Master 1 - Economie-Gestion Mention Ingénierie économique et statistique 2016 - 2017

Formulaire - Techniques de Sondage

Ce document est un formulaire accompagnant le cours 'Techniques de Sondage' du Master 1 IES, il est autorisé pour l'examen.

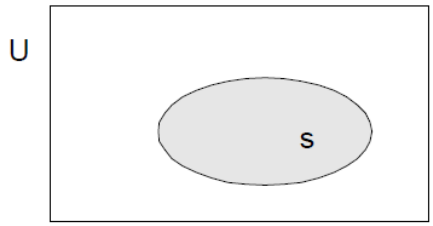
Philippe Périé

Contents

1. Plans aléatoires simples

Principe

Tirage d'un échantillon de n unités sans remise et à probabilités égales dans une population de taille N finie



Notations

1. ***Dans la population (ou univers)*** $U = \{1,2,...,k,...,N\}$
   - Variable d'intérêt : Y de caractéristique individuelle Y_k
   - Total : $T_Y = \sum_{k\in U} Y_k$
   - Moyenne: $\overline{Y} = \frac{T_Y}{N} = \frac{1}{N}\sum_{k\in U} Y_k$
   - Variance : $\sigma_y^2 = \frac{1}{N}\sum_{k\in U}\left(Y_k - \overline{Y}\right)^2$
   - Dispersion (variance modifiée) : $S_y^2 = \frac{1}{N-1}\sum_{k\in U}\left(Y_k - \overline{Y}\right)^2 = \frac{N}{N-1}\sigma_y^2$

2. ***Dans l'échantillon*** *s* : sous-ensemble de *U* de taille *n(s)*
   - Ensemble des échantillons possibles : S
   - Plan de sondage probabiliste : loi de probabilité sur S

     $$p(s) \geq 0, \forall s \in S, \text{ et } \sum_{s\in S} p(s) = 1.$$

   - Moyenne : $\hat{\overline{y}} = \frac{1}{n}\sum_{k\in S} Y_k$
   - Dispersion empirique : $\hat{s}_y^2 = \frac{1}{n-1}\sum_{k\in S}\left(Y_k - \hat{\overline{y}}\right)^2$
   - Probabilité d'inclusion d'ordre un de *k* : $\pi_k = P(k \in s) = \sum_{s\in S / k\in s} p(s)$
   - Probabilité d'inclusion ou double de *k* et *l* : $\pi_{kl} = P(k \in s, l \in s) = \sum_{s\in S / k,l\in s} p(s)$
   - $\Delta_{kl} = \pi_{kl} - \pi_k\pi_l$

Formulaire sondage aléatoire simple

Probabilité de sélectionner l'échantillon s : $p(s) = 1/C_N^n$

Probabilité de sélectionner l'individu k : $\forall k \in U,\ \pi_k = P(k \in s) = \frac{n}{N} = f$ (taux de sondage)

Paramètre d'intérêt / Statistique	Moyenne	Proportion $p = N_0/N$	Total
Estimateur du paramètre d'intérêt	$\hat{\bar{y}} = \frac{1}{n}\sum_{k\in S} Y_k = \hat{\bar{y}}(s)$	$\hat{p} = \frac{1}{n}\sum_{k\in s} y_k = \frac{n_0}{n}$	$\hat{ty} = N \times \hat{\bar{y}} = \frac{N}{n}\sum_{k\in S} Y_k$
Vraie variance d'échantillonnage de cet estimateur	$Var(\hat{\bar{y}}) = \left(1 - \frac{n}{N}\right)\frac{S_y^2}{n}$	$Var(\hat{p}) = \left(1 - \frac{n}{N}\right)\frac{N}{N-1}\frac{p(1-p)}{n}$	$Var(\hat{t}_y) = N^2\left(1 - \frac{n}{N}\right)\frac{S_y^2}{n}$
Estimateur de la variance d'échantillonnage	$\hat{Var}(\hat{\bar{y}}) = \left(1 - \frac{n}{N}\right)\frac{\hat{s}_y^2}{n}$	$\hat{Var}(\hat{p}) = \left(1 - \frac{n}{N}\right)\frac{\hat{p}(1-\hat{p})}{n-1}$	$\hat{Var}(\hat{t}_y) = N^2\left(1 - \frac{n}{N}\right)\frac{\hat{s}_y^2}{n}$

Intervalle au niveau de confiance 95% pour la moyenne :

$$IC_{95\%}(\bar{Y}) = \left[\hat{\bar{y}} - 1{,}96\sqrt{\hat{Var}(\hat{\bar{y}})},\ \hat{\bar{y}} + 1{,}96\sqrt{\hat{Var}(\hat{\bar{y}})}\right]$$

sous hypothèse que *n* est grand $\frac{\hat{\bar{y}} - \bar{Y}}{\sqrt{Var(\hat{\bar{y}})}} \to N(0, 1)$

2. Plans a probabilités inégales

Principe

Retenir les unités les plus porteuses d'information : sélection des unités avec des probabilités proportionnelles à une variable X de taille corrélée positivement à une variable d'intérêt Y

Notations

Formulaire

Probabilité de sélectionner l'individu k :

- Pour un plan à probabilités proportionnelles à une variable X de taille (corrélée positivement à Y)

$$\forall k \in U,\ \pi_k = P(k \in S) = n\frac{X_k}{\sum_{k \in U} X_k}$$

- Pour un plan de taille fixe, $\sum_{k \in U} \pi_k = n$

Paramètre d'intérêt / Statistique	Moyenne	Total
Estimateur d'Horvitz-Thompson du paramètre d'intérêt (π-estimateur)	Si la taille N est connue : $\hat{\mu}_{y\pi} = \frac{1}{N}\sum_{k \in s}\frac{Y_k}{\pi_k} = \frac{\hat{t}_{y\pi}}{N}$ Sinon, estimateur de Hájek : $\hat{\mu}_{yH} = \frac{1}{\hat{N}_\pi}\sum_{k \in s}\frac{Y_k}{\pi_k} = \frac{\sum_{k \in s}\frac{Y_k}{\pi_k}}{\sum_{k \in s}\frac{1}{\pi_k}} = \frac{\hat{t}_{y\pi}}{\hat{N}_\pi}$	$\hat{t}_{y\pi} = \sum_{k \in s}\frac{Y_k}{\pi_k}$ En particulier : $\hat{N}_\pi = \sum_{k \in s}\frac{1}{\pi_k}$
Vraie variance d'échantillonnage de cet estimateur	Cas général $Var(\hat{\mu}_{y\pi}) = \frac{1}{N^2}\sum_{k \in U}\sum_{l \in U}\frac{Y_k}{\pi_k}\frac{Y_l}{\pi_l}\Delta_{kl}$ Si la taille de l'échantillon est fixe $Var(\hat{\mu}_{y\pi}) = -\frac{1}{2N^2}\sum_{k \in U}\sum_{l \in U}\left(\frac{Y_k}{\pi_k} - \frac{Y_l}{\pi_l}\right)^2\Delta_{kl}$	Cas général : $Var(\hat{t}_{y\pi}) = \sum_{k \in U}\sum_{l \in U}\frac{Y_k}{\pi_k}\frac{Y_l}{\pi_l}\Delta_{kl}$ Si la taille de l'échantillon est fixe $Var(\hat{t}_{y\pi}) = -\frac{1}{2}\sum_{k \in U}\sum_{l \in U}\left(\frac{Y_k}{\pi_k} - \frac{Y_l}{\pi_l}\right)^2\Delta_{kl}$
Estimateur de la variance d'échantillonnage	Cas général $\hat{Var}(\hat{\mu}_{y\pi}) = \frac{1}{N^2}\sum_{k \in s}\sum_{l \in s}\frac{Y_k}{\pi_k}\frac{Y_l}{\pi_l}\frac{\Delta_{kl}}{\pi_{kl}}$ Si la taille de l'échantillon est fixe $\hat{Var}_2(\hat{\mu}_{y\pi}) = -\frac{1}{2N^2}\sum_{k \in s}\sum_{l \in s}\left(\frac{Y_k}{\pi_k} - \frac{Y_l}{\pi_l}\right)^2\frac{\Delta_{kl}}{\pi_{kl}}$	Cas général $\hat{Var}(\hat{t}_{y\pi}) = \sum_{k \in s}\sum_{l \in s}\frac{Y_k}{\pi_k}\frac{Y_l}{\pi_l}\frac{\Delta_{kl}}{\pi_{kl}}$ Si la taille de l'échantillon est fixe $\hat{Var}_2(\hat{t}_{y\pi}) = -\frac{1}{2}\sum_{k \in s}\sum_{l \in s}\left(\frac{Y_k}{\pi_k} - \frac{Y_l}{\pi_l}\right)^2\frac{\Delta_{kl}}{\pi_{kl}}$

Si *n* est grand, **l'intervalle de confiance pour la moyenne au niveau de confiance 1-** α est :

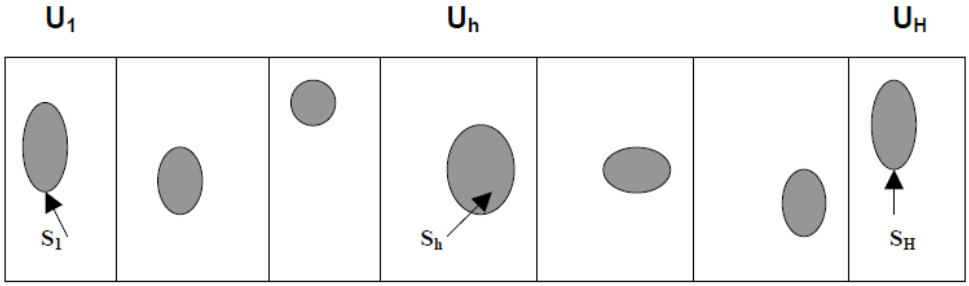
$$IC_{1-\alpha}(\mu_y) = \left[\hat{\mu}_{y\pi} - u_{1-\frac{\alpha}{2}}\sqrt{\hat{V}ar(\hat{\mu}_{y\pi})};\hat{\mu}_{y\pi} + u_{1-\frac{\alpha}{2}}\sqrt{\hat{V}ar(\hat{\mu}_{y\pi})}\right]$$

où $u_{1-\frac{\alpha}{2}}$ désigne le fractile d'ordre 1- α/2 de la loi N(0,1)

3. Plans stratifiés

Principe

Partitionner la population en sous-groupes homogènes appelés strates, et tirage d'autant d'échantillons indépendants qu'il y a de strates. H strates notées 1, …h, …H



Notations

1. *Dans la population*

- $U=\bigcup_{h=1}^{H}U_h$ et $N=\sum_{h=1}^{H}N_h$
- Total : $t_y=\sum_{h=1}^{H}t_{yh}=\sum_{h=1}^{H}N_h\bar{y}_h$
- Moyenne: $\bar{y}=\frac{t_y}{N}=\sum_{h=1}^{H}\frac{N_h}{N}\bar{y}_h$ avec $\bar{y}_h=\frac{1}{N_h}\sum_{k\in U_h}y_k$
- Variance : $\sigma_y^2=\frac{1}{N}\sum_{k\in U}(y_k-\bar{y})^2=\sum_{h=1}^{H}\frac{N_h}{N}\sigma_{yh}^2+\sum_{h=1}^{H}\frac{N_h}{N}(\bar{y}_h-\bar{y})^2=\sigma_{y\text{intra}}^2+\sigma_{y\text{inter}}^2=\frac{N-1}{N}S_y^2$

  avec $\sigma_{yh}^2=\frac{1}{N_h}\sum_{k\in U_h}(y_k-\bar{y}_h)^2$

2. *Dans l'échantillon*

- $S=\bigcup_{h=1}^{H}S_h$ et $n=\sum_{h=1}^{H}n_h$
- Moyenne dans S_h : $\hat{\bar{y}}_h=\frac{1}{n_h}\sum_{k\in S_h}y_k$
- Dispersion dans S_h : $\hat{S}_{yh}^2=\frac{1}{n_h-1}\sum_{k\in S_h}(y_k-\hat{\bar{y}}_h)^2$

Formulaire

Paramètre d'intérêt / Statistique	Moyenne	Proportion	Total
Estimateur du paramètre d'intérêt	$\hat{\bar{y}}=\sum_{h=1}^{H}\frac{N_h}{N}\hat{\bar{y}}_h$	$\hat{p}=\sum_{h=1}^{H}\frac{N_h}{N}\hat{p}_h$	$\hat{t}_y=N\hat{\bar{y}}=\sum_{h=1}^{H}\hat{t}_{yh}=\sum_{h=1}^{H}N_h\hat{\bar{y}}_h$
Vraie variance d'échantillonnage de cet estimateur	$Var\left[\hat{\bar{y}}\right]=Var\left[\sum_{h=1}^{H}\frac{N_h}{N}\hat{\bar{y}}_h\right]=\sum_{h=1}^{H}Var\left[\frac{N_h}{N}\hat{\bar{y}}_h\right]$ Si plan simple dans chaque strate : $Var\left[\hat{\bar{y}}\right]=\sum_{h=1}^{H}\left(\frac{N_h}{N}\right)^2\left(1-\frac{n_h}{N_h}\right)\frac{S_{yh}^2}{n_h}$	Si plan simple dans chaque strate $Var\left[\hat{p}\right]=\sum_{h=1}^{H}\left(\frac{N_h}{N}\right)^2\left(1-\frac{n_h}{N_h}\right)\frac{N_h}{N_h-1}\frac{p_h(1-p_h)}{n_h}$	$Var\left[\hat{t}_y\right]=Var\left[N\hat{\bar{y}}\right]=N^2Var\left[\hat{\bar{y}}\right]$ Si plan simple dans chaque strate : $Var\left[\hat{t}_y\right]=\sum_{h=1}^{H}N_h^2\left(1-\frac{n_h}{N_h}\right)\frac{S_{yh}^2}{n_h}$
Estimateur de la variance d'échantillonnage	Si plan simple dans chaque strate $\hat{Var}\left[\hat{\bar{y}}\right]=\sum_{h=1}^{H}\left(\frac{N_h}{N}\right)^2\left(1-\frac{n_h}{N_h}\right)\frac{\hat{S}_{yh}^2}{n_h}$	Si plan simple dans chaque strate $\hat{Var}\left[\hat{p}\right]=\sum_{h=1}^{H}\left(\frac{N_h}{N}\right)^2\left(1-\frac{n_h}{N_h}\right)\frac{\hat{p}_h(1-\hat{p}_h)}{n_h-1}$	Si plan simple dans chaque strate $\hat{Var}\left[\hat{t}_y\right]=\sum_{h=1}^{H}N_h^2\left(1-\frac{n_h}{N_h}\right)\frac{\hat{S}_{yh}^2}{n_h}$

Intervalle au niveau de confiance 95% pour la moyenne :

$$IC_{95\%}(\bar{Y})=\left[\hat{\bar{y}}-1{,}96\sqrt{\hat{Var}(\hat{\bar{y}})},\ \hat{\bar{y}}+1{,}96\sqrt{\hat{Var}(\hat{\bar{y}})}\right]$$

sous hypothèse que *n* est grand $\frac{\hat{\bar{y}}-\bar{Y}}{\sqrt{Var(\hat{\bar{y}})}}\rightarrow N(0,1)$

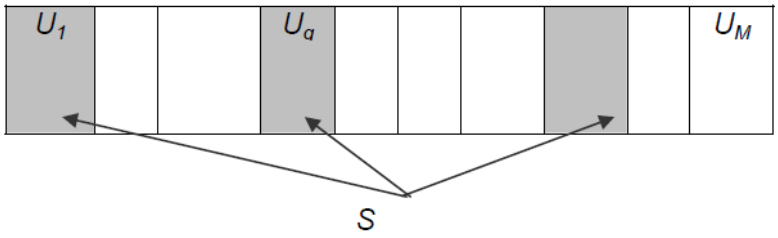
Choix des allocations

- Allocations proportionnelles : $\frac{n_h}{n}=\frac{N_h}{N}\quad\forall h\in\{1,\dots H\}$
- Allocations optimales de Neyman (sans contrainte de budget) : $n_h=n\frac{N_hS_{yh}}{\sum_{l=1}^{H}N_lS_{yl}}$
- Allocations optimales sous contrainte budgétaires : $n_h=C\frac{N_hS_{yh}}{\sqrt{C_h}\sum_{l=1}^{H}N_lS_{yl}\sqrt{C_l}}$

4. Plans par grappes

Principe

Partition de la population en sous-groupes appelés grappes, échantillonner un certain nombre de grappes, puis recenser tous les individus d'une grappe selectionnée



Notations

1. Dans la population U constituée de M grappes et N individus

- $U=\bigcup_{g=1}^{M} U_g$ et $N=\sum_{g=1}^{M} N_g$
- $t_y=\sum_{g=1}^{M} t_{yg}=\sum_{g=1}^{M} N_g \bar{y}_g$
- $\bar{y}=\frac{t_y}{N}=\sum_{g=1}^{M} \frac{N_g}{N}\bar{y}_g$ avec $\bar{y}_g=\frac{1}{N_g}\sum_{k\in U_g} y_k$
- $S_G^2=\frac{1}{M-1}\sum_{g=1}^{M}\left(t_{yg}-\frac{t_y}{M}\right)^2$

2. Dans l'échantillon S constitué de m grappes et n_s individus

- $S=\bigcup_{g\in S_G} U_g$ et $n_s=\sum_{g\in S_G} N_g$

Formulaire

Paramètre d'intérêt / Statistique	Total	Moyenne
Estimateur du paramètre d'intérêt	$\hat{t}_y = \frac{M}{m} \sum_{g \in S_G} t_{yg}$	$\hat{\bar{y}} = \frac{1}{N} \hat{t}_y = \frac{M}{Nm} \sum_{g \in S_G} N_g \bar{y}_g$
Vraie variance d'échantillonnage de cet estimateur	$Var[\hat{t}y] = M^2 \left(1 - \frac{m}{M}\right) \frac{1}{m} \frac{1}{M-1} \sum_{g=1}^{M} \left(t_{yg} - \frac{t_y}{M} \right)^2$	$Var[\hat{\bar{y}}] = \frac{1}{N^2} Var[\hat{t}y]$
Estimateur de la variance d'échantillonnage	$\hat{V}ar[\hat{t}y] = M^2 \left(1 - \frac{m}{M}\right) \frac{1}{m} \frac{1}{m-1} \sum_{g \in S_G} \left(t_{yg} - \frac{\hat{t}_y}{M} \right)^2$	$\hat{V}ar[\hat{\bar{y}}] = \frac{1}{N^2} \hat{V}ar[\hat{t}y]$

Intervalle au niveau de confiance 95% pour la moyenne :

$$IC_{95\%}(\bar{Y}) = \left[\hat{\bar{y}} - 1{,}96\sqrt{\hat{V}ar(\hat{\bar{y}})},\ \hat{\bar{y}} + 1{,}96\sqrt{\hat{V}ar(\hat{\bar{y}})} \right]$$

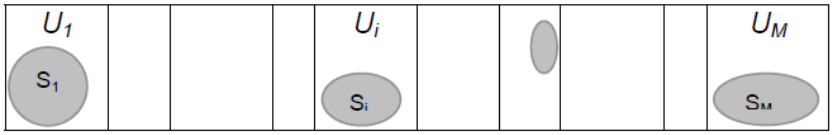
sous hypothèse que la taille de l'échantillon est assez grande.

5. Plans a deux degrés

Principe

Dans une population partitionnées en sous-groupes appelés unités primaires, eux-mêmes constitués d'unités secondaires :

1er degré : tirer un échantillon d'unités primaires
2ème degré : tirage d'unités secondaires dans chaque unité primaire retenue, de manière indépendante



Notations

1. *Dans la population U constituée de M unités primaires et N individus*

- $U=\bigcup_{i=1}^{M} U_i$ et $N=\sum_{i=1}^{M} N_i$
- $t_y=\sum_{i=1}^{M} t_{yi}=\sum_{i=1}^{M} N_i \bar{y}_i$
- $\bar{y}=\frac{t_y}{N}=\sum_{i=1}^{M} \frac{N_i}{N}\bar{y}_i$ avec $\bar{y}_i=\frac{1}{N_i}\sum_{k\in Ui} y_k$
- $S_I^2=\frac{1}{M-1}\sum_{i=1}^{M}\left(t_{yi}-\frac{t_y}{M}\right)^2$ et $S_i^2=\frac{1}{N_i-1}\sum_{k=1}^{N_i}\left(y_k-\bar{y}_i\right)^2$

2. *Dans l'échantillon S constitué de m unités primaires et n_s individus*

- $S=\bigcup_{i\in SUP} S_i$ et $n_s=\sum_{i\in Si} n_i$
- $\hat{s}_I^2=\frac{1}{m-1}\sum_{i\in SUP}\left(\hat{t}_{yi}-\frac{\hat{t}_y}{M}\right)^2$ et $\hat{s}_i^2=\frac{1}{n_i-1}\sum_{k\in Si}^{i}\left(y_k-\hat{\bar{y}}_i\right)^2$

Formulaire

Paramètre d'intérêt / Statistique	Total	Moyenne
Estimateur du paramètre d'intérêt	$\hat{t}_y = \sum_{k \in S} \frac{y_k}{\pi_k} = \frac{M}{m} \sum_{i \in SUP} \hat{t}_{yg} = \frac{M}{m} \sum_{i \in SUP} \frac{N_i}{n_i} \sum_{k \in Si} y_k$	$\hat{\bar{y}} = \frac{1}{N} \hat{t}_y = \frac{M}{Nm} \sum_{g \in SG} N_g \hat{\bar{y}}_g$
Vraie variance d'échantillonnage de cet estimateur	$Var[\hat{t}y] = M^2 \left(1 - \frac{m}{M}\right) \frac{S_t^2}{m} + \frac{M}{m} \sum_{i=1}^{M} N_i^2 \left(1 - \frac{n_i}{N_i}\right) \frac{S_i^2}{n_i}$	$Var[\hat{\bar{y}}] = \frac{1}{N^2} Var[\hat{t}y]$
Estimateur de la variance d'échantillonnage	$\hat{Var}[\hat{t}y] = M^2 \left(1 - \frac{m}{M}\right) \frac{\hat{S}_t^2}{m} + \frac{M}{m} \sum_{i \in SUP} N_i^2 \left(1 - \frac{n_i}{N_i}\right) \frac{\hat{S}_i^2}{n_i}$	$\hat{Var}[\hat{\bar{y}}] = \frac{1}{N^2} \hat{Var}[\hat{t}y]$

Intervalle au niveau de confiance 95% pour la moyenne :

$$IC_{95\%}(\bar{Y}) = \left[\hat{\bar{y}} - 1{,}96 \sqrt{\hat{Var}(\hat{\bar{y}})},\ \hat{\bar{y}} + 1{,}96 \sqrt{\hat{Var}(\hat{\bar{y}})} \right]$$

sous hypothèse que la taille de l'échantillon est assez grande

6. Redressements

Principe

Après la collecte, intégrer de l'information auxiliaire dans la construction de l'estimateur

Formulaire (estimateur Post Stratifié et par la régression vus en cours)

En notant X l'information auxiliaire,

Méthode	Estimateur du total	Vraie erreur quadratique moyenne de cet estimateur
Estimateur d'Horvitz-Thompson	$\hat{t}_{y_\pi}=\sum_{k\in S}\frac{y_k}{\pi_k}=\frac{N}{n}\sum_{k\in S}y_k$	$Var\left(\hat{t}_{y_\pi}\right)=N^2\left(1-\frac{n}{N}\right)\frac{S_y^2}{n}$
Estimateur par la différence	$\hat{t}_{yD}=\hat{t}_{y_\pi}+t_x-\hat{t}_{x\pi}$	$Var\left(\hat{t}_{y_D}\right)=N^2\left(1-\frac{n}{N}\right)\frac{\left(S_y^2+S_x^2-2S_{xy}\right)}{n}$
Estimateur par le ratio (ou le quotient)	$\hat{t}_{yQ}=\hat{t}_{y_\pi}\frac{t_x}{\hat{t}_{x\pi}}$	$Var\left(\hat{t}_{yQ}\right)=N^2\left(1-\frac{n}{N}\right)\frac{\left(S_y^2+R^2S_x^2-2RS_{xy}\right)}{n}$ avec $R=\frac{t_y}{t_x}=\frac{\bar{Y}}{\bar{X}}$
Estimateur par la régression	$\hat{t}_{yD}=\hat{t}_{y_\pi}+\hat{b}\left(t_x-\hat{t}_{x\pi}\right)$ avec $\hat{b}=\frac{\hat{s}_{xy}}{\hat{s}_x^2}$	$Var\left(\hat{t}_{yQ}\right)=N^2\left(1-\frac{n}{N}\right)\frac{S_y^2}{n}\left(1-\rho^2\right)$ avec $\rho=\frac{s_{xy}}{s_x s_y}$
Estimateur post-stratifié	$\hat{t}_{y\,post}=\sum_{h=1}^{H}N_h\hat{\bar{y}}_h=\sum_{h=1}^{H}\frac{N_h}{n_h}\sum_{k\in S_h}y_k$	$Var\left(\hat{t}_{y\,post}\right)=\frac{N-n}{n}\sum_{h=1}^{H}N_hS_{yh}^2+\frac{N-n}{N-1}\frac{N^2}{n^2}\sum_{h=1}^{H}\frac{N-N_h}{N}S_{yh}^2$

Estimateur par substitution de l'erreur quadratique moyenne

Intervalle au niveau de confiance 95% pour la moyenne :

$$IC_{95\%}\left(\bar{Y}\right)=\left[\hat{\bar{y}}-1{,}96\sqrt{\hat{Var}\left(\hat{\bar{y}}\right)},\ \hat{\bar{y}}+1{,}96\sqrt{\hat{Var}\left(\hat{\bar{y}}\right)}\right]$$

sous hypothèse que la taille de l'échantillon est assez grande.