LES DESSOUS DU SONDAGE INDIRECT

(ou **Les grandes idées de Jean-Claude Deville**)

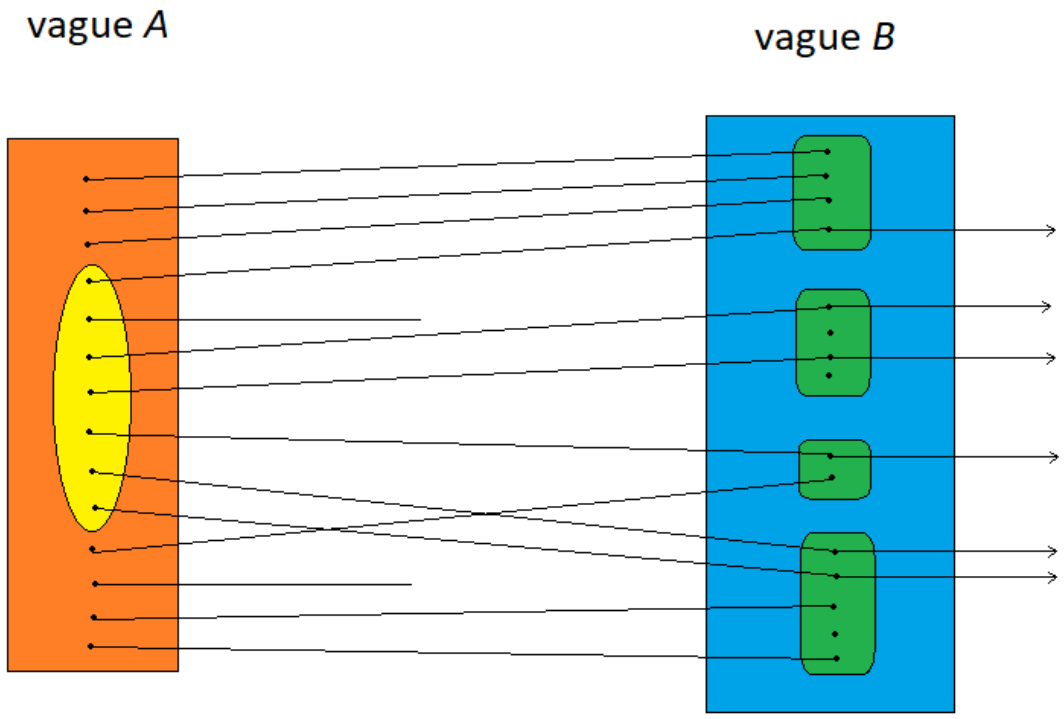
Pierre Lavallée

Colloque francophone sur les sondages
Lyon
Octobre 2018

CONTENU

1. LES TOUT DÉBUTS…

Au début : un problème de pondération relié aux enquêtes longitudinales…



<u>Question</u> : comment associer un poids (sans biais) aux individus des ménages enquêtés de la vague *B*?

Solution : le Partage des poids

Développé, notamment, par Huang (1984), Judkins et coll. (1984), Ernst, Hubble et Judkins (1984) et Ernst (1989)

Description :

Échantillon s^A contenant m^A individus tiré de la population de la vague A contenant M^A individus.

$\pi_k^A > 0$: Probabilité de sélection de l'individu k.

À la vague B, la population contient lors M^B individus répartis dans N^B ménages, où le ménage i contient M_i^B individus.

Processus de l'enquête longitudinale :

1. Pour chaque individu k de s^A, on établit la liste des M_i^B individus de la grappe i de la vague B contenant cet individu.

s^B : Ensemble des n^B ménages identifiées par les individus $k \in s^A$.

2. On enquête auprès de tous les individus k des grappes $i \in s^B$ pour mesurer la variable d'intérêt y.

La méthode du Partage des poids attribue un poids d'estimation w_{ik} à chaque individu k d'un ménage enquêté i.

Étapes du Partage des poids :

Étape 1 : Pour chaque individu k des ménages i de s^B, on calcule le poids initial $w'_{ik} = \frac{t_k}{\pi_k^A}$,

où $t_k = 1$ si $k \in s^A$, et 0 sinon.

Étape 2 : Pour chaque ménage i de s^B, on obtient le nombre total d'individus M_i^{AB} du ménage i présents à la vague A (mais pas nécessairement contenu dans s^A).

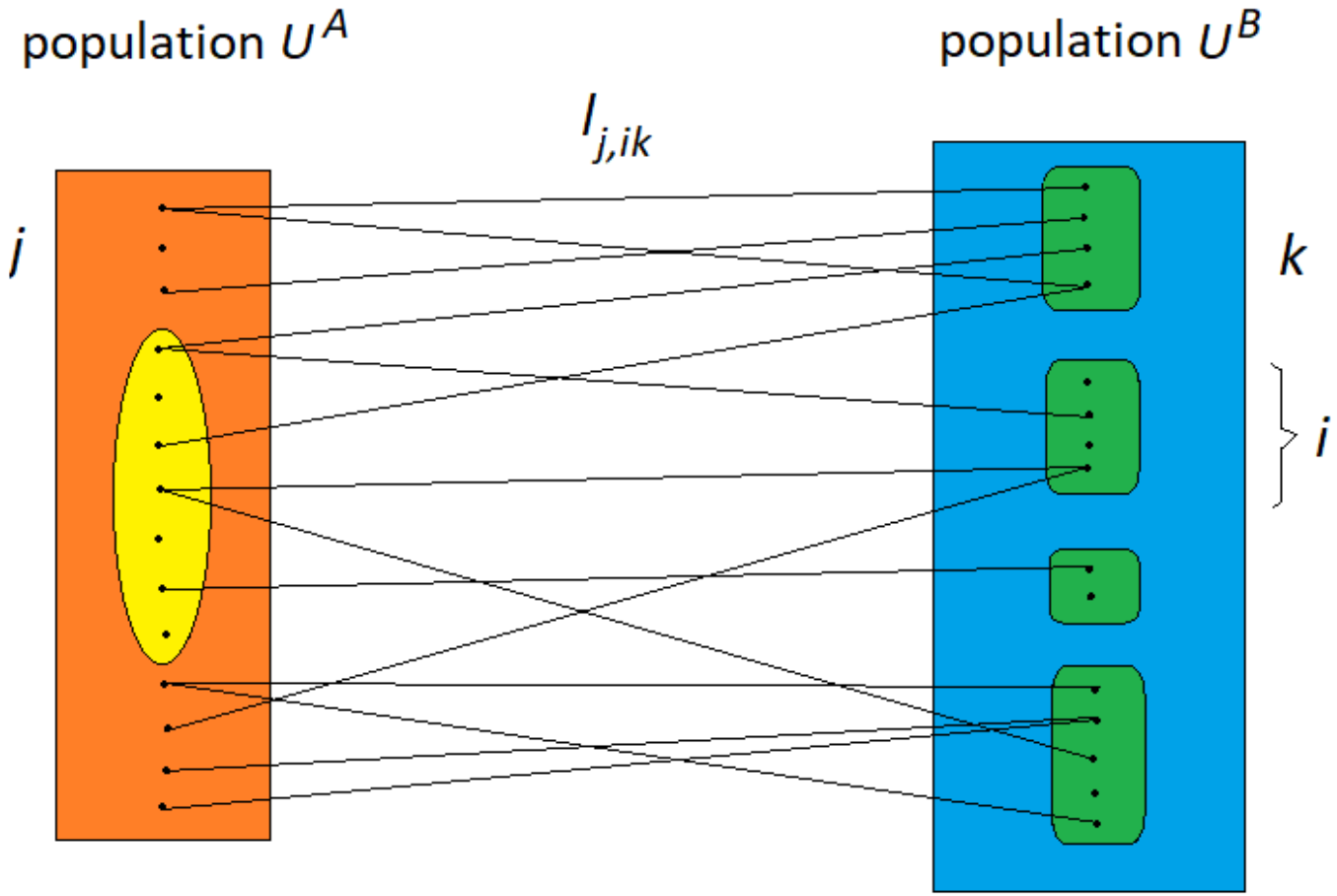
Étape 3 : On calcule le poids final $w_i = \frac{\sum_{k=1}^{M_i^B} w'_{ik}}{M_i^{AB}}$.

Étape 4 : Enfin, nous posons $w_{ik} = w_i$ pour tous les $k \in i$.

Idée de Jean-Claude :

Pourquoi ne pas généraliser les liens?

Au lieu de liens « un à un », pourquoi pas « plusieurs à plusieurs »?



Question remaniée : comment associer un poids (sans biais) aux unités de U^B enquêtées à la suite de la sélection d'unités dans U^A?

Nouveau problème étudié :

Deux populations U^A et U^B **reliées entre elles.**
On désire produire une estimation pour U^B (population cible).
Base de sondage disponible pour U^A seulement.

Solution :

Tirage d'un échantillon de U^A
afin de produire une estimation pour U^B
en se servant de la correspondance (liens)
existante entre les deux populations.

Terme proposé par Jean-Claude :

Sondage indirect

Processus d'enquête du sondage indirect :

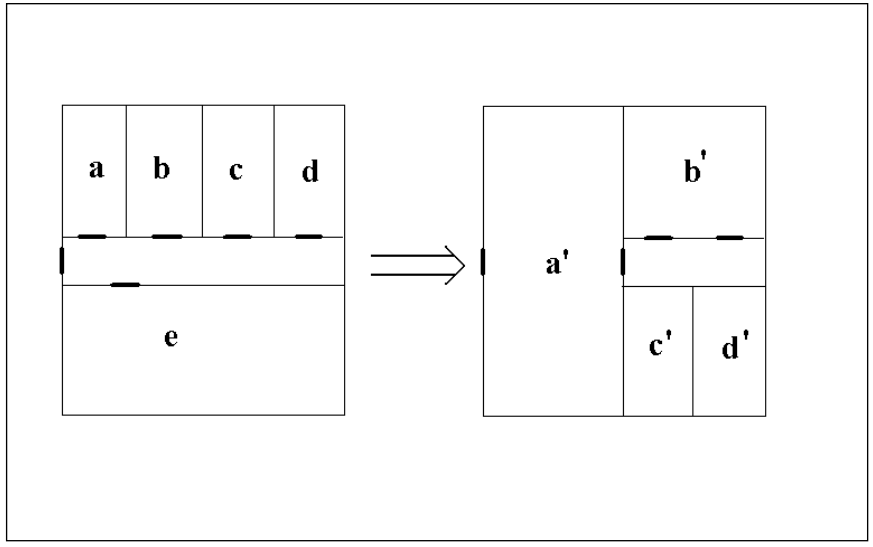
1. Pour chaque unité *j* de s^A, on identifie les unités *ik* de U^B qui ont un lien avec *j*.

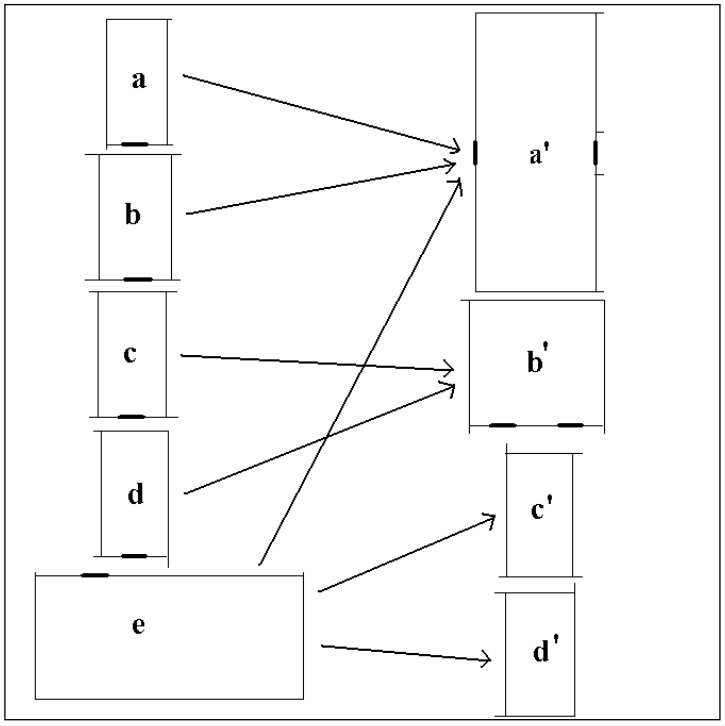
2. Pour chaque unité *ik* identifiée, on établit la liste des M_i^B unités de la grappe *i* contenant cette unité.

3. On enquête auprès de toutes les unités *k* des grappes $i \in s^B$ pour mesurer la variable d'intérêt *y*.

Exemple proposé par Jean-Claude:

Sondage auprès de logements





Estimation du total Y^B en se servant de s^A tiré de U^A :

Défi de taille si les liens entre les unités de U^A et U^B ne sont pas bijectifs.

Difficulté d'associer une probabilité de sélection, ou un poids d'estimation, aux unités enquêtées dans U^B.

Solution :

Méthode généralisée du partage des poids (MGPP)

Permet d'obtenir un poids d'estimation pour chaque unité enquêtée de la population cible U^B.

Description de la MGPP :

Échantillon s^A contenant m^A unités tiré de U^A contenant M^A unités.

$\pi_j^A > 0$: Probabilité de sélection de l'unité j.

Population cible U^B contient M^B unités.

U^B divisée en N^B grappes, où la grappe i contient M_i^B unités.

Liens (ou correspondance) entre les unités j de U^A et les unités k des grappes i de U^B.

Lien est identifié par $l_{i,jk}$, où $l_{j,ik} = 1$ s'il existe un lien entre l'unité $j \in U^A$ et l'unité $ik \in U^B$, et 0 sinon.

Étapes de la MGPP :

Étape 1 : Pour chaque unité k des grappes i de s^B, on calcule le poids initial $w'_{ik} = \sum_{j=1}^{M^A} l_{j,ik} \frac{t_j}{\pi_j^A}$,

où $t_j = 1$ si $j \in s^A$, et 0 sinon.

Étape 2 : Pour chaque unité k des grappes i de s^B, on obtient le nombre total de liens $L_{ik}^B = \sum_{j=1}^{M^A} l_{j,ik}$.

Étape 3 : On calcule le poids final $w_i = \frac{\sum_{k=1}^{M_i^B} w'_{ik}}{\sum_{k=1}^{M_i^B} L_{ik}^B}$.

Étape 4 : Enfin, nous posons $w_{ik} = w_i$ pour tous les $k \in U_i^B$.

Utilité de la MGPP :

Solution simple à des problèmes de sondage et de pondération complexes.

En général, donne les mêmes résultats que la théorie classique pour les problèmes simples.

Solution intéressante, même si la MGPP n'est pas toujours la plus précise (variance minimale) par rapport à une autre méthode d'estimation plus complexe.

1995…

Je demande à Jean-Claude réduire le rythme pour que je puisse faire ma thèse de doctorat sur le sondage indirect…

> *(Demande insistante, mais toujours amicale, de **Jean-Jacques Droesbeke** pour que je fasse un doctorat à l'Université libre de Bruxelles…)*

Le travail avance lentement, très lentement, sujet par sujet…

2. PROPRIÉTÉS

Théorème 1 : Dualité de la forme de $\hat{Y}^B$ par rapport à U^A et U^B

L'estimateur $\hat{Y}^B$ *peut s'écrire sous les deux formes :*

$$\hat{Y}^B = \sum_{i=1}^{n^B} \sum_{k=1}^{M_i^B} w_{ik} y_{ik} \qquad \textit{(poids de la MGPP)}$$

ET

$$\hat{Y}^B = \sum_{j=1}^{M^A} \frac{t_j}{\pi_j^A} Z_j$$

$$\text{où } Z_j = \sum_{i=1}^{N^B} \sum_{k=1}^{M_i^B} l_{j,ik} \frac{Y_i}{L_i^B}$$

On a donc un simple estimateur de Horwitz-Thompson.

Corollaire 1 : Biais de $\hat{Y}^B$

L'estimateur $\hat{Y}^B$ est sans biais pour l'estimation de Y^B, par rapport au plan de sondage.

*(Démonstration discutée avec **Carl Särndal**…)*

Corollaire 2 : Variance de $\hat{Y}^B$

La formule de la variance de l'estimateur $\hat{Y}^B$, par rapport au plan de sondage, est donnée par

$$Var(\hat{Y}^B) = \sum_{j=1}^{M^A} \sum_{j'=1}^{M^A} \frac{(\pi_{jj'}^A - \pi_j^A \pi_{j'}^A)}{\pi_j^A \pi_{j'}^A} Z_j Z_{j'}$$

3. CALAGE SUR MARGES

On désire corriger les poids de la MGPP pour que les estimations produites correspondent à des totaux connus (information auxiliaire).

Deux sources possibles d'information auxiliaire :

U^A :

- Vecteur colonne $\mathbf{x}_j^A$.
- Total $\mathbf{X}^A = \sum_{j=1}^{M^A} \mathbf{x}_j^A$ supposé connu.

U^B :

- Vecteur colonne $\mathbf{x}_{ik}^B$.
- Total $\mathbf{X}^B = \sum_{i=1}^{N} \sum_{k=1}^{M_i^B} \mathbf{x}_{ik}^B$ supposé connu.

Contraintes de calage sur marges associées à la MGPP :

$$\hat{\mathbf{X}}^{CAL,A} = \sum_{j=1}^{m^A} w_j^{CAL,A} \mathbf{x}_j^A = \mathbf{X}^A$$

ET $$\hat{\mathbf{X}}^{CAL,B} = \sum_{i=1}^{n} \sum_{k=1}^{M_i^B} w_{ik}^{CAL,B} \mathbf{x}_{ik}^B = \mathbf{X}^B$$

$w_j^{CAL,A}$: Poids de calage obtenu à partir des $d_j^A = 1/\pi_j^A$.

$w_{ik}^{CAL,B}$: Poids de calage de l'unité *k* de la grappe enquêtée *i* où on a appliqué la MGPP.

À partir du théorème 1, on peut réécrire la contrainte sous la forme :

$$\hat{\mathbf{X}}^{CAL,B} = \sum_{j=1}^{m^A} w_j^{CAL,A} \boldsymbol{\Gamma}_j = \mathbf{X}^B$$

où $\boldsymbol{\Gamma}_j = \sum_{i=1}^{N^B} \sum_{k=1}^{M_i^B} l_{j,ik} \frac{X_i^B}{L_i^B}$

$\Rightarrow$ Contrainte exprimée en fonction des unités $j \in s^A$.

Soient $\mathbf{x}_j^{AB} = \begin{bmatrix} \mathbf{x}_j^A \\ \boldsymbol{\Gamma}_j \end{bmatrix}$ et $\mathbf{X}^{AB} = \begin{bmatrix} \mathbf{X}^A \\ \mathbf{X}^B \end{bmatrix}$.

Contrainte unique englobant U^A et U^B:

$$\hat{\mathbf{X}}^{CAL,AB} = \sum_{j=1}^{m^A} w_j^{CAL,A} \mathbf{x}_j^{AB} = \mathbf{X}^{AB}$$

Formulation de la détermination de l'estimateur

$\hat{Y}^{CAL,B} = \sum_{j=1}^{m^A} w_j^{CAL,A} Z_j$ associé à la MGPP :

Déterminer $w_j^{CAL,A}$, pour $j=1,...,m^A$, afin de minimiser

$$\sum_{j=1}^{m^A} G_j(w_j^{CAL,A}, d_j^A)$$

sous la contrainte unique $\hat{\mathbf{X}}^{CAL,AB} = \sum_{j=1}^{m^A} w_j^{CAL,A} \mathbf{x}_j^{AB} = \mathbf{X}^{AB}$

Correspond à la formulation du calage de Deville et Särndal (1992).

Travail décisif pour mon doctorat...

Mais après tout ce travail, découverte de l'article de ***Jean-Claude*** *(1998) :*

« Les enquêtes par panel : en quoi diffèrent-elles des autres enquêtes ? suivi de : comment attraper une population en se servant d'une autre »

où on retrouve la solution du calage sur marges associé à la MGPP…

4. OPTIMISATION DES LIENS

Utilisation de liens pondérés :

Variable indicatrice $l_{j,ik}$:

- Indique s'il y a un lien ou non entre les unités *j* et *ik* des populations U^A et U^B.
- N'indique pas l'importance relative que pourraient avoir certains liens par rapport à d'autres.

Possible de remplacer $l_{j,ik}$ par une variable quantitative $\theta_{j,ik}$ représentant l'importance qu'on veut donner au lien $l_{j,ik}$.

$\theta_{j,ik}$ définie sur $[0,+\infty$

$\theta_{j,ik} = 0$ équivaut à $l_{j,ik} = 0$

Si le processus d'assignation des valeurs de $\theta_{j,ik}$ est indépendant du tirage de s^A, la MGPP reste sans biais.

Nouvel estimateur (sans bias) : $\hat{Y}_\theta^B$

où les liens $l_{j,ik}$ sont remplacés par $\theta_{j,ik}$

Problème : déterminer des valeurs optimales de $\theta_{j,ik}$ de manière à minimiser la variance de $\hat{Y}_\theta^B$.

Utilisation des liens pondérés optimaux :

Deville et Lavallée (2006) : valeurs de $\theta_{j,ik}$ telles que la variance de l'estimateur $\hat{Y}_{\theta}^{B}$ soit (presque) minimale.

La solution optimale n'est pas simple à écrire, et elle dépend souvent de la variable d'intérêt *y*.

Idée de Jean-Claude :

Optimalité faible

Minimiser la variance de $\hat{Y}_{\theta}^{B}$ pour un choix très précis d'une variable d'intérêt : $y_{ik} = 1$ pour une unité ik de U^B et $y_{i'k'} = 0$ pour toutes les autres unités de U^B.

Les liens pondérés faiblement optimaux résultants ne font pas intervenir la valeur de y.

Relativement facile à calculer.

Autre idée de Jean-Claude :

Optimalité forte indépendante de y

Obtention d'un critère pour vérifier si l'optimalité faible correspond à l'optimalité forte (variance minimale de $\hat{Y}_\theta^B$).

Si c'est le cas, l'optimalité forte ne dépend pas de la variable d'intérêt *y*!

5. CONCLUSION

Sondage indirect (et MGPP) :

> Solution viable pour des problèmes de sondage (de grappes) complexes.

Thomas Edison : « *Le génie est fait d'un pour cent d'inspiration et de quatre-vingt-dix-neuf pour cent de transpiration.* »

Si Jean-Claude a été bien inspiré, j'ai beaucoup transpiré!

Résultats : une thèse, deux livres, plusieurs articles, solution utilisée par plusieurs statisticiens…

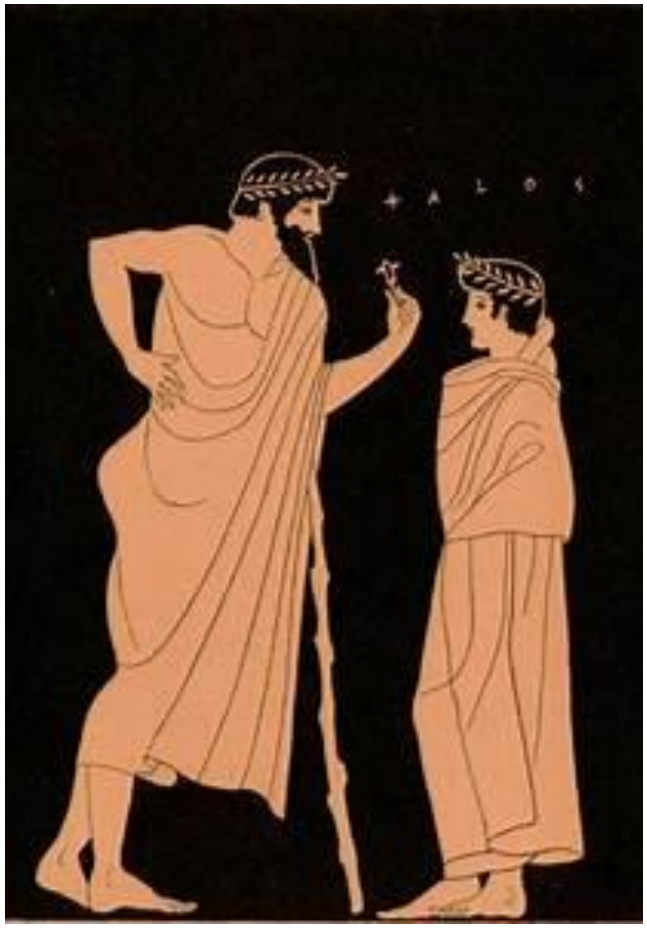
Merci Jean-Claude pour toutes ces idées!

Si Jean-Claude est…

Ulysse

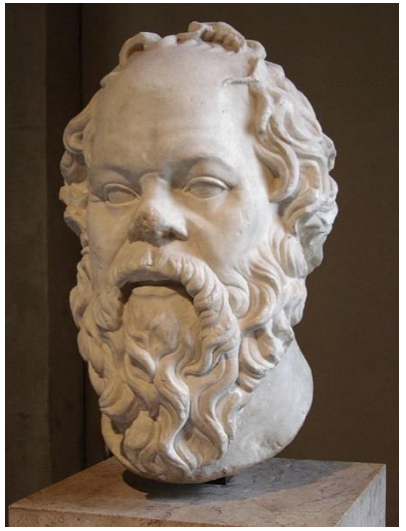
Alors, je ne suis que…

Télémaque



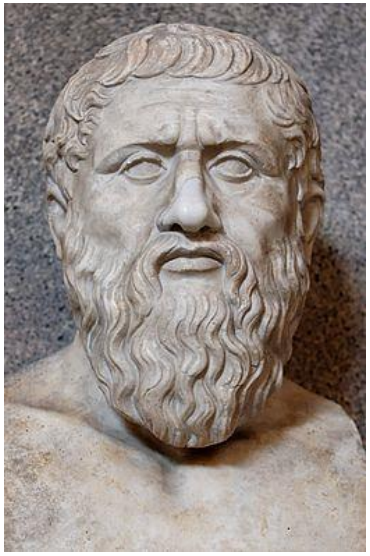
Si Jean-Claude est…

Socrate



Alors, je ne suis que…

Platon



Si Jean-Claude est…

Alexandre Le Grand

Alors, je ne suis que…

Ptolémée

Si Jean-Claude est…

Don Quichotte

Alors, je ne suis que…

Sancho Panza

Si Jean-Claude est…

Robinson Crusoé

Alors, je ne suis que…

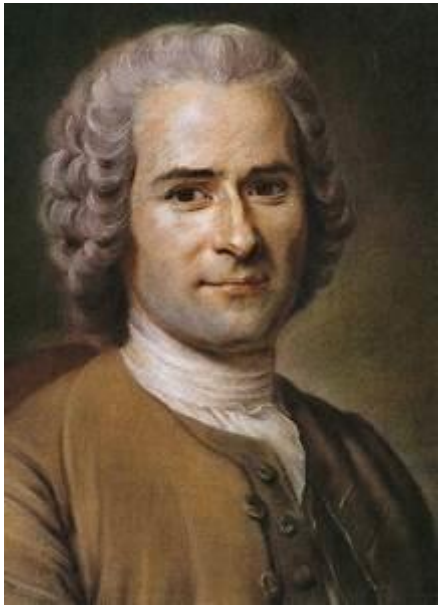
Vendredi

Si Jean-Claude est…

Voltaire

Alors, je ne suis que…

Rousseau



Si Jean-Claude est…

Sherlock Holmes

Alors, je ne suis que…

le Docteur Watson

Si Jean-Claude est…

Astérix

Alors, je ne suis que…

Obélix

Si Jean-Claude est…

Gaston Lagaffe

Alors, je ne suis que…

Bertrand Labévue

Si Jean-Claude est…

Rocky Balboa

Alors, je ne suis que…

Apollo Creed

Si Jean-Claude est…

Kasparov

Alors, je ne suis que…

Karpov

Finalement, si Jean-Claude est…

Jean-Claude Deville

Alors, je ne suis et ne resterai que…

Pierre Lavallée