Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Département fédéral de l'intérieur DFI
Office fédéral de la statistique OFS

Une introduction à la théorie des sondages

Paul-André Salamin
Service de méthodes statistiques METH
Office fédéral de la statistique

Corso di aggiornamento di matematica
Lugano, 6 mars 2009

Table des matières

Buts

- Donner un aperçu de la théorie des sondages.
- Montrer les liens entre la théorie des sondages et la théorie des probabilités.
- Illustrer par des simulations les propriétés de différents types d'estimateurs utilisés dans les enquêtes par échantillonnage.

Concepts de base

POPULATION $\qquad$ CARACTERISTIQUE

$\downarrow \qquad\qquad\qquad \uparrow$

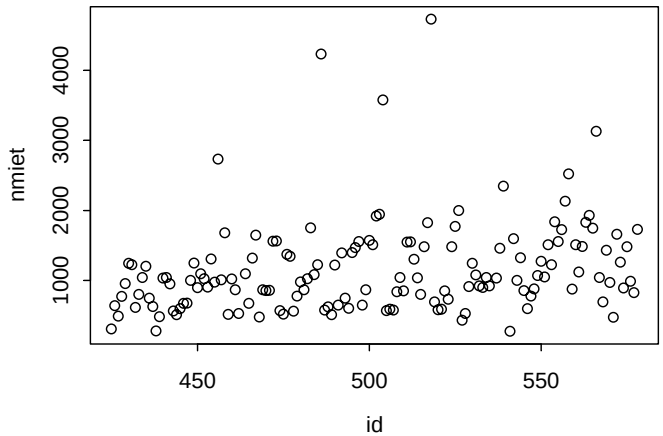
ECHANTILLON $\longrightarrow$ DONNEES

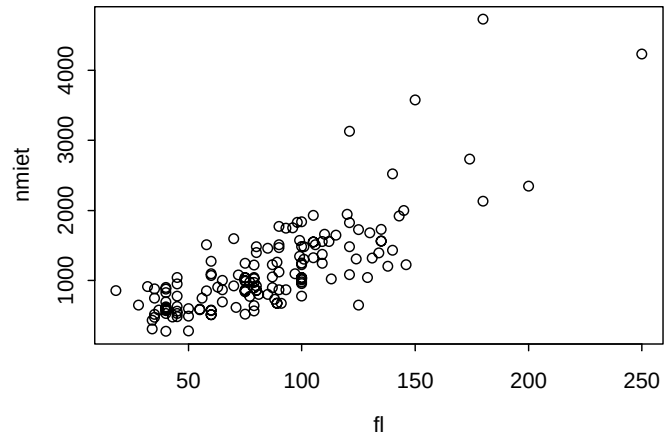
Appartements dans une commune

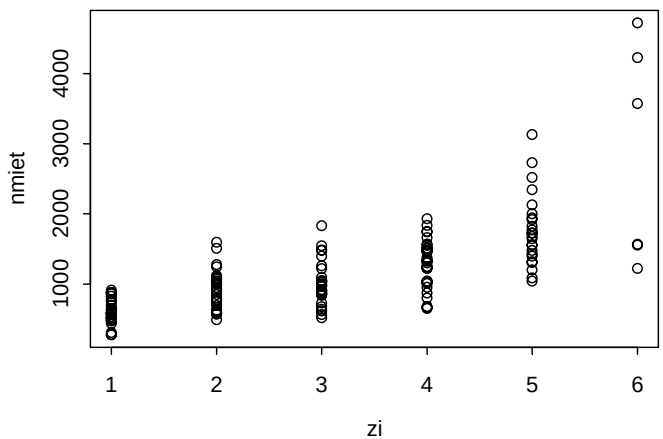
Population de 151 logements dans une commune. On s'intéresse aux loyers.

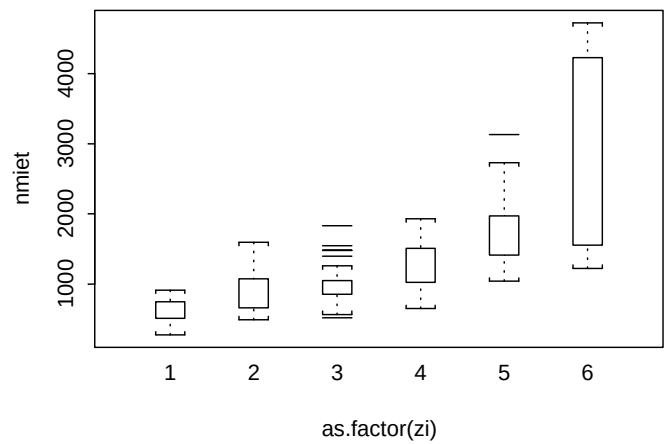
Moyenne	1158	Ecart-type	657
		Coefficient de variation	57%
Médiane	1025		
Minimum	274	Etendue	4451
Maximum	4725		
Q1	740	Intervalle interquartile	735
Q3	1475	Coefficient interquartile	72%


Statistique suisse


nmiet
1000
2000
3000
4000
450
500
550
id


nmiet
1000
2000
3000
4000
50
100
150
200
250
fl


nmiet
1000
2000
3000
4000
1
2
3
4
5
6
zi


nmiet
1000
2000
3000
4000
1
2
3
4
5
6
as.factor(zi)

Population (1)

- On s'intéresse à une population U bien précise, la *population cible*, $U = \{1, \ldots, i, \ldots, N\}$, de taille N.
- Les éléments $i \in U$ de la population sont les *unités d'observation*.

Population (2)

- La liste des unités de la population est établie à partir de recensements et/ou de registres administratifs.
- Le *cadre de sondage F* est la liste des unités d'échantillonnage.
- Une *unité d'échantillonnage* est l'unité qui est effectivement sélectionnés dans un échantillon.

Population (3)

Idéalement, le cadre de sondage F est identique à la population cible U. En général, il peut arriver que

- les unités d'observation ne soient pas les mêmes que les unités d'échantillonnage,
- $F \setminus U \neq \emptyset$: il y a des unités du cadre de sondage qui ne sont pas dans la population cible (sur-couverture),
- $U \setminus F \neq \emptyset$: il y a des unités de la population cible qui ne sont pas dans le cadre de sondage (sous-couverture).

Population (4)

- *Variable d'intérêt* = variable d'enquête = y.
- Valeur de la variable y pour l'unité $i \in U$: y_i.
- *Variable auxiliaire* = variable du cadre de sondage = x.
- Valeur de la variable x pour l'unité $i \in U$: x_i.

i	x_1	x_2	y_1	y_2
1	1	34	310	1
2	3	79	639	0
3	2	50	490	0
4	2	77	772	0
5	4	100	955	1
6	4	109	1245	0
7	4	87	1226	1
8	3	71	614	1
9	4	85	800	0
10	5	129	1040	1
11	5	138	1204	1
12	1	35	749	0

Caractéristiques d'une population (1)

Caractéristique ou *paramètre* = fonction des y_i pour $i \in U$.

Variable quantitative

- Total $Y = \sum_{i \in U} y_i$
- Moyenne $\overline{Y} = \frac{1}{N} \sum_{i \in U} y_i = \frac{1}{N} Y$

Caractéristiques d'une population (2)

Variable qualitative de modalités $a = 1, \ldots, A$.

- Effectifs N_a, $a = 1, \ldots, A$
- Proportions $p_a = \frac{N_a}{N}$, $a = 1, \ldots, A$.

Pour chaque modalité $a = 1, \ldots, A$ on définit la variable

$$y_{ai} = \begin{cases} 1 & \text{si } i \text{ est de modalité } a \\ 0 & \text{autrement} \end{cases}$$

Alors $p_a = \frac{1}{N} \sum_{i \in U} y_{ai}$.

Caractéristiques d'une population (3)

Pour une décomposition de la population U en domaines, $U = \bigcup_{k \in K} U_k$, on a

- Taille du domaine $N_k = |U_k|$
- Total dans un domaine $Y_k = \sum_{i \in U_k} y_i$
- Moyenne dans un domaine $\overline{Y}_k = \frac{1}{N_k} \sum_{i \in U_k} y_i = \frac{1}{N_k} Y_k$
- Proportion dans un domaine $p_{ak} = \frac{N_{ak}}{N_k}$

Caractéristiques d'une population (4)

On définit les variables indicatrices des domaines

$$z_{ki} = \begin{cases} 1 & \text{si } i \in U_k \\ 0 & \text{si } i \notin U_k \end{cases}$$

Alors

$$N_k = \sum_{i \in U} z_{ki}$$

$$Y_k = \sum_{i \in U} z_{ki} y_i$$

$$\overline{Y}_k = \frac{Y_k}{N_k} = \frac{\sum_{i \in U} z_{ki} y_i}{\sum_{i \in U} z_{ki}}$$

$$p_{ak} = \frac{N_{ak}}{N_k} = \frac{\sum_{i \in U} z_{ki} y_{ai}}{\sum_{i \in U} z_{ki}}$$

Caractéristiques d'une population (5)

Caractéristiques décrivant la dispersion d'une variable y dans la population.

- Variance $D^2 = D_y^2 = \frac{1}{N-1} \sum_{i \in U} (y_i - \overline{Y})^2$
- Ecart-type $D = D_y$
- Coefficient de variation $\mathrm{CV} = \mathrm{CV}_y = D_y / \overline{Y}$

Echantillon

- Un échantillon $S \subseteq U$ est un sous-ensemble de la population
- $S =$ échantillon brut
- Taille de l'échantillon $|S| = n$, taille brute

i	x_1	x_2	y_1	y_2	brut
1	1	34	.	.	1
2	3	79	.	.	0
3	2	50	.	.	0
4	2	77	.	.	1
5	4	100	.	.	1
6	4	109	.	.	1
7	4	87	.	.	0
8	3	71	.	.	1
9	4	85	.	.	1
10	5	129	.	.	0
11	5	138	.	.	1
12	1	35	.	.	0

Données

- L'échantillon des répondants $R \subseteq S$ est un sous-ensemble de l'échantillon brut
- $R =$ échantillon net
- Nombre de répondants $|R| = m$, taille nette
- Fichier avec m lignes et $p + q$ colonnes, où $p =$ nombre de variables d'enquête et $q =$ nombre de variables auxiliaires

i	x_1	x_2	y_1	y_2	brut	net
1	1	34	310	1	1	1
2	3	79	.	.	0	.
3	2	50	.	.	0	.
4	2	77	772	0	1	1
5	4	100	.	.	1	0
6	4	109	1245	0	1	1
7	4	87	.	.	0	.
8	3	71	614	1	1	1
9	4	85	.	.	1	0
10	5	129	.	.	0	.
11	5	138	.	.	1	0
12	1	35	.	.	0	.

Les flèches

- Population $\longrightarrow$ Echantillon
  - Comment choisir l'échantillon ?
  - Plan de sondage
- Echantillon $\longrightarrow$ Données
  - Comment obtenir des données de bonne qualité ?
  - Gestion d'une enquête, vérification des données
- Données $\longrightarrow$ Caractéristiques
  - Comment estimer les caractéristiques de la population sur la base d'un échantillon ?
  - Estimateur

Stratégie (1)

- Un *plan de sondage* est une probabilité sur un ensemble d'échantillons.
- *Echantillon aléatoire* : chaque échantillon S a une probabilité connue $P(S)$ d'être sélectionné.
- Un *estimateur* est une fonction des y_i pour $i \in S$.
- Une caractéristique $\theta\left(y_i, i \in U\right)$ est estimée par $\widehat{\theta}\left(y_i, i \in S\right)$.
- Si S est un échantillon aléatoire, alors $\widehat{\theta}$ est une variable aléatoire dont on peut calculer l'espérance et la variance.

Stratégie (2)

Une *stratégie* est le choix d'un plan de sondage et d'un estimateur.

POPULATION — CARACTÉRISTIQUE

plan de sondage ↘ stratégie ↗ estimateur

ÉCHANTILLON

Bonne stratégie : le biais et la variance de l'estimateur $\widehat{\theta}$ sont petits (biais $= \mathrm{E}(\widehat{\theta}) - \theta$).

Stratégie de Horvitz-Thompson (1)

- *Estimateur de Horvitz-Thompson* pour un total $Y = \sum_{i \in U} y_i$

$$\widehat{Y} = \sum_{i \in S} \frac{y_i}{\pi_i} = \sum_{i \in S} w_i y_i.$$

- *Probabilités d'inclusion* $\pi_i = P(\{S \subseteq U \mid S \ni i\})$.
- *Poids de sondage* $w_i = 1/\pi_i$.
- L'estimateur de Horvitz-Thompson est sans biais : $\mathrm{E}(\widehat{Y}) = Y$.

Stratégie de Horvitz-Thompson (2)

- La variance de $\widehat{Y}$ est donnée par

$$\text{var}(\widehat{Y}) = \sum_{i,j \in U} (\pi_{ij} - \pi_i \pi_j) \left(\frac{y_i}{\pi_i}\right) \left(\frac{y_j}{\pi_j}\right),$$

où les $\pi_{ij} = P(\{S \subset U \mid S \ni i,j\})$ sont les probabilités d'inclusion d'ordre deux.

- On estime $\text{var}(\widehat{Y})$ par

$$\widehat{\text{var}}(\widehat{Y}) = \sum_{i,j \in S} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \left(\frac{y_i}{\pi_i}\right) \left(\frac{y_j}{\pi_j}\right).$$

Estimateur de Horvitz-Thompson / compléments (1)

- Population $U = \{1, \ldots, i, \ldots, N\}$
- Ensemble d'échantilllons $\Omega \subseteq \mathcal{P}(U)$
- Plan de sondage : probabilité P sur Ω
- On veut estimer un total $Y = \sum_{i \in U} y_i$
- Estimateur

$$\widehat{Y}(S) = \sum_{i \in S} w_i(S) y_i$$

- Stratégie de Horvitz-Thompson : déterminer des poids $w_i(S)$ tels que l'estimateur $\widehat{Y}$ soit sans biais :

$$E(\widehat{Y}) = \sum_{S \in \Omega} \widehat{Y}(S) P(S) = Y$$

Estimateur de Horvitz-Thompson / compléments (2)

- Pour chaque unité $i \in U$ on définit une variable aléatoire

$$\begin{aligned} I_i : \Omega &\rightarrow \{0,1\} \\ S &\mapsto I_i(S) = \begin{cases} 1 & i \in S \\ 0 & i \notin S \end{cases} \end{aligned}$$

- On a

$$E(I_i) = 1 \times P(I_i = 1) + 0 \times P(I_i = 0) = P\{S \in \Omega; S \ni i\} = \pi_i$$

On note que

$$P\{S \in \Omega; S \ni i\} = \sum\nolimits_{\{S \in \Omega; S \ni i\}} P(S)$$

Estimateur de Horvitz-Thompson / compléments (3)

- Critère d'invariance : les poids $w_i(S)$ ne dépendent de l'échantillon que dans la mesure où $i \in S$:

$$w_i(S) = w_i \, I_i(S) = \begin{cases} w_i & i \in S \\ 0 & i \notin S \end{cases}$$

- On peut alors écrire

$$\widehat{Y}(S) = \sum_{i \in S} w_i(S) y_i = \sum_{i \in U} w_i y_i I_i(S)$$

- Il suit que

$$E(\widehat{Y}) = \sum_{i \in U} w_i y_i E(I_i) = \sum_{i \in U} (w_i \pi_i) y_i$$

Si $w_i = 1/\pi_i$, alors l'estimateur $\widehat{Y}$ est sans biais

Estimateur de Horvitz-Thompson / compléments (4)

- Variance de $\widehat{Y}$

$$\text{var}(\widehat{Y}) = \text{var}\left(\sum_{i\in U} w_i y_i I_i\right) = \sum_{i,j\in U} \text{cov}(I_i, I_j)(\underbrace{w_i y_i}_{\check{y}_i})(\underbrace{w_j y_j}_{\check{y}_j})$$

- Pour $i \neq j$,

$$\text{cov}(I_i, I_j) = \text{E}(I_i, I_j) - \text{E}(I_i)\,\text{E}(I_j) = \pi_{ij} - \pi_i\pi_j$$

où $\pi_{ij} = E(I_i, I_j) = P\{S \in \Omega; S \ni i, j\}$

Estimateur de Horvitz-Thompson / compléments (5)

- On pose

$$\Delta_{ij} = \mathrm{cov}(I_i, I_j) = \begin{cases} \pi_i(1-\pi_i) & i = j \\ \pi_{ij} - \pi_i\pi_j & i \neq j \end{cases}$$

alors

$$\mathrm{var}(\widehat{Y}) = \sum_{i,j \in U} \Delta_{ij} \check{y}_i \check{y}_j = \check{y}^t \Delta \check{y}$$

Echantillonnage aléatoire simple (1)

- La sélection d'un échantillon S de taille n dans une population U de taille N est faite par *échantillonnage aléatoire simple* si chaque sous-ensemble $S \subseteq U$ de taille $n \leq N$ a la même probabilité d'être sélectionné.
- TAS = tirage aléatoire simple

Echantillonnage aléatoire simple (2)

- Ensemble des échantillons

$$\Omega = \{S \subseteq U; |S| = n\}, \quad |\Omega| = \frac{N!}{n!(N-n)!} = \binom{N}{n},$$

où $N! = N(N-1)(N-2)\cdots 3 \times 2 \times 1$.

- Plan de sondage

$$P(S) = \frac{1}{|\Omega|} = \binom{N}{n}^{-1}.$$

Algorithme

TAS de taille n dans une population de taille N.

1. On génère, pour chaque unité $i \in U$ de la population, des nombres aléatoires u_i indépendants et de loi uniforme sur l'intervalle $(0, 1)$.
2. On ordonne les unités de la population par nombres aléatoires u_i croissants.
3. On sélectionne dans l'échantillon $S \subset U$ les n premières unités.

Estimateur de Horvitz-Thompson (1)

- Les probabilités d'inclusion pour un sondage aléatoire simple sont données par

$$\pi_i = \frac{n}{N} = f,$$

où f est le *taux de sondage* (sampling fraction).

- On appelle $1 - f$ le facteur de correction pour une population finie.

Estimateur de Horvitz-Thompson (2)

- L'estimateur de HT pour un *total* $Y = \sum_{i \in U} y_i$ est alors donné par

$$\widehat{Y} = \sum_{i \in S} \frac{y_i}{\pi_i} = \sum_{i \in S} w_i y_i = \frac{N}{n} \sum_{i \in S} y_i, \text{ où } w_i = \frac{1}{\pi_i} = \frac{N}{n}.$$

- L'estimateur de HT pour une *moyenne* $\overline{Y} = \frac{1}{N} Y$ est la moyenne de y sur l'échantillon

$$\widehat{\overline{Y}} = \frac{1}{N} \widehat{Y} = \frac{1}{N} \left(\frac{N}{n} \right) \sum_{i \in S} y_i = \frac{1}{n} \sum_{i \in S} y_i =: \overline{y}_S.$$

Estimateur de Horvitz-Thompson (3)

L'estimateur de HT pour la *proportion* d'unités de modalité a, $a = 1, \ldots, A$,

$$p_a = \frac{N_a}{N} = \frac{1}{N} \sum_{i \in U} y_{ai}, \text{ où } y_{ai} = \begin{cases} 1 & \text{si } i \text{ est de modalité } a \\ 0 & \text{autrement} \end{cases}$$

est donné par la proportion dans l'échantillon

$$\widehat{p}_a = \frac{1}{N} \left(\frac{N}{n} \right) \sum_{i \in S} y_{ai} = \frac{1}{n} \sum_{i \in S} y_{ai} = \frac{n_a}{n}.$$

Estimateur de HT pour un TAS / Compléments

- Pour $i \in U$ donné, $\pi_i = E(I_i) = P\{|S| = n, S \ni i\}$. Donc

$$\pi_i = \sum_{\{|S|=n, S \ni i\}} P(S) = \frac{|\{|S| = n, S \ni i\}|}{\binom{N}{n}}$$

On a

$$|\{|S| = n, S \ni i\}| = \binom{N-1}{n-1} = \frac{n}{N}\binom{N}{n}$$

et donc finalement $\pi_i = n/N$

- Pour les probabilités d'inclusion d'ordre 2 on a

$$\pi_{ij} = \frac{|\{|S| = n, S \ni i, j\}|}{\binom{N}{n}} = \frac{\binom{N-2}{n-2}}{\binom{N}{n}} = \frac{n(n-1)}{N(N-1)}$$

Estimation sur un domaine (1)

- Etant donné une décomposition de la population U en domaines, $U = \bigcup_{k \in K} U_k$, on définit les variables indicatrices des domaines

$$z_{ki} = \begin{cases} 1 & \text{si } i \in U_k \\ 0 & \text{si } i \notin U_k \end{cases}, \; k = 1, \ldots, K.$$

- Les estimateurs de HT des différents types de paramètres utilisent systématiquement les variables indicatrices z_k, $k = 1, \ldots, K$.

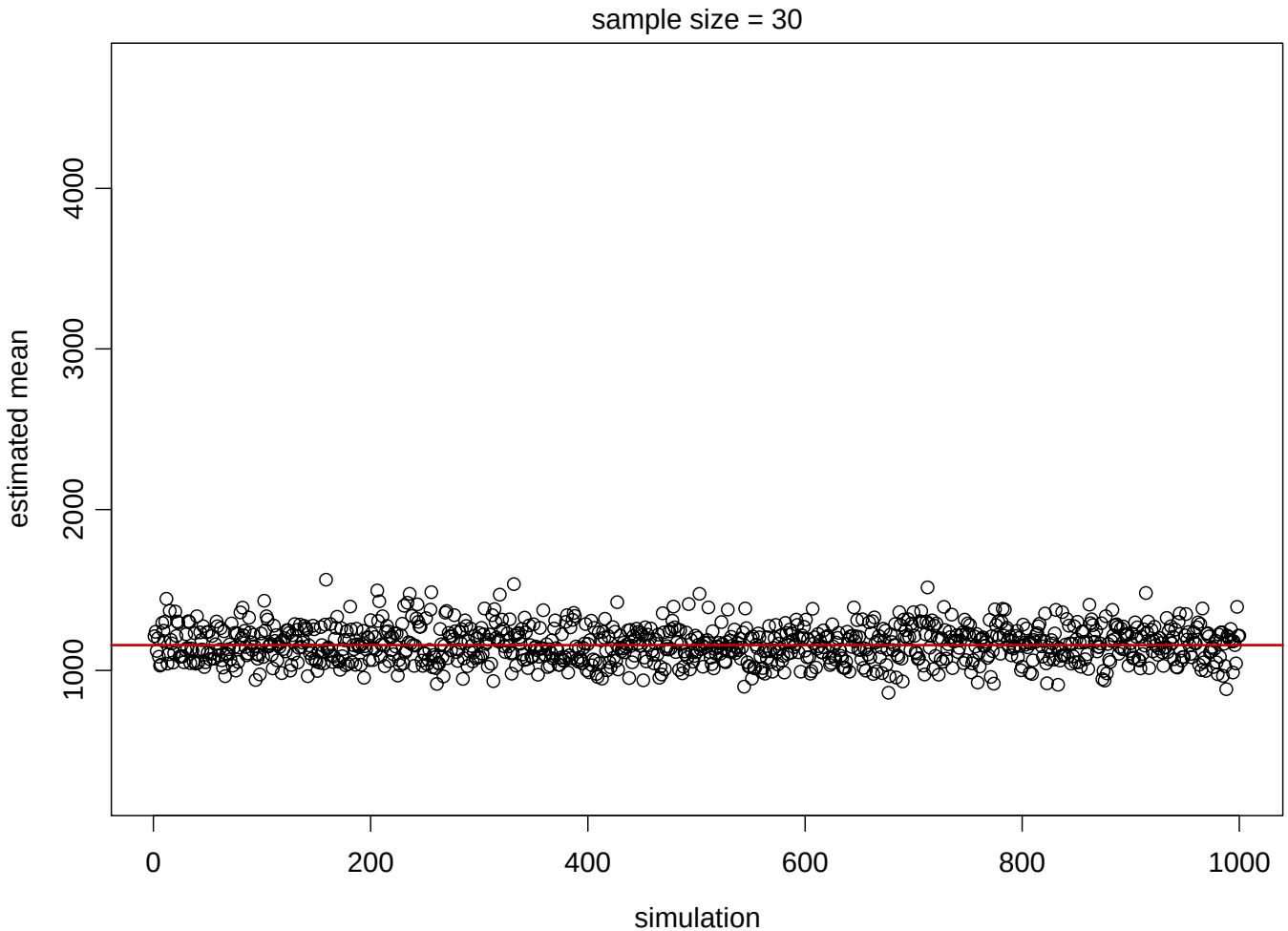
Estimation sur un domaine (2)

Paramètre	Estimateur
Taille	$\widehat{N}_k = \frac{N}{n} \sum_{i \in S} z_{ki}$
Total	$\widehat{Y}_k = \frac{N}{n} \sum_{i \in S} z_{ki} y_i$
Moyenne	$\widehat{\overline{Y}}_k = \frac{(N/n) \sum_{i \in S} z_{ki} y_i}{(N/n) \sum_{i \in S} z_{ki}} = \frac{\widehat{Y}_k}{\widehat{N}_k}$
Proportion	$\widehat{p}_{ak} = \frac{(N/n) \sum_{i \in S} z_{ki} y_{ai}}{(N/n) \sum_{i \in S} z_{ki}} = \frac{\widehat{N}_{ak}}{\widehat{N}_k}$

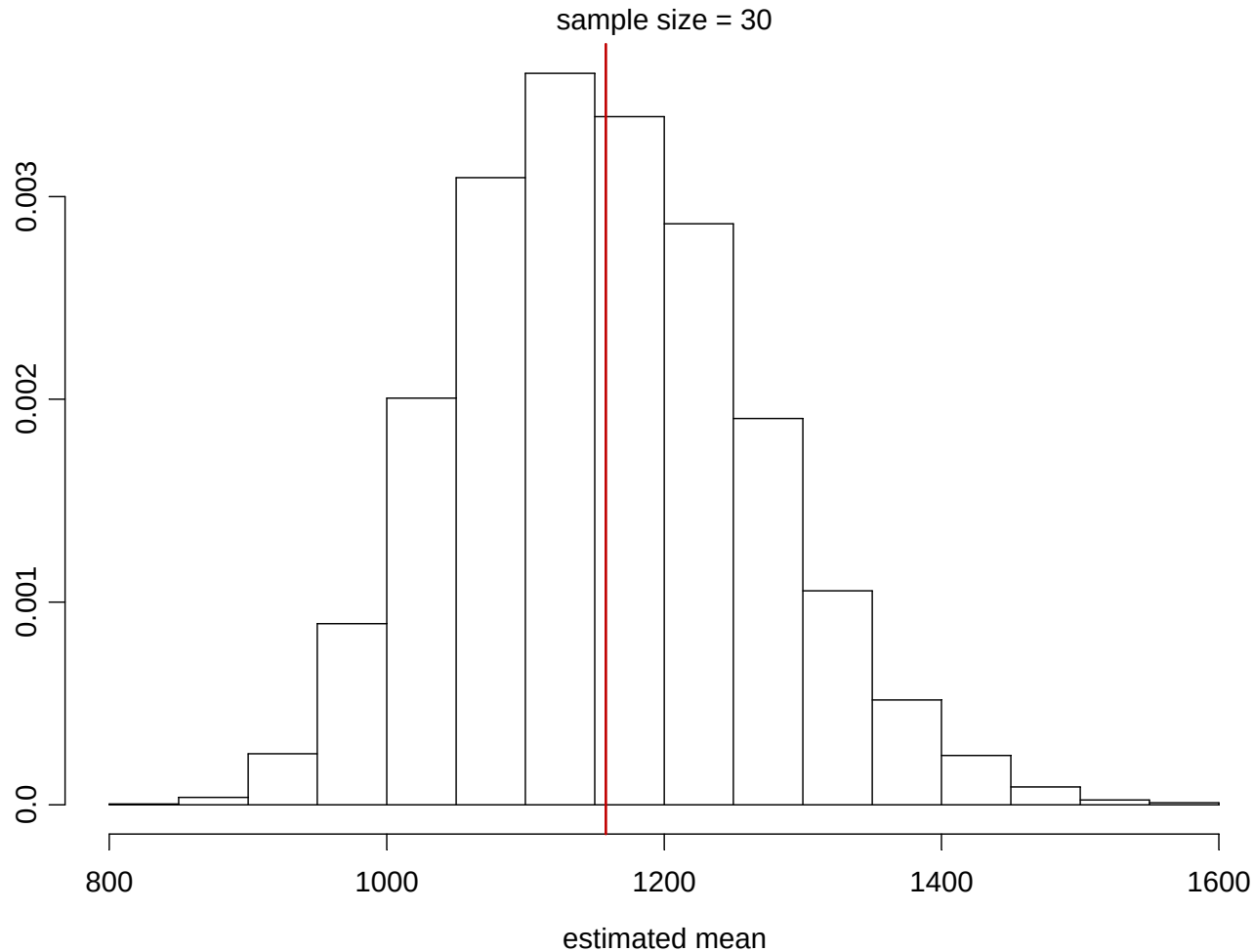
Variance (1)

- Si le tirage de l'échantillon se fait selon un plan de sondage bien déterminé, un estimateur est une variable aléatoire dont on peut en principe calculer la densité de probabilité, l'espérance, la variance, etc.
- On considère ici un TAS de taille $n = 30$ dans la population des logements ($N = 151$). On estime le loyer moyen $\overline{Y} = 1158$ par la moyenne des loyers dans l'échantillon $\overline{y}_S$.
- L'estimateur $\overline{y}_S$ est une variable aléatoire.

1000 simulations de $\overline{y}_S$



Histogramme de 10000 simulations de $\overline{y}_S$



Variance (2)

- La variance de l'estimateur de HT dépend des probabilités d'inclusion d'ordre deux.
- Pour un TAS on a

$$\pi_{ij} = \frac{n(n-1)}{N(N-1)}.$$

Variance (3)

On montre que la variance de l'estimateur de HT pour une moyenne est donné par

$$\text{var}(\overline{y}_S) = \left(1 - \frac{n}{N}\right) \frac{1}{n} D^2,$$

où

$$D^2 = \frac{1}{N-1} \sum_{i \in U} (y_i - \overline{Y})^2$$

est la variance de la variable y dans la population U.

Variance (4)

- Le coefficient de variation de $\overline{y}_S$ est donné par

$$\mathrm{CV}(\overline{y}_S) = \frac{\sqrt{\mathrm{var}(\overline{y}_S)}}{\mathrm{E}(\overline{y}_S)} = \frac{\sqrt{\mathrm{var}(\overline{y}_S)}}{\overline{Y}}.$$

- Variance de l'estimateur de HT pour un total

$$\mathrm{var}(\widehat{Y}) = N^2 \left(1 - \frac{n}{N}\right) \frac{1}{n} D^2 = N^2 (1 - f) \frac{1}{n} D^2.$$

- Variance de l'estimateur de HT d'une proportion

$$\mathrm{var}(\widehat{p}_a) = \frac{N}{N-1} (1 - f) \frac{1}{n} p_a (1 - p_a).$$

Variance / Compléments (1)

- Pour $i \neq j$,

$$\Delta_{ij} = \pi_{ij} - \pi_i \pi_j = \frac{n(n-1)}{N(N-1)} - \left(\frac{n}{N}\right)^2 = -\frac{1}{N-1} f(1-f)$$

où $f = n/N$ est la fraction de sondage. Par ailleurs

$$\Delta_{ii} = \pi_i(1-\pi_i) = \frac{n}{N}\left(1 - \frac{n}{N}\right) = f(1-f)$$

- Donc, pour tous $i, j \in U$,

$$\Delta_{ij} = \frac{N\delta_{ij} - 1}{N-1} f(1-f) = \frac{N}{N-1} f(1-f) \underbrace{\left(\delta_{ij} - \frac{1}{N}\right)}_{=H_{ij}}$$

où $H = Id_N - 1_N(1_N^t 1_N)^{-1} 1_N^t$.

Variance / Compléments (2)

- Finalement, la matrice de variance-covariance des fonctions indicatrices d'appartenance à un échantillon est donnée par

$$\Delta = \frac{N}{N-1} f(1-f)H$$

où $H = Id_N - 1_N(1_N^t 1_N)^{-1} 1_N^t$.

- La matrice H centre les observations

$$Hy = y - \frac{1}{N} 1_N (1_N^t y) = y - \overline{Y}\, 1_N = (y_i - \overline{Y}, i \in U)$$

Variance / Compléments (3)

- La variance de

$$\widehat{Y} = \sum_{i \in S} w_i y_i = \frac{N}{n} \sum_{i \in S} y_i$$

est donnée par

$$\begin{aligned} \text{var}(\widehat{Y}) &= \check{y}^t \Delta \check{y} = \left(\frac{N}{n}\right)^2 \frac{N}{N-1} f(1-f) \left(y^t H y\right) \\ &= N^2 (1-f) \frac{1}{n} D_y^2 \end{aligned}$$

Estimation d'une proportion / Compléments (1)

- Population U décomposée en sous-populations $U = \bigcup_{a \in A} U_a$. Un échantillon $S \subseteq U$ s'en trouve lui aussi décomposé : $S = \bigcup_{a \in A} S_a$, où $S_a = S \cap U_a$
- Distribution de $(n_a, a \in A)$

$$P(|S \cap U_a| = n_a, a \in A) = \frac{\prod_{a \in A} \binom{N_a}{n_a}}{\binom{N}{n}}$$

Loi hypergéométrique multiple

Estimation d'une proportion / Compléments (2)

- Si on a 2 modalités seulement, on note $U = D \cup (U \setminus D)$ et $|S \cap D| = d$. Alors

$$P(d) = \frac{\binom{D}{d}\binom{N-D}{n-d}}{\binom{N}{n}}$$

- Preuve bijective de

$$\binom{N}{n} = \sum_{d=0}^{n} \binom{D}{d}\binom{N-D}{n-d}$$

Estimation d'une proportion / Compléments (3)

- Puisque d suit une loi hypergéométrique on a

$$\begin{aligned} E(d) &= n\frac{D}{N} \\ \text{var}(d) &= \frac{n(N-n)}{N-1}\frac{D}{N}\left(1-\frac{D}{N}\right) \\ &= \frac{N}{N-1}(1-f)\,n\,\frac{D}{N}\left(1-\frac{D}{N}\right) \end{aligned}$$

- Alors

$$\begin{aligned} E(\widehat{p}) &= E\left(\frac{d}{n}\right) = \frac{D}{N} = p \\ \text{var}(\widehat{p}) &= \text{var}\left(\frac{d}{n}\right) = \frac{N}{N-1}(1-f)\,\frac{1}{n}\,p(1-p) \end{aligned}$$

Estimation d'une moyenne $\overline{Y}$ par $\overline{y}_S$.

$$\text{var}(\overline{y}_S) = \left(1 - \frac{n}{N}\right)\frac{1}{n}D^2 = (1-f)\frac{1}{n}D^2.$$

1. Précision de l'estimation dans le cas d'un recensement ?
2. Un échantillon de taille $n = 1'000$ tiré dans une population de taille $N = 50'000$ est-il plus précis qu'un échantillon de taille $n = 1'000$ tiré dans une population de taille $N = 5'000'000$?
3. Un échantillon tiré dans une population homogène est-il plus précis qu'un échantillon tiré dans une population hétérogène ?
4. Comment peut-on influencer la précision d'un échantillon aléatoire simple ?

Estimation de la variance (1)

Estimation de la variance de l'estimateur de HT d'une moyenne par

$$\widehat{\text{var}}(\overline{y}_S) = \left(1 - \frac{n}{N}\right) \frac{1}{n} d^2,$$

où

$$d^2 = \frac{1}{n-1} \sum_{i \in S} (y_i - \overline{y}_S)^2$$

est la variance de la variable y dans l'échantillon S.

Estimation de la variance (2)

- Estimation du coefficient de variation de $\overline{y}_S$ par

$$\widehat{\mathrm{CV}}(\overline{y}_S) = \frac{\sqrt{\widehat{\mathrm{var}}(\overline{y}_S)}}{\overline{y}_S}.$$

- Estimation de la variance de l'estimateur de HT d'un total

$$\widehat{\mathrm{var}}(\widehat{Y}) = N^2(1 - \frac{n}{N})\frac{1}{n}d^2.$$

- Estimation de la variance de l'estimateur de HT d'une proportion

$$\widehat{\mathrm{var}}(\widehat{p}_a) = \frac{n}{n-1}(1-f)\frac{1}{n}\widehat{p}_a(1-\widehat{p}_a).$$

Intervalle de confiance (1)

- Echantillon aléatoire simple S de taille n tiré dans une population U de taille N.
- Estimation d'une moyenne $\overline{Y} = Y/N$ par $\overline{y}_S$.
- L'intervalle de confiance pour $\overline{Y}$ au niveau $1-\alpha$ est donné par

$$\overline{y}_S \pm z_{1-\alpha/2}\sqrt{\text{var}(\overline{y}_S)} = \overline{y}_S \pm z_{1-\alpha/2}\sqrt{(1-f)\frac{1}{n}D^2},$$

où $z_{1-\alpha/2}$ est le quantile $1-\alpha/2$ pour la loi normale $\mathcal{N}(0,1)$.

Intervalle de confiance (2)

- Dans la pratique, on estime $\text{var}(\overline{y}_S)$ par $\widehat{\text{var}}(\overline{y}_S) = (1-f)\frac{1}{n}d^2$, et l'intervalle de confiance pour $\overline{Y}$ est alors donné par

$$\overline{y}_S \pm z_{1-\alpha/2}\sqrt{\widehat{\text{var}}(\overline{y}_S)} = \overline{y}_S \pm z_{1-\alpha/2}\sqrt{(1-f)\frac{1}{n}d^2},$$

- Pour tenir compte de l'estimation de D par d, `proc surveymeans` calcule l'intervalle de confiance comme

$$\overline{y}_S \pm t_{n-1,1-\alpha/2}\sqrt{\widehat{\text{var}}(\overline{y}_S)},$$

où $t_{n-1,1-\alpha/2}$ est le quantile $1-\alpha/2$ de la distribution t avec $n-1$ degrés de liberté.

Intervalle de confiance (3)

Quelques valeurs de $z_{1-\alpha/2}$ et $t_{n-1,1-\alpha/2}$ pour $n = 30$

$1-\alpha$	$\alpha/2$	$z_{1-\alpha/2}$	$t_{29,1-\alpha/2}$
0.90	0.050	1.645	1.697
0.95	0.025	1.960	2.042
0.99	0.005	2.576	2.750

Remarque : Pour $n \to \infty$, $t_{n,\alpha} \to z_\alpha$.

Taille de l'échantillon : estimation d'une moyenne

- Estimation d'une moyenne avec un coefficient de variation désiré CV_0

$$\begin{aligned} \mathrm{CV}^2(\overline{y}_S) &= \frac{\mathrm{var}(\overline{y}_S)}{\mathrm{E}(\overline{y}_S)^2} = \frac{\mathrm{var}(\overline{y}_S)}{\overline{Y}^2} = (1-f)\frac{1}{n}\frac{D^2}{\overline{Y}^2} \\ &= (1-f)\frac{1}{n}\,\mathrm{CV}_y^2 =: \mathrm{CV}_0^2 \\ \Rightarrow \quad n_0 &:= \frac{n}{1-f} = \left(\frac{\mathrm{CV}_y}{\mathrm{CV}_0}\right)^2. \end{aligned}$$

- Comme $f = n/N$

$$n_0 = \frac{n}{1-f} = \frac{n}{1-n/N} \Rightarrow n = \frac{n_0}{1+n_0/N}.$$

Taille de l'échantillon : estimation d'une proportion

- Estimation d'une proportion avec une variance désirée V_0

$$\begin{aligned} \text{var}(\widehat{p}_a) &= \frac{N}{N-1}(1-f)\frac{1}{n}p_a(1-p_a) \\ &\approx (1-f)\frac{1}{n}p_a(1-p_a) =: V_0 \\ \Rightarrow \quad n_0 &:= \frac{n}{1-f} = \frac{p_a(1-p_a)}{V_0}. \end{aligned}$$

- Puisque $f = n/N$

$$n_0 = \frac{n}{1-f} = \frac{n}{1-n/N} \Rightarrow n = \frac{n_0}{1+n_0/N}.$$

Taille de l'échantillon : estimation d'une proportion

Comme $p_a(1-p_a)$ est maximal pour $p_a = 0.5$, au pire $n_0 = (1/4)(1/V_0)$

$\pm 2\sqrt{V_0}$	V_0	$1/V_0$	$n_0 = (1/4)(1/V_0)$
$\pm$ 10 %	$(0.05)^2$	400	100
$\pm$ 5 %	$(0.025)^2$	1'600	400
$\pm$ 2 %	$(0.01)^2$	10'000	2'500
$\pm$ 0.5 %	$(0.0025)^2$	160'000	40'000

Utilisation d'informations auxiliaires (1)

Variables d'intérêt : $y_1, y_2, \ldots, y_p$. Variables auxiliaires : $x_1, x_2, \ldots, x_q$.

i	x_1	$\ldots$	x_q	y_1	$\ldots$	y_p	brut
1	x_{11}	$\ldots$	x_{1q}	y_{11}	$\ldots$	y_{1p}	1
2	x_{21}	$\ldots$	x_{2q}				0
3	x_{31}	$\ldots$	x_{3q}				0
4	x_{41}	$\ldots$	x_{4q}	y_{41}	$\ldots$	y_{4p}	1
5	x_{51}	$\ldots$	x_{5q}	y_{51}	$\ldots$	y_{5p}	1
6	x_{61}	$\ldots$	x_{6q}	y_{61}	$\ldots$	y_{6p}	1
7	x_{71}	$\ldots$	x_{7q}				0
8	x_{81}	$\ldots$	x_{8q}	y_{81}	$\ldots$	y_{8p}	1
9	x_{91}	$\ldots$	x_{9q}	y_{91}	$\ldots$	y_{9p}	1
	X_1	$\ldots$	X_q				

Utilisation d'informations auxiliaires (2)

- Plan de sondage
  - Plan stratifié
  - Plan par grappe
  - Plan à plusieurs degrés
- Estimation
  - Estimateur par le quotient
  - Estimateur par la regression
  - Post-stratification
  - Estimateur par calage

Utilisation d'informations auxiliaires (3)

Buts

- Gain de précision
- Calage sur des caractéristiques de population connues
- Correction pour la non-réponse

Echantillonage stratifié

- Décomposition de la population en H strates : $U = \bigcup_{h=1}^{H} U_h$
- On tire un échantillon aléatoire simple $S_h \subseteq U_h$ au sein de chaque strate
- TASST = tirage aléatoire simple stratifié

Pourquoi stratifier ? (1)

- Décomposition de la variance d'une variable y sur la population en *variance dans les strates* et *variance entre les strates*.
- Comme les tirages dans les strates sont indépendants les uns des autres, la précision de l'estimateur va dépendre des variances dans les strates.
- Si les strates sont homogènes, on peut s'attendre à un gain de précision.

Pourquoi stratifier ? (2)

$$D^2 \approx \sum_h W_h D_h^2 + \sum_h W_h (\overline{Y}_h - \overline{Y})^2$$

$W_h = N_h/N$, $N_h =$ nombre d'unités dans la strate U_h

$\overline{Y} =$ moyenne de la variable y dans la population U

$\overline{Y}_h =$ moyenne de y dans la strate U_h

$D^2 =$ variance de y dans la population U

$D_h^2 =$ variance de y dans la strate U_h

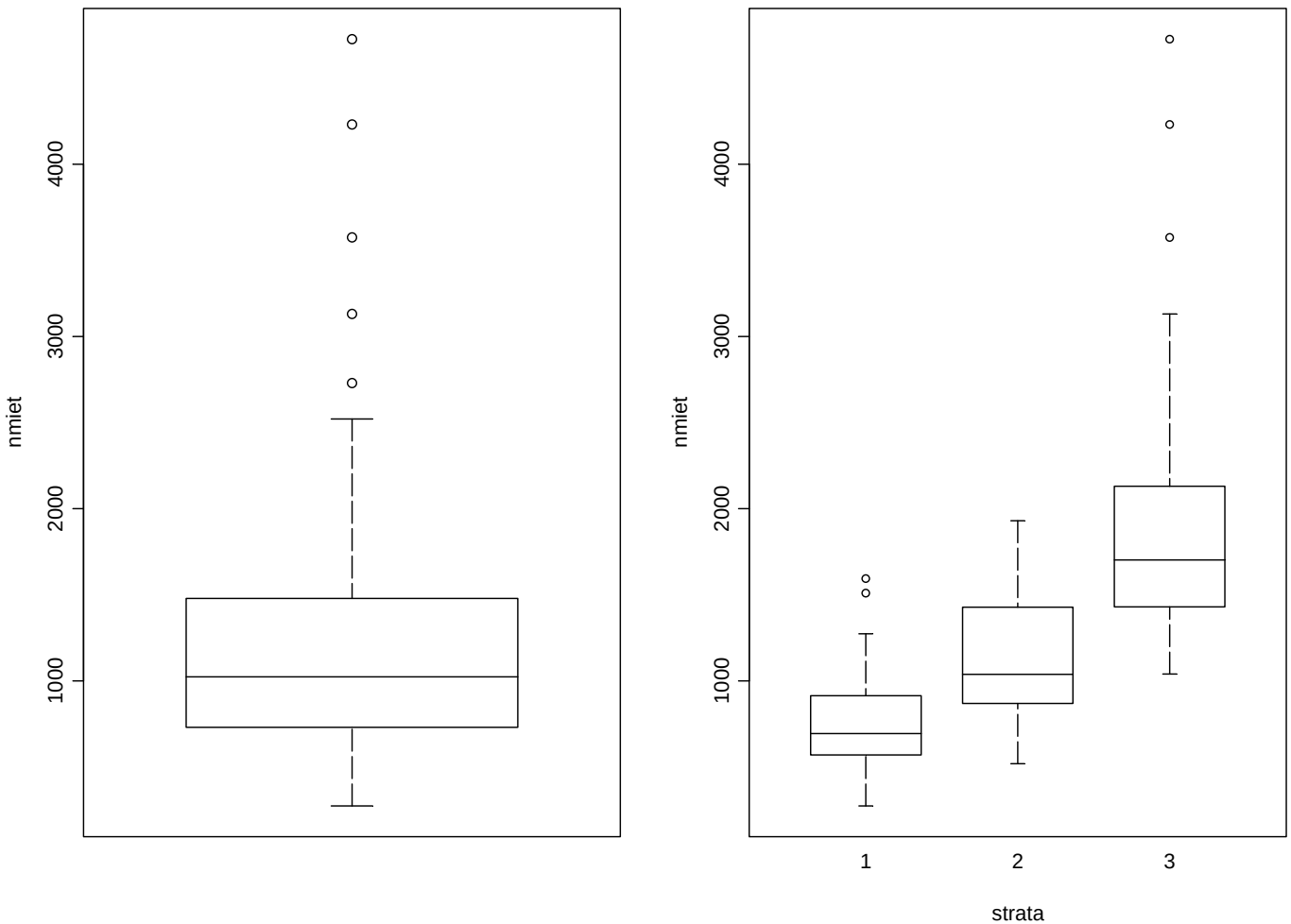
Pourquoi stratifier ? (3)

$$D^2 \approx \sum_h W_h D_h^2 + \sum_h W_h (\overline{Y}_h - \overline{Y})^2$$

Population des logements dans une commune

h	zi	N_h	$\overline{Y}_h$	D_h
1	1,2	57	755.2	280.2
2	3,4	64	1132.4	348.9
3	5,6	30	1978.1	895.0
		151	1158.0	657.1

Loyers : population et strates



Avantages et inconvénients de la stratification (1)

- Meilleure précision ou des coûts plus faibles
- Flexibilité plus grande
  - augmenter la taille de l'échantillon dans les strates intéressantes ou dans les petites strates
  - utiliser des plans de sondage différents dans les différentes strates
- Définition des strates
  - dépend de l'information disponible
  - souvent déterminée par les besoins de l'enquête, p.ex. on définit des sous-populations intéressantes comme strates
  - dispersion des variables d'intérêt au sein des strates devrait être faible : on essaie de construire des strates homogènes

Avantages et inconvénients de la stratification (2)

- Nombre de strates
  - les strates ne doivent pas être trop petites (non-réponse, estimation sur des domaines, estimation de la variance)
  - le gain en précision devient moins important quand on augmente le nombre de strates
- Détermination de la taille de l'échantillon est plus complexe
- Allocation de l'échantillon aux strates
- Procédure d'estimation est plus complexe

Estimateur de Horvitz-Thompson (1)

- Stratification de la population en H strates $U = \bigcup_{h=1}^{H} U_h$ de tailles N_h.
- On tire dans chaque strate un échantillon aléatoire simple $S_h \subseteq U_h$ de taille n_h.
- On veut estimer un total $Y = \sum_{i \in U} y_i$.

Estimateur de Horvitz-Thompson (2)

- On note que

$$Y = \sum_{i \in U} y_i = \sum_h \sum_{i \in U_h} y_i = \sum_h Y_h.$$

- L'estimateur de HT pour un *total* $Y = \sum_{i \in U} y_i$ est alors donné par

$$\widehat{Y} = \sum_h \widehat{Y}_h = \sum_h \frac{N_h}{n_h} \sum_{i \in S_h} y_i = \sum_{i \in S} w_i y_i$$

où $w_i = N_h / n_h$ pour $i \in S_h$.

Estimateur de Horvitz-Thompson (3)

Comparaison TAS et TASST

$$\widehat{Y} = \sum_{i \in S} \frac{y_i}{\pi_i} = \sum_{i \in S} w_i y_i, \text{ avec } w_i = \frac{1}{\pi_i}.$$

$$w_i = \begin{cases} \frac{N}{n} = f^{-1} & \text{pour } i \in U : \text{TAS} \\ \frac{N_h}{n_h} = f_h^{-1} & \text{pour } i \in U_h : \text{TASST} \end{cases}$$

Estimateur de Horvitz-Thompson (4)

- L'estimateur de HT pour une *moyenne* $\overline{Y} = \frac{1}{N} Y$

$$\widehat{\overline{Y}} = \frac{1}{N} \widehat{Y} = \sum_h \left(\frac{N_h}{N} \right) \frac{1}{n_h} \sum_{i \in S_h} y_i = \sum_h W_h \overline{y}_{S_h}.$$

- L'estimateur de HT pour une *proportion* $p_a = N_a / N$

$$\widehat{p}_a = \sum_h \left(\frac{N_h}{N} \right) \frac{1}{n_h} \sum_{i \in S_h} y_{ai} = \sum_h W_h \widehat{p}_{ah},$$

où

$$\widehat{p}_{ah} = \frac{1}{n_a} \sum_{i \in S_h} y_{ai} = \frac{n_{ah}}{n_h}.$$

Estimation sur un domaine (1)

- Etant donné une décomposition de la population U en domaines, $U = \bigcup_{k \in K} U_k$, on définit les variables indicatrices des domaines

$$z_{ki} = \begin{cases} 1 & \text{si } i \in U_k \\ 0 & \text{si } i \notin U_k \end{cases}, \; k = 1, \ldots, K.$$

- Les estimateurs de HT des différents types de paramètres utilisent systématiquement les variables indicatrices z_k, $k = 1, \ldots, K$.

Estimation sur un domaine (2)

Paramètre	Estimateur
Taille	$\widehat{N}_k = \sum_h \frac{N_h}{n_h} \sum_{i \in S_h} z_{ki}$
Total	$\widehat{Y}_k = \sum_h \frac{N_h}{n_h} \sum_{i \in S_h} z_{ki} y_i$
Moyenne	$\widehat{\overline{Y}}_k = \widehat{Y}_k / \widehat{N}_k$
Proportion	$\widehat{p}_{ak} = \widehat{N}_{ak} / \widehat{N}_k$

Variance (1)

- Dans un plan stratifié, l'estimateur pour un total $Y = \sum_{i \in U} y_i$ est donné par $\widehat{Y} = \sum_h \widehat{Y}_h$.
- Comme les tirages dans les strates sont des TAS indépendants les uns des autres, nous avons

$$\text{var}(\widehat{Y}) = \sum_h \text{var}(\widehat{Y}_h) = \sum_h N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{1}{n_h} D_h^2,$$

Variance (2)

- Variance de l'estimateur de HT pour une moyenne

$$\text{var}(\widehat{\overline{Y}}) = \sum_h W_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{1}{n_h} D_h^2.$$

- Variance de l'estimateur de HT pour une proportion

$$\text{var}(\widehat{p}_a) = \sum_h W_h^2 \left(1 - \frac{n_h}{N_h}\right) \left(\frac{N_h}{N_h - 1}\right) \frac{1}{n_h} p_{ah}(1 - p_{ah}),$$

où

$$p_{ah} = \frac{1}{N_h} \sum_{i \in U_h} y_{ai} = \frac{N_{ah}}{N_h}$$

est la proportion de la modalité a dans la strate U_h.

Estimation de la variance (1)

- Estimation de la variance de l'estimateur de HT d'un total

$$\widehat{\text{var}}(\widehat{Y}) = \sum_h N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{1}{n_h} d_h^2,$$

où

$$d_h^2 = \frac{1}{n_h - 1} \sum_{i \in S_h} (y_i - \overline{y}_{S_h})^2$$

est la variance de la variable y dans l'échantillon $S_h \subseteq U_h$.

Estimation de la variance (2)

- Estimation de la variance de l'estimateur de HT d'une moyenne

$$\widehat{\text{var}}(\widehat{\overline{Y}}) = \sum_h W_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{1}{n_h} d_h^2.$$

- Estimation de la variance de l'estimateur de HT d'une proportion

$$\widehat{\text{var}}(\widehat{p}_a) = \sum_h W_h^2 \left(1 - \frac{n_h}{N_h}\right) \left(\frac{n_h}{n_h - 1}\right) \frac{1}{n_h} \widehat{p}_{ah}(1 - \widehat{p}_{ah}).$$

Intervalle de confiance (1)

- Stratification de la population $U = \bigcup_{h \in H} U_h$ en H strates.
- On tire dans chaque strate un échantillon aléatoire simple $S_h \subseteq U_h$.
- Le taux de sondage dans la strate h est $f_h = n_h / N_h$.
- Estimation d'une moyenne $\overline{Y} = \frac{1}{N} Y$ par

$$\widehat{\overline{Y}} = \sum_h \left(\frac{N_h}{N} \right) \frac{1}{n_h} \sum_{i \in S_h} y_i = \sum_h W_h \overline{y}_{S_h}.$$

Intervalle de confiance (2)

- Estimation de la précision de $\widehat{\overline{Y}}$ par

$$\widehat{\text{var}}(\widehat{\overline{Y}}) = \sum_h W_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{1}{n_h} d_h^2.$$

- Intervalle de confiance pour $\overline{Y}$ au niveau $1 - \alpha$ calculé par `proc surveymeans`

$$\widehat{\overline{Y}} \pm t_{n-H,1-\alpha/2} \sqrt{\widehat{\text{var}}(\widehat{\overline{Y}})}$$

où $t_{n-H,1-\alpha/2}$ est le quantile $1 - \alpha/2$ de la distribution t avec $n - H$ degrés de liberté.

Allocation de l'échantillon

On suppose que la taille brute n de l'échantillon est connue.

Allocation proportionnelle $\quad n_h = n(N_h/N) = nW_h = n\frac{N_h}{\sum_j N_j}$

Allocation optimale $\quad n_h = n(N_h D_h / \sum_j N_j D_j)$

Comparaison des précisions

$\text{var(TASST, alloc. opt.)} \leq \text{var(TASST, alloc. prop.)} \leq \text{var(TAS)}$

Est-ce que ça marche ?

- Estimation d'une moyenne $\overline{Y} = Y/N$ par $\widehat{\overline{Y}} = \sum_h W_h \overline{y}_{S_h}$.
- Précision de l'estimation

$$\text{var}(\widehat{\overline{Y}}) = \sum_h W_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{1}{n_h} D_h^2, \quad \text{CV}(\widehat{\overline{Y}}) = \frac{\text{std}(\widehat{\overline{Y}})}{\overline{Y}}$$

- On considère la population de logements ($N = 151$) dans laquelle on tire un échantillon de taille $n = 30$.
- Plans de sondage : TAS, TASST avec une allocation proportionnelle, TASST avec une allocation optimale.
- Stratification selon le nombre de pièces.

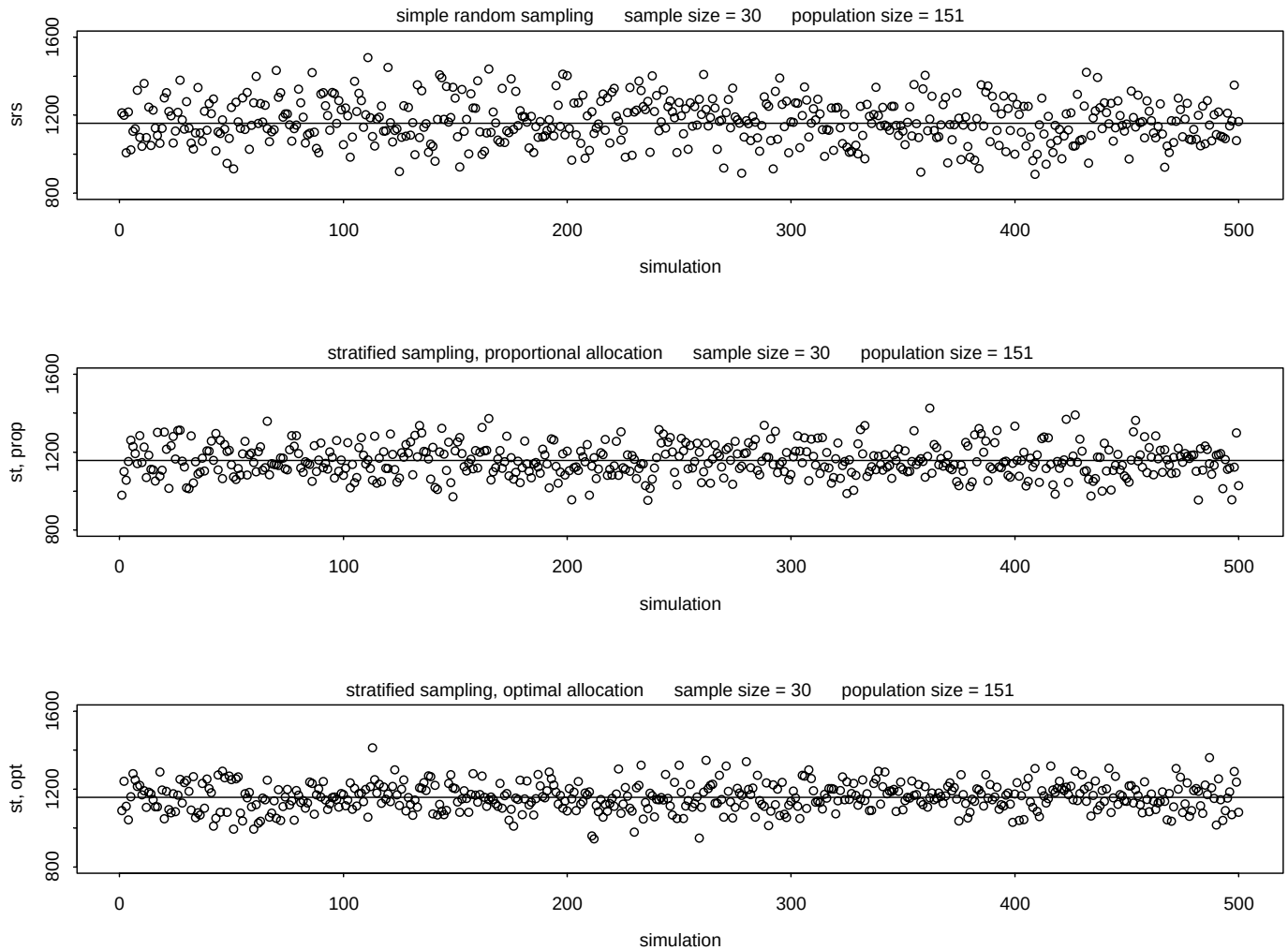
Allocations

h	zi	N_h	D_h	W_h	$n_{h,prop}$	$n_{h,opt}$
1	1,2	57	280.2	0.38	11	7
2	3,4	64	348.9	0.42	13	10
3	5,6	30	895.0	0.20	6	13
		151	657.1		30	30

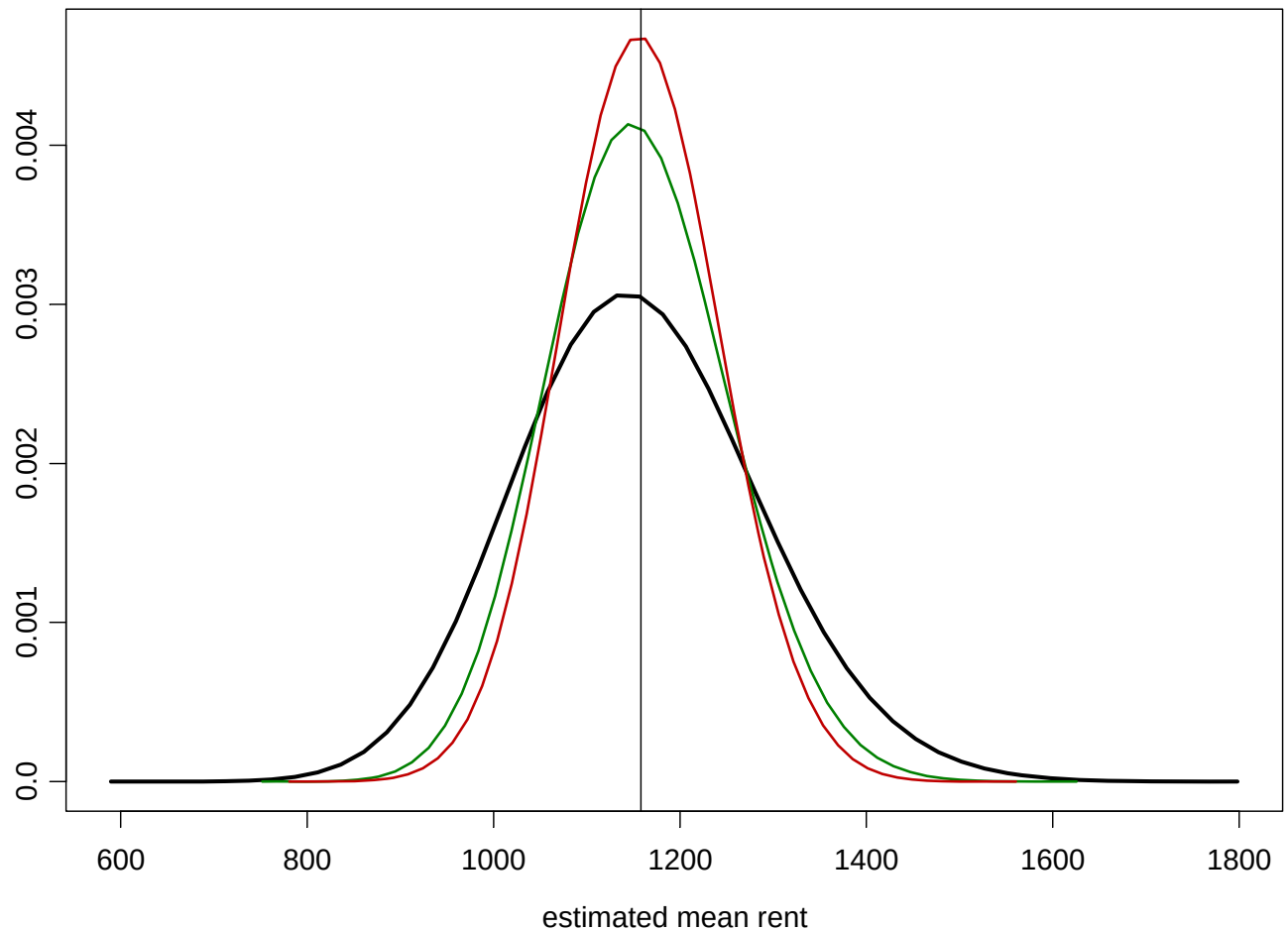
Précision des plans de sondage

Plan de sondage	$CV(\widehat{\overline{Y}})$
TAS	9.3%
TASST, allocation proportionnelle	6.9%
TASST, allocation optimale	5.9%

TAS, TASST + allocation prop., TASST + allocation opt.



TAS, TASST + allocation prop., TASST + allocation opt.



Taille de l'échantillon pour un TASST (1)

- Estimation d'un total $Y = \sum_{i \in U} y_i$ par

$$\widehat{Y} = \sum_h \frac{N_h}{n_h} \sum_{i \in S_h} y_i.$$

- La variance de $\widehat{Y}$ est donnée par

$$\text{var}(\widehat{Y}) = \sum_h N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{1}{n_h} D_h^2.$$

Taille de l'échantillon pour un TASST (2)

- La variance de $\widehat{Y}$ sous une allocation proportionnelle est donnée par

$$\begin{aligned} \text{var}(\widehat{Y}_{prop}) &= N^2(1-f)\frac{1}{n}\sum_h W_h D_h^2 \\ &= \frac{1}{n}\left(N^2\sum_h W_h D_h^2\right) - \sum_h N_h D_h^2. \end{aligned}$$

Taille de l'échantillon pour un TASST (3)

- La variance de $\widehat{Y}$ sous une allocation optimale est donnée par

$$\text{var}(\widehat{Y}_{opt}) = \frac{1}{n}\left(N\sum_h W_h D_h\right)^2 - \sum_h N_h D_h^2.$$

Taille de l'échantillon pour un TASST (4)

- On note que var($\widehat{Y}_{prop}$) et var($\widehat{Y}_{opt}$) sont de la forme $A/n - B$. Si on pose $A/n - B = V_0$, où V_0 est la variance de que l'on désire obtenir, alors $n = A/(V_0 + B)$.
- Nous avons donc

$$n_{prop} = \frac{N^2 \sum_h W_h D_h^2}{V_0 + \sum_h N_h D_h^2}$$

et

$$n_{opt} = \frac{N^2 (\sum_h W_h D_h)^2}{V_0 + \sum_h N_h D_h^2}.$$

Estimateur par le quotient

- Echantillon $S \subseteq U$, poids de sondage d_i, $i \in S$. On dispose de y_i et x_i pour les unités de l'échantillon $i \in S$.
- On veut estimer $Y = \sum_{i \in U} y_i$ et on connaît le total $X = \sum_{i \in U} x_i$.
- On soupçonne que la variable y est proportionnelle à la variable x

$$y_i \approx R x_i \text{ où } R = \frac{Y}{X} = \frac{\overline{Y}}{\overline{X}}$$

- Comment mettre à profit cette information auxiliaire afin de construire un estimateur de Y qui tienne compte de la connaissance de X ?

Estimateur par le *quotient* (ou par le *ratio*) pour un total

$$\widehat{Y}_R = \left(\frac{X}{\widehat{X}}\right)\widehat{Y} = X\,\frac{\sum_{i\in S} d_i y_i}{\sum_{i\in S} d_i x_i}$$

- Si on observe que $\widehat{X} > X$ et si on a bien $y_i \approx R x_i$, alors il est probable que $\widehat{Y} > Y$: l'estimation de Y est trop haute. L'estimateur par le ratio corrige l'estimation initiale $\widehat{Y}$ vers le bas, grâce au facteur $X/\widehat{X} < 1$.
- Si on observe que $\widehat{X} < X$ et si on a bien $y_i \approx R x_i$, alors il est probable que $\widehat{Y} < Y$: l'estimation de Y est trop basse. L'estimateur par le ratio corrige l'estimation initiale $\widehat{Y}$ vers le haut, grâce au facteur $X/\widehat{X} > 1$.

Exemples

- Estimateur par le quotient pour un TAS : $d_i = N/n$

$$\widehat{Y}_{R,TAS} = X \frac{(N/n)\sum_{i\in S} y_i}{(N/n)\sum_{i\in S} x_i} = X \frac{\sum_{i\in S} y_i}{\sum_{i\in S} x_i}$$

- Estimateur par le quotient pour un TASST : $d_i = N_h/n_h$ si $i \in S_h$. On obtient l'estimateur par le *quotient combiné*

$$\widehat{Y}_{R,TASST} = X \frac{\sum_h \frac{N_h}{n_h} \sum_{i\in S_h} y_i}{\sum_h \frac{N_h}{n_h} \sum_{i\in S_h} x_i}$$

Ajustement des poids

Représentation de $\widehat{Y}_R$ comme somme pondérée

$$\widehat{Y}_R = X \frac{\sum_{i \in S} d_i y_i}{\sum_{i \in S} d_i x_i} = \sum_{i \in S} \left(\frac{X}{\widehat{X}} \right) d_i y_i = \sum_{i \in S} w_i y_i .$$

Pondération initiale $\longrightarrow$ Pondération finale

$$d_i \longrightarrow w_i = g_i d_i, \; g_i = \frac{X}{\widehat{X}} = \frac{\sum_{i \in U} x_i}{\sum_{i \in S} d_i x_i}$$

On remarque que g_i dépend de l'échantillon S.

Calage

- En général, avec les poids de sondage,

$$\widehat{X} = \sum_{i \in S} d_i x_i \neq X.$$

- Avec les poids ajustés

$$\widehat{X}_R = \sum_{i \in S} w_i x_i = \left(\frac{X}{\widehat{X}} \right) \sum_{i \in S} d_i x_i = \left(\frac{X}{\widehat{X}} \right) \widehat{X} = X.$$

Variance et biais (1)

- La variance de $\widehat{Y}_R$ ne peut pas se calculer directement :

$$\text{var}(\widehat{Y}_R) = X^2 \,\text{var}\left(\frac{\widehat{Y}}{\widehat{X}}\right) = \ ?$$

- On doit recourir à une approximation linéaire de $\widehat{Y}_R$

$$\widehat{Y}_R \approx Y + (\widehat{Y} - R\widehat{X}) \text{ où } R = \frac{Y}{X} = \frac{\overline{Y}}{\overline{X}}$$

Variance et biais (2)

- On a

$$\widehat{Y} - R\widehat{X} = \sum_{i \in S} d_i y_i - R \sum_{i \in S} d_i x_i = \sum_{i \in S} d_i (y_i - R x_i)$$

- Si on définit les résidus $v_i = y_i - R x_i$,

$$\widehat{Y} - R\widehat{X} = \sum_{i \in S} d_i (y_i - R x_i) = \sum_{i \in S} d_i v_i.$$

- On remarque que $\sum_{i \in S} d_i v_i$ est un estimateur sans biais de $V = \sum_{i \in U} v_i$ et que

$$V = \sum_{i \in U} v_i = \sum_{i \in U} (y_i - R x_i) = Y - RX = Y - \frac{Y}{X} X = 0.$$

Variance et biais (3)

- Ainsi, $\widehat{Y}_R$ est donné en première approximation par la vraie valeur Y à laquelle on ajoute un terme $\widehat{V}$ d'espérance nulle :

$$\begin{aligned} \widehat{Y}_R &\approx Y + (\widehat{Y} - R\widehat{X}) = Y + \widehat{V}, \\ \widehat{Y}_R = \sum_{i \in S} w_i y_i &\approx Y + \sum_{i \in S} d_i v_i. \end{aligned}$$

Variance et biais (4)

- L'estimateur $\widehat{Y}_R$ est donc approximativement sans biais :

$$\mathrm{E}(\widehat{Y}_R) \approx Y + \mathrm{E}(\widehat{V}) = Y,$$

- La variance de $\widehat{Y}_R$ est donnée approximativement par

$$\mathrm{var}(\widehat{Y}_R) \approx \mathrm{var}(\widehat{V}) = \mathrm{var}\left(\sum_{i \in S} d_i v_i\right),$$

ce que l'on peut calculer à partir du plan de sondage pour S.

Estimation par le quotient pour un TAS (1)

- Echantillon aléatoire simple $S \subseteq U$, $|S| = n$, $|U| = N$. Les poids de sondage sont $d_i = N/n = 1/f$.
- On suppose que l'on connaît de y_i et x_i pour les unités de l'échantillon $i \in S$ et que le total $X = \sum_{i \in U} x_i$ est connu.
- Estimateur par le quotient pour un *total* $Y = \sum_{i \in U} y_i$

$$\widehat{Y}_R = X \frac{\sum_{i \in S} y_i}{\sum_{i \in S} x_i}$$

Estimation par le quotient pour une *moyenne* $\overline{Y} = \frac{1}{N} Y$

$$\widehat{\overline{Y}}_R = \frac{1}{N} \widehat{Y}_R = \overline{X} \frac{\overline{y}_S}{\overline{x}_S}$$

Estimation par le quotient pour un TAS (2)

- La variance de l'estimateur par le quotient d'un total est donnée approximativement par

$$\text{var}(\widehat{Y}_R) \approx \text{var}\left(\sum_{i\in S} d_i v_i\right) = \text{var}\left(\sum_{i\in S} d_i \left(y_i - \frac{Y}{X}x_i\right)\right)$$

- Pour un TAS, on obtient

$$\text{var}(\widehat{Y}_R) \approx N^2(1-f)\frac{1}{n}D_v^2,$$

où

$$D_v^2 = \frac{1}{N-1}\sum_{i\in U}(v_i - \overline{V})^2.$$

Estimation par le quotient pour un TAS (3)

- Puisque $\overline{V} = 0$, on a

$$D_v^2 = \frac{1}{N-1}\sum_{i\in U} v_i^2 = \frac{1}{N-1}\sum_{i\in U}(y_i - Rx_i)^2 \text{ avec } R = \frac{Y}{X}$$

- Ainsi, pour un TAS, la variance de l'estimateur par le quotient d'un total est donnée approximativement par

$$\text{var}(\widehat{Y}_R) \approx N^2(1-f)\frac{1}{n}\left\{\frac{1}{N-1}\sum_{i\in U}(y_i - Rx_i)^2\right\}$$

- Estimation de la variance

$$\widehat{\text{var}}(\widehat{Y}_R) = (1-f)\frac{1}{n}\left\{\frac{1}{n-1}\sum_{i\in S}(y_i - \widehat{R}x_i)^2\right\} \text{ avec } \widehat{R} = \frac{\widehat{Y}}{\widehat{X}}$$

Estimation par le quotient pour un TAS (4)

- Variance de l'estimateur par le quotient pour une moyenne

$$\text{var}(\widehat{\overline{Y}}_R) \approx (1-f)\frac{1}{n}\left\{\frac{1}{N-1}\sum_{i\in U}(y_i - Rx_i)^2\right\}$$

- Estimation de la variance

$$\widehat{\text{var}}(\widehat{\overline{Y}}_R) = (1-f)\frac{1}{n}\left\{\frac{1}{n-1}\sum_{i\in S}(y_i - \widehat{R}x_i)^2\right\}$$

- On remarque que

$$\widehat{R} = \frac{\widehat{Y}}{\widehat{X}} = \frac{(N/n)\sum_{i\in S} y_i}{(N/n)\sum_{i\in S} x_i} = \frac{\sum_{i\in S} y_i}{\sum_{i\in S} x_i} = \frac{\overline{y}_S}{\overline{x}_S}$$

Estimation par le quotient pour un TAS (5)

- Pour un sondage aléatoire simple

$$\theta = \frac{var(\widehat{Y}_R)}{var(\widehat{Y})} \approx 1 + \left(R\frac{D_x}{D_y}\right)^2 - 2\rho\left(R\frac{D_x}{D_y}\right)$$
$$= 1 + \left(\frac{CV(x)}{CV(y)}\right)^2 - 2\rho\frac{\mathrm{CV}(x)}{\mathrm{CV}(y)}$$

puisque

$$R\frac{D_x}{d_y} = \frac{Y}{X}\frac{D_x}{D_y} = \frac{\mathrm{CV}(x)}{CV(y)}.$$

Estimation par le quotient pour un TAS (6)

- Thus, in large sample, with simple random sample, the ratio estimator has smaller variance than the Horvitz-Thompson estimator if

$$\rho > \frac{1}{2}\frac{CV(x)}{CV(y)}.$$

Est-ce que ça marche ? (1)

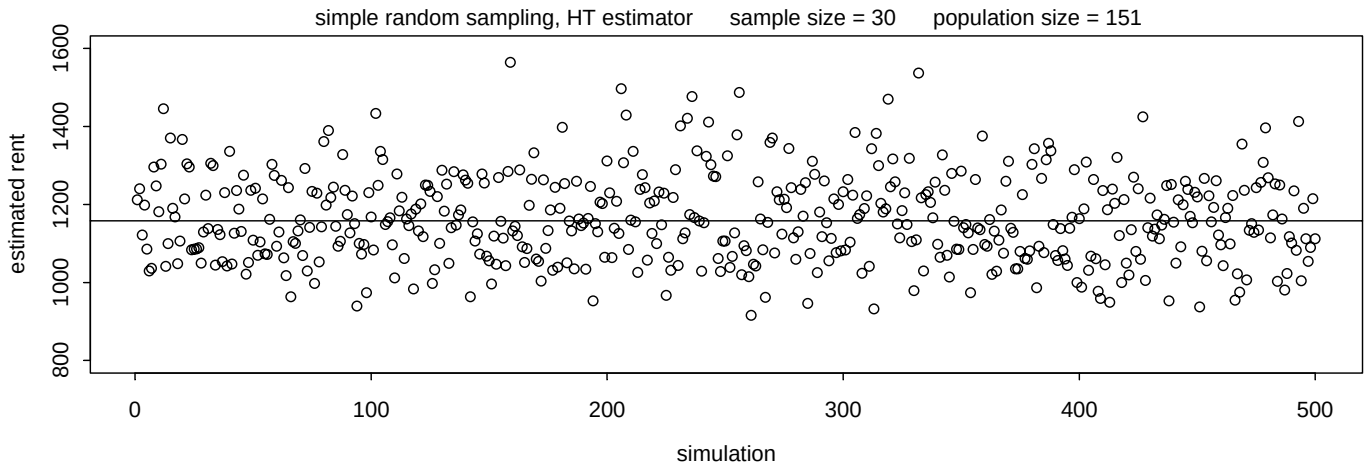
- Estimation du loyer moyen pour la population de logements ($N = 151$).
- Echantillons de taille $n = 30$.
- Stratification selon le nombre de pièce (cf. exemple stratification).
- Estimateur par le quotient avec la surface habitable.

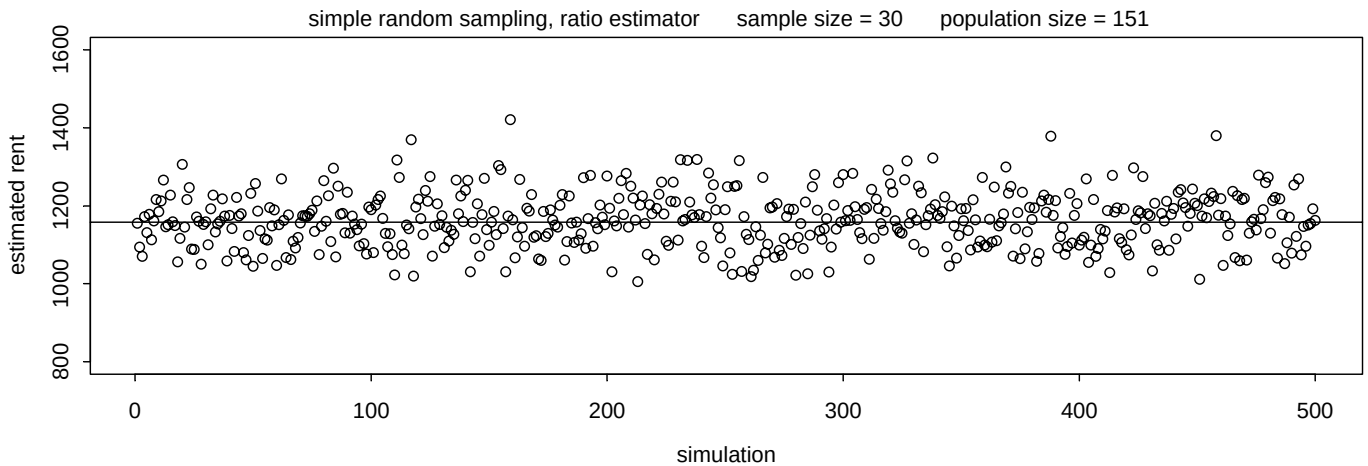
Est-ce que ça marche ? (2)

Précision des stratégies (stratégie = plan de sondage + estimateur)

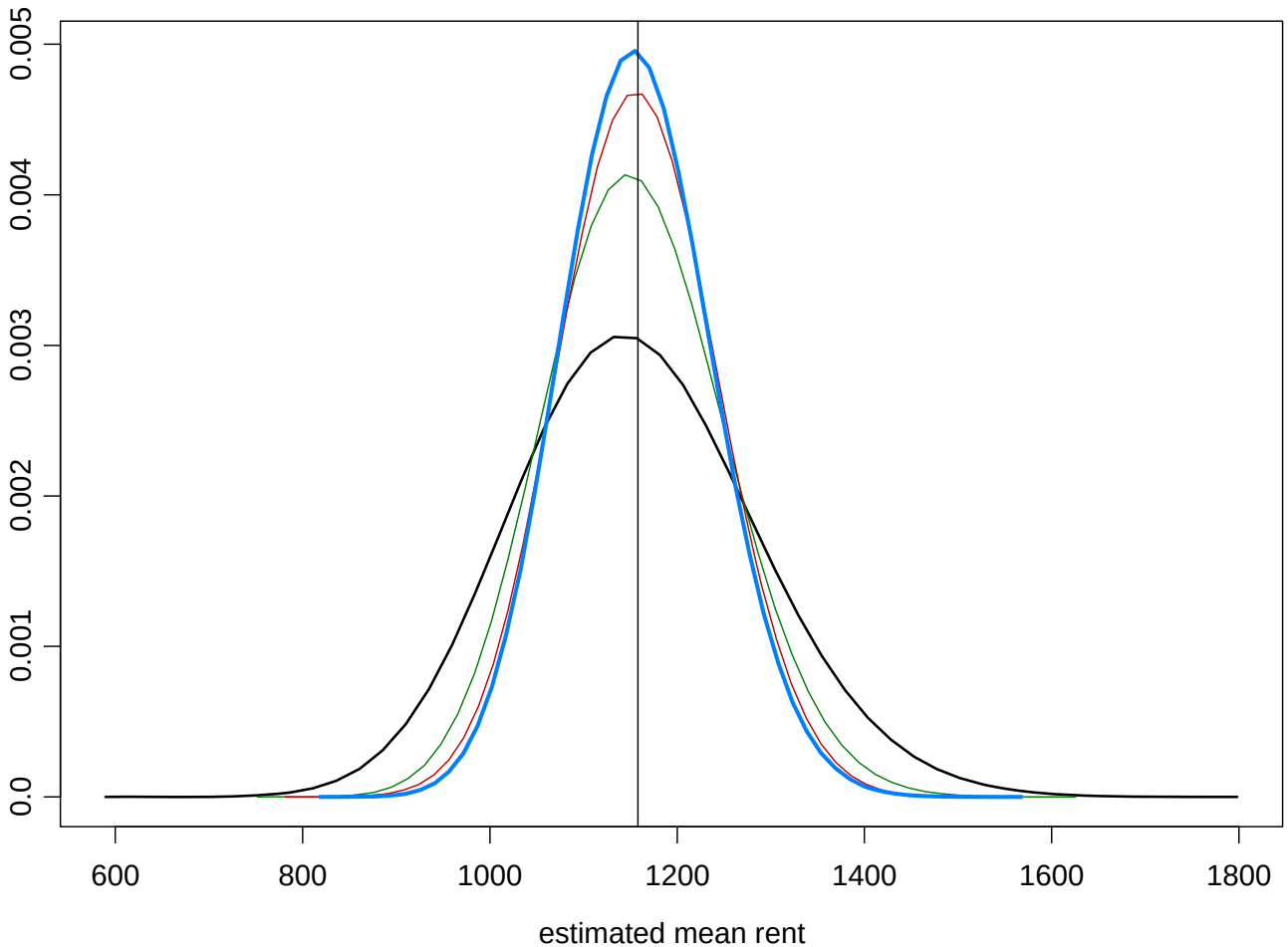
Plan de sondage	Estimateur	CV($\widehat{Y}$)
TAS	HT	9.3%
TASST, allocation proportionnelle	HT	6.9%
TASST, allocation optimale	HT	5.9%
TAS	quotient	5.8%

HT pour TAS, Ratio pour TAS





HT pour TAS, TASST (prop. + opt.). Ratio pour TAS



Post-stratification (1)

- Echantillon $S \subseteq U$, poids de sondage d_i, $i \in S$.
- Stratification $U = \bigcup_{h=1}^{H} U_h$. On connaît les tailles des strates $|U_h| = N_h$.
- Pour les unités de l'échantillon $i \in S$, l'appartenance aux strates $i \in S_h = S \cap U_h$ n'est connue qu'après l'enquête.
- La taille de l'échantillon dans les strates $n_h = |S \cap U_h|$ est donc une variable aléatoire.

Post-stratification (2)

- Estimateur *post-stratifié* pour un total

$$\widehat{Y}_{post} = \sum_h \left(\frac{N_h}{\widehat{N}_h} \right) \widehat{Y}_h = \sum_h N_h \frac{\sum_{i \in S_h} d_i y_i}{\sum_{i \in S_h} d_i}$$

- $\widehat{Y}_{post}$ n'est pas défini si $n_h = |S \cap U_h| = 0$.

Exemple : estimateur post-stratifié pour un TAS

Comme les poids de sondage sont donnés par $d_i = N/n$, on a

$$\widehat{Y}_h = \sum_{i \in S_h} d_i y_i = \frac{N}{n} \sum_{i \in S_h} y_i,$$

$$\widehat{N}_h = \sum_{i \in S_h} d_i = \frac{N}{n} n_h,$$

et donc

$$\widehat{Y}_{post} = \sum_h N_h \frac{\sum_{i \in S_h} d_i y_i}{\sum_{i \in S_h} d_i} = \sum_h \frac{N_h}{n_h} \sum_{i \in S_h} y_i.$$

Ajustement des poids

Représentation comme une somme pondérée

$$\widehat{Y}_{post} = \sum_h \sum_{i \in S_h} \left(\frac{N_h}{\widehat{N}_h} \right) d_i y_i = \sum_{i \in S} w_i y_i$$

Pondération initiale $\longrightarrow$ Pondération après post-stratification

d_i $\longrightarrow$ $w_i = g_i d_i$, $g_i = \frac{N_h}{\widehat{N}_h}$ pour $i \in S_h$

Calage

- En général, avec les poids de sondage,

$$\widehat{N}_h = \sum_{i \in S_h} d_i \neq N_h.$$

- Avec les poids ajustés

$$\widehat{N}_{h,post} = \sum_{i \in S_h} w_i = \left(\frac{N_h}{\widehat{N}_h}\right) \sum_{i \in S_h} d_i = \left(\frac{N_h}{\widehat{N}_h}\right) \widehat{N}_h = N_h.$$

Estimateur post-stratifié pour un TAS (1)

- Echantillon aléatoire simple $S \subseteq U$, $|S| = n$, $|U| = N$. Les poids de sondage sont $d_i = N/n = 1/f$.
- Stratification $U = \bigcup_{h=1}^{H} U_h$. On connaît les tailles des strates $|U_h| = N_h$.
- Pour les unités de l'échantillon $i \in S$, l'appartenance aux strates $i \in S_h = S \cap U_h$ n'est connue qu'après l'enquête.

Estimateur post-stratifié pour un TAS (2)

- La taille de l'échantillon dans les strates $n_h = |S \cap U_h|$ est une variable aléatoire de loi hypergéométrique. En particulier
$$\mathrm{E}(n_h) = n \frac{N_h}{N}.$$
- L'estimateur post-stratifié n'est pas défini si $n_h = 0$.

Estimateur post-stratifié pour un TAS (3)

- Estimateur post-stratifié pour un *total* $Y = \sum_{i \in U} y_i$

$$\widehat{Y}_{post} = \sum_h \frac{N_h}{n_h} \sum_{i \in S_h} y_i = \sum_{i \in S} w_i y_i, \text{ avec } w_i = \frac{N_h}{n_h} \text{ pour } i \in S_h.$$

- Estimateur post-stratifié pour une *moyenne* $\overline{Y} = \frac{1}{N} Y$

$$\widehat{\overline{Y}}_{post} = \frac{1}{N} \widehat{Y}_{post} = \sum_h \left(\frac{N_h}{N} \right) \frac{1}{n_h} \sum_{i \in S_h} y_i = \sum_h W_h \overline{y}_{S_h}.$$

- Estimateur post-stratifié pour une *proportion* $p_a = N_a / N$

$$\widehat{p}_{a,post} = \sum_h \left(\frac{N_h}{N} \right) \frac{n_{ah}}{n_h} = \sum_h W_h \widehat{p}_{ah}.$$

Estimateur post-stratifié pour un TAS (4)

- Conditionnellement aux tailles d'échantillon $n_1, \ldots, n_H$ dans les post-strates, un TAS post-stratifié est équivalent à un TASST.
- Ainsi, les formules de variance pour un TASST donnent les variances conditionnelles pour un TAS post-stratifié.

Estimateur post-stratifié pour un TAS (5)

- Variance conditionnelle de $\widehat{\overline{Y}}_{post}$

$$\text{var}(\widehat{\overline{Y}}_{post} \mid n_1 \ldots, n_H) = \sum_h W_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{1}{n_h} D_h^2.$$

- Estimation de la variance conditionnelle de $\widehat{\overline{Y}}_{post}$

$$\widehat{\text{var}}(\widehat{\overline{Y}}_{post}) = \sum_h W_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{1}{n_h} d_h^2.$$

- D_h^2 et d_h^2 sont définis comme pour le sondage stratifié.

Estimateur post-stratifié pour un TAS (6)

On montre que la variance non-conditionnelle de $\widehat{\overline{Y}}_{post}$ est donnée par

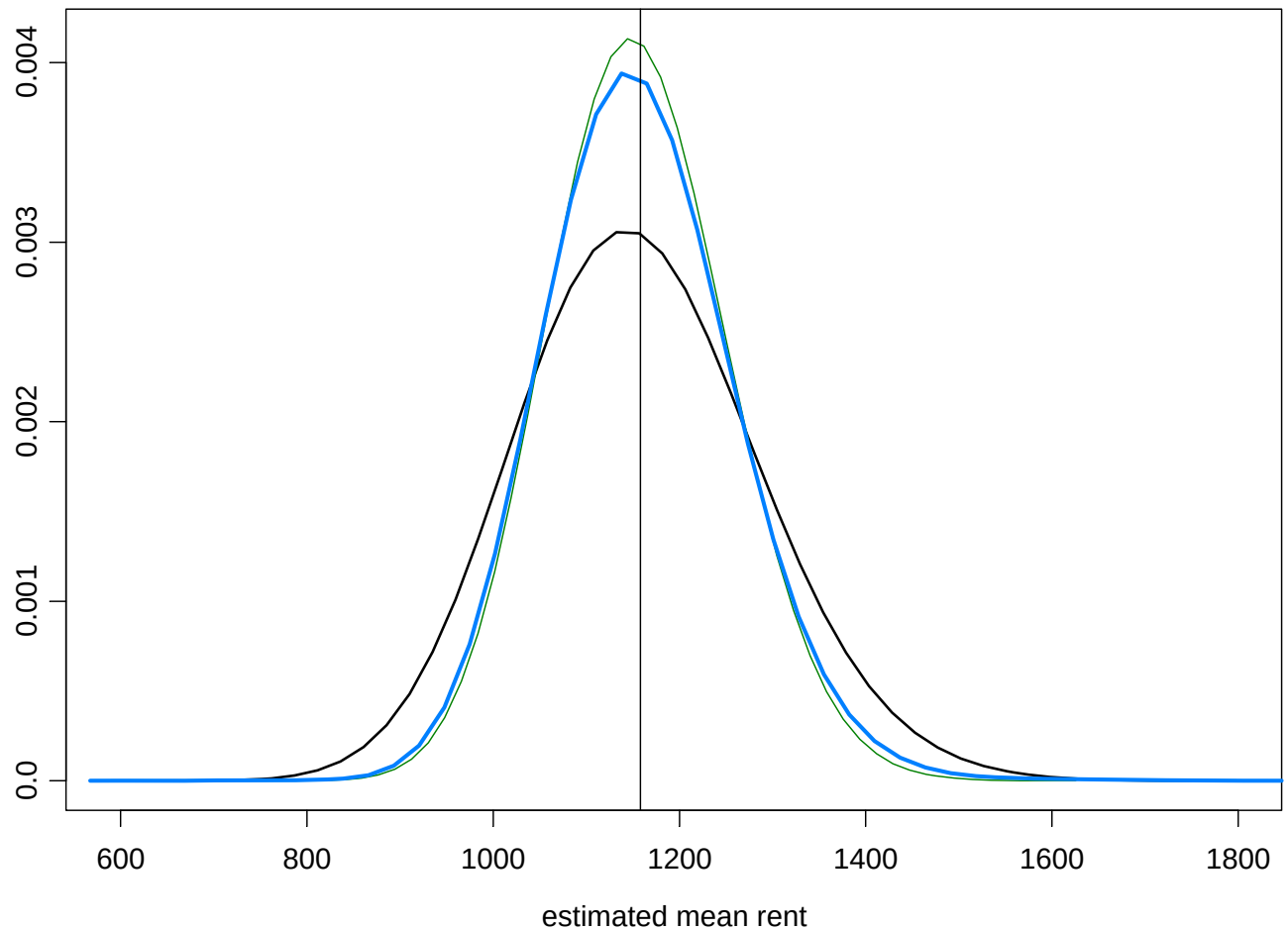
$$\begin{aligned} \text{var}(\widehat{\overline{Y}}_{post}) &= \text{E}(\text{var}(\widehat{\overline{Y}}_{post} \mid n_1, \ldots, n_H)) \\ &= \text{var}(\widehat{\overline{Y}}_{TASST,\,prop}) + O\left(\frac{1}{n^2}\right). \end{aligned}$$

Est-ce que ça marche ?

- Estimation du loyer moyen pour la population de logements ($N = 151$).
- TAS de taille $n = 30$.
- Post-stratification selon le nombre de pièce.

Post-strate	`zi`	N_h
1	1,2	57
2	3,4	64
3	5,6	30
		151

HT pour TAS et TASST (prop.). Post-stratification pour TAS.



Non-réponse

- Dans presque toutes les enquêtes
- Effet de la non-réponse difficile à prévoir
- Mesures lors de la planification et de l'exécution d'une enquête
- Procédures d'extrapolation spéciales

Types de non-réponse

		y_1	y_2	y_3
Réponse		$*$	$*$	$*$
Non-réponse totale	(unit nonresponse)			
Non-réponse partielle	(item nonresponse)	$*$		$*$

- Traitement de la non-réponse partielle par *imputation*
- Traitement de la non-réponse totale par *re-pondération*

Biais : un modèle simple

- Population $U = U_r \cup U_{nr}$: population de répondants ($|U_r| = N_r$), population de non-répondants ($|U_{nr}| = N_{nr}$).
- On a

$$\overline{Y} = \frac{N_r}{N}\overline{Y}_r + \left(1 - \frac{N_r}{N}\right)\overline{Y}_{nr}.$$

- Soit $\widehat{\overline{Y}}_r$ un estimateur sans biais de $\overline{Y}_r$. Alors

$$\textit{Biais} = \mathrm{E}(\widehat{\overline{Y}}_r) - \overline{Y} = \overline{Y}_r - \overline{Y} = \left(1 - \frac{N_r}{N}\right)(\overline{Y}_r - \overline{Y}_{nr}).$$

Biais : un modèle plus réaliste (1)

$$\begin{array}{ccccc} \text{Population} & \longrightarrow & \text{Echantillon} & \longrightarrow & \text{Répondants} \\ U & \xrightarrow{\pi_i} & S & \xrightarrow{\theta_{S,i}} & R \end{array}$$

Si la probabilité de réponse

$$\theta_{S,i} = P(\text{unité } i \text{ répond} \mid \text{échantillon } S \text{ est sélectionné})$$

était connue, alors on pourrait estimer $Y = \sum_{i \in U} y_i$ sans biais par

$$\widehat{Y} = \sum_{i \in R} \frac{y_i}{\pi_i \, \theta_{S,i}} = \sum_{i \in R} \left(\frac{1}{\pi_i}\right) \left(\frac{1}{\theta_{S,i}}\right) y_i.$$

Biais : un modèle plus réaliste (2)

- Mais la probabilité de réponse

$$\theta_{S,i} = P(\text{unité } i \text{ répond} \mid \text{échantillon } S \text{ est sélectionné})$$

n'est pas connue, et $\theta_{S,i}$ doit être estimée sur la base d'un modèle, d'où un risque de biais si le modèle ne prend pas en compte tous les facteurs qui influencent le comportement de réponse.

- Estimateur du total $Y = \sum_{i \in U} y_i$ sous le modèle de non-réponse

$$\widehat{Y} = \sum_{i \in R} \frac{y_i}{\pi_i \, \widehat{\theta}_{S,i}} = \sum_{i \in Ri} \left(\frac{1}{\pi_i}\right) \left(\frac{1}{\widehat{\theta}_{S,i}}\right) y_i.$$

Mesures pour réduire la non-réponse

- Thème de l'enquête
- Date de l'enquête
- Enquêteur
- Mode d'enquête
- Questionnaire
- Charge
- Introduction de l'enquête
- Récompenses
- Essais de contact et rappels

Mécanismes de réponse (1)

U	S	x	y
1	1	x_1	y_1
2	0	x_2	$*$
3	0	x_3	$*$
4	1	x_4	y_4
5	1	x_5	$*$
6	1	x_6	y_6
7	0	x_7	$*$
8	1	x_8	y_8
9	1	x_9	y_9

Mécanismes de réponse (2)

- La probabilité de réponse θ_i *ne dépend pas* de S, x_i et y_i (missing completely at random, MCAR) : pas de biais, uniquement une perte de précision.
- La probabilité de réponse θ_i *ne dépend que* de S et x_i (missing at random, MAR) : on peut estimer θ_i à l'aide de la variable auxiliaire x_i (ignorable nonresponse).
- La probabilité de réponse θ_i *dépend* de S, x_i et y_i : l'information auxiliaire x_i ne suffit pas pour estimer θ_i (nonignorable nonresponse).

Traitement de la non-réponse par re-pondération (1)

- Population U, $|U| = N$
- Paramètre $\overline{Y} = (1/N) \sum_i y_i$
- Echantillon aléatoire simple $S \subseteq U$, $|S| = n$: échantillon brut
- Réponse $R \subseteq S$, $|R| = m$: échantillon net

Traitement de la non-réponse par re-pondération (2)

Modèle des *classes de non-réponse uniforme*

- Les unités de la population répondent indépendamment les unes des autres
- Décomposition de la population en classes $U = \bigcup_{h=1}^{H} U_h$ telles que les unités de la classe h répondent avec la même probabilité θ_h :

$$U \overset{\pi_i = n/N}{\longrightarrow} S, \quad S_h = U_h \cap S, \quad S_h \overset{\theta_h}{\longrightarrow} R_h$$

Traitement de la non-réponse par re-pondération (3)

- Estimateur pour la probabilité de réponse dans la classe U_h

$$\widehat{\theta}_i = \frac{|R_h|}{|S_h|} = \frac{m_h}{n_h}, \text{ pour } i \in U_h.$$

- Estimateur par re-pondération de $\overline{Y} = (1/N)\sum_i y_i$ (weighting classes estimator)

$$\begin{aligned}\widehat{\overline{Y}}_{wcl} &= \frac{1}{N}\sum_{i\in R} d_i \widehat{\theta}_i^{-1} y_i \\ &= \frac{1}{N}\sum_h \left(\frac{N}{n}\right)\left(\frac{n_h}{m_h}\right)\sum_{i\in R_h} y_i = \sum_h \left(\frac{n_h}{n}\right)\overline{y}_{R_h}.\end{aligned}$$

Traitement de la non-réponse par re-pondération (4)

- Les poids

$$w_i^{(1)} = d_i \widehat{\theta}_i^{-1} = \left(\frac{N}{n}\right)\left(\frac{n_h}{m_h}\right)$$

représentent un ajustement pour la non-réponse sous le modèle des classes de non-réponse uniforme.

Traitement de la non-réponse par re-pondération (5)

- Si $|U_h| = N_h$ est connu, on peut utiliser l'estimateur post-stratifié, en prenant comme poids initiaux les poids ajustés pour la non-réponse

$$\widehat{\bar{Y}}_{post} = \frac{1}{N} \sum_h \left(\frac{N_h}{\widehat{N}_h} \right) \widehat{Y}_h = \frac{1}{N} \sum_h N_h \frac{\sum_{i \in S_h} w_i^{(1)} y_i}{\sum_{i \in S_h} w_i^{(1)}}.$$

- On a

$$\widehat{N}_h = \sum_{i \in S_h} w_i^{(1)} = \left(\frac{N}{n} \right) \left(\frac{n_h}{m_h} \right) m_h = \frac{N}{n} n_h.$$

Traitement de la non-réponse par re-pondération (6)

- Les poids après post-stratification sont donnés, pour $i \in R_h$, par

$$w_i^{(2)} = \left(\frac{N_h}{\widehat{N}_h}\right) w_i^{(1)} = N_h \left(\frac{n}{N n_h}\right) \left(\frac{N}{n}\right) \left(\frac{n_h}{m_h}\right) = \frac{N_h}{m_h}$$

- On obtient finalement

$$\begin{aligned} \widehat{\overline{Y}}_{post} &= \frac{1}{N} \sum_h \frac{N_h}{m_h} \sum_{i \in R_h} y_i = \sum_h \frac{N_h}{N} \frac{1}{m_h} \sum_{i \in R_h} y_i \\ &= \sum_h \left(\frac{N_h}{N}\right) \overline{y}_{R_h} \end{aligned}$$

Sondage en deux phases et prob. conditionnelle (1)

- Probabilité P sur Ω
- Probabilité conditionnelle à $A \subset \Omega$ $(P(A) \neq 0)$

$$P(\omega \mid A) = \begin{cases} P(\omega)/P(A) & \omega \in A \\ 0 & \omega \notin A \end{cases}$$

- Sondage en deux phases : probabilité sur $\Omega = \Omega_1 \times \Omega_2$

Sondage en deux phases et prob. conditionnelle (2)

- Probabilité conditionnelle

$$P(\omega_2 \mid \omega_1) = \begin{cases} P(\omega_1, \omega_2)/P(\omega_1) & (\omega_1, \omega_2) \in \omega_1 \times \Omega_2 \\ 0 & (\omega_1, \omega_2) \notin \omega_1 \times \Omega_2 \end{cases}$$

où

$$\begin{aligned} P(\omega_2 \mid \omega_1) &= P((\omega_1, \omega_2) \mid \omega_1 \times \Omega_2) \\ P(\omega_1) &= P(\omega_1 \times \Omega_2) \end{aligned}$$

- Espérance conditionnelle d'une variable aléatoire
  $X : \Omega_1 \times \Omega_2 \to \mathbb{R}$

$$E(X \mid \omega_1) = E(X \mid \omega_1 \times \Omega_2) = \sum_{\omega_2 \in \Omega_2} X(\omega_1, \omega_2) P(\omega_2 \mid \omega_1)$$

Classes de non-réponse uniforme / Compléments (1)

- Probabilité sur

$$\Omega_1 \times \Omega_2 = \mathcal{P}_n(U) \times \bigcup_{m=0}^{n} \mathcal{P}_m(U)$$

donnée par

$$P(S, R) = P(R \mid S)P(S)$$

où, si $R \subseteq S$,

$$P(R \mid S) = \prod_{h \in H} \theta_h^{m_h}(1 - \theta_h)^{n_h - m_h}$$

Classes de non-réponse uniforme / Compléments (2)

- Décomposition de $\Omega_1 \times \Omega_2$ adaptée à la décomposition $U = \bigcup_{h \in H} U_h$. On a

$$\Omega_1 = \bigcup_{\substack{n_1 + \cdots + n_H = n \\ n_h \geq 0, h \in H}} \Omega(n_h, h \in H)$$

où $\Omega(n_h, h \in H) = \{S \in \Omega_1; |S \cap U_h| = n_h, h \in H\}$. On a une décomposition similaire de Ω_2.

- Alors

$$\Omega_1 \times \Omega_2 = \bigcup_{m=0}^{n} \bigcup_{\mathbf{n}} \bigcup_{\mathbf{m}} \Omega(n_h, h \in H) \times \Omega(m_h, h \in H)$$

Classes de non-réponse uniforme / Compléments (3)

- Estimateur par classe de repondération

$$\begin{aligned}\widehat{Y}_{wcl} &= \frac{N}{n}\sum_{h\in H}\frac{n_h}{m_h}\sum_{i\in R_h} y_i = \sum_{h\in H}\widehat{Y}_h \\ &= \frac{N}{n}\left(\frac{\sum_{i\in U} I_{i1}z_{hi}}{\sum_{i\in U} I_{i2}z_{hi}}\right)\sum_{i\in U} I_{i2}z_{hi}y_i\end{aligned}$$

- Les tailles $n_h = |S \cap U_h|$ suivent une loi hypergéométrique $n_h \sim H(N, N_h, n)$
- Conditionnellement à n_h, les tailles $m_h = |R \cap U_h|$ suivent une loi binomiale $m_h \mid n_h \sim B(n_h, \theta_h)$
- Conditionnellement à n_h et m_h, R_h est un échantillon aléatoire simple

Classes de non-réponse uniforme / Compléments (4)

- Une espérance conditionnelle

$$E(\widehat{Y}_h \mid n_h, m_h) = \frac{N}{n}\frac{n_h}{m_h}\sum_{i\in U}\underbrace{E(I_{i2} \mid n_h, m_h)}_{=m_h/N_h} z_{hi}y_i$$

$$= \frac{N}{n}\frac{n_h}{N_h}\sum_{i\in U} z_{hi}y_i = \underbrace{\left(\frac{N}{n}n_h\right)}_{=\widehat{N}_h}\frac{Y_h}{N_h} = \frac{\widehat{N}_h}{N_h}Y_h$$

- Il suit que

$$E(\widehat{Y}_h) = \frac{N}{n}E(n_h)\frac{Y_h}{N_h} = \frac{N}{n}\left(n\frac{N_h}{N}\right)\frac{Y_h}{N_h} = Y_h$$

Est-ce que ça marche ? (1)

- Estimation du loyer moyen pour la population de logements ($N = 151$). TAS de taille $n = 30$. Le modèle de non-réponse est donné par

Classes	zi	N_h	θ_h
1	1,2,3	90	0.8
2	4,5,6	61	0.5
		151	

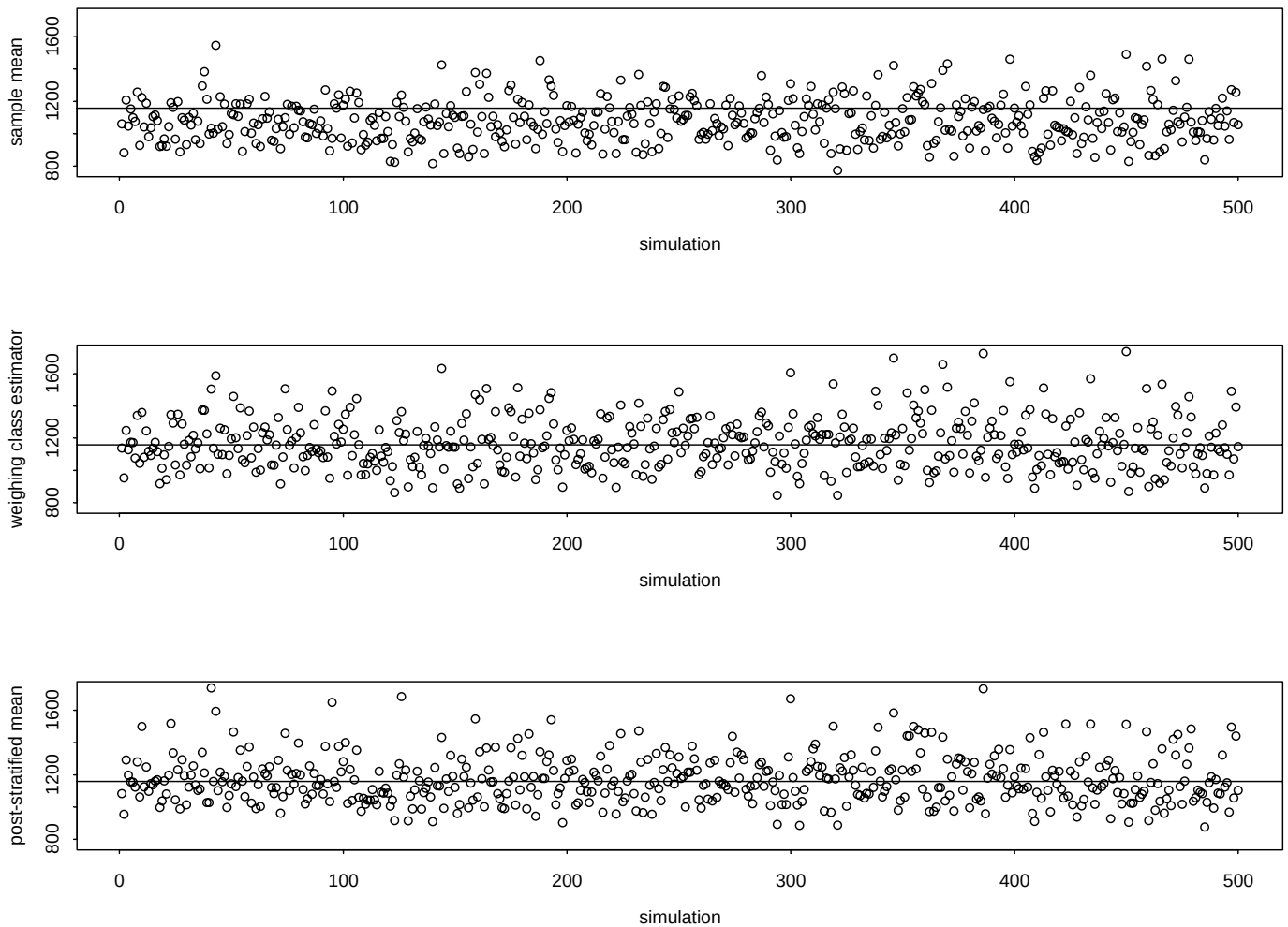
Est-ce que ça marche ? (2)

- On compare l'estimateur naïf $\overline{y}_R$ avec les estimateurs par re-pondération

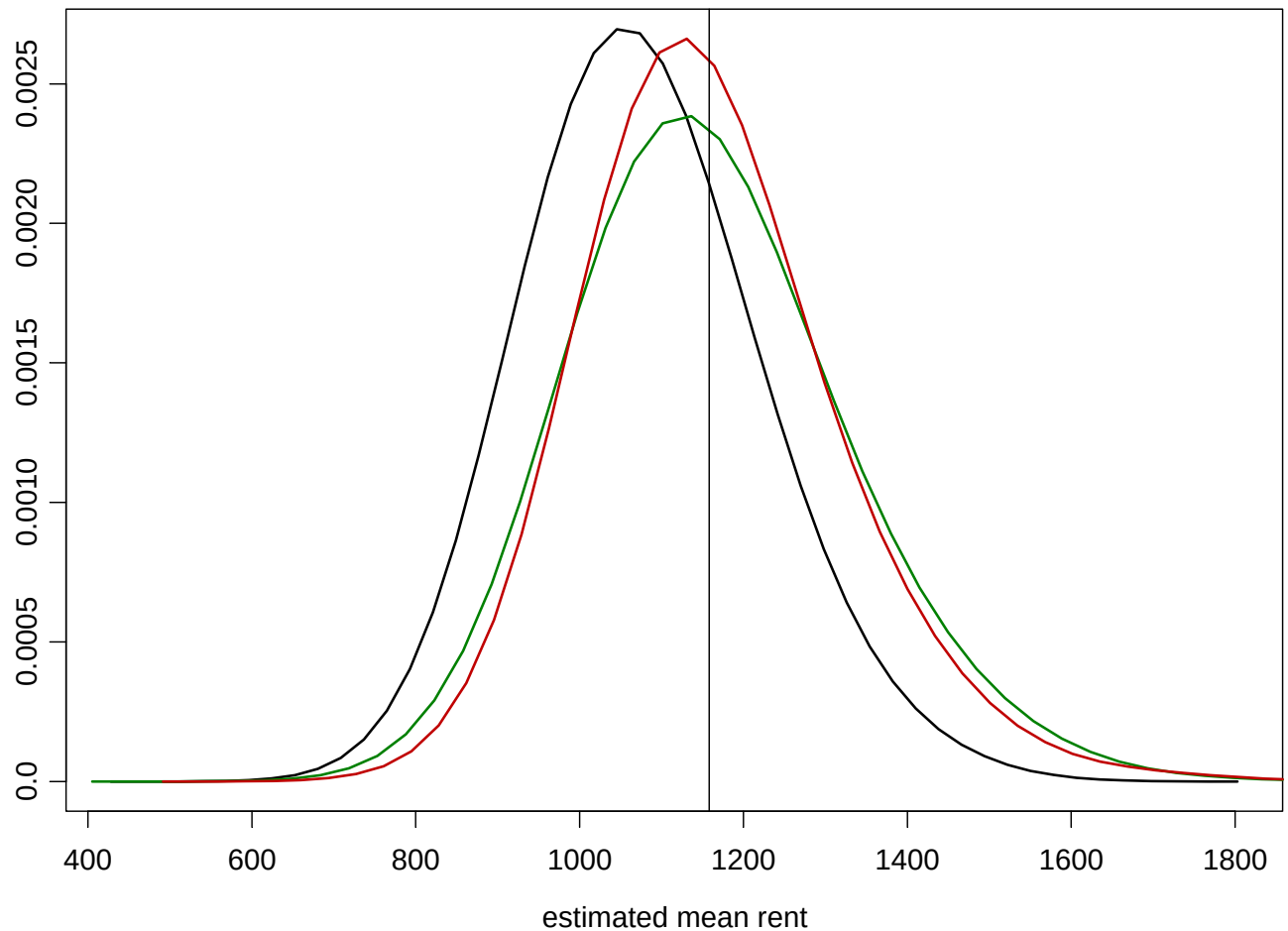
$$\widehat{\overline{Y}}_{wcl} = \frac{1}{N}\sum_{i\in R} w_i^{(1)} y_i = \sum_h \left(\frac{n_h}{n}\right)\overline{y}_{R_h},$$

$$\widehat{\overline{Y}}_{post} = \frac{1}{N}\sum_{i\in R} w_i^{(2)} y_i = \sum_h \left(\frac{N_h}{N}\right)\overline{y}_{R_h}.$$

Traitement de la non-réponse par re-pondération



Traitement de la non-réponse par re-pondération



Références

- Ardilly, P. (2006). Les techniques de sondages. Editions Technip, Paris.
- Cochran, W.G. (1977). Sampling Techniques. Wiley, New York.
- Särdnal, C.-E., Swensson, B., Wretman J. (1992). Model Assisted Survey Sampling. Springer, New York.
- Lohr, S.L. (1999). Sampling : Design and Analysis. Duxbury Press, Pacific Grove, USA.
- Tillé, Y. (2001). Théorie des sondages. Dunod, Paris.
- Ardilly, P., Tillé Y. (2003). Exercices corrigés de méthodes de sondages. Ellipses, Paris.