sum_{\substack{k,\ell\in\mathcal{U}\\ k\neq\ell}}\left(\frac{y_k}{\pi_k}-\frac{y_\ell}{\pi_\ell}\right)^2\frac{\Delta_{k\ell}}{\pi_{k\ell}}1_{\{k\in S,\ell\in S\}}.$$

Ce dernier estimateur est appelé l'estimateur de Sen–Yates–Grundy, noms des personnes l'ayant trouvé.

Remarque. Cet estimateur est sans biais uniquement lorsque le plan est de taille fixe et il n'est pas difficile de voir que l'estimateur sera toujours positif dès lors que $\Delta_{k\ell} \leq 0$ pour tout $k, \ell \in \mathcal{U}$, $k \neq \ell$. C'est la condition de Sen–Yates–Grundy.

1.7 L'estimateur de Hájek

Bien que le π–estimateur soit très largement utilisé, il existe certaines situations où ce dernier se comporte pas très bien... Afin d'illustrer nos propos, supposons que

$$\mathrm{Var}\left(\sum_{k\in\mathcal{U}}\frac{1}{\pi_k}1_{\{k\in S\}}\right)\neq 0.$$

Remarque. Ceci est par exemple le cas lorsque la taille de l'échantillon est aléatoire.

Supposons de plus que $y_k = c$ pour tout $k \in \mathcal{U}$. Alors le π–estimateur de la moyenne μ_y est alors

$$\hat{\mu}_{y,\pi} = \frac{c}{N}\sum_{k\in\mathcal{U}}\frac{1}{\pi_k}1_{\{k\in S\}},$$

et nous concluons que $\hat{\mu}_{y,\pi}$ n'est pas égale à c mais est une variable aléatoire d'espérance c. Avouons que c'est une propriété assez embarrassante.

L'estimateur de Hájek a été introduit afin de remédier à ce problème et est donné par

$$\hat{\mu}_{y,H} = \left(\sum_{k\in\mathcal{U}}\frac{1}{\pi_k}1_{\{k\in S\}}\right)^{-1}\sum_{k\in\mathcal{U}}\frac{y_k}{\pi_k}1_{\{k\in S\}}.$$

Remarque. L'estimateur de Hájek correspond à un ratio de deux variables aléatoires. Le calcul de ses moments est alors compliqué voire impossible.

Évidemment on peut étendre cet estimateur pour l'estimation d'un total t_y en posant

$$\hat{t}_{y,H} = N\left(\sum_{k\in\mathcal{U}}\frac{1}{\pi_k}1_{\{k\in S\}}\right)^{-1}\sum_{k\in\mathcal{U}}\frac{y_k}{\pi_k}1_{\{k\in S\}},$$

dès lors que N est connu bien évidemment.

Chapitre 2

Les plans simples

Ce chapitre traite exclusivement des plans simples. Il est important de bien maîtriser ces plans car ils forment souvent la base de plans de sondage plus complexes, tel que les plans stratifiés ou par grappes. A bien connaître donc !

2.1 Plans simples sans remise

2.1.1 Plan de sondage et probabilités d'inclusion

Un plan est dit **simple** si tous les échantillons **de même taille** ont la même probabilité d'être sélectionnés. En conséquence, il n'existe qu'un seul plan simple de **taille fixe** n.

Définition 2.1.1. Un plan de **taille fixe** n est dit **simple sans remise** si

$$p(s) = \begin{cases} \binom{N}{n}^{-1}, & |s| = n, \\ 0, & \text{sinon}, \end{cases}$$

avec $n \in \{1, \dots, N\}$.

Comme vous le savez maintenant, il est souvent utile pour nous statisticiens de connaître les probabilités d'inclusion—afin de pouvoir établir le π–estimateur et sa variance par exemple.

Ces probabilités d'inclusions se calculent facilement. En effet

$$\pi_k = \sum_{s \ni k} p(s) = \underbrace{\binom{N-1}{n-1}}_{\substack{\text{nb. d'échantillons} \\ \text{contenant } k}} \binom{N}{n}^{-1} = \frac{(N-1)!}{(n-1)!(N-n)!} \frac{n!(N-n)!}{N!} = \frac{n}{N}$$

$$\pi_{k\ell} = \sum_{s:\, k,\ell \in s} p(s) = \underbrace{\binom{N-2}{n-2}}_{\substack{\text{nb. échantillons} \\ \text{contenant } k \text{ et } \ell}} \binom{N}{n}^{-1} = \frac{(N-2)!}{(n-2)!(N-n)!} \frac{n!(N-n)!}{N!} = \frac{n(n-1)}{N(N-1)}$$

Remarque. Notons que $\pi_{k\ell} \neq \pi_k \pi_\ell$, indiquant une dépendance entre les unités choisies dû au tirage sans remise.

Des deux expressions précédentes, on en déduit

$$\Delta_{k\ell} = \begin{cases} \frac{n(n-1)}{N(N-1)} - \frac{n^2}{N^2} = -\frac{n(N-n)}{N^2(N-1)}, & k \neq \ell, \\ \frac{n}{N}\left(1 - \frac{n}{N}\right) = \frac{n(N-n)}{N^2}, & k = \ell. \end{cases}$$

2.1.2 Le π–estimateur pour ces plans

A l'aide des probabilités d'inclusions de la Section 2.1.1, nous pouvons donner une version plus explicite du π–estimateur. Le π–estimateur d'une moyenne μ_y

$$\hat{\mu}_{y,\pi} = \frac{1}{N}\sum_{k\in\mathcal{U}} \frac{y_k}{\pi_k} 1_{\{k\in S\}} = \frac{N}{nN}\sum_{k\in\mathcal{U}} y_k 1_{\{k\in S\}} = \overline{y},$$

et le π–estimateur du total t_y est évidemment $\hat{t}_{y,\pi} = N\overline{y}$, avec

$$\overline{y} = \frac{1}{n}\sum_{k\in\mathcal{U}} y_k 1_{\{k\in S\}}.$$

Rappelons que puisque le plan est à taille fixe et que les probabilités d'inclusions des deux premiers ordres sont strictement positives, on peut utiliser la formule de la variance de $\hat{\mu}_{y,\pi}$ trouvée par Sen–Yates–Grundy, i.e.,

$$\begin{aligned}\mathrm{Var}(\hat{\mu}_{y,\pi}) &= -\frac{1}{2N^2}\sum_{\substack{k,\ell\in\mathcal{U}\\ k\neq\ell}} \left(\frac{y_k}{\pi_k} - \frac{y_\ell}{\pi_\ell}\right)^2 \Delta_{k\ell} = \frac{1}{2N^2}\times\frac{N-n}{n(N-1)}\sum_{\substack{k,\ell\in\mathcal{U}\\ k\neq\ell}} (y_k - y_\ell)^2 \\ &= \frac{N-n}{nN} S_y^2,\end{aligned}$$

avec

$$S_y^2 = \frac{1}{N-1}\sum_{k\in\mathcal{U}} (y_k - \mu_y)^2 = \frac{1}{2N(N-1)}\sum_{\substack{k,\ell\in\mathcal{U}\\ k\neq\ell}} (y_k - y_\ell)^2.$$

Remarque. La variance précédente peut également s'écrire

$$\mathrm{Var}(\hat{\mu}_{y,\pi}) = \left(1 - \frac{n}{N}\right)\frac{S_y^2}{n}.$$

Le terme S_y^2/n correspond à la variance d'une moyenne empirique pour les statistiques inférentielles classique alors que le premier terme $(1 - n/N)$ correspond au **facteur de correction en population finie**. On appelle également le ratio $f = n/N$ le **taux de sondage**.

De l'expression précédente, on en déduit directement la variance du π–estimateur du total t_y

$$\mathrm{Var}(\hat{t}_{y,\pi}) = N(N-n)\frac{S_y^2}{n}.$$

Théorème 2.1.1. *Pour un plan de taille fixe n, simple et sans remise la variance corrigée de la population S_y^2 est estimée sans biais par*

$$\widehat{S_y^2} = \frac{1}{n-1}\sum_{k\in\mathcal{U}} (y_k - \bar{y})^2 1_{\{k\in S\}}.$$

Démonstration.

□

Au final, on peut estimer sans biais la variance de $\hat{\mu}_{y,\pi}$ pour ces plans particuliers par

$$\widehat{\mathrm{Var}}(\hat{\mu}_{y,\pi}) = \frac{N-n}{N}\frac{\widehat{S_y^2}}{n},$$

et pour le π–estimateur du total $\hat{t}_{y,\pi}$

$$\widehat{\mathrm{Var}}(\hat{t}_{y,\pi}) = N(N-n)\frac{\widehat{S_y^2}}{n}.$$

2.2 Plans simples avec remise

Le plan de taille fixe n, simple et avec remise correspond au cadre de la statistique inférentielle usuelle. En effet le plan de sondage consiste à sélectionner une unité aléatoire avec probabilités égales $1/N$ et de recommencer l'opération n fois indépendemment. On se ramène donc au cadre de variable aléatoire indépendantes et identiquement distribuées de moyenne

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

et de variance

$$\sigma_y^2 = \frac{1}{N}\sum_{k\in\mathcal{U}}(y_k-\mu_y)^2.$$

Nous le savons déjà mais la moyenne sur la population μ_y est estimée sans biais par

$$\hat{\mu}_y = \frac{1}{n}\sum_{k\in\mathcal{U}} y_k 1_{\{k\in S\}} = \bar{y}.$$

En effet

$$\mathbb{E}(\hat{\mu}_y) = \frac{1}{n}\sum_{k\in\mathcal{U}} y_k \frac{n}{N} = \mu_y.$$

De plus, puisque les y_k de l'échantillon sont sélectionnées indépendemment et sont de même loi,

$$\mathrm{Var}(\hat{\mu}_y) = \frac{\sigma_y^2}{n}.$$

Théorème 2.2.1. *Pour un plan de taille fixe n, simple et sans remise, la variance non corrigée de la population*

$$\sigma_y^2 = \frac{1}{N}\sum_{k\in\mathcal{U}}(y_k-\mu_y)^2,$$

est estimée sans biais par

$$\widehat{\sigma_y^2} = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2.$$

Démonstration.

□

Au final la variance de $\hat{\mu}_y$ est estimée sans biais par

$$\widehat{\mathrm{Var}}(\hat{\mu}_y) = \frac{\widehat{\sigma_y^2}}{n}.$$

2.3 Comparaison des plans simples avec et sans remise

Table 2.1: *Récapitulatif des résultats pour les plans simples de taille fixe n.*

	Sans remise	Avec remise
Estimateur de la moyenne	$\frac{1}{n}\sum_{k\in\mathcal{U}} y_k 1_{\{k\in S\}}$	$\frac{1}{n}\sum_{k\in S} y_k$
Variance de l'estimateur de la moyenne	$\frac{N-n}{N} \times \frac{S_y^2}{n}$	$\frac{\sigma_y^2}{n}$
Estimateur de la variance de l'estimateur de la moyenne	$\frac{N-n}{N}\frac{\widehat{S_y^2}}{n}$	$\frac{\widehat{\sigma_y^2}}{n}$

Le sondage simple et sans remise est **toujours** préférable à celui avec remise. En effet si l'on appelle $\hat{\mu}_{y,\pi}$ et $\tilde{\mu}_{y,\pi}$ les π–estimateurs de la moyenne avec et sans remise, alors pour tout $n \geq 2$

$$\frac{\mathrm{Var}(\tilde{\mu}_{y,\pi})}{\mathrm{Var}(\hat{\mu}_{y,\pi})} = \frac{(N-n)}{N} \times \frac{S_y^2}{\sigma_y^2} = \frac{N-n}{N} \times \frac{N}{N-1} = \frac{N-n}{N-1} < 1.$$

Voilà pourquoi nous allons essentiellement nous concentrer sur les plan simples sans remise.

2.4 Plans simples sans remise et fonction d'intérêt

Jusqu'à présent nous avons essentiellement parlé de l'estimation d'un total t_y ou d'une moyenne μ_y. Parfois l'étude porte sur d'autres grandeurs et donc d'autres fonctions d'intérêt.

2.4.1 Estimation d'une proportion

Il est fréquent qu'une étude porte sur l'estimation d'une proportion p. Avec notre terminologie estimer une proportion revient à compter le nombre d'unités y_k, $k \in \mathcal{U}$, possédant une certaine caractéristique. À partir du caractère y_k, on introduit alors un nouveau caractère

$$z_k = \begin{cases} 1, & \text{si } y_k \text{ possède la caractéristique,} \\ 0, & \text{sinon,} \end{cases} \qquad k \in \mathcal{U},$$

ce qui nous permettra généralement de nous servir des fonctions d'intérêt déjà rencontrées :

$$\begin{aligned} \mu_z &= \frac{1}{N}\sum_{k\in\mathcal{U}} z_k = \frac{\#\{z_k \in \mathcal{U} \colon z_k = 1\}}{N} = p \\ t_z &= \#\{z_k \in \mathcal{U} \colon z_k = 1\} = Np \\ \sigma_z^2 &= \frac{1}{N}\sum_{k\in\mathcal{U}} z_k^2 - \mu_z^2 = p - p^2 = p(1-p) \\ S_z^2 &= \frac{N}{N-1}p(1-p). \end{aligned}$$

Nous voyons donc qu'estimer une proportion n'est rien d'autre qu'estimer une moyenne. En revanche, pour des proportions, les expressions pour la variance se voient considérablement simplifiées du fait que $z_k^2 = z_k$ pour tout $k \in \mathcal{U}$. Ainsi pour un plan simple sans remise, nous avons

$$\begin{aligned} \hat{p} &= \frac{1}{n}\sum_{k\in\mathcal{U}} z_k 1_{\{k\in S\}} \\ s_z^2 &= \frac{1}{n-1}\sum_{k\in\mathcal{U}} (z_k - \hat{p})^2 1_{\{k\in S\}} = \frac{n}{n-1}\hat{p}(1-\hat{p}) \\ \operatorname{Var}(\hat{p}) &= \frac{N-n}{N} \times \frac{S_p^2}{n} = \frac{N-n}{N-1} \times \frac{p(1-p)}{n} \\ \widehat{\operatorname{Var}}(\hat{p}) &= \frac{N-n}{N} \times \frac{\hat{p}(1-\hat{p})}{n-1}. \end{aligned}$$

Remarque. Une fois p estimé par $\hat{p}$, nous obtenons directement une estimation de $\operatorname{Var}(\hat{p})$. Merci les proportions !

2.4.2 Estimation d'un ratio

Considérons cette fois deux caractères y et x. On sera souvent intéressé par l'estimation du ratio

$$R = \frac{\sum_{k\in\mathcal{U}} y_k}{\sum_{k\in\mathcal{U}} x_k} = \frac{\mu_y}{\mu_x}.$$

Pour un plan simple sans remise, on estimera ce ratio par le rapport des moyennes empiriques, i.e.,

$$\hat{R} = \frac{\sum_{k\in\mathcal{U}} y_k 1_{\{k\in S\}}}{\sum_{k\in\mathcal{U}} x_k 1_{\{k\in S\}}} = \frac{\bar{y}}{\bar{x}}.$$

Toutefois l'étude des propriétés de cet estimateur, comme le biais ou l'erreur quadratique, s'avère compliquée puisque nous sommes en présence d'un rapport de deux variables aléatoires ! Lors de tels cas, une technique à retenir est la **linéarisation** du ratio.

$$\hat{R} - R = \frac{\bar{y} - R\bar{x}}{\bar{x}} = \frac{\bar{y} - R\bar{x}}{\mu_x(1+\varepsilon)}, \qquad \varepsilon = \frac{\bar{x} - \mu_x}{\mu_x}$$

Commençons par noter que $\mathbb{E}(\varepsilon) = 0$ et que $\varepsilon \to 0$ lorsque $n \to N$. Ainsi un développement limité de $(1+\varepsilon)^{-1}$ en 0 à l'ordre 1 donne

$$(1+\varepsilon)^{-1} = 1 - \varepsilon + o(\varepsilon^2),$$

et donc

$$\begin{aligned}
\mathbb{E}(\hat{R} - R) &\approx \mathbb{E}\left\{\frac{\bar{y} - R\bar{x}}{\mu_x}(1-\varepsilon)\right\} \\
&= -\mathbb{E}\left(\frac{\bar{y} - R\bar{x}}{\mu_x}\varepsilon\right), && \text{car } \mathbb{E}\left(\frac{\bar{y} - R\bar{x}}{\mu_x}\right) = 0 \\
&= \mathbb{E}\left\{\frac{(R\bar{x} - \bar{y})(\bar{x} - \mu_x)}{\mu_x^2}\right\} \\
&= \mathbb{E}\left\{\frac{(R\bar{x} - R\mu_x + \mu_y - \bar{y})(\bar{x} - \mu_x)}{\mu_x^2}\right\}, && \text{car } \mu_y = R\mu_x \\
&= \frac{R\mathbb{E}\{(\bar{x} - \mu_x)^2\} - \mathrm{Cov}(\bar{x}, \bar{y})}{\mu_x^2}.
\end{aligned}$$

Au final on a donc que le biais de $\hat{R}$ est approximativement

$$\mathrm{Biais}(\hat{R}) \approx \frac{1}{\mu_x^2}\left(R\frac{N-n}{N} \times \frac{S_x^2}{n} - \frac{N-n}{N} \times \frac{S_{xy}}{n}\right) = \frac{1}{\mu_x^2} \times \frac{N-n}{N} \times \frac{1}{n}\left(RS_x^2 - S_{xy}\right),$$

avec

$$S_{xy} = \frac{1}{N-1}\sum_{k \in \mathcal{U}} (x_k - \mu_x)(y_k - \mu_y).$$

Remarque. Le biais est donc approximativement nul dès lors que la taille de l'échantillon est grande.

On procède de même pour approcher l'erreur quadratique moyenne de $\hat{R}$.

$$\begin{aligned}
\mathbb{E}\{(\hat{R} - R)^2\} &\approx \mathbb{E}\left\{\left(\frac{\bar{y} - R\bar{x}}{\mu_x}\right)^2\right\} \\
&= \mathbb{E}\left\{\left(\frac{\bar{y} - \mu_y + R\mu_x - R\bar{x}}{\mu_x}\right)^2\right\} \\
&= \frac{1}{\mu_x^2}\left\{\mathrm{Var}(\bar{y}) + R^2\,\mathrm{Var}(\bar{x}) - 2R\,\mathrm{Cov}(\bar{x}, \bar{y})\right\} \\
&= \frac{1}{\mu_x^2} \times \frac{N-n}{N} \times \frac{1}{n}\left(S_y^2 + R^2S_x^2 - 2RS_{xy}\right).
\end{aligned}$$

Cette erreur quadratique étant naturellement estimée par

$$\hat{\mathbb{E}}\{(\hat{R} - R)^2\} = \frac{1}{\bar{x}^2} \times \frac{N-n}{N} \times \frac{1}{n}\left(\widehat{S_y^2} + \hat{R}^2\widehat{S_x^2} - 2\hat{R}\widehat{S_{xy}}\right),$$

avec

$$\widehat{S_{xy}} = \frac{1}{n-1}\sum_{k \in \mathcal{U}} (x_k - \bar{x})(y_k - \bar{y})1_{\{k \in S\}}.$$

2.5 Détermination de la taille de l'échantillon

Avant de commencer un sondage, il est toujours souhaitable de se poser la question des incertitudes liées à nos futures estimations. Généralement les limites budgétaires fixeront la taille de l'échantillon et on se contentera alors de répondre si le budget alloué est suffisant pour une précision donnée—et de demander une rallonge à son chef le cas échéant...

Par précision donnée nous entendons que le paramètre d'intérêt θ sera contenu dans un intervalle de confiance centré en $\hat{\theta}$ avec une probabilité d'au moins $1-\alpha$, i.e., trouver $\ell > 0$ tel que

$$\Pr\left\{\theta \in \left[\hat{\theta}-\ell, \hat{\theta}+\ell\right]\right\} \geq 1-\alpha$$

En supposant que notre estimateur $\hat{\theta}$ suit approximativement une loi normale (ce qui sera souvent le cas), on sait que

$$\Pr\left\{\theta \in \left[\hat{\theta}-z_{1-\alpha/2}\sqrt{\widehat{\mathrm{Var}}(\hat{\theta})}, \hat{\theta}+z_{1-\alpha/2}\sqrt{\widehat{\mathrm{Var}}(\hat{\theta})}\right]\right\} = 1-\alpha,$$

où $z_{1-\alpha/2}$ le quantile d'une loi normale centrée réduite de probabilité au non dépassement $1-\alpha/2$, i.e., $\Pr(Z \leq z_{1-\alpha/2}) = 1-\alpha/2$, $Z \sim N(0,1)$.

Remarque. Puisque $\widehat{\mathrm{Var}}(\hat{\theta})$ dépend de la taille de l'échantillon n, on cherchera donc la taille minimale n_0 induisant la précision requise.

Pour illustrer nos propos prenons le cas de l'estimation de la moyenne μ_y pour un plan simple sans remise. On a donc

$$\Pr\left\{\mu_y \in \left[\bar{y}-z_{1-\alpha/2}\sqrt{\frac{N-n}{nN}S_y^2}, \bar{y}+z_{1-\alpha/2}\sqrt{\frac{N-n}{nN}S_y^2}\right]\right\} = 1-\alpha,$$

et il faut donc nécessairement

$$\begin{aligned}
\ell^2 \geq z_{1-\alpha/2}^2\frac{N-n}{nN}S_y^2 &\iff nN\ell^2 \geq z_{1-\alpha/2}^2(N-n)S_y^2 \\
&\iff n(N\ell^2 + z_{1-\alpha/2}^2 S_y^2) \geq NS_y^2 z_{1-\alpha/2}^2 \\
&\iff n \geq \frac{NS_y^2 z_{1-\alpha/2}^2}{N\ell^2 + z_{1-\alpha/2}^2 S_y^2}
\end{aligned}$$

Malheureusement cette expression n'est pas si utile en pratique car si notre objectif initial était d'estimer μ_y, il est fort à parier que nous connaissions la variance corrigée S_y^2... En pratique on pourra prendre par exemple une estimation de S_y^2 basée sur des études antérieures.

Lorsque notre paramètre d'intérêt est une proportion, nous pouvons tout de même déterminer la taille minimale. Dans ce contexte, nous avons alors

$$n \geq \frac{N\frac{n}{n-1}\hat{p}(1-\hat{p})z_{1-\alpha/2}^2}{N\ell^2 + z_{1-\alpha/2}^2\frac{n}{n-1}\hat{p}(1-\hat{p})},$$

et nous pouvons considérer le pire cas possible qui est atteint lorsque $\hat{p} = 0.5$. En effet puisque $\widehat{\mathrm{Var}}(\hat{p})$ est proportionnel à $\hat{p}(1-\hat{p})$ la variance est maximale lorsque $\hat{p} = 1/2$.

Chapitre 3

Plans à probabilités inégales

Ce chapitre explique comment nous pouvons bénéficier de la connaissance d'un caractère auxiliaire pour obtenir des estimations plus précises.

3.1 Caractère auxiliaire et probabilités d'inclusion

Soit x_k, $k \in \mathcal{U}$, les valeurs prises par le caractère auxiliaire. Notons tout de suite que cela implique donc sa connaissance sur toute la population ! Notre étude portant toujours sur une fonction d'intérêt telle que la moyenne ou le total d'un caractère y. Le principe d'un plan à probabilités inégales consiste à définir des probabilités d'inclusion du premier ordre proportionnelles aux x_k.

Rappelons que pour un plan de taille fixe, la variance du π–estimateur du total t_y est

$$\mathrm{Var}(\hat{t}_y) = \frac{1}{2} \sum_{\substack{k,\ell \in \mathcal{U} \\ k \neq \ell}} \left(\frac{y_k}{\pi_k} - \frac{y_\ell}{\pi_\ell} \right)^2 \Delta_{k\ell}. \tag{3.1}$$

Si nous souhaitons minimiser (3.1) en jouant seulement sur les probabilités d'inclusions du premier ordre π_k, il est clair que prendre

$$\pi_k = \frac{y_k}{\sum_{\ell \in \mathcal{U}} y_\ell} \propto y_k, \qquad k \in \mathcal{U},$$

est un choix judicieux puisque $\mathrm{Var}(\hat{t}_y)$ est alors nulle. Bien évidemment cette approche est impossible puisqu'elle suppose connaître les valeurs prise par le caractère y sur toute la population $\mathcal{U}$ — inutile alors de faire un sondage !

En revanche si nous disposons d'un caractère auxiliaire x connu sur toute la population et dont on pense qu'il est approximativement proportionnel au caractère y, alors on gagnera à définir les probabilités d'inclusion du premier ordre proportionnellement aux x_k.

Remarque. Si au contraire le caractère x n'est pas du tout proportionnel à y, le plan de sondage sera alors catastrophique et il sera préférable de prendre un plan simple. A méditer donc !

Puisque pour un plan de taille fixe n, cf. Section 1.5,

$$\sum_{k \in \mathcal{U}} \pi_k = n, \tag{3.2}$$

pour obtenir des probabilités d'inclusion proportionnelles aux x_k, i.e., $\pi_k = c x_k$ avec

$$c = \frac{n}{\sum_{\ell \in \mathcal{U}} x_\ell} = \frac{n}{t_x}.$$

Attention toutefois, il n'y a aucune garantie que les $\pi_k \in [0,1]$ et il sera fréquent que certaines "probabilités d'inclusion" soient supérieure à 1. Pour de telles situations, on sélectionnera d'office les unités correspondantes, i.e., $\pi_k = 1$, et l'on recommencera la procédure avec les unités restantes en prenant soin de diminuer la taille de l'échantillon n dans (3.2).

Exemple 3.1.1. Considérons la population $\mathcal{U} = \{1, 2, \ldots, 6\}$ avec une variable auxiliaire x telle que

$$x_1 = 1, \quad x_2 = 9, \quad x_3 = 10, \quad x_4 = 70, \quad x_5 = 90, \quad x_6 = 120.$$

On a donc $t_x = 300$. Si l'on souhaite obtenir un plan de taille fixe $n = 3$, alors les "probabilités d'inclusions temporaires".

$$\pi_1 = \frac{3 \times 1}{300}, \qquad \pi_2 = \frac{3 \times 9}{300}, \qquad \pi_3 = \frac{3 \times 10}{300},$$
$$\pi_4 = \frac{3 \times 70}{300}, \qquad \pi_5 = \frac{3 \times 90}{300}, \qquad \pi_6 = \frac{3 \times 104}{300} > 1.$$

L'unité 6 est alors sélectionnée d'office, le total sans la 6ème unité est

$$\sum_{k \in \mathcal{U} \setminus \{6\}} x_k = t_x - 120 = 180,$$

et les "probabilités d'inclusions" deviennent

$$\pi_1 = \frac{(3-1) \times 1}{180}, \qquad \pi_2 = \frac{(3-1) \times 9}{180}, \qquad \pi_3 = \frac{(3-1) \times 10}{180},$$
$$\pi_4 = \frac{(3-1) \times 70}{180}, \qquad \pi_5 = \frac{(3-1) \times 90}{180}, \qquad \pi_6 = 1.$$

On arrête ici la procédure et les "vraies" probabilités d'inclusion sont

$$\pi_1 = \frac{1}{90}, \quad \pi_2 = \frac{1}{10}, \quad \pi_3 = \frac{1}{9}, \quad \pi_4 = \frac{7}{9}, \quad \pi_5 = \pi_6 = 1.$$

Les unités 5 et 6 sont donc sélectionnées d'office et il restera donc à choisir une unité parmi $\{1, 2, 3, 4\}$. Notons que

$$\sum_{k=1}^{6} \pi_k = \frac{1 + 9 + 10 + 70}{90} + 2 = 3,$$

comme souhaité.

Rappelons qu'un plan de sondage est défini par les $p(s)$ et non par les π_k. Pour avoir un plan à probabilités inégales, il faut donc définir un plan de sondage $p(\cdot)$ tel que pour tout $k \in \mathcal{U}$,

$$\sum_{\substack{s \ni k \\ s \in \mathscr{S}_n}} p(s) = \pi_k, \qquad \mathscr{S}_n = \{s \subset \mathcal{U} \colon |s| = n\}.$$

Remarque. Il existe une infinité de plans de sondage vérifiant ces conditions. Nous allons donc par la suite introduire quelques plans de sondage à probabilités inégales à taille fixe n couramment utilisés.

Algorithme 1 : Algorithme pour un plan de Poisson.

Entrée : les probabilités d'inclusions π_k, la taille de la population N
Sortie : Un échantillon s

```
s = ∅;
pour k ← 1 a N faire
    U ∼ U(0, 1);
    si U < π_k alors
        s ← s ∪ {k};
    fin
fin
retourner s;
```

3.2 Plan de Poisson†

Le plan de Poisson a de très bonnes qualités mais également un gros défaut : il n'est pas de taille fixe. Néanmoins nous allons l'introduire car il va nous servir afin d'en déduire des plans de taille fixe.

Le plan de Poisson se programme très facilement et est décrit par l'algorithme 1. Il est clair que cet algorithme n'est pas de taille fixe : impossible de connaître la taille de l'échantillon avant d'avoir terminé l'exécution de l'algorithme. Il y a même une probabilité non nulle de sélectionner un échantillon de taille nulle ! ! ! Il a cependant de bonnes qualités.

Puisque les unités sont sélectionnées indépendemment

$$\pi_{k\ell} = \Pr(k \in S, \ell \in S) = \Pr(k \in S)\Pr(\ell \in S) = \pi_k \pi_\ell,$$

et donc

$$\Delta_{k\ell} = \pi_{k\ell} - \pi_k\pi_\ell = 0, \qquad k \neq \ell.$$

Clairement le plan de sondage est donné pour tout $s \subset \mathcal{U}$

$$p(s) = \underbrace{\prod_{k\in s} \pi_k}_{\substack{\text{proba. de sélectionner} \\ \text{les unités chosies}}} \times \underbrace{\prod_{k\in\mathcal{U}\setminus\{s\}} (1-\pi_k)}_{\substack{\text{proba. de ne pas sélectionner} \\ \text{les unités non retenues}}}$$

Puisque $\Delta_{k\ell} = 0$, $k \neq \ell$, la variance du π–estimateur du total t_y est

$$\mathrm{Var}(\hat{t}_y) = \sum_{k,\ell\in\mathcal{U}} \frac{y_k y_\ell}{\pi_k\pi_\ell}\Delta_{k\ell} = \sum_{k\in\mathcal{U}} \frac{y_k^2\pi_k(1-\pi_k)}{\pi_k^2} = \sum_{k\in\mathcal{U}} \frac{y_k^2(1-\pi_k)}{\pi_k},$$

et peut être estimée par

$$\widehat{\mathrm{Var}}(\hat{t}_y) = \sum_{k\in\mathcal{U}} \frac{y_k^2(1-\pi_k)}{\pi_k^2} 1_{\{k\in S\}}.$$

Le plan de Poisson est intéressant car il est simple à mettre en oeuvre mais également car il maximise l'entropie. Nous introduisons maintenant un mesure du "désordre".

Définition 3.2.1. On appelle entropie d'un plan $p(\cdot)$ la quantité

$$I(p) = -\sum_{s\subset\mathcal{U}} p(s)\ln p(s),$$

avec la convention que $0\ln 0 = 0$.

Clairement l'entropie est toujours positive. De plus, comme mesure du désordre, plus $I(p)$ sera grand plus le plan $p(\cdot)$ sera "aléatoire". Pour des probabilités d'inclusion fixées, on cherchera donc le plan le plus aléatoire ou désordonné, i.e., celui maximisant l'entropie.

Lemme 3.2.1.

$$\sum_{s \subset \mathcal{U}} \prod_{k \in s} x_k = \prod_{k \in \mathcal{U}} (1 + x_k).$$

Démonstration.

□

Théorème 3.2.2. *Étant donné des probabilités d'inclusions fixées π_k, $k \in \mathcal{U}$, le plan de Poisson est le plan d'entropie maximale sur $\mathscr{S} = \{s \colon s \subset \mathcal{U}\}$.*

Démonstration.

□

On retiendra donc que le plan de Poisson est un plan de sondage respectant les probabilités d'inclusions d'ordre un fixée a priori et étant le « plus aléatoire possible » (au sens de l'entropie). Il a toutefois l'inconvénient de ne pas être à taille fixe.

3.3 Sondage systématique à probabilités inégales

Ce plan de sondage a été introduit vers 1950 et est toujours largement utilisé puisqu'il a le mérite d'être simple et exact ! Contrairement au plan de Poisson, elle a également le bon goût d'être de taille fixe.

Comme depuis le début de ce Chapitre, on désire tirer des échantillons dont les probabilités d'inclusion d'ordre un sont fixées a priori et telles que $0 < \pi_i < 1$, $k \in \mathcal{U}$ et

$$\sum_{k \in \mathcal{U}} \pi_k = n.$$

Définissons les probabilités d'inclusion cumulées

$$C_k = \sum_{\ell=1}^{k} \pi_\ell, \qquad k \in \mathcal{U}, \qquad C_0 = 0.$$

L'approche consiste à générer $U \sim U(0,1)$ et de sélectionner les unités à partir de cette unique réalisation. La première unité sélectionnée, appelons là k_1, sera celle telle que

$$C_{k_1 - 1} \leq U < C_{k_1};$$

la deuxième unité sélectionnée, notons la k_2, sera cette fois ci

$$C_{k_2 - 1} \leq 1 + U < C_{k_2};$$

et ainsi de suite... De manière générale, la $j$ème unité sélectionnée, notée k_j, sera alors

$$C_{k_j - 1} \leq j - 1 + U < C_{k_j}.$$

Exercice 1. Prenons la situation où $N = 6$, $n = 3$, $\pi_1 = 0.2, \pi_2 = 0.7, \pi_3 = 0.8, \pi_4 = 0.5$ et $\pi_5 = \pi_6 = 0.4$. Déterminer l'échantillon sélectionné sachant que $U = 0.3658$.

Solution.

□

Nous venons de voir que cette méthode est en effet très simple. Elle a quand même quelques défauts ; notamment les probabilités d'inclusions d'ordre deux sont souvent nulles.

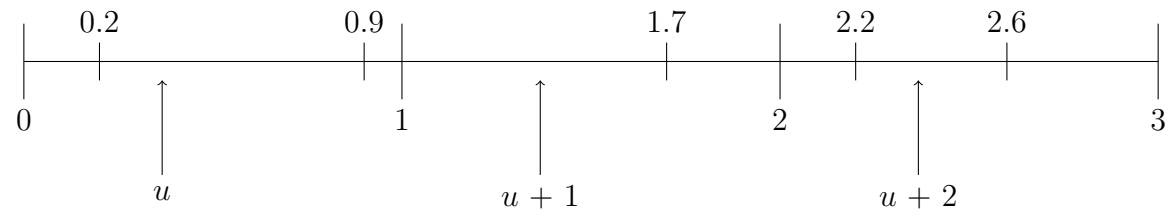

Figure 3.1: *Illustration du tirage systématique de l'exercice 1.*

Exercice 2. Montrez que la matrice $P = (\pi_{k\ell})_{k,\ell}$ des probabilités d'inclusion d'ordre deux de l'exercice 1 est

$$P = \begin{bmatrix} — & 0.0 & 0.2 & 0.2 & 0.0 & 0 \\ 0.0 & — & 0.5 & 0.2 & 0.4 & 0.3 \\ 0.2 & 0.5 & — & 0.3 & 0.4 & 0.2 \\ 0.2 & 0.2 & 0.3 & — & 0.0 & 0.3 \\ 0.0 & 0.4 & 0.4 & 0.0 & — & 0 \\ 0.0 & 0.3 & 0.2 & 0.3 & 0.0 & — \end{bmatrix}$$

Solution.

□

Chapitre 4

Stratification

La technique de stratification est largement utilisée en sondage car elle permet facilement d'introduire de l'information auxiliaire pour la construction d'un plan de sondage adéquat.

4.1 Population et strates

Supposons que la population $\mathcal{U}$ soit partitionnée en H sous-ensembles $\mathcal{U}_1, \ldots, \mathcal{U}_H$ appelés strates et tels que

$$\bigcup_{i=1}^{H} \mathcal{U}_i = \mathcal{U}, \qquad \mathcal{U}_i \cap \mathcal{U}_h = \emptyset, \quad i \neq h.$$

Chaque strate $\mathcal{U}_h$ admet une taille N_h et l'on a bien évidemment

$$\sum_{h=1}^{H} N_h = N,$$

où N est la taille de la population $\mathcal{U}$.

Remarque. Les tailles des strates N_h sont ici supposées connues et constituent l'information auxiliaire.

Notre but étant toujours d'estimer un total ou une moyenne, remarquons que le total (resp. la moyenne) s'écrit à l'aide des strates

$$t_y = \sum_{k \in \mathcal{U}} y_k = \sum_{h=1}^{H} \sum_{k \in \mathcal{U}_h} y_k = \sum_{h=1}^{H} t_{y,h},$$

où $t_{y,h}$ est le total des valeurs prises par le caractère y sur la strate $\mathcal{U}_h$, i.e.,

$$t_{y,h} = \sum_{k \in \mathcal{U}_h} y_k.$$

De même la moyenne sur la population s'écrit

$$\mu_y = \frac{1}{N} \sum_{k \in \mathcal{U}} y_k = \frac{1}{N} \sum_{h=1}^{H} \sum_{k \in \mathcal{U}_h} y_k = \frac{1}{N} \sum_{h=1}^{H} N_h \mu_{y,h},$$

où $\mu_{y,h}$ est la moyenne des valeurs prises par le caractère y sur la strate $\mathcal{U}_h$, i.e.,

$$\mu_{y,h} = \frac{1}{N_h} \sum_{k \in \mathcal{U}_h} y_k.$$

On définit également la variance et la variance corrigée sur une strate $\mathcal{U}_h$ par

$$\sigma_{y,h}^2 = \frac{1}{N_h} \sum_{k \in \mathcal{U}_h} (y_k - \mu_{y,h})^2,$$

et

$$S_{y,h}^2 = \frac{1}{N_h - 1} \sum_{k \in \mathcal{U}_h} (y_k - \mu_{y,h})^2.$$

Remarque. La variance sur la population (totale) σ_y^2 s'écrit

$$\begin{aligned}
\sigma_y^2 &= \frac{1}{N} \sum_{k \in \mathcal{U}} (y_k - \mu_y)^2 \\
&= \frac{1}{N} \sum_{h=1}^{H} \sum_{k \in \mathcal{U}_h} \{(y_k - \mu_{y,h}) + (\mu_{y,h} - \mu_y)\}^2 \\
&= \frac{1}{N} \sum_{h=1}^{H} \left\{ \sum_{k \in \mathcal{U}_h} (y_k - \mu_{y,h})^2 + 2(\mu_{y,h} - \mu_y) \underbrace{\sum_{k \in \mathcal{U}_h} (y_k - \mu_{y,h})}_{=0} + N_h (\mu_{y,h} - \mu_y)^2 \right\} \\
&= \frac{1}{N} \sum_{h=1}^{H} N_h \sigma_{y,h}^2 + \frac{1}{N} \sum_{h=1}^{H} N_h (\mu_{y,h} - \mu_y)^2 \\
&= \sigma_{y,\text{intra}}^2 + \sigma_{y,\text{inter}}^2,
\end{aligned}$$

où $\sigma_{y,\text{intra}}^2$ est la variance intra–srates, i.e.,

$$\sigma_{y,\text{intra}}^2 = \frac{1}{N} \sum_{h=1}^{H} N_h \sigma_{y,h}^2,$$

et $\sigma_{y,\text{inter}}^2$ est la variance inter–srates, i.e.,

$$\sigma_{y,\text{inter}}^2 = \frac{1}{N} \sum_{h=1}^{H} N_h (\mu_{y,h} - \mu_y)^2.$$

4.2 Échantillons, probabilités d'inclusion et estimation

Définition 4.2.1. Un sondage est dit **stratifié** si, pour chaque strate, on tire un échantillon selon un sondage aléatoire simple sans remise de taille fixe n_h et que les tirages au sein de chaque strate sont mutuellement indépendant.

Soit S_h l'échantillon aléatoire tiré dans la strate $\mathcal{U}_h$ à l'aide d'un plan de sondage $p_h(\cdot)$. L'échantillon aléatoire S obtenu au final est donc

$$S = \bigcup_{h=1}^{H} S_h.$$

Le plan de sondage associé $p(\cdot)$ n'est rien d'autre que

$$p(s) = \prod_{h=1}^{H} p_h(s_h), \qquad s = \bigcup_{h=1}^{H} s_h,$$

et la taille de l'échantillon S est

$$n = \sum_{h=1}^{H} n_h.$$

Le calcul des probabilités d'inclusion pour un sondage stratifié n'est pas difficile mais il faut tout de même faire attention. Pour les probabilités d'inclusion d'ordre un et si l'unité k appartient à la strate $\mathcal{U}_h$ alors

$$\pi_k = \frac{n_h}{N_h},$$

puisqu'on a effectué un plan simple sans remise de taille n_h pour cette strate.

Pour les probabilités d'inclusion d'ordre deux, c'est un peu plus difficile et le résultat dépend du fait ou non que les unités k et ℓ appartiennent à la même strate ou non.

— Si k et ℓ appartiennent à la même strate $\mathcal{U}_h$ alors

$$\pi_{k\ell} = \frac{n_h(n_h-1)}{N_h(N_h-1)}.$$

— Si k et ℓ appartiennent à deux strates différentes $\mathcal{U}_{h_1}$ et $\mathcal{U}_{h_2}$ alors (par indépendance entre les strates)

$$\pi_{kl} = \pi_k\pi_\ell = \frac{n_{h_1}}{N_{h_1}}\frac{n_{h_2}}{N_{h_2}}.$$

En conséquence on a

$$\Delta_{k\ell} = \begin{cases} \frac{n_h}{N_h}\left(1-\frac{n_h}{N_h}\right), & k=\ell,\ k\in\mathcal{U}_h, \\ -\frac{n_h(N_h-n_h)}{N_h^2(N_h-1)}, & k\neq\ell,\ k,\ell\in\mathcal{U}_h, \\ 0, & k\in\mathcal{U}_h,\ \ell\notin\mathcal{U}_h. \end{cases}$$

Du coup les π–estimateurs du total t_y et de la moyenne μ_y sont

$$\hat{t}_{y,\text{strat}} = \sum_{k\in S}\frac{y_k}{\pi_k} = \sum_{h=1}^{H}\sum_{k\in S_h}\frac{N_h y_k}{n_h} = \sum_{h=1}^{H}\hat{t}_{y,h},$$

et

$$\hat{\mu}_{y,\text{strat}} = \frac{1}{N}\sum_{h=1}^{H}\frac{N_h}{n_h}\sum_{k\in S_h} y_k = \frac{1}{N}\sum_{h=1}^{H} N_h\bar{y}_h,$$

où $\hat{t}_{y,h}$ est l'estimateur du total pour la strate h, i.e.,

$$\hat{t}_{y,h} = \frac{N_h}{n_h}\sum_{k\in S_h} y_k,$$

et $\bar{y}_h$ est la moyenne de l'échantillon prélevé sur la strate h, i.e.,

$$\bar{y}_h = \frac{1}{n_h}\sum_{k\in S_h} y_k.$$

Puisque les strates sont indépendantes, la variance de ces estimateurs se calcule facilement

$$\text{Var}(\hat{t}_{y,\text{strat}}) \overset{\text{ind}}{=} \sum_{h=1}^{H}\text{Var}(\hat{t}_{y,h}) \overset{\text{simple}}{=} \sum_{h=1}^{H} N_h(N_h-n_h)\frac{S_{y,h}^2}{n_h},$$

variance qui s'estime sans biais par

$$\widehat{\text{Var}}(\hat{t}_{y,\text{strat}}) = \sum_{h=1}^{H} N_h(N_h-n_h)\frac{\widehat{S_{y,h}^2}}{n_h},$$

avec

$$\widehat{S_{y,h}^2} = \frac{1}{n_h-1}\sum_{k\in S_h}(y_k-\bar{y}_h)^2, \qquad h=1,\ldots,H.$$

4.3 Plan stratitifié et allocation proportionnelle

Définition 4.3.1. Un plan stratifié est dit à **allocation proportionnelle** si

$$\frac{n_h}{N_h} = \frac{n}{N}, \qquad h = 1, \dots, H,$$

c'est à dire que les « strates de tailles importantes » auront plus d'unité dans l'échantillon que celles de « tailles plus petites ».

Remarque. Généralement la taille d'échantillon pour chaque strate

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

ne sera pas entière mais afin de simplifier les développements théoriques qui viennent nous allons tout de même le supposer... No comment !

Les π–estimateur du total et de la moyenne sont alors

$$\hat{t}_{y,\text{strat. prop.}} = \sum_{h=1}^{H} \hat{t}_{y,h} = \frac{N}{n} \sum_{k \in S} y_k,$$
$$\hat{\mu}_{y,\text{strat. prop.}} = \frac{1}{n} \sum_{k \in S} y_k.$$

La variance du total est alors

$$\begin{aligned}\operatorname{Var}(\hat{t}_{y,\text{strat. prop.}}) &= \sum_{h=1}^{H} N_h(N_h - n_h)\frac{S_{y,h}^2}{n_h} \\ &= \sum_{h=1}^{H} N_h \left(\frac{N}{n} - 1\right) S_{y,h}^2 \\ &= \frac{N-n}{n} \sum_{h=1}^{H} N_h S_{y,h}^2.\end{aligned}$$

Remarque. Lorsque les tailles des strates N_h sont suffisamment grandes, alors $S_{y,h}^2 \approx \sigma_{y,h}^2$ et donc

$$\operatorname{Var}(\hat{t}_{y,\text{strat. prop.}}) \approx \frac{N-n}{n} \sum_{h=1}^{H} N_h \sigma_{y,h}^2 = N(N-n)\frac{\sigma_{y,\text{intra}}^2}{n},$$

alors que la variance de l'estimateur du total pour un plan simple sans remise vérifie

$$\operatorname{Var}(\hat{t}_{y,\pi}) \approx N(N-n)\frac{\sigma_y^2}{n}.$$

Les deux expressions sont quasiment identiques mais puisque

$$\sigma_y^2 = \sigma_{y,\text{intra}}^2 + \sigma_{y,\text{inter}}^2,$$

la première expression est plus petite, i.e., on obtient de meilleurs résultat avec un plan stratifié avec allocation proportionnelle qu'avec un plan simple sans remise !

Ceci est bien entendu d'autant plus vrai que la variance inter-strate sera grande, ce qui est le cas lorsque le caractère d'intérêt y dépend fortement du caractère servant à la stratification, ici les tailles N_h.

Bien entendu on estimera sans biais cette variance par

$$\widehat{\operatorname{Var}}(\hat{t}_{y,\text{strat. prop.}}) = \frac{N-n}{n} \sum_{h=1}^{H} N_h \widehat{S_{y,h}^2},$$

avec

$$\widehat{S_{y,h}^2} = \frac{1}{n_h - 1} \sum_{k \in S_h} (y_k - \bar{y}_h)^2, \qquad h = 1, \dots, H.$$

4.4 Plan stratifié optimal pour le total

Si notre intérêt est d'estimer un total ou une moyenne alors il existe une taille optimale pour les strates. On cherche donc les tailles d'échantillon $n_1, \dots, n_h$ minimisant la variance du π–estimateur du total t_y pour une taille d'échantillon fixée n, i.e., minimiser

$$\operatorname{Var}(\hat{t}_{y,\text{strat}}) = \sum_{h=1}^{H} N_h(N_h - n_h)\frac{S_{y,h}^2}{n_h}$$

par rapport aux n_h et sous la contrainte

$$\sum_{h=1}^{H} n_h = n.$$

Le Lagrangien de ce problème de minimisation est

$$\mathscr{L}(n_1, \dots, n_H, \lambda) = \sum_{h=1}^{H} N_h(N_h - n_h)\frac{S_{y,h}^2}{n_h} + \lambda\left(\sum_{h=1}^{H} n_h - n\right).$$

On a donc

$$\begin{aligned}\frac{\partial \mathscr{L}}{\partial n_h} = 0 &\iff -\frac{N_h^2}{n_h^2}S_{y,h}^2 + \lambda = 0\\ &\iff n_h = \frac{N_h S_{y,h}}{\sqrt{\lambda}}.\end{aligned}$$

Mais puisque $\sum_h n_h = n$ on a

$$\lambda^{-1/2}\sum_{h=1}^{H} N_h S_{y,h} = n,$$

et il vient

$$n_h = n\frac{N_h S_{y,h}}{\sum_{j=1}^{H} N_j S_{y,j}}, \qquad h = 1, \dots, H.$$

Remarque. La taille optimale pour une strate $\mathcal{U}_h$ est donc proportionnelle au produit de la taille de cette strate et de l'écart-type du caractère y sur cette strate.

Bien entendu en pratique on ne connaîtra pas $S_{y,h}$ et donc la formule précédente n'est pas d'un grand intérêt. Elle est cependant assez instructive—et intuitive ! Instructive puisqu'elle indique qu'il faut surreprésenter les strates qui ont une forte variabilité ; ce qui est intuitif non ?

Remarque. En pratique les tailles $n_h \notin \mathbb{N}$ et on arrondira les résultats. De plus il peut arriver également que $n_h > N_h$ pour un $h \in \{1, \dots, H\}$. Pour de tels cas, on posera alors $n_h = N_h$ et on déterminera les tailles optimales sur les strates restantes — en itérant le procédé si nécessaire.

Supposons que nos tailles optimales soient des entiers et telles que $n_h < N_h$ pour tout h. Alors la variance du π–estimateur est alors

$$\begin{aligned}\operatorname{Var}(\hat{t}_{y,\text{opt}}) &= \sum_{h=1}^{H} N_h(N_h - n_h)\frac{S_{y,h}^2}{n_h}\\ &= \sum_{h=1}^{H} N_h^2 \frac{\sum_{\ell=1}^{H} N_\ell S_{y,\ell}}{n N_h S_{y,h}} S_{y,h}^2 - \sum_{h=1}^{H} N_h S_{y,h}^2\\ &= \left(\frac{\sum_{\ell=1}^{H} N_\ell S_{y,\ell}}{n}\right)\sum_{h=1}^{H} N_h S_{y,h} - \sum_{h=1}^{H} N_h S_{y,h}^2\\ &= \frac{1}{n}\left(\sum_{h=1}^{H} N_h S_{y,h}\right)^2 - \sum_{h=1}^{H} N_h S_{y,h}^2.\end{aligned}$$

4.5 Prise en compte du coût

Faire une enquête est bien souvent coûteux de sorte que l'allocation optimale présentée dans la Section 4.4 sera bien souvent déconnectée de la réalité. Bien souvent on visera plutôt une allocation optimale pour un budget fixé C. Nous allons donc minimiser la variance de l'estimateur du total

$$\mathrm{Var}(\hat{t}_{y,\mathrm{strat}}) = \sum_{h=1}^{H} N_h(N_h - n_h)\frac{S_{y,h}^2}{n_h},$$

sous la contrainte

$$\sum_{h=1}^{H} n_h C_h = C,$$

o C_h représente le coût d'interroger une unité dans la strate $\mathcal{U}_h$.

Exercice 3. Montrez que la taille optimale est alors

$$n_h = \frac{C N_h S_{y,h}}{\sqrt{C_h}\sum_{\ell=1}^{H} N_\ell S_{y,\ell}\sqrt{C_\ell}}, \qquad h = 1, \ldots, H.$$

Solution.

□

Remarque. De manière assez logique nous constatons que nous sélectionnons moins les strates les plus « coûteuses ».

Chapitre 5

Plans par grappes et à plusieurs degrés

Dans ce chapitre nous allons voir comment une variable auxiliaire peut être utilisée non pas pour améliorer la précision de nos estimations mais le déroulement d'une enquête !

5.1 Plans par grappes

Les plans par grappes ressemblent (aux premiers abords) fortement aux plans stratifiés. Ce n'est pourtant pas du tout le cas ! ! !

Supposons que la population $\mathcal{U}$ soit partitionnée en M sous-ensembles $\mathcal{U}_1, \ldots, \mathcal{U}_M$ appelés grappes et tels que

$$\bigcup_{i=1}^{M} \mathcal{U}_i = \mathcal{U}, \qquad \mathcal{U}_i \cap \mathcal{U}_j = \emptyset, \quad i \neq j.$$

Chaque grappe $\mathcal{U}_i$ admet une taille N_i et l'on a bien évidemment

$$\sum_{i=1}^{M} N_i = N,$$

où N est la taille de la population $\mathcal{U}$.

Notre but étant toujours d'estimer un total ou une moyenne, remarquons que le total (resp. la moyenne) s'écrit à l'aide des grappes

$$t_y = \sum_{k \in \mathcal{U}} y_k = \sum_{i=1}^{M} \sum_{k \in \mathcal{U}_i} y_k = \sum_{i=1}^{M} t_{y,i},$$

où $t_{y,i}$ est le total des valeurs prises par le caractère y sur la grappe $\mathcal{U}_i$, i.e.,

$$t_{y,i} = \sum_{k \in \mathcal{U}_i} y_k.$$

De même la moyenne sur la population s'écrit

$$\mu_y = \frac{1}{N} \sum_{k \in \mathcal{U}} y_k = \frac{1}{N} \sum_{i=1}^{M} \sum_{k \in \mathcal{U}_i} y_k = \frac{1}{N} \sum_{i=1}^{M} N_i \mu_{y,i},$$

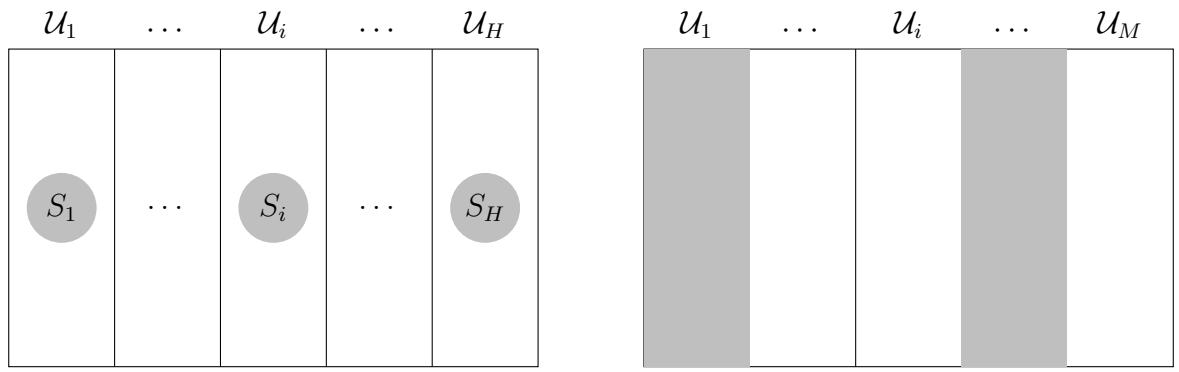

Figure 5.1: *Illustration de la différence entre un plan de sondage stratifié (gauche) et par grappes (droite). Pour l'un, un échantillon aléatoire est prélevé dans chaque strate. Pour l'autre un échantillon aléatoire sur les grappes est prélevé et chaque grappe ainsi piochée est entièrement retenue.*

où $\mu_{y,i}$ est la moyenne des valeurs prises par le caractère y sur la grappe $\mathcal{U}_i$, i.e.,

$$\mu_{y,i} = \frac{1}{N_i} \sum_{k \in \mathcal{U}_i} y_k.$$

On définit également la variance et la variance corrigée sur une grappe $\mathcal{U}_i$ par

$$\sigma^2_{y,i} = \frac{1}{N_i} \sum_{k \in \mathcal{U}_i} (y_k - \mu_{y,i})^2,$$

et

$$S^2_{y,i} = \frac{1}{N_i - 1} \sum_{k \in \mathcal{U}_i} (y_k - \mu_{y,i})^2.$$

Jusque là rien de bien nouveau par rapport à la manière dont nous avions introduit les plans stratifiés me direz vous. C'est à ce moment bien précis que les deux approches divergent ! ! !

Définition 5.1.1. Un plan est dit **par grappes** si l'on procède comme suit :

1. On sélectionne un échantillon aléatoire de grappes S_g selon un plan de sondage $p_g(\cdot)$ défini sur les parties non vides de $\mathcal{U}_g = \{1, \ldots, M\}$;
2. Toutes les unités des grappes sélectionnées sont alors retenues.

La Figure 5.1 illustre la différence entre ces deux plans de sondages. Nous voyons clairement que le plan stratifié utilise un échantillons dans chaque strates alors que le plan par grappes sélectionne soit totalement une grappe soit pas du tout. Ainsi un échantillon aléatoire S issu d'un plan par grappes s'écrit

$$S = \bigcup_{i \in S_g} \mathcal{U}_i,$$

et sa taille n_S est

$$n_S = \sum_{i \in S_g} N_i.$$

Remarque. La taille de l'échantillon n_S sera le plus souvent aléatoire même si le plan de sondage sur les grappes $p_g(\cdot)$ est à taille fixe — les grappes n'ayant pas forcément des tailles identiques.

Les probabilités d'inclusion d'ordre un et deux découlent des probabilités de sélection des grappes (et donc du plan de sondage $p_g(\cdot)$). Ainsi si l'unité k appartient à la grappe i, on a

$$\pi_k = \sum_{\substack{s\in\mathscr{S}_g\\ i\in s}} p_g(s) \stackrel{\text{def}}{=} \pi_{g,i}, \qquad k\in\mathcal{U}_i, \quad i\in\mathcal{U}_g,$$

où $\mathscr{S}_g$ est l'ensemble des échantillons possibles de $\mathcal{U}_g$.

Les probabilités d'inclusions d'ordre deux s'écrivent de manière analogue

$$\pi_{k\ell} = \begin{cases} \pi_{g,i}, & k,\ell\in\mathcal{U}_i \\ \pi_{g,ij}, & k\in\mathcal{U}_i,\ \ell\in\mathcal{U}_j, \end{cases}$$

avec

$$\pi_{g,ij} = \sum_{\substack{s\in\mathscr{S}_g\\ i,j\in s}} p_g(s), \qquad i,j\in\mathcal{U}_g, \quad i\neq j.$$

Exercice 4. Montrez que les conditions de Sen–Yates–Grundy, i.e., $\Delta_{k\ell}<0$, ne sont pas satisfaites lorsque k et ℓ appartiennent à la même grappe.

Solution.

□

Les π–estimateurs du total et de la moyenne sont

$$\hat{t}_{y,\pi} = \sum_{i\in S_g} \frac{t_{y,i}}{\pi_{g,i}}, \qquad \hat{\mu}_{y,\pi} = \frac{1}{N}\sum_{i\in S_g}\frac{N_i\mu_{y,i}}{\pi_{g,i}}.$$

Notons toutefois que, pour les plans par grappes, il est rare que la taille de la population N soit connue. On utilisera plutôt le ratio de Hájek de la Section 1.7 pour estimer la moyenne μ_y.

La variance du π–estimateur du total t_y est, cf. Section 1.6.2,

$$\begin{aligned}\mathrm{Var}(\hat{t}_{y,\pi}) &= \sum_{k,\ell\in\mathcal{U}} \frac{y_k y_\ell}{\pi_k\pi_\ell}\Delta_{k\ell} \\ &= \sum_{i,j=1}^{M} \frac{t_{y,i}t_{y,j}}{\pi_{g,i}\pi_{g,j}}\Delta_{g,ij} \\ &= \sum_{i=1}^{M}\frac{t_{y,i}^2}{\pi_{g,i}^2}\pi_{g,i}(1-\pi_{g,i}) + \sum_{i\neq j}\frac{t_{y,i}t_{y,j}}{\pi_{g,i}\pi_{g,j}}(\pi_{g,ij}-\pi_{g,i}\pi_{g,j}),\end{aligned}$$

que l'on estimera classiquement par le π–estimateur.

Si le nombre de grappe sélectionné est fixe, alors on peut écrire cette variance sous une autre forme (cf. Section 1.6.3),

$$\mathrm{Var}(\hat{t}_{y,\pi}) = -\frac{1}{2}\sum_{\substack{i,j=1\\ i\neq j}}^{M}\left(\frac{t_{y,i}}{\pi_{g,i}} - \frac{t_{y,j}}{\pi_{g,j}}\right)^2\Delta_{g,ij}. \tag{5.1}$$

5.2 Choix sur le plan de sondage $p_g(\cdot)$

5.2.1 Tirage des grappes à probabilités égales

La première idée venant à l'esprit pour le choix de $p_g(\cdot)$ est de faire un plan de sondage sans remise et à taille fixe m. Pour ce choix nous avons alors

$$\pi_{g,i} = \frac{m}{M}, \qquad \pi_{g,ij} = \frac{m(m-1)}{M(M-1)}, \qquad i,j = 1,\dots,M, \quad i \neq j.$$

Cela dit bien que $p_g(\cdot)$ soit de taille fixe, la taille n_S de l'échantillon S obtenue est comme nous l'avons déjà dit aléatoire et vaut en espérance

$$\mathbb{E}(n_S) = \mathbb{E}\left(\sum_{i\in S_g} N_i\right) = \sum_{i=1}^{M} N_i \mathbb{E}(1_{\{i\in S_g\}}) = \sum_{i=1}^{M} N_i \pi_{g,i} = \frac{mN}{M}.$$

Les π–estimateurs du total et de la moyenne se simplifient en

$$\hat{t}_{y,\pi} = \frac{M}{m}\sum_{i\in S_g} t_{y,i}, \qquad \hat{\mu}_{y,\pi} = \frac{M}{mN}\sum_{i\in S_g} N_i \mu_{y,i}.$$

Puisque $p_g(\cdot)$ est à taille fixe, la variance s'écrit d'après (5.1)

$$\begin{aligned}
\mathrm{Var}(\hat{t}_{y,\pi}) &= -\frac{1}{2}\sum_{\substack{i,j=1\\ i\neq j}}^{M}\left(\frac{t_{y,i}}{\pi_{g,i}} - \frac{t_{y,j}}{\pi_{g,j}}\right)^2 \Delta_{g,ij} \\
&= -\frac{M^2}{2m^2}\sum_{i\neq j}(t_{y,i}-t_{y,j})^2\left\{\frac{m(m-1)}{M(M-1)} - \frac{m^2}{M^2}\right\} \\
&= -\frac{M^2}{2m^2}\frac{m(m-M)}{M^2(M-1)}\sum_{i\neq j}(t_{y,i}-t_{y,j})^2 \\
&= \frac{M-m}{M-1}\frac{M}{m}\sum_{i=1}^{M}\left(t_{y,i} - \frac{t_y}{M}\right)^2,
\end{aligned}$$

où on a utilisé pour la dernière équation le fait que

$$\sum_{i,j=1}^{n}(x_i - x_j)^2 = 2n\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Il est peut-être plus parlant d'écrire cette dernière expression de la manière suivante

$$\mathrm{Var}(\hat{t}_{y,\pi}) = M(M-m)\frac{\frac{1}{M-1}\sum_{i\in S_g}(t_{y,i} - t_y/M)^2}{m},$$

qui nous fait furieusement penser à l'expression vue à maintes reprises

$$\mathrm{Var}(\hat{t}_{y,\pi}) = N(N-n)\frac{S_y^2}{n},$$

mais où la variance corrigée est maintenant calculée sur les sous–totaux des grappes—ce qui est logique non ?

On estimera bien entendu cette variance par

$$\widehat{\mathrm{Var}}(\hat{t}_{y,\pi}) = \frac{M-m}{m-1}\frac{M}{m}\sum_{i\in S_g}\left(t_{y,i} - \frac{\hat{t}_y}{M}\right)^2.$$

5.2.2 Tirage proportionnel aux tailles des grappes

On peut également effectuer un plan sans remise de taille fixe m dont les probabilités de sélection sont proportionnelles à la taille de chacune des grappes comme nous l'avons vu lors de la Section 3.1.

Pour simplifier les choses on supposera que $mN_i \leq N$ pour tout $i = 1, \dots, M$ — sinon les grappes ne vérifiant pas cela seront systématiquement choisies. Les probabilités de sélection des grappes sont alors

$$\pi_{g,i} = \frac{mN_i}{N}, \qquad i = 1, \dots, M.$$

La taille n_S de l'échantillon S est toujours aléatoire et vaut en moyenne

$$\mathbb{E}(n_S) = \mathbb{E}\left(\sum_{i \in S_g} N_i\right) = \sum_{i=1}^{M} N_i \pi_{g,i} = \frac{m}{N} \sum_{i=1}^{M} N_i^2.$$

Les π–estimateurs du total et de la moyenne sont

$$\hat{t}_{y,\pi} = \frac{N}{m} \sum_{i \in S_g} N_i t_{y,i} = \frac{N}{m} \sum_{i \in S_g} \mu_{y,i}, \qquad \hat{\mu}_{y,\pi} = \frac{1}{m} \sum_{i \in S_g} \mu_{y,i},$$

et puisque le plan de sondage $p_g(\cdot)$ est à taille fixe, la variance s'écrit d'après (5.1)

$$\begin{aligned}
\mathrm{Var}(\hat{t}_{y,\pi}) &= -\frac{1}{2} \sum_{\substack{i,j=1 \\ i \neq j}}^{M} \left(\frac{t_{y,i}}{\pi_i} - \frac{t_{y,j}}{\pi_j}\right)^2 \Delta_{g,ij} \\
&= -\frac{1}{2} \sum_{\substack{i,j=1 \\ i \neq j}}^{M} \left(\frac{N t_{y,i}}{m N_i} - \frac{N t_{y,j}}{m N_j}\right)^2 \left(\pi_{g,ij} - \frac{mN_i}{N} \frac{mN_j}{N}\right) \\
&= -\frac{N^2}{2M^2} \sum_{\substack{i,j=1 \\ i \neq j}}^{M} (\mu_{y,i} - \mu_{y,j})^2 \left(\pi_{g,ij} - \frac{m^2 N_i N_j}{N^2}\right).
\end{aligned}$$

Remarque. Nous ne pouvons pas aller plus loin dans le calcul de cette variance car de manière générale les probabilités d'inclusion d'ordre 2 ne sont pas connues pour les tirages proportionnels, cf. Chapitre 3.

5.3 Plans à deux degrés

Les plans à deux degrés portent bien leurs noms puisqu'ils consistent en un double échantillonnage :

1. sur les unités primaires ;
2. puis les unités secondaires.

En exemple valant mille mots, pour un sondage sur les ménages, les unités primaires seraient les communes alors que les unités secondaires seraient les ménages. Un plan à deux degrés consisterait donc à échantillonner les communes puis à prélever, pour chaque commune retenue, un échantillon de ménages.

Remarque. C'est un peu la stratégie de diviser pour mieux régner et cela permet parfois de réduire les coût de l'enquête. En effet pour notre exemple sur les ménages, les unités (secondaires) seront forcément proches car issues de la même commune. Imaginez la facture d'essence si l'on avait échantillonné directement sur les ménages français !

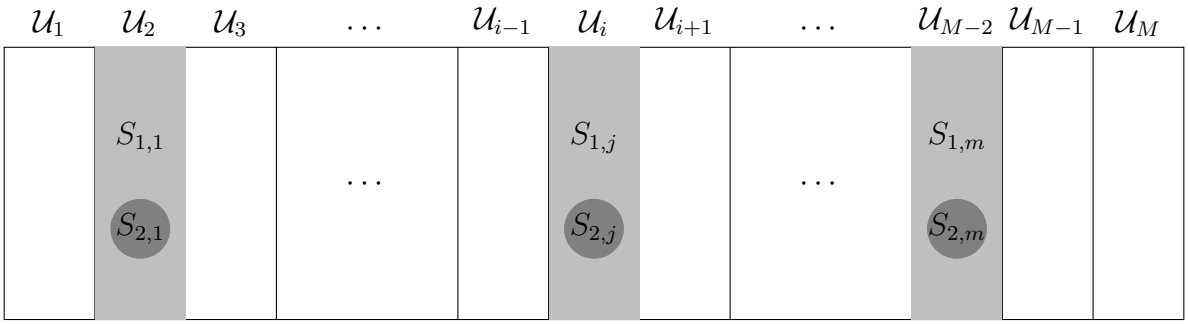

Figure 5.2: *Illustration du concept de plan à deux degrés. L'échantillon du premier degré est de taille m et est $S_1 = \cup_{j=1}^{m} S_{1,j}$. L'échantillon « final » obtenu par un plan à deux degrés est alors $S = \cup_{j \in S_1} S_{2,j}$.*

5.3.1 Population, untiés primaires et secondaires

Comme pour les sections précédentes, on supposera que la population $\mathcal{U} = \{1, \ldots, N\}$ est subdivisée en M sous–populations $\mathcal{U}_i$, $i = 1, \ldots, M$, que l'on appellera **unités primaires**. Les unités primaires sont composées de N_i **unités secondaires** et l'on a bien entendu

$$\sum_{i=1}^{M} N_i = N.$$

Pour effectuer un plan à deux degrés, il faut donc

— construire un échantillon S_1 d'unités primaires à partir d'un plan de sondage $p_1(\cdot)$ sur $\{1, \ldots, M\}$;

— pour chaque unité primaire sélectionnée, construire un échantillon S_2 sur les unités secondaires à partir d'un plan de sondage $p_2(\cdot)$.

Il est souhaitable que les plans à deux degrés possèdent les deux propriétés suivantes :

Invariance : le plan du second degré $p_2(\cdot)$ est indépendant du premier plan $p_1(\cdot)$, i.e., $\Pr(S_2 = s_2 \mid S_1 = s_1) = \Pr(S_2 = s_2)$;

Indépendance : les tirages du second degré sont mutuellement indépendants.

La Figure 5.2 essaye d'illustrer le principe de fonctionnement d'un plan de sondage à deux degrés. Clairement l'échantillon obtenu par de tels plan s'écrit

$$S = \bigcup_{i \in S_1} S_{2,i},$$

et sa taille (aléatoire) est

$$n_S = \sum_{i \in S_1} n_i, \qquad n_i = |S_{2,i}|.$$

Notons $\pi_{1,i}$ et $\pi_{1,ij}$ les probabilités d'inclusion d'ordre 1 et 2 pour le premier degré, i.e.,

$$\pi_{1,i} = \Pr(\mathcal{U}_i \in S_1), \qquad \pi_{1,ij} = \Pr(\mathcal{U}_i \in S_1, \mathcal{U}_j \in S_1).$$

Notons également $\pi_{k|i}$ la probabilité de sélectionner l'unité (secondaire) k sachant que l'unité (primaire) $\mathcal{U}_i$ a été choisie. De manière analogue on notera $\pi_{k\ell|i}$ la probabilité d'inclusion d'ordre 2 sachant que $\mathcal{U}_i$ a été retenue.

Avec ces notations, pour un $k \in \mathcal{U}_i$, la probabilité d'inclusion (usuelle) π_k s'écrit

$$\pi_k = \Pr(k \in S_{2,i}, i \in S_1) = \Pr(k \in S_{2,i} \mid i \in S_1) \Pr(i \in S_1) = \pi_{k|i} \pi_{1,i}.$$

Un même raisonnement nous conduit aux expressions pour les probabilités d'inclusion d'ordre 2,

$$\pi_{k\ell} = \begin{cases} \pi_{k\ell|i}\pi_{1,i}, & k,\ell \in \mathcal{U}_i, \\ \pi_{k|i}\pi_{\ell|j}\pi_{1,ij}, & k \in \mathcal{U}_i,\ \ell \in \mathcal{U}_j,\ i \neq j, \end{cases}$$

où pour le deuxième cas nous nous sommes servit de l'hypothèse d'indépendance pour le deuxième tirage.

5.3.2 Le π–estimateur

Rappelons que dans ce contexte le total t_y s'écrit

$$t_y = \sum_{k\in\mathcal{U}} y_k = \sum_{i=1}^{M} \sum_{k\in\mathcal{U}_i} y_k = \sum_{i=1}^{M} t_{y,i},$$

où $t_{y,i}$ est le total pour l'unité primaire $\mathcal{U}_i$, i.e.,

$$t_{y,i} = \sum_{k\in\mathcal{U}_i} y_k.$$

Le π–estimateur de ce total est donc

$$\hat{t}_{y,\pi} = \sum_{i\in S_1} \sum_{k\in S_{2,i}} \frac{y_k}{\pi_{k|i}\pi_{1,i}} = \sum_{i\in S_1} \frac{\hat{t}_{y,i}}{\pi_{1,i}},$$

où $\hat{t}_{y,i}$ est bien entendu le π–estimateur du (sous) total $t_{y,i}$, i.e.,

$$\hat{t}_{y,i} = \sum_{k\in S_{2,i}} \frac{y_k}{\pi_{k|i}}.$$

Remarque. On peut tout a fait calculer la variance du π–estimateur, mais nous ne le ferons pas...

Chapitre 6

Utilisation d'une information auxiliaire

Dans ce chapitre nous allons voir comment nous pouvons bénéficier de l'utilisation d'une information auxiliaire qui était non disponible lors de la mise en oeuvre du sondage. Le but étant bien entendu d'obtenir de meilleures estimation du paramètre d'intérêt.

6.1 Post-stratification

6.1.1 Notations

Lorsque l'on parle d'utilisation d'une information auxiliaire, il faut à tout prix connaître l'approche dite de **post-stratification**. Cette méthode fait office de référence et a en plus le bon goût d'être particulièrement simple !

On suppose que le caractère auxiliaire est **qualitatif** et peut prendre H valeurs distinctes disons $\{1, \ldots, H\}$. Ce caractère auxiliaire nous permet ainsi de former une partition de la population $\mathcal{U}$, i.e.,

$$\mathcal{U} = \bigcup_{h=1}^{H} \mathcal{U}_h, \qquad \mathcal{U}_h = \{i \in \mathcal{U} \colon y_i = h\}.$$

Remarque. Le terme post-stratification vient du fait que cette partition de la population $\mathcal{U}$ ressemble à s'y méprendre à la technique de stratification introduite au Chapitre 4. Puisque cette stratification intervient après le sondage, on parlera naturellement de **post**-stratification et de post-strates $\mathcal{U}_h$.

Le nombre d'unités N_h de la post-strate $\mathcal{U}_h$ est appelé la taille de la post-strate et bien entendu

$$N = \sum_{h=1}^{H} N_h.$$

Notons que nous supposons que les N_h sont connus et constituent notre fameuse information auxiliaire.

Comme pour le sondage stratifié, le total et la moyenne s'écrivent

$$t_y = \sum_{k \in \mathcal{U}} y_k = \sum_{h=1}^{H} \sum_{k \in \mathcal{U}_h} y_k = \sum_{h=1}^{H} N_h \mu_h$$

$$\mu_y = \frac{1}{N} \sum_{h=1}^{H} N_h \mu_h,$$

où μ_h est la moyenne sur la post-strate $\mathcal{U}_h$, i.e.,

$$\mu_h = \frac{1}{N_h} \sum_{k \in \mathcal{U}_h} y_k, \qquad h = 1, \ldots, H.$$

Nous pouvons également s'intéresser à la variance (corrigée) pour chaque post-strate

$$\sigma^2_{y,h} = \frac{1}{N_h} \sum_{k \in \mathcal{U}_h} (y_k - \mu_h)^2, \qquad S^2_{y,h} = \frac{1}{N_h - 1} \sum_{k \in \mathcal{U}_h} (y_k - \mu_h)^2, \qquad h = 1, \ldots, h.$$

Exercice 5. Montrez que l'on peut décomposer la variance totale σ^2_y à l'aide des variances des post-strates, i.e.,

$$\sigma^2_y = \frac{1}{N} \sum_{h=1}^{H} N_h \sigma^2_{y,h} + \frac{1}{N} \sum_{h=1}^{H} N_h (\mu_{y,h} - \mu_y)^2.$$

Solution.

□

6.1.2 L'estimateur post-stratifié

Supposons qu'un échantillon alétoire S de taille n ait été tiré au sein d'une population $\mathcal{U}$ de taille N à l'aide d'un plan simple sans remise. Le π–estimateur du total t_y est donc

$$\hat{t}_{y,\pi} = \sum_{k \in S} \frac{y_k}{n/N} = \frac{N}{n} \sum_{k \in S} y_k = \frac{N}{n} \sum_{\substack{h=1 \\ n_h > 0}}^{H} n_h \hat{\mu}_{y,h},$$

où n_h est la taille des post-strates et

$$\hat{\mu}_{y,h} = \frac{1}{n_h} \sum_{k \in S_h} y_k.$$

L'estimateur post-stratifié s'écrit alors

$$\hat{t}_{y,\text{post}} = \sum_{\substack{h=1 \\ n_h>0}}^{H} N_h \hat{\mu}_{y,h}.$$

Remarque. La connaissance des tailles N_h est nécessaire afin d'utiliser cet estimateur.

6.1.3 Propriété de l'estimateur

Le calcul de l'espérance de $\hat{t}_{y,\text{post}}$ est quelque peu compliqué du fait que les tailles n_h des échantillons des post-strates sont aléatoires. Commençons par calculer cette espérance sachant les tailles n_h. Nous avons

$$\begin{aligned}
\mathbb{E}(\hat{t}_{y,\text{post}} \mid n_1, \ldots, n_H) &= \sum_{\substack{h=1 \\ n_h>0}}^{H} N_h \mathbb{E}(\hat{\mu}_{y,h} \mid n_1, \ldots, n_H) \\
&= \sum_{\substack{h=1 \\ n_h>0}}^{H} t_{y,h} \\
&= t_y - \sum_{\substack{h=1 \\ n_h=0}}^{H} t_{y,h}.
\end{aligned}$$

Puisque $\mathbb{E}\{\mathbb{E}(X \mid Y)\} = \mathbb{E}(X)$, nous avons

$$\mathbb{E}(\hat{t}_{y,\text{post}}) = t_y - \sum_{h=1}^{H} t_{y,h} \Pr(n_h = 0).$$

Or puisque

$$\Pr(n_h = 0) = \Pr(\nexists k \in S \colon k \in \mathcal{U}_h) = \frac{\binom{N-N_h}{n}}{\binom{N}{n}} = \frac{(N-N_h)!(N-n)!}{(N-N_h-n)!N!},$$

on a donc

$$\mathbb{E}(\hat{t}_{y,\text{post}}) - t_y = \sum_{h=1}^{H} t_{y,h} \frac{(N-N_h)!(N-n)!}{(N-N_h-n)!N!}.$$

Remarque. L'estimateur post-stratifié n'est donc pas sans biais mais est approximativement sans biais dès lors que $\Pr(n_h = 0)$ est suffisamment faible pour tout $h = 1, \ldots, H$.

Une règle de pouce consiste à ce que les post-strates soient suffisamment grandes, i.e., que les tailles N_h des post-strates vérifient

$$n\frac{N_h}{N} \geq 30, \qquad h = 1, \ldots, H.$$

On peut également calculer la variance de l'estimateur post-stratifié en utilisant la célèbre formule

$$\text{Var}(\hat{t}_{y,\text{post}}) = \text{Var}\{\mathbb{E}(\hat{t}_{y,\text{post}} \mid n_1, \ldots, n_H)\} + \mathbb{E}\{\text{Var}(\hat{t}_{y,\text{post}} \mid n_1, \ldots, n_H)\}.$$

Mais puisque nous avons montré que

$$\mathbb{E}(\hat{t}_{y,\text{post}} \mid n_1, \ldots, n_H) = t_y - \sum_{\substack{h=1 \\ n_h=0}}^{H} t_{y,h},$$

cela implique que $\mathrm{Var}\{\mathbb{E}(\hat{t}_{y,\mathrm{post}} \mid n_1, \ldots, n_H)\} \approx 0$ dès lors que $\Pr(n_h = 0)$ est suffisamment faible. On a donc

$$\begin{aligned}\mathrm{Var}(\hat{t}_{y,\mathrm{post}}) &= \mathbb{E}\{\mathrm{Var}(\hat{t}_{y,\mathrm{post}} \mid n_1, \ldots, n_H)\} \\ &= \mathbb{E}\left\{ \sum_{\substack{h=1 \\ n_h>0}}^{H} N_h(N_h - n_h)\frac{S_{y,h}^2}{n_h} \right\} \\ &\approx \sum_{h=1}^{H} N_h \left\{ N_h \mathbb{E}(n_h^{-1}) - 1 \right\} S_{y,h}^2.\end{aligned}$$

Il reste donc à calculer $\mathbb{E}(n_h^{-1})$ ce qui n'est pas évident. En fait on calculera une approximation de cette espérance en ayant recours à la linéarisation. Ceci est un peu long mais reste tout à fait faisable...

6.2 Caractère auxiliaire quantitatif

Dans la section précédente, nous avons introduit la technique de post-stratification ; mais cette dernière supposait que l'information auxiliaire était qualitative. Parfois cette information auxiliaire sera **quantitative**.

Soit x le caractère auxiliaire (qui est quantitatif rappelons le encore une fois) dont le total

$$t_x = \sum_{k \in \mathcal{U}} x_k$$

est supposé connu.

Si l'on soupçonne que le caractère x soit lié au caractère d'intérêt y, alors on aimerait bien bénéficier de la connaissance de x pour estimer une fonction d'intérêt sur y. Dans cette section, nous allons voir différentes approches de ce type et nous supposerons qu'un **plan de sondage simple est réalisé**. Avant d'introduire ces différentes techniques, posons quelques notations.

6.2.1 Notations

Comme d'habitude on appelera

$$\mu_x = \frac{1}{N}\sum_{k \in \mathcal{U}} x_k, \qquad \mu_y = \frac{1}{N}\sum_{k \in \mathcal{U}} y_k,$$

les moyennes des caractères x et y sur la population et

$$S_x^2 = \frac{1}{N-1}\sum_{k=1}^{N}(x_k - \mu_x)^2, \qquad S_y^2 = \frac{1}{N-1}\sum_{k=1}^{N}(y_k - \mu_y)^2,$$

les variances corrigées des caractères x et y sur la population. On introduit également la nouvelle notation

$$S_{xy} = \frac{1}{N-1}\sum_{k=1}^{N}(x_k - \mu_x)(y_k - \mu_y),$$

i.e., la covariance entre le caractère x et le caractère y sur la population.

En ce qui concerne les quantités échantillonnées, on notera

$$\widehat{S_x^2} = \frac{1}{n-1}\sum_{k \in S}(x_k - \hat{\mu}_x)^2, \quad \widehat{S_y^2} = \frac{1}{n-1}\sum_{k \in S}(y_k - \hat{\mu}_y)^2, \quad \widehat{S_{xy}} = \frac{1}{n-1}\sum_{k \in S}(x_k - \hat{\mu}_x)(y_k - \hat{\mu}_y),$$

les variances et la covariance calculées à partir de l'échantillon S de taille n.

6.2.2 Estimation par la différence

L'estimateur par la différence du total t_y, noté $\hat{t}_{y,D}$, est

$$\hat{t}_{y,D} = \hat{t}_{y,\pi} + t_x - \hat{t}_{x,\pi},$$

où $\hat{t}_{x,\pi}$ et $\hat{t}_{y,\pi}$ sont les π–estimateurs des totaux t_x et t_y.

En quelque sorte l'idée de cet estimateur est de reporter l'erreur du π–estimateur commise sur l'estimation de t_x sur l'estimation de t_y.

Exercice 6. Montrez que cet estimateur est sans biais.

Solution.

□

La variance (et donc l'erreur quadratique puisque c'est un estimateur sans biais) se calcule également aisément :

$$\begin{aligned}\mathrm{Var}(\hat{t}_{y,D}) &= \mathrm{Var}(\hat{t}_{y,\pi}) + \mathrm{Var}(\hat{t}_{x,\pi}) - 2\mathrm{Cov}(\hat{t}_{x,\pi}, \hat{t}_{y,\pi}) \\ &= \frac{N(N-n)}{n}\left(S_y^2 + S_x^2 - 2S_{xy}\right).\end{aligned}$$

Cette variance sera bien entendue estimée par

$$\widehat{\mathrm{Var}}(\hat{t}_{y,D}) = \frac{N(N-n)}{n}\left(\widehat{S_y^2} + \widehat{S_x^2} - 2\widehat{S_{xy}}\right).$$

6.2.3 Estimation par le quotient

L'estimateur par le quotient du total t_y, noté $\hat{t}_{y,Q}$, est

$$\hat{t}_{y,Q} = \frac{\hat{t}_{y,\pi}}{\hat{t}_{x,\pi}} t_x.$$

En quelque sorte l'idée de cet estimateur est similaire à celle de l'estimateur par la différence mais cette fois ci l'erreur est reportée de manière multiplicative plutôt qu'additive.

Le biais de cet estimateur n'est pas calculable de manière explicite du fait de la présence d'un quotient. On aura donc recours comme d'habitude à la technique de **linéarisation**.

Puisque

$$\hat{t}_{y,Q} - t_y = \frac{\hat{t}_{y,\pi} - R\hat{t}_{x,\pi}}{\hat{t}_{x,\pi}} t_x = \frac{\hat{t}_{y,\pi} - R\hat{t}_{x,\pi}}{1+\varepsilon},$$

avec $R = t_y/t_x$ et

$$\varepsilon = \frac{\hat{t}_{x,\pi} - t_x}{t_x}.$$

A l'aide d'un développement limité de $(1+\varepsilon)^{-1}$ en $\varepsilon = 0$ et d'ordre 1, on obtient

$$\hat{t}_{y,Q} - t_y \approx (\hat{t}_{y,\pi} - R\hat{t}_{x,\pi})(1-\varepsilon).$$

Au final on peut donc avoir une approximation du biais

$$\begin{aligned}\mathbb{E}(\hat{t}_{y,Q} - t_y) &\approx -\mathbb{E}\left\{(\hat{t}_{y,\pi} - R\hat{t}_{x,\pi})\varepsilon\right\} \\ &= -\frac{\mathbb{E}(\hat{t}_{x,\pi}\hat{t}_{y,\pi}) - t_x t_y - R\mathbb{E}(\hat{t}_{x,\pi}^2) + Rt_x^2}{t_x} \\ &= \frac{R\operatorname{Var}(\hat{t}_{x,\pi}) - \operatorname{Cov}(\hat{t}_{x,\pi}, \hat{t}_{y,\pi})}{t_x} \\ &= \frac{N(N-n)}{n}\frac{RS_x^2 - S_{xy}}{t_x}.\end{aligned}$$

Remarque. Le biais devient négligeable dès lors que n est grand.

Exercice 7. Calculez une approximation de l'erreur quadratique de l'estimateur par quotient.

Solution.

□

6.2.4 Estimation par la régression

L'estimateur du total t_y par la régression est

$$\hat{t}_{y,R} = \hat{t}_{y,\pi} + \hat{a}(t_x - \hat{t}_{x,\pi}), \qquad \hat{a} = \frac{\widehat{S_{xy}}}{\widehat{S_x^2}}.$$

L'idée de cet estimateur est de supposer qu'il existe une relation linéaire de la forme $y = ax + b$ entre les caractères x et y et donc que

$$t_y \approx \hat{a}t_x + \hat{b}, \qquad \hat{t}_{y,\pi} \approx \hat{a}\hat{t}_{x,\pi} + \hat{b}.$$

On estime alors le total par

$$\hat{t}_{y,\pi} + (t_y - \hat{t}_{y,\pi}) = \hat{t}_{y,\pi} + \hat{a}(t_x - \hat{t}_{x,\pi}).$$

Comme pour les estimateurs précédents, le calcul de l'espérance de $\hat{t}_{y,R}$ ne peut être qu'approché. Puisque

$$\hat{t}_{y,R} = \hat{t}_{y,\pi} + a(t_x - \hat{t}_{x,\pi}) + (\hat{a} - a)(t_x - \hat{t}_{x,\pi}), \qquad a = \frac{S_{xy}}{S_x^2},$$

Table 6.1: *Récapitulatif des différentes méthodes de redressement à l'aide d'une variable quantitative.*

Estimateur	Définition	$\left\{\frac{N(N-n)}{n}\right\}^{-1} \times$ EQM
π–estimateur	$\hat{t}_{y,\pi} = n^{-1}N\sum_{k\in S} y_k$	S_y^2
par la différence	$\hat{t}_{y,D} = \hat{t}_{y,\pi} + t_x - \hat{t}_{x,\pi}$	$S_y^2 + S_x^2 - 2S_{xy}$
par le quotient	$\hat{t}_{y,Q} = \hat{t}_{y,\pi} t_x / \hat{t}_{x,\pi}$	$S_y^2 + R^2 S_x^2 - 2RS_{xy}$
par la régression	$\hat{t}_{y,R} = \hat{t}_{y,\pi} + \hat{a}(t_x - \hat{t}_{x,\pi})$	$S_y^2(1-\rho^2)$

et où l'on peut montrer (admis) que le dernier terme est négligeable, on a donc

$$\mathbb{E}(\hat{t}_{y,R}) \approx \mathbb{E}\{\hat{t}_{y,\pi} + a(t_x - \hat{t}_{x,\pi})\} = t_y.$$

L'erreur quadratique est approchée par

$$\begin{aligned}
\mathrm{EQM}(\hat{t}_{y,R}) &\approx \mathrm{Var}(\hat{t}_{y,\pi}) + a^2\,\mathrm{Var}(\hat{t}_{x,\pi}) - 2a\mathrm{Cov}(\hat{t}_{x,\pi}, \hat{t}_{y,\pi}) \\
&= \frac{N(N-n)}{n}\left(S_y^2 + a^2 S_x^2 - 2aS_{xy}\right) \\
&= \frac{N(N-n)}{n}\left(S_y^2 + \frac{S_{xy}^2}{S_x^2} - 2\frac{S_{xy}^2}{S_x^2}\right) \\
&= \frac{N(N-n)}{n}\left(S_y^2 - \frac{S_{xy}^2}{S_x^2}\right) \\
&= \frac{N(N-n)}{n} S_y^2\left(1-\rho^2\right), \qquad \rho = \frac{S_{xy}}{S_x S_y}.
\end{aligned}$$

On estimera cette dernière par

$$\frac{N(N-n)}{n}\widehat{S_y^2}\left(1-\hat{\rho}^2\right), \qquad \hat{\rho} = \frac{\widehat{S_{xy}}}{\widehat{S_x}\widehat{S_y}}.$$

6.2.5 Comparaison

Le Tableau 6.1 donne l'expression des erreurs quadratiques moyennes pour les différents estimateurs par redressement introduit précédemment ainsi, qu'à titre de référence, celle du π–estimateur. Nous allons donc maintenant comparer ces estimateurs deux à deux afin d'établir une "règle de décision" afin de choisir le meilleur estimateur — au sens de l'erreur quadratique bien entendu.

— Estimateur par la différence vs. π–estimateur :

$$\begin{aligned}
\mathrm{EQM}(\hat{t}_{y,\pi}) - \mathrm{EQM}(\hat{t}_{y,D}) &= \frac{N(N-n)}{n}S_y^2 - \frac{N(N-n)}{n}(S_y^2 + S_x^2 - 2S_{xy}) \\
&= \frac{N(N-n)}{n}\left(2S_{xy} - S_x^2\right).
\end{aligned}$$

L'estimateur par la différence est donc meilleur lorsque

$$2S_{xy} - S_x^2 > 0 \iff a > \frac{1}{2}.$$

— Estimateur par quotient vs. π–estimateur :

$$\begin{aligned}
\mathrm{EQM}(\hat{t}_{y,\pi}) - \mathrm{EQM}(\hat{t}_{y,Q}) &\approx \frac{N(N-n)}{n}S_y^2 - \frac{N(N-n)}{n}\left(S_y^2 + R^2 S_x^2 - 2RS_{xy}\right) \\
&= \frac{N(N-n)}{n}\left(2RS_{xy} - R^2 S_x^2\right).
\end{aligned}$$

L'estimateur par le quotient est donc (approximativement ! ! !) meilleur lorsque

$$2RS_{xy} - R^2S_x^2 > 0 \iff \begin{cases} a > \frac{R}{2}, & R > 0, \\ a < \frac{R}{2}, & R \leq 0. \end{cases}$$

— Estimateur par le quotient vs. estimateur par la différence :

$$\begin{aligned} \mathrm{EQM}(\hat{t}_{y,D}) - \mathrm{EQM}(\hat{t}_{y,Q}) &\approx \frac{N(N-n)}{n}\left(S_y^2 + S_x^2 - 2S_{xy}\right) - \\ &\quad \frac{N(N-n)}{n}\left(S_y^2 + R^2S_x^2 - 2RS_{xy}\right) \\ &= \frac{N(N-n)}{n}\left\{(1-R^2)S_x^2 + 2(1-R)S_{xy}\right\}. \end{aligned}$$

L'estimateur par le quotient est donc (approximativement ! ! !) meilleur lorsque

$$(1-R^2)S_x^2 + 2(1-R)S_{xy} > 0 \iff 2(1-R)a > 1 - R^2.$$

— Estimateur par régression vs. "les autres" : Cet estimateur est (approximativement ! ! !) le meilleurs de tous. En effet

$$\begin{aligned} \mathrm{EQM}(\hat{t}_{y,\pi}) - \mathrm{EQM}(\hat{t}_{y,R}) &\approx \frac{N(N-n)}{n}S_y^2\rho^2 \\ &= \rho^2\mathrm{EQM}\left(\hat{t}_{y,\pi}\right) \geq 0 \\ \mathrm{EQM}(\hat{t}_{y,D}) - \mathrm{EQM}(\hat{t}_{y,R}) &\approx \frac{N(N-n)}{n}\left(\rho^2S_y^2 + S_x^2 - 2S_{xy}\right) \\ &= \frac{N(N-n)}{n}\left(\frac{S_{xy}^2}{S_x^2} + S_x^2 - 2S_{xy}\right) \\ &= \frac{N(N-n)}{n}\left(\frac{S_{xy}^2}{S_x^2} - S_x\right)^2 \geq 0 \\ \mathrm{EQM}(\hat{t}_{y,Q}) - \mathrm{EQM}(\hat{t}_{y,R}) &\approx \frac{N(N-n)}{n}\left(\rho^2S_y^2 + R^2S_x^2 - 2RS_{xy}\right) \\ &= \frac{N(N-n)}{n}\left(\frac{S_{xy}^2}{S_x^2} + R^2S_x^2 - 2RS_{xy}\right) \\ &= \frac{N(N-n)}{n}\left(\frac{S_{xy}}{S_x} - RS_x\right)^2 \geq 0 \end{aligned}$$

Remarque. Il faut tout de même nuancer le fait que l'estimateur par régression soit toujours meilleur que les autres estimateurs, puisque ce n'est que du calcul approché. De plus l'estimateur par régression requiert l'estimation de la "pente" a ; et la variabilité de l'estimation de a n'a pas été prise en compte dans nos calculs.

Conclusion

Ce cours est maintenant terminé ; j'espère qu'il vous aura plu et que vous aurez appris beaucoup de choses. J'espère qu'avec un peu de recul maintenant sur la théorie des sondages, vous remarquez que les éléments théoriques ne sont pas en fait si nombreux et qu'ainsi la plupart des formules peuvent se retrouver facilement...

Feuilles d'exercices

Vous trouverez plus bas, les feuilles d'exercices que nous allons utiliser tout au long de ce cours. Puisque vous les avez sous la main profitez en pour travailler à la maison...

TD 1 : Se familiariser avec les bases/notations

Exercice 1. Soient une population $\mathcal{U} = \{1, 2, 3\}$ et le plan $p(\cdot)$ suivant

$$p(\{1,2\}) = \frac{1}{2}, \qquad p(\{1,3\}) = \frac{1}{4}, \qquad p(\{2,3\}) = \frac{1}{4}.$$

1. Donnez les probabilités d'inclusion d'ordre un.
2. Pourquoi la somme de ces probabilités d'inclusion vaut nécessairement 2 ?
3. Donnez la matrice de variance–covariance des variables indicatrices $1_{\{k \in S\}}$, $k \in \mathcal{U}$.

Exercice 2. Soient une population $\mathcal{U} = \{1, 2, 3\}$ et le plan $p(\cdot)$ suivant

$$p(\{1,2\}) = \frac{1}{2}, \qquad p(\{1,3\}) = \frac{1}{4}, \qquad p(\{2,3\}) = \frac{1}{4}.$$

1. Donnez la distribution de probabilité du π–estimateur de la moyenne μ_y pour un caractère d'intérêt y.
2. En déduire le biais de cet estimateur.

Exercice 3. Soit la matrice de variance–covariance $\Delta = (\Delta_{k\ell})_{k,\ell}$ des indicatrices $1_{\{k \in S\}}$ pour un plan $p(\cdot)$ donné. On sait que

$$\Delta = \frac{6}{25} \begin{bmatrix} 1 & 1 & 1 & -1 & -1 \\ 1 & 1 & 1 & -1 & -1 \\ 1 & 1 & 1 & -1 & -1 \\ -1 & -1 & -1 & 1 & 1 \\ -1 & -1 & -1 & 1 & 1 \end{bmatrix}.$$

1. Le plan $p(\cdot)$ est-il de taille fixe ?
2. Satisfait il aux conditions de Sen–Yates–Grundy ?
3. Sachant que $\pi_1 = \pi_2 = \pi_3 > \pi_4 = \pi_5$, calculer les probabilités d'inclusions d'ordre 1.
4. Donnez la matrice des probabilités d'inclusions d'ordre deux.
5. Donnez les probabilités associées à tous les échantillons possibles.

Exercice 4. On considère un plan sans remise effectué sur une population de taille N. On suppose que les probabilités d'inclusions d'ordre 1 et 2 π_k et $\pi_{k\ell}$ sont strictement positives. A partir d'un échantillon aléatoire S, on s'intéresse à l'estimateur suivant

$$\hat{\theta} = \frac{1}{N^2} \sum_{k \in S} \frac{y_k}{\pi_k} + \frac{1}{N^2} \sum_{\substack{k,\ell \in S \\ k \neq \ell}} \frac{y_\ell}{\pi_{k\ell}}.$$

1. Pour quelle fonction d'intérêt cet estimateur est il sans biais ?

Exercice 5. 1. Montrez que

$$\frac{1}{N}\sum_{k\in\mathcal{U}}(y_k-\mu_y)^2=\frac{1}{2N^2}\sum_{\substack{k,\ell\in\mathcal{U}\\k\neq\ell}}(y_k-y_\ell)^2.$$

2. Pour un plan sans remise quelconque mais dont les probabilités d'inclusion d'ordre 1 et 2 sont strictement positives, construisez un estimateur sans biais de σ_y^2.

TD 2 : Plans simples

Exercice 6. On souhaite estimer la surface moyenne cultivée dans les fermes d'un canton rural donné. Sur les $N = 2010$ fermes de ce canton, on en tire 100 par sondage aléatoire simple. On mesure y_k la surface cultivée dans la ferme k en hectares, et l'on trouve

$$\sum_{k\in S} y_k = 2907\text{ha}, \qquad \sum_{k\in S} y_k^2 = 154593\text{ha}^2.$$

1. Donnez l'estimateur sans biais classique de la moyenne

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

2. Donnez un intervalle de confiance à 95% pour μ_y.

Exercice 7. On s'intéresse à la proportion d'hommes atteints par une maladie professionnelle dans une entreprise de 1500 travailleurs. On sait par ailleurs que trois travailleurs sur dix sont ordinairement touchés par cette maladie dans des entreprises du même type. On se propose de sélectionner un échantillon au moyen d'un sondage aléatoire simple.

Quelle taille d'échantillon faut–il sélectionner pour que la longueur totale d'un intervalle de confiance avec un niveau de confiance 0.95 soit inférieure à 0.02 pour les plans simples avec et sans remise ?

Exercice 8. Un échantillon de 100 étudiants est constitué au moyen d'un plan aléatoire simple sans remise dans une population de 1000 étudiants. Le résultat obtenu est présenté au sein du Tableau 2.

Table 2: *Nombre de succès/échec selon le sexe pour un échantillon de 100 étudiants pris parmi 1000.*

	Hommes	Femmes	Total
Réussite	$n_{11} = 35$	$n_{12} = 25$	$n_{1\cdot} = 60$
Échec	$n_{21} = 20$	$n_{22} = 20$	$n_{2\cdot} = 40$
Total	$n_{\cdot 1} = 55$	$n_{\cdot 2} = 45$	$n = 100$

— Estimez le taux de réussite des hommes et des femmes.
— Calculez le biais (approché) des taux de réussite.
— Estimez l'erreur quadratique moyenne de ces taux de réussite.

TD 3 : Plans à probablités inégales

Exercice 9. Une population est composée de 6 ménages de tailles respectives 2, 4, 3, 9, 1 et 2 (la taille x_k d'un ménage k est le nombre de personnes physiques qu'il comprend). On tire 3 ménages sans remise, avec une probabilité proportionnelle à leur taille.

1. Donnez les probabilités d'inclusion des 6 ménages de la base de sondage—soyez prudent...
2. Réalisez effectivement le tirage par une méthode systématique.
3. A partir de l'échantillon obtenu en 2, donnez une estimation de la taille moyenne $\bar{x}$ des ménages ; le résultat était-il prévisible ?

Exercice 10. On a répertorié dans une petite municipalité 6 entreprises dont les chiffres d'affaires (variable x_k) sont respectivement de 40, 10, 8, 1, 0.5 et 0.5 millions d'euros. Dans le but d'estimer l'emploi salarié total, sélectionnez trois entreprises au hasard et sans remise, à probabilités inégales selon le chiffre d'affaires, par la méthode du tirage systématique (en justifiant votre démarche). Pour ce faire, on utilise la réalisation suivante d'une variable aléatoire $U(0,1)$: 0.83021. Que se passe-t-il si on modifie l'ordre du fichier ?

Exercice 11. Soit une population de 5 unités. On veut sélectionner par un tirage systématique à probabilités inégales un échantillon de deux unités avec des probabilités d'inclusion proportionnelles aux valeurs x_i suivantes

$$1, 1, 6, 6, 6.$$

1. Calculez les probabilités d'inclusion d'ordre un.
2. Considérant les deux unités dont la valeur x_i vaut 1, calculez leurs probabilités d'inclusion d'ordre deux pour chacune des permutations possibles du fichier. Conséquence ?

Exercice 12. Soit une population $\mathcal{U}$ composée de 6 unités. On connaît les valeurs prises par un caractère auxiliaire x sur toutes les unités de la population :

$$x_1 = 200, \quad x_2 = 80, \quad x_3 = 50, \quad x_4 = 50, \quad x_5 = 10, \quad x_6 = 10.$$

1. Calculez les probabilités d'inclusion d'ordre un proportionnelles aux x_k pour une taille d'échantillon $n = 4$. Soit 0.48444 une réalisation d'une $U(0,1)$. Sélectionnez un échantillon à probabilités inégales sans remise de taille 4 au moyen d'un tirage systématique, en gardant l'ordre initial du fichier.
2. Donnez la matrice des probabilités d'inclusions d'ordre deux (ordre initial du fichier fixé).
3. On suppose qu'une variable d'intérêt y prend les valeurs suivantes :

$$y_1 = 80, \quad y_2 = 50, \quad y_3 = 30, \quad y_4 = 25, \quad y_5 = 10, \quad y_6 = 5.$$

Constituez un tableau avec, en ligne chaque échantillon s possible, et en colonne les probabilités de tirage $p(s)$, les estimateurs respectifs du total $\hat{Y}(s)$ et de la variance $\widehat{\text{Var}}[\hat{Y}]$. Calculez, sur la base de ce tableau, les espérances $\mathbb{E}[\hat{Y}]$ et $\mathbb{E}[\widehat{\text{Var}}[\hat{Y}]]$. Commentez.

TD 4 : Plans stratifiés

Exercice 13. Dans une population $\mathcal{U} = \{1, 2, 3, 4, 5\}$, on considère le plan de sondage suivant :

$$p(\{1,2,4\}) = p(\{1,2,5\}) = p(\{1,4,5\}) = p(\{2,3,4\}) = p(\{2,3,5\}) = p(\{3,4,5\}) = \frac{1}{6}.$$

Calculez les probabilités d'inclusion d'ordre un et deux ainsi que les $\Delta_{k\ell}$. Montrez qu'il s'agit d'un plan stratifié.

Exercice 14. On considère une population $\mathcal{U}$ de taille N partitionnée en H strates notées $\mathcal{U}_1, \ldots, \mathcal{U}_H$, de tailles respectives $N_1, \ldots, N_H$. On note également $\mu_{y,1}, \ldots, \mu_{y,_H}$ les moyennes de chaque strates.

Pour chaque strate, on sélectionne un échantillon selon un plan aléatoire simple sans remise de taille n_h, $h = 1, \ldots, H$. Les tirages sont indépendants d'une strate à l'autre. Un jeune statisticien propose d'estimer μ_Y par

$$\hat{\mu}_Y = \frac{1}{n} \sum_{k \in S} y_k, \qquad n = \sum_{h=1}^{H} n_h.$$

1. Calculez $\mathbb{E}(\hat{\mu}_Y)$ et en déduire le biais de $\hat{\mu}_Y$.
2. Calculez $\mathrm{Var}(\hat{\mu}_Y)$.
3. Calculez le *ratio du biais*, i.e., le rapport entre le biais et l'écart-type de $\hat{\mu}_Y$.
4. Pourquoi ne faut il pas utiliser cet estimateur.

Exercice 15. Sur les 7500 employés de l'INSEE, on souhaite connaître la proportion P d'entre eux qui possèdent au moins un véhicule.

Pour chaque individu de la base de sondage, on dispose de la valeur de son revenu. On décide alors de constituer trois strates dans la population : individus de revenu faible (strate 1), de revenu moyen (strate 2), de revenu élevé (strate 3).

Table 3: *Sondage stratifié sur les 7500 employés de l'INSEE.*

	$h=1$	$h=2$	$h=3$
N_h	3500	2000	2000
n_h	500	300	200
p_h	0.13	0.45	0.50

On note

N_h :	la taille de la strate h
n_h :	la taille de l'échantillon dans la strate h
p_h :	l'estimateur de la proportion d'individus possédant au moins un véhicule dans la strate h

Les détails du sondage sont donnés au sein du Tableau 3.

1. Quel estimateur, noté $\hat{P}$, de P proposez vous ? Que peut–on dire de son biais ?
2. Calculez la précision de $\hat{P}$ et donnez un intervalle de confiance à 95% pour P.
3. Estimez–vous que le critère de stratification est adéquat ? Justifiez votre réponse.

Exercice 16. Un directeur de cirque possède 100 éléphants classés en 2 catégories : « mâles et femmelles ». Le directeur veut estimer le poids total de son troupeau car il veut traverser un fleuve en bateau. Cependant, l'année précédente, ce même directeur de cirque avait fait peser tous les éléphants de son troupeau et avait obtenu les résultats présentés dans le Tableau 4.

Table 4: *Poids moyens (tonnes) et dispersions selon les strates pour les 100 éléphants du cirque l'année précédente.*

	Effectifs N_h	Moyennes $\mu_{y,h}$	Dispersions $S^2_{y,h}$
Mâles	60	6	4
Femelles	40	4	2.25

1. Calculez la dispersion dans la population de la variable « poids de l'éléphant » pour l'année précédente.
2. Le directeur suppose désormais que les dispersions de poids n'évoluent pas sensiblement d'une année sur l'autre. Si le directeur procède à un tirage aléatoire simple sans remise de 10 éléphants, quelle est la variance de l'estimateur du poids total du troupeau ?
3. Si le directeur procède à un tirage stratifié avec allocation proportionnelle de 10 éléphants, quelle est la variance de l'estimateur du poids total du troupeau ?
4. Si le directeur procède à un tirage stratifié optimal de 10 éléphants, quels sont les effectifs de l'échantillon dans chacune des 2 strates et quelle est la variance de l'estimateur total ?

TD 5 : Plans par grappes et à plusieurs degrés

Exercice 17. L'objectif est d'estimer le revenu moyen des ménages. Dans un arrondissement d'une ville composée de 60 îlots de maisons, on sélectionne 3 îlots à probabilités égales et sans remise. On sait, en outre, que 5000 ménages résident dans cet arrondissement. Le résultat est donné dans le tableau suivant :

Table 5: *Étude du revenu moyen des ménages par un sondage simple sans remise sur 3 îlots pris parmi 60.*

Numéro de l'îlot	Nombre de ménages dans l'îlot	Revenu total des ménages dans l'îlot
1	120	1500
2	100	2000
3	80	2100

1. Estimez le revenu moyen des ménages de l'arrondissement et les revenus totaux des ménages de l'arrondissement par le π–estimateur.
2. Estimez la variance du π–estimateur de la moyenne. (Il est demandé de retrouver les formules du cours et pas de les appliquer).
3. Estimez le revenu moyen des ménages dans l'arrondissement par le ratio de Hájek. Commentez.

Exercice 18. Soit la population $\mathcal{U} = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$ et le plan de sondage suivant :

$$
\begin{array}{lll}
p(\{1,2\}) = 1/6, & p(\{1,3\}) = 1/6, & p(\{2,3\}) = 1/6, \\
p(\{4,5\}) = 1/12, & p(\{4,6\}) = 1/12, & p(\{5,6\}) = 1/12, \\
p(\{7,8\}) = 1/12, & p(\{7,9\}) = 1/12, & p(\{8,9\}) = 1/12.
\end{array}
$$

1. Donnez les probabilités d'inclusion d'ordre un.
2. Ce plan est–il simple, stratifié, en grappes, à deux degrés ou aucun de ces plans particuliers ? Justifiez votre réponse.

Exercice 19. On veut faire une enquêtes sur les hôtels hors région parisienne (90 départements). On met en place un plan à deux degrés en tirant dans un premier temps 10 départements (simple, sans remise) puis en interrogeant 20% des hôtels de chaque département sur la proportion d'étrangers dans sa clientèle (encore simple, sans remise). Les résultats sont résumés dans le tableau suivant :

Table 6: *Résultat du sondage sur la proportion d'étrangers dans les hôtels.*

N° département	18	43	25	45	3	57	23	32	28	68
Nb. hôtels	50	65	45	48	52	58	42	66	40	56
Nb. hôtels sondés	10	13	9	10	10	12	8	13	8	11
Proportion	0.40	0.38	0.22	0.30	0.50	0.25	0.38	0.31	0.25	0.36

1. Montrez qu'un estimateur de la proportion moyenne p sur toute la France (sauf la région parisienne) est donné tout simplement par

$$\hat{p} = \frac{\sum_{i \in S_1} N_i \hat{p}_i}{\sum_{i \in S_1} N_i},$$

où vous prendrez soin de définir les quantités N_i, S_1 et $\hat{p}_i$.

2. Selon vous, cet estimateur est–il sans biais ? (pas de calcul demandé)
3. Estimez la proportion d'étranger sur les 90 départements.

Exercice 20. On considère un plan à deux degrés pour lequel le π–estimateur du total t_y s'écrit

$$\hat{t}_{y,\pi} = \sum_{i \in S_1} \frac{\hat{t}_{y,i}}{\pi_{1,i}}, \qquad \hat{t}_{y,i} = \sum_{k \in S_{2,i}} \frac{y_k}{\pi_{k|i}}.$$

1. Montrez que la variance de cet estimateur peut s'écrire sous la forme $V_{\text{up}} + V_{\text{us}}$ où V_{up} est un terme de variance relié aux unités primaires et V_{us} aux unités secondaires.
2. Que devient cette expression lorsque les unités primaires et secondaires sont choisies selon un plan simple sans remise ?

Astuce : On se rappellera (ou pas !) que $Var(Y) = Var\{\mathbb{E}(Y \mid X)\} + \mathbb{E}\{Var(Y \mid X)\}$.

TD 6 : Utilisation d'une information auxiliaire

Exercice 21. Soit un plan stratifié composé de H strates de taille N_h. On souhaite estimer la moyenne de la population μ_y d'un caractère y. Notons $\mu_{x,h}$, $h = 1, \ldots, H$, les moyennes dans les strates d'un caractère auxiliaire x qui sont supposées connues.

Votre chef propose d'estimer μ_y via l'estimateur suivant

$$\tilde{\mu}_y = \hat{\mu}_y + \mu_x - \hat{\mu}_x,$$

où $\hat{\mu}_y$ et $\hat{\mu}_x$ sont les π–estimateurs μ_y et μ_x respectivement.

On réalise un sondage aléatoire simple, sans remise dans chaque strate.

1. Montrez que $\tilde{\mu}_y$ est un estimateur sans biais de μ_y.
2. Donnez sa variance.
3. Quelle est l'allocation optimale des n_h pour minimiser sa variance ? On négligera le facteur de correction en population finie.
4. Quand est-ce que $\tilde{\mu}_y$ est préférable à $\hat{\mu}_y$?

Exercice 22. Deux dentistes font une enquête sur l'état des dents des 200 enfants d'un village. Le premier dentiste sélectionne selon un sondage aléatoire simple 20 enfants parmi les 200, et comptabilise les effectifs dans l'échantillon selon le nombre de dents cariées. Les résultats sont présentés au sein du tableau ci-dessous.

Nombre de dents cariées	0	1	2	3	4	5	6	7	8
Nombre d'enfants	8	4	2	2	1	2	0	0	1

Le second dentiste examine les 200 enfants, mais dans le seul but de déterminer ceux qui n'ont aucune carie. Il constate que 50 enfants sont dans ce cas.

1. Estimez le nombre moyen de dents cariées par enfant dans le village en utilisant seulement les résultats du premier dentiste. Estimez la précision de l'estimateur obtenu et l'intervalle de confiance associé.
2. Proposez un autre estimateur du nombre moyen de dents cariées par enfant en utilisant les résultats des deux dentistes. Calculez la nouvelle estimation et appréciez le gain d'efficacité obtenu.

Exercice 23. Le directeur d'une entreprise de confection de chaussure veut estimer la longueur moyenne des pieds droits des hommes adultes d'une ville. Soient y le caractère « longueur du pied droit » (en cm) et x la taille de l'individu (en cm).

Le directeur sait en outre par les résultats d'un recensement que la taille moyenne des hommes adultes de cette ville est de 168cm. Pour estimer la longueur des pieds, le directeur effectue un sondage aléatoire simple sans remise de 100 hommes adultes. Les résultats sont les suivants :

$$\bar{x} = 169, \quad \bar{y} = 24, \quad s_{xy} = 15, \quad s_x^2 = 100, \quad s_y^2 = 4.$$

Sachant que 400000 hommes adultes vivent dans cette ville,

1. Calculez le π–estimateur, l'estimateur par le quotient, l'estimateur par différence et l'estimateur par la régression.
2. Estimez les variances de ces 4 estimateurs.
3. Quel estimateur conseilleriez–vous au directeur ?
4. Exprimez la différence entre la variance estimée de l'estimateur par le quotient et la variance estimée de l'estimateur par la régression, en fonction de $\bar{x}, \bar{y}$, de la pente $\hat{a}$ de la régression de y sur x. Commentez.