Stratification des Populations Asymétriques

M.A. Hidiroglou
Statistique Canada
12 Octobre, 2016

Notes pour Pierre

- Je tiens premièrement à remercier le comité du programme du 9 ieme colloque sur le sondage de m'avoir invité à cette session en l'honneur de Pierre Lavallée.

  Je débute en parlant un peu de ma collaboration avec mon ami Pierre durant les 31 dernières années. Je connais Pierre depuis 1985, l'année ou j'ai participé avec Marie-France Germain au recrutement universitaire pour le groupe de la méthodologie de Statistique Canada. Nous l'avons recruté à l'université de Carleton si je me rappelle bien. Nous avions été très impressionnés par ses connaissances du sondage et l'aise avec laquelle il avait répondu à toutes nos questions. Ce fut donc un bon début.
- En 1986, j'avais publié un article sur la stratification d'une population en deux strates : une à tirage complet et l'autre à tirage partiel. Sachant que le sujet de maitrise de Pierre à l'université de Carleton était la stratification, je lui ai demandé si on pouvait développer un algorithme résultant aux meilleures bornes possible pour un plan de sondage avec une strate à tirage complet et plusieurs à tirage partiel. Il dériva la solution en 4 pages le 20 février 1986. Pierre présenta la méthode au JSM 1987, et elle apparaissait en 1988 dans techniques d'enquêtes. Ceci fut le début de notre collaboration : nous avons collaboré à l'écriture d'environ 10 articles ensemble.

Notes pour Pierre

- Ceci dit, Pierre et moi avons passé une grande partie de notre carrière à Statistique Canada sur les enquêtes entreprises. Nous avons travaillé ensemble sur pas mal de ces enquêtes : de 1989 à 1990 sur les enquêtes transport. Ce fut l'année que Pierre me disait qu'il me donnait un an de sa vie professionnelle. On était loin de se douter à l'époque qu'il m'en donnerait 6 autres entre 1997 et 2003. Ce fut un plaisir de travailler avec Pierre. Son côté pratique et fortes connaissances en sondage ont beaucoup facilité le développement de ces enquêtes. Pierre s'est souvent retrouvé consulté par le méthodologistes dans plusieurs domaines de la Division des Méthodes Enquêtes Entreprises. Et grâce à son expertise technique, il a relevé différents défis avec brio. D'où découle sa célèbre réponse à la question « Puis-je te déranger quelques minutes ? » « Tu ne me déranges jamais ! ». On ne peut passer sous silence la méthode du partage des poids qu'il a développée de même que la publication en 2002 de son premier roman intitulé Le sondage indirect ou la méthode du

  partage des poids.

- Nous avons aussi participé à un bon nombre de conférences francophones ensemble. Ceci débuta en 1991 en présentant à la première Journées de Méthodologie Statistique qui a eu lieu à Paris les 13 et 14 mars 1991. A cette occasion, le titre de la présentation de Pierre était « Plan de sondage pour la sélection de panels d'entreprises ». Cette première conférence donna suite à plusieurs d'autres (jusqu'en 2012) ou Statistique Canada présentait son travail en méthodologie. QQ photos INSEE ici.

INSEE 1991

Anne-Marie Dussaix et Pierre

INSEE 1991

Jean-Claude Deville, Pierre, et Michel Hidiroglou

INSEE 1991

Pascal Ardilly et Pierre

Notes pour Pierre

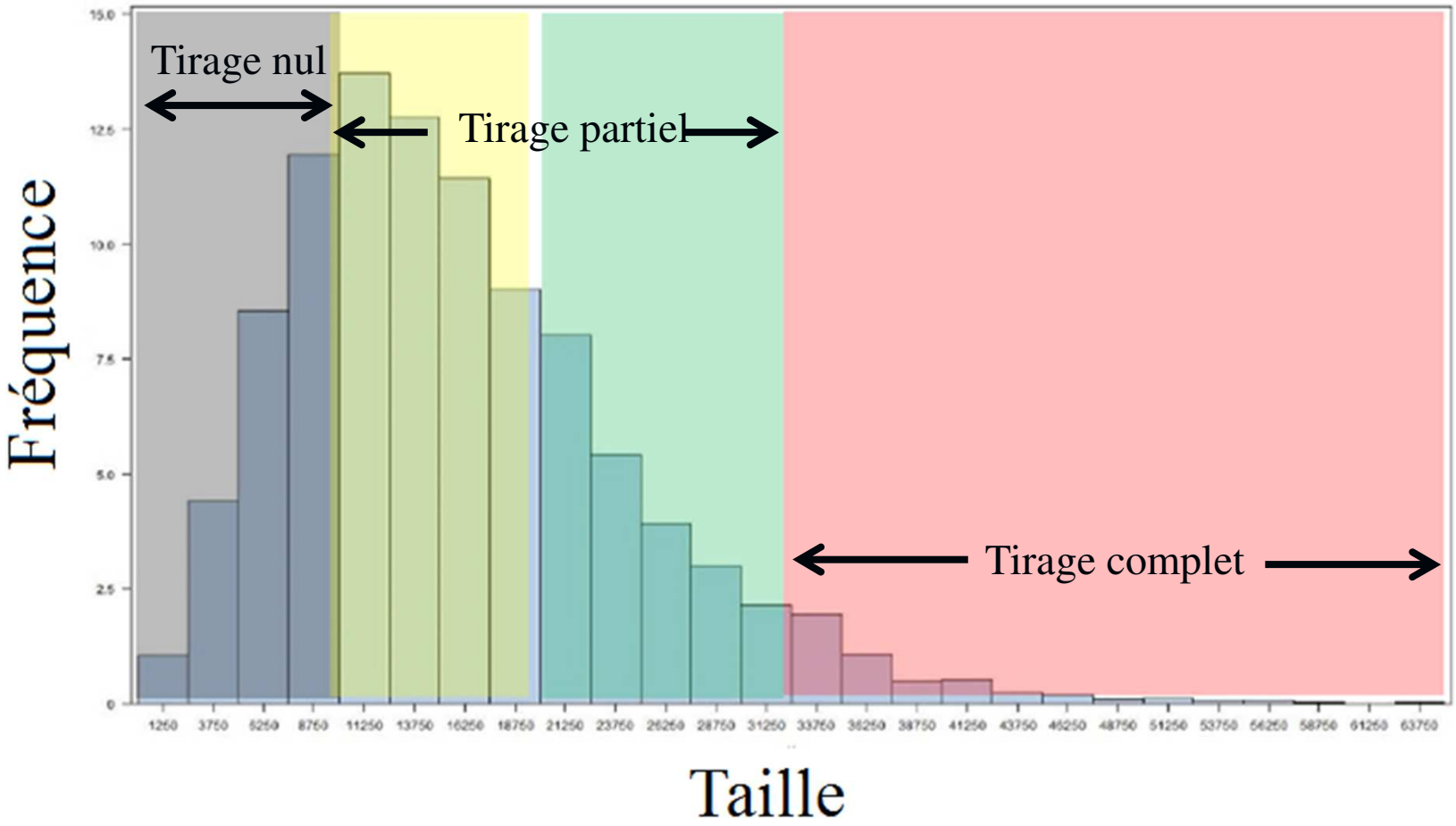
- Pierre a aussi beaucoup contribué aux Colloques sur les sondages qui ont débuté en 1997. Il l'a fait en de présentateur, de modérateur et plusieurs rôles du Comité d'organisation. Il est comme vous le savez le Président du Comité d'organisation de ce colloque. Il a aussi édité avec L.-P. Rivest l'ouvrage publié chez Dunod du quatrième Colloque qui a eu lieu à Québec du 25 au 27 mai 2005.

  Même si ma liste de ses accomplissements est loin d'être exhaustive, par manque de temps, il n'en demeure pas moins que nous devons exprimer notre reconnaissance à l'égard de Pierre pour sa carrière remarquable à Statistique Canada, sa fidélité aux amis, et sa générosité sincère et profonde.

1. Introduction

- La taille est critique pour un plan de sondage efficace
  - Répartition des variables d'intérêt est asymétrique
  - Peu de grandes unités représentent une grande partie du total
  - Stratifier en strates en fonction de la variable asymétrique (emploi, ventes): x_1 , x_2,…, x_N
  - Types de strate:
    - Strate à tirage complet
    - Strate (s)à tirage partiel
    - Strate à tirage nul

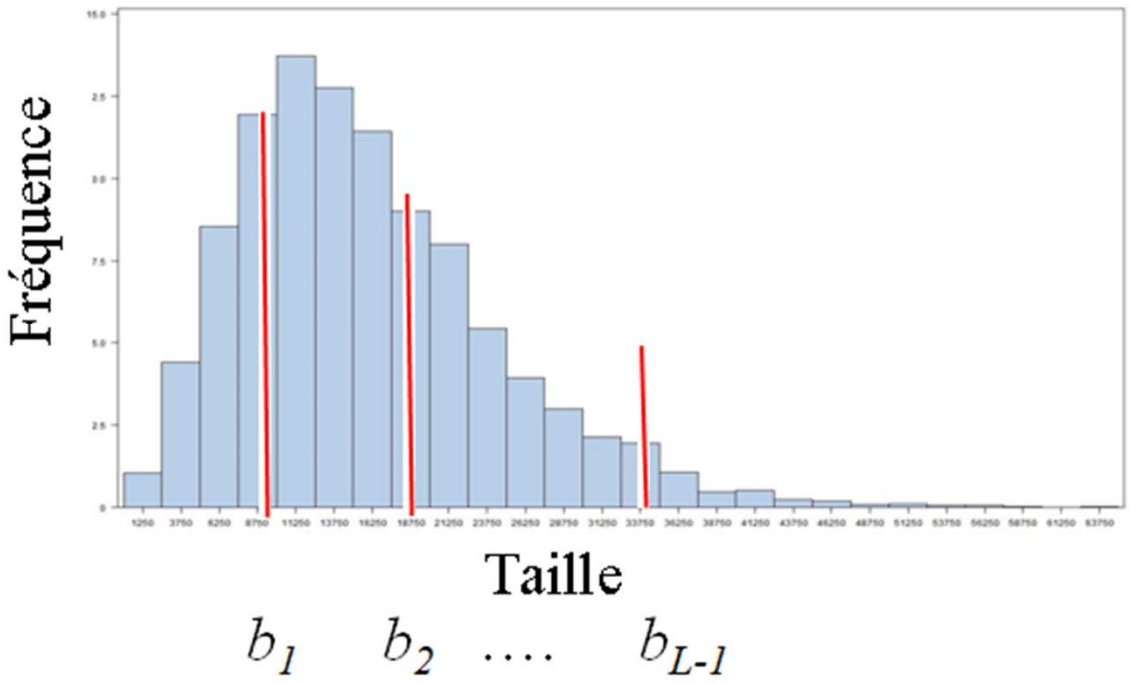
1. Introduction



2. Bornes pour la stratification

- Définir les bornes des strates

$$b_1, b_2, \ldots, b_{L-1} \quad (b_1 < b_2 < \ldots < b_{L-1})$$



2. Bornes pour la stratification

- Cumrootf: Dalenius et Hodges (1959)
  - Bornes basées sur l'allocation de Neyman
- Géométrique: Gunning et Horgan (2004)
  - Les bornes nécessitent que les coefficients de variation soient égaux dans chaque strate
- LH: Lavallée et Hidiroglou (1988)
  - Exige une strate à tirage complet
  - Les bornes dépendent de l'allocation des unités aux strates à tirage partiel, la taille de l'échantillon ou des exigences relative à la fiabilité

2. Bornes pour la stratification

Allocation $\mathring{A}_h = \dfrac{N_h^{2q_1}\,\bar{X}_h^{2q_2}\,S_h^{2q_3}}{\sum_{h=1}^{L} N_h^{2q_1}\,\bar{X}_h^{2q_2}\,S_h^{2q_3}} \quad h = 1, 2, \ldots, L$

ou $0 \le 2\,q_1 \le 1,\ 0 \le 2\,q_2 \le 1,$ **et** $0 \le 2\,q_3 \le 1$

Allocations	$\mathring{A}_h, h = 1, 2, \ldots, L$	$2q_1$	$2q_2$	$2q_3$
Neyman	$N_h\,S_h \Big/ \left(\sum_{h=1}^{L} N_h\,S_h\right)$	1	0	1
X-proportionnel	$X_h \Big/ \sum_{h=1}^{L} X_h$	1	1	0
N-proportionnel	$N_h \Big/ \sum_{h=1}^{L} N_h$	1	0	0

2. Bornes pour la stratification

- Étant donnée une allocation $\mathring{A}_h$
- Options afin d'allouer les unités aux strates à tirage partiel

Scenario-*n*

- Les bornes $b_1, b_2, \ldots, b_{L-1}$ sont calculées afin de minimiser

$$n = N W_L + \frac{N \sum_{h=1}^{L-1} W_h^2 S_h^2 / \mathring{A}_h}{N\left(c\,\bar{X}\right)^2 + \sum_{h=1}^{L-1} W_h S_h^2}$$

Scenario-*c*

- Les bornes $b_1, b_2, \ldots, b_{L-1}$ sont calculées afin de minimiser

$$V\left(\hat{X}\right) = \sum_{h=1}^{L} \frac{N_h^2}{n\,\mathring{A}_h} S_h^2 - \sum_{h=1}^{L} N_h S_h^2$$

2. Bornes pour la stratification

- **Deux possibilités pour la solution**
  - Méthode de Sethi (1962): Résoudre les équations récursives quadratiques de la forme

$$\alpha_h b_h^2 + \beta_h b_h + \gamma_h = 0, \quad 1 \le h \le L-1$$

    - Les $\alpha_h, \beta_h,$ et γ_h sont les termes provenant des dérivés des versions continues de n ou de $V\left(\hat{X}\right)$ avec b_h

  - Méthode de Kozak (2004): Recherche aléatoire ou on choisit les " $L-1$ " bornes des strates parmi les valeurs ordonnées des x, avec élimination des doublons.

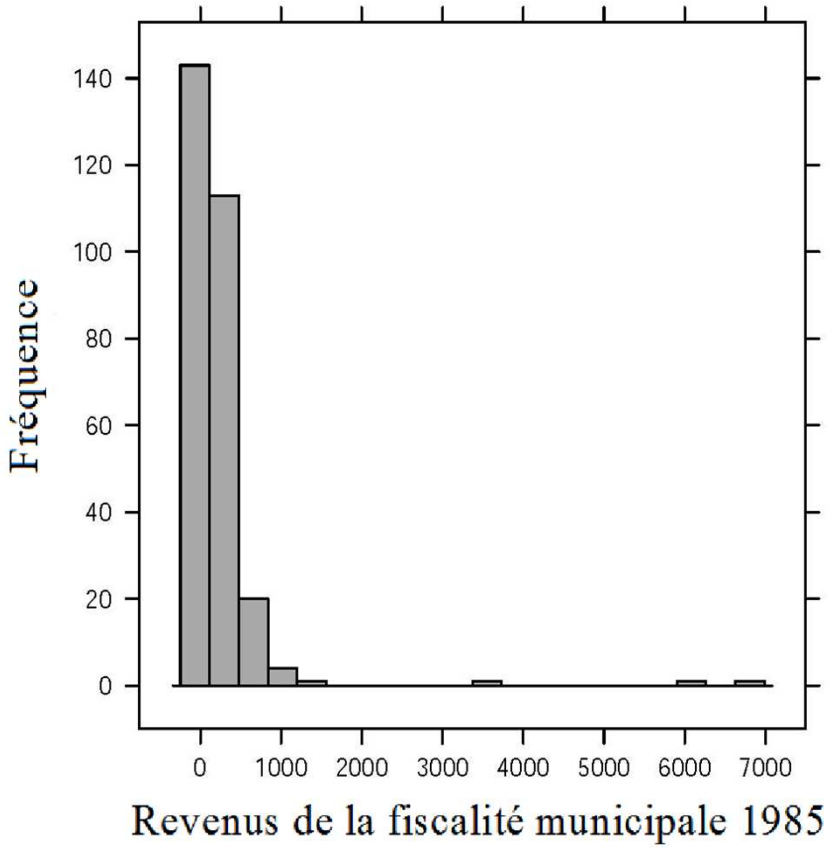
3. Propriétés des méthodes

- Cumrootf est invariant à la location
  - Ajouter une constante à chaque valeur: les frontières se déplaceront par cette constante.
- LH et géométrique ne sont pas invariants à la location
  - Le décalage est plus grand pour la méthode géométrique
- Cumrootf, LH et géométrique sont invariants à l'échelle (scale)
  - Multiplier chaque valeur par une constante: les frontières seront multipliées par cette constante

4. Population des Municipalités en Suède

- Ensemble de données
  - 11 variables qui décrivent 284 municipalités en Suède.
  - Avons choisi les revenus de la fiscalité municipale 1985 (en millions de couronnes).
  - Baillargeon et Rivest:

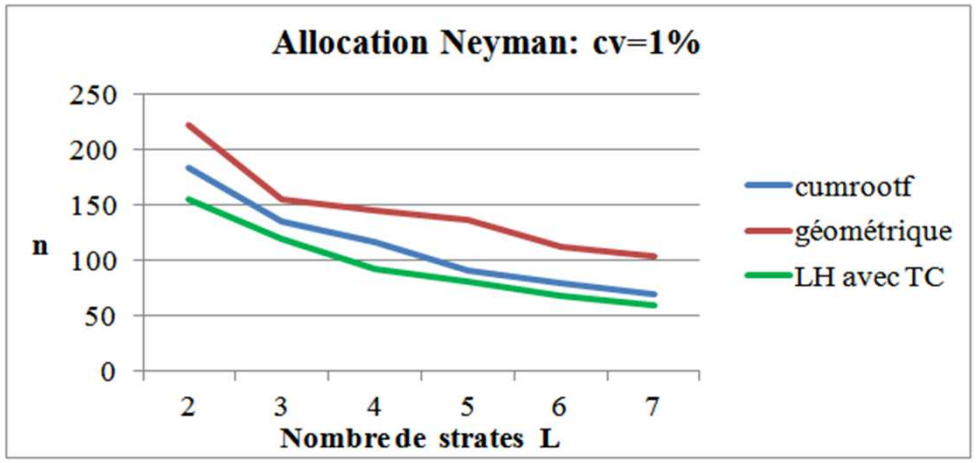
    CRAN package Univariate Stratification of Survey Populations (2014)

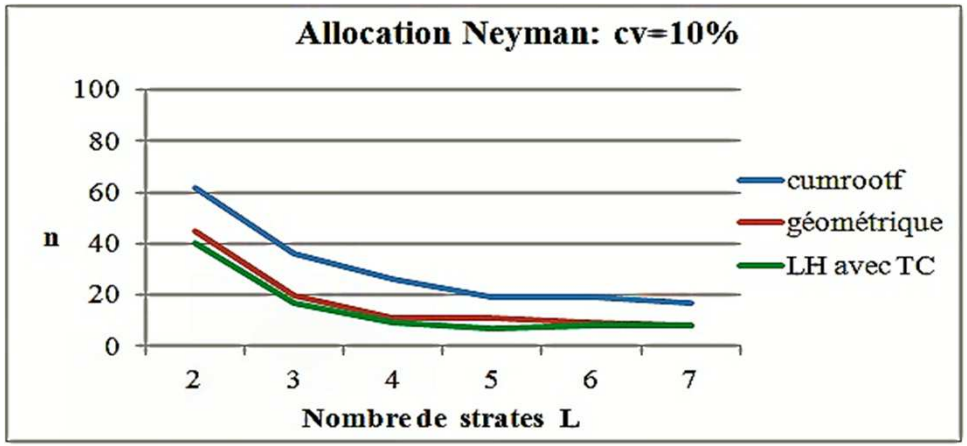


4. Population des Municipalités en Suède

Allocation Neyman

- Meilleure: LH avec strate à tirage complet (TC)
- Géométrique meilleure que cumrootf quand le cv=1%
- Inférence: Géométrique meilleure que cumrootf pour un seuil de CV?

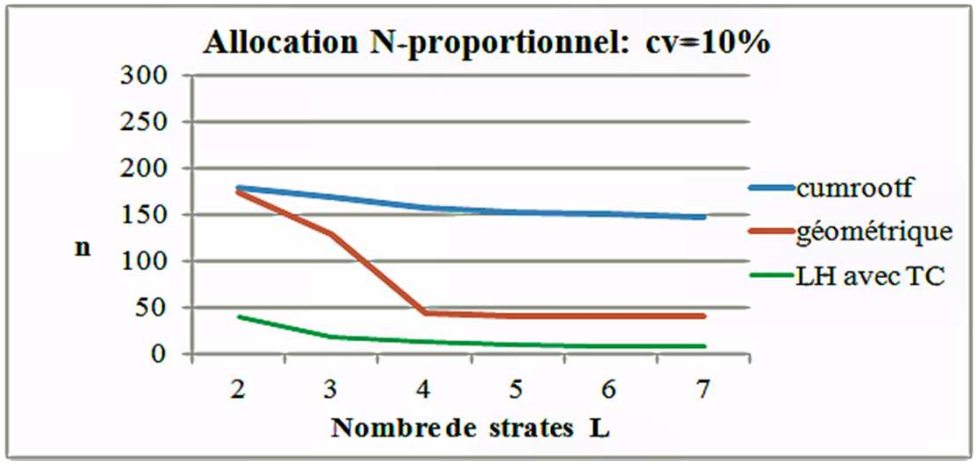


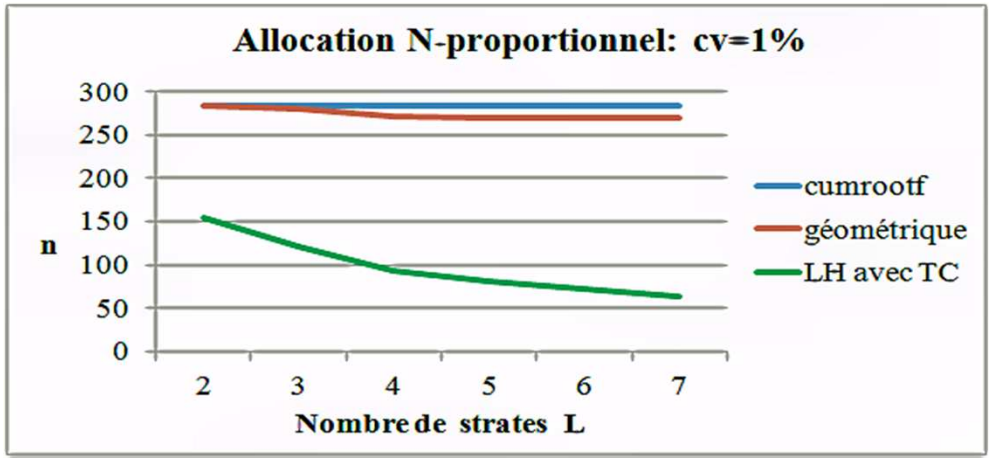


4. Population des Municipalités en Suède

Allocation *N*-Proportionnel

- Meilleure: LH avec strate à tirage complet (TC)
- LH ne souffre pas avec cette allocation: ajuste automatiquement les bornes
- Cumrootf paye un gros pris
- Géométrique meilleurs avec un cv élevé (10%) et plus de strates (3+)





5. Conclusions

- Solution complète pour le calcul des bornes stratification
  - Utilise soit l'algorithme de Kozak (2004) ou de Sethi (1963).
  - Les deux scénarios *c* et *n* sont équivalents
- Les méthodes basées sur des algorithmes d'optimisation résultent à une stratification plus efficace que celle basée sur de simples méthodes (géométrique/ cumrootf)
- L'application de ces algorithmes est facile grâce au programme *R* de Baillargeon et Rivest (2014)
  - Tient compte de la non-réponse
  - Incorpore des modèles entre la variable d'enquête *y* étant donné la variable de stratification *x*
  - Incorpore aussi la strate à tirage nul

Il ne faut pas oublier les amis

Paneforte