UNIVERSITÉ LAVAL

UNIVERSITY of INFORMATION TECHNOLOGY and MANAGEMENT in Rzeszow, POLAND

La mise en oeuvre d'algorithmes de stratification univariée

Louis-Paul Rivest avec Sophie Baillargeon & Marcin Kozak (University of Information Technology and Management, Rzeszow)

Sommaire

1- Développement du package `R stratification` pour la stratification univariée;

2- Impact de l'arrondissement des tailles d'échantillon;

3- Mise en oeuvre de l'algorithme de Kozak pour la stratification univariée;

4- Quelques exemples.

Stratification univariée

- La valeur X d'une variable associée à la variable d'intérêt Y est connue pour chaque unité de la population.
- La strate h se compose des unités pour lesquelles $b_{h-1} \leq X < b_h$;
- Une stratification est donnée par un ensemble de bornes $b_1, b_2, \ldots, b_{L-1}$
- Critère d'optimalité: minimiser le CV à n fixe ou vice-versa

Stratification univariée

- Méthodes de base: cumrootf (Dalenius & Hodges 1959), géométrique (Gunning & Horgan, 2004)
- Lavallée & Hidiroglou (1988) utilise un algorithme itératif pour déterminer la “meilleure stratification” pour la population à l’étude + strate à tirage complet
- Rivest (2002) tient compte du fait que $Y \neq X$ dans la détermination de la meilleure stratification
- Kozak (2004) suggère un algorithme de recherche aléatoire pour trouver la meilleure stratification

Stratification univariée

- Baillargeon & Rivest (2009):
  - strate à tirage nulle,
  - tient compte de la non-réponse anticipée
  - modèles pour $Y \neq X$
  - compare les algorithmes disponibles pour trouver la meilleure stratification
- Baillargeon & Rivest (2011) présentent le package `stratification` qui met en oeuvre toutes ces possibilités

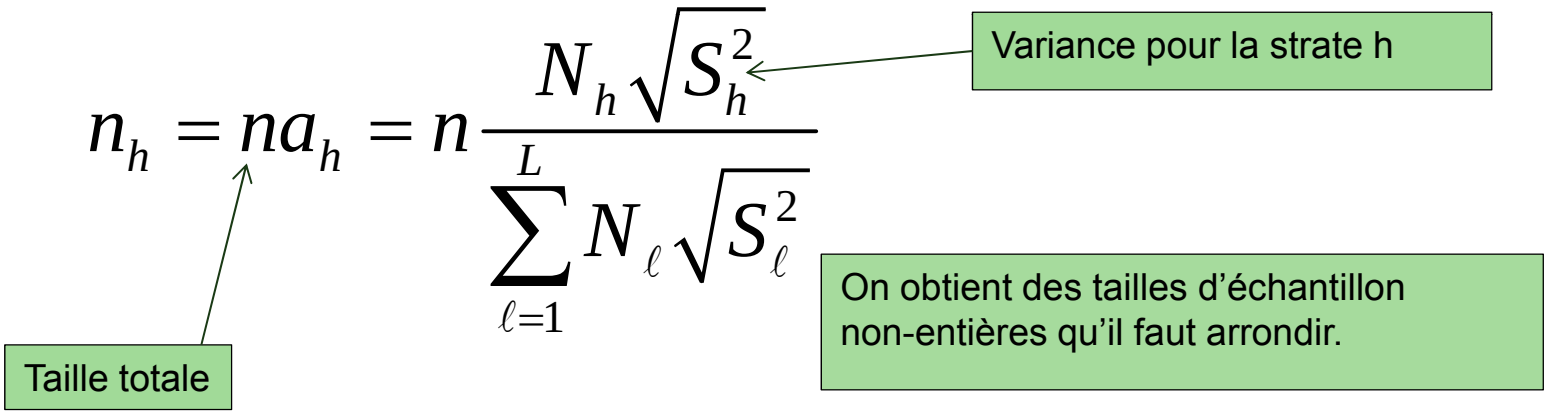
Techniques d'enquête, juin 2011
Vol. 37, N° 1, pp. 59-72
Statistique Canada, N° 12-001-X au catalogue

Élaboration de plans stratifiés en R à l'aide du programme *stratification*

Sophie Baillargeon et Louis-Paul Rivest [1]

Méthode d'arrondissement des tailles d'échantillon

- Calcul de tailles d'échantillon pour un plan stratifié selon l'allocation de Neyman



- D'autres règles d'allocation (proportionnelle, égale...) sont aussi utilisées

All. Neyman: Exemple L=4, n=40

Allocation de Neyman

h	1	2	3	4
S_h^2	5.8	10.6	12.9	39.7
N_h	4000	3000	2000	1000
n_h (n.e.)	11.72	11.88	8.74	7.66
n_h (e.)	12	12	9	7

Les tailles d'échantillon (12, 12, 9,7) obtenues en arrondissant ne sont pas optimales. Elles donnent une variance de 0.269676 alors que les tailles (11,12,9,8) donnent une variance de 0.269617.

(*la variance dans la strate 4 étant supérieure à celle de la strate 1 il est plus payant d'arrondir n_4 vers le haut)*

All. Neyman: Exemple H=4, n=40

La stratification proposée est-elle meilleure qu'une autre pour estimer la moyenne de y?

Pour évaluer la performance de cette stratification on peut utiliser trois variances "minimales":

i. la variance obtenue avec des tailles d'échantillon non entières (11.72, 11.88, 8.74,7.66): 0.269145

ii. la variance obtenue avec les tailles d'échantillon entières arrondies (12, 12, 9,7): 0.269677

iii. la variance minimale avec des tailles d'échantillon entières (11,12,9,8): 0.269617.

Stratification univariée: n fixe

On a

$$\mathrm{Var}(\bar{y}_{str})(b_1,\ldots,b_{L-1}) = \sum_{h=1}^{L}\left(\frac{N_h}{N}\right)^2\left(\frac{1}{na_h}-\frac{1}{N_h}\right)S_{yh}^2$$

On cherche les bornes b_h qui minimise la variance.

Peut représenter une taille d'échantillon entière ou réelle selon que l'on arrondisse ou non.

Stratification univariée: n fixe

Mise en oeuvre:

1-Sethi (1963) traite le cas N=∞; Lavallée & Hidiroglou (1988) estime cette solution optimale avec les données pour la population finie

2- Kozak (2004) traite le cas discret à l'aide d'un algorithme de recherche aléatoire des bornes optimales

Tailles d'échantillon discrètes dans l'algorithme?

Impossible car l'algorithme utilise des dérivées partielles par rapport aux b_h

Possible car l'algorithme essaie un très grand nombre de stratifications.

Stratification univariée: Kozak

L'algorithme choisit au hasard une des bornes b_h et crée une nouvelle stratification en la déplaçant.

Cette nouvelle stratification est conservée si elle est "meilleure" que la précédente.

La performance de l'algorithme dépend (i) des bornes initiales (ii) de la règle pour changer le b_h choisi (iii) du # essais infructueux après lesquels l'algorithme se termine (iv) du nombre de répétitions.

Stratification univariée: Kozak

Choix de la meilleure stratification sur la base de

les variances de $\overline{y}_{str}$ calculées à l'aide des tailles d'échantillon entières

ou bien

les variances de $\overline{y}_{str}$ calculées à l'aide des tailles d'échantillon non-entières

Est-ce que ça fait une différence?

Stratification univariée: Expérience

On considère 7 jeux de données de la littérature (Gunning & Horgan, 2004, TE et Keskinturk & Er 2007, CSDA), un nombre de strates L allant de 2 à 6, l'allocation de Neyman, proportionnelle ou égale et n=80, 100.

Il y a en tout 20 cas et on applique Kozak avec les 2 critères à chacun.

Dans tous les cas Y=X, la variable de stratification et la variable à l'étude sont les mêmes.

Résultats: n fixé

L'algorithme qui utilise les n_h entiers pour discriminer entre 2 stratifications possibles a permis une réduction de variance allant de 0 à 40%

Résultats

$(v^{(1)} - v^{(2)})/v^{(1)}$	0%	De 0% à 5%	Plus de 5%	Total
fréq	7	9	4	20

On obtient les réductions les plus importantes pour l'allocation proportionnelle.

Le choix du critère de comparaison avec n_h entiers ou non entiers a un impact sur le résultats.

Stratification univariée: CV fixe

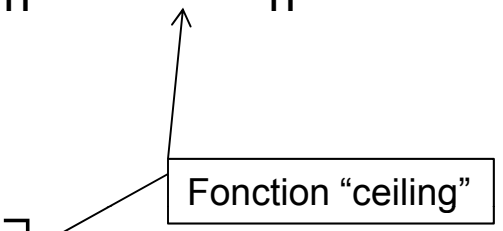
Etant donné un CV cible (et une règle d'allocation a_h), on cherche les bornes qui minimisent la taille d'échantillon requise

$$n^{(1)}(b_1,\ldots,b_{L-1}) = \frac{\sum_{h=1}^{L} N_h{}^2 S_{yh}^2 / a_h}{(CV)^2 T_y^2 + \sum_{h=1}^{L} N_h \, S_{yh}^2}$$

Comment comparer deux stratifications?

Stratification univariée: CV fixe

Méthode 1: sur la base du $n^{(1)}$ non entier. Une fois que l'algorithme a convergé on prend $n_h = \lceil n^{(1)} a_h \rceil$



Méthode 2: sur la base de $n^{(2)} = \Sigma \lceil n^{(1)} a_h \rceil$ ou de $n^{(1)}$ si les deux stratifications ont des $n^{(2)}$ égaux.

Est-ce que ça fait une différence?

Stratification univariée: Expérience 2

On considère 7 jeux de données de la littérature, un nombre de strates L allant de 2 à 6, l'allocation de Neyman et de puissance et des CV cibles allant de 1% à 10%.

Souvent Y=X; parfois Y≠X et les moments de Y sont prédits à l'aide d'un modèle.

Il y a en tout 49 cas.

Résultats 2: CV fixé

L'algorithme qui utilise les n_h entiers pour discriminer entre 2 stratifications a permis une réduction de n allant de 0 à 2 pour des n allant de 7 à 700.

$n^{(1)} - n^{(2)}$	0	1	2	total
fréq	20	27	2	49

L'impact sur le résultat est faible, plus faible que pour la première expérience. La réduction moyenne est de 2%.

Conclusion

Pour trouver une stratification optimale il faut utiliser les valeurs entières des tailles d'échantillon dans l'algorithme de stratification.

Exemple 1: n=80; Y=P75 de SUÈDE

La base de données SUÈDE contient N=284 municipalités. On veut la diviser en 4 strates sur la base de la variable P75. On optimise la stratification pour estimer la moyenne de P75 avec une allocation proportionnelle:

Algo avec n_h entiers

```
strata.LH(Sweden$P75,n=80,Ls=4,
alloc=c(0.5,0,0))

Strata information:
          |  bh     Nh nh   fh
stratum 1 |  31.5   218 61 0.28
stratum 2 |  76.5   51  14 0.27
stratum 3 | 346.5   13   4 0.31
stratum 4 | 672.0    2   1 0.50
Total               284 80 0.28
Anticipated CV: 0.043
```

Algo avec n_h non-entiers

```
strata.LH(Sweden$P75,n=80,Ls=4,
alloc=c(0.5,0,0)
algo.control=list(idopti = "nhnonint"))
Strata information:
          |  bh     Nh nh   fh
stratum 1 |  31.5   218 61 0.28
stratum 2 |  81.5   52  15 0.29
stratum 3 | 346.5   12   3 0.25
stratum 4 | 672.0    2   1 0.50
Total               284 80 0.28
Anticipated CV: 0.045
```

L'algo avec des n_h non-entiers ne trouve pas la borne optimale entre les deux premières strates ce qui augmente le CV de 8%.

Exemple 2: CV=5%; X=REV84, Y=RMT85

On veut stratifier les 284 municipalités de SUÈDE à l'aide de la variable X=REV84. L'objectif est d'estimer la moyenne de la variable Y=RMT85. Les valeurs de Y pour les 284 municipalités sont inconnues cependant on dispose d'un modèle liant Y à X:

$$\log Y = c + 1.058 \times \log X + \varepsilon.$$

$\varepsilon \sim N(0, 0.257^2)$

Les instructions (allocation de Neyman)

```
plan=strata.LH(Sweden$REV84,CV=.05,Ls=5,alloc=c(0.5,0,0.5), model =
"loglinear",model.control=(list(beta = 1.058355, sig2 = 0.25677^2))
```

comparent les stratifications possibles en utilisant les moments anticipés de Y sachant X.

Exemple 2: CV=5%; X=REV84, Y=RMT85

Plus la variance résiduelle σ^2 est grande, plus la variance anticipée de Y dans les strates est grande. Pour un CV fixe, n croît avec σ^2.

Algo avec n_h entiers

```
Strata information:
           |  bh      Nh nh   fh
stratum 1 | 1683    130  6 0.05
stratum 2 | 3174     76  6 0.08
stratum 3 | 5824     41  6 0.15
stratum 4 |11619     32 10 0.31
stratum 5 |59878      5  5 1.00

Total               284 33 0.12
Anticipated CV: 0.049
```

Algo avec n_h non-entiers

```
Strata information:
           |  bh      Nh nh   fh
stratum 1 | 1568    121  6 0.05
stratum 2 | 3004     81  7 0.09
stratum 3 | 5509     44  7 0.16
stratum 4 |11619     33 11 0.33
stratum 5 |59878      5  5 1.00

Total               284 36 0.12
Anticipated CV: 0.046
```

Il faut augmenter n par environ 10% pour remplir l'objectif de précision avec la stratification trouvée avec des n_h non-entiers.

Exemple 2: CV=5%; X=REV84, Y=RMT85

Le plan précédent atteint l'objectif de précésion fixé:

`var.strata(plan,y=Sweden$RMT85) donne un CV de 4.9%`

avec n=33.

Si on stratifie en supposant Y=X

`plan2<-strata.LH(Sweden$REV84,CV=.05,Ls=5,alloc=c(0.5,0,0.5))`

on obtient un plan avec n=17 qui donne une précision de

`var.strata(plan2,y=Sweden$RMT85) =` **`8.2%`**

pour estimer la moyenne de la variable RMT85.

Strata information: Y≠X

	bh	Nh	nh	fh
stratum 1	1683	130	6	0.05
stratum 2	3174	76	6	0.08
stratum 3	5824	41	6	0.15
stratum 4	11619	32	10	0.31
stratum 5	59878	5	5	1.00
Total		284	33	0.12

Anticipated CV: 0.049

Strata information: Y=X

	bh	Nh	nh	fh
stratum 1	1599	122	3	0.02
stratum 2	3174	84	3	0.04
stratum 3	5509	40	2	0.05
stratum 4	12658	34	5	0.15
stratum 5	59878	4	4	1.00
Total		284	17	0.06

Anticipated CV: 0.082

Exemple 3: Enquête mensuelle sur le commerce de détail (données MRTS)

Statistique Canada Statistics Canada

- X = revenu annuel de N=2000 entreprises estimé à l'aide de données fiscales
- Y = vente mensuelle annualisée
- Allocation de Neyman
- CVcible=1%
- L=4
- Param. loglinéaires: β_1=.90 σ^2=.015 (ρ=.979)
- Taux de survie: p_h=(.95,.95,.975,1)
- Non réponse r_h=(.9,.9,.95,1)
- Travail réalisé en collaboration avec Michel Ferland

Exemple 3: Enquête mensuelle sur le commerce de détail (données MRTS)

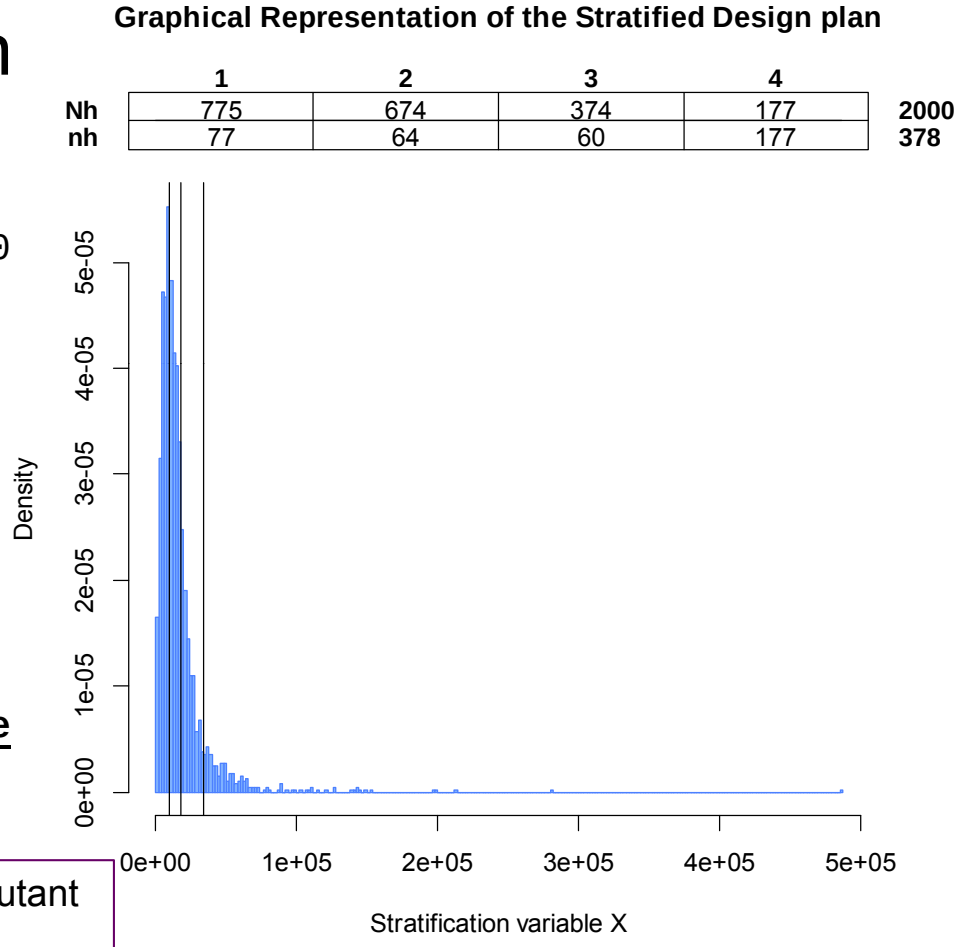
Stratification obtenue en posant Y=X

```
plan<-
strata.LH(MRTS,CV=.01,Ls=4,alloc=c(0
.5,0,0.5))
```

La strate 1 comprend de nombreux établissements. Une strate à tirage nulle pourrait-elle être utile?

```
plan2<-
strata.LH(....,takenone=TRUE,bias.pe
nalty=0.5)
```

Calcule un EQM prédit en ajoutant $(\text{biais}/2)^2$ à la variance



Exemple 3: Enquête mensuelle sur le commerce de détail (données MRTS)

La strate à tirage nul permet de faire baisser la taille d'échantillon de 10% (378 à 342).

Cette strate ne représente que 0.6% du total de X.

```
Strata information:
           |      type rh |         bh      E(Y)        Var(Y)   Nh  nh   fh
stratum 1 | take-none  - |    2751.94  1697.55      525099.3  110   0 0.00
stratum 2 | take-some  1 |   11052.50  7206.21     5366226.0  797  72 0.09
stratum 3 | take-some  1 |   19523.50 14908.30     5525050.4  611  56 0.09
stratum 4 | take-some  1 |   35457.86 25097.57    17391557.1  320  52 0.16
stratum 5 |  take-all  1 | 486367.49 66020.52 2482025995.1  162 162 1.00
Total                                                       2000 342 0.17
```

Évidemment les tailles d'échantillon ne sont pas réalistes car Y et X diffèrent.

```
plan3<-strata.LH(....,model="loglinear", model.control=(list(beta = 0.9,
sig2 = 0.015)) )
```

Exemple 3: Enquête mensuelle sur le commerce de détail (données MRTS)

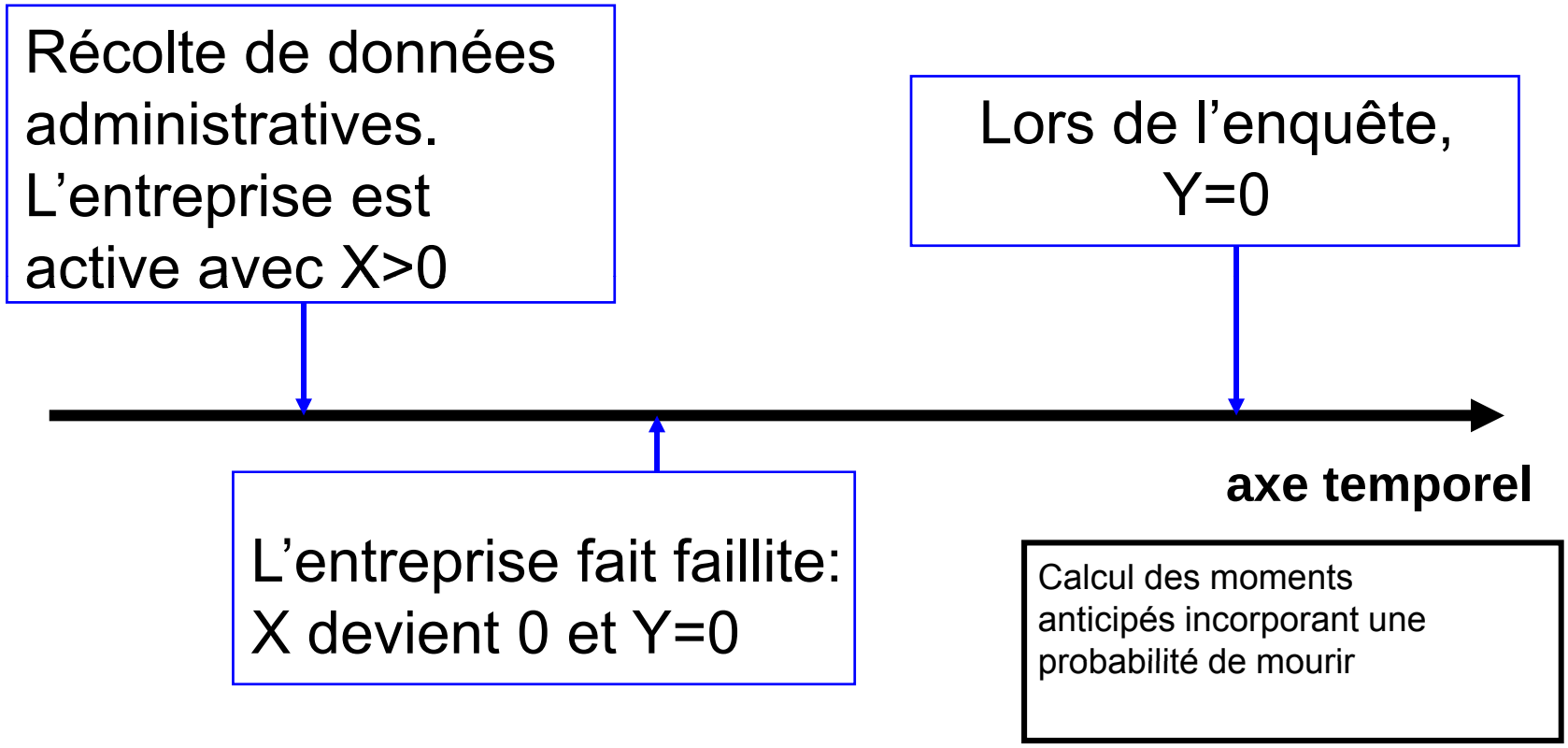
Plan avec Y=X

```
Strata information:
           |      type rh |         bh     E(Y)       Var(Y)   Nh  nh   fh
stratum 1 | take-none  - |   2751.94  1697.55     525099.3  110   0 0.00
stratum 2 | take-some  1 |  11052.50  7206.21    5366226.0  797  72 0.09
stratum 3 | take-some  1 |  19523.50 14908.30    5525050.4  611  56 0.09
stratum 4 | take-some  1 |  35457.86 25097.57   17391557.1  320  52 0.16
stratum 5 |  take-all  1 | 486367.49 66020.52 2482025995.1  162 162 1.00
Total                                                     2000 342 0.17
```

Plan avec Y≠X: n croît par plus de 20%

```
Strata information:
           |      type ph rh |         bh     E(Y)       Var(Y)   Nh  nh   fh
stratum 1 | take-none  1  - |   2296.46   680.49     78434.47   84   0 0.00
stratum 2 | take-some  1  1 |   9801.28  2656.91    737103.51  691  71 0.10
stratum 3 | take-some  1  1 |  18176.38  5254.97   1125514.19  674  85 0.13
stratum 4 | take-some  1  1 |  34940.01  8695.60   3211063.72  385  82 0.21
stratum 5 |  take-all  1  1 | 486367.49 21208.95 194024681.72  166 166 1.00
Total                                                     2000 404 0.20
```

Exemple 3: Décès d'une entreprise



On va supposer des taux de mortalités de 5% pour les petites commerces et de 2.5% pour les plus gros.

Exemple 3: Enquête mensuelle sur le commerce de détail, mortalité

Pour tenir compte de la mortalité dans le calcul des moments anticipés on utilise

```
plan4<-strata.LH(.... model="loglinear", model.control=(list(beta = 0.9,
sig2 = 0.015,ph= c(.95,.95,.975,1)))
```

```
strata information:
           |      type   ph rh |         bh     E(Y)       Var(Y)   Nh  nh   fh
stratum 1 | take-none 1.00  - |    1992.40   594.60      64167.2   65   0 0.00
stratum 2 | take-some 0.95  1 |    7676.27  2056.92     650589.2  478  52 0.11
stratum 3 | take-some 0.95  1 |   14461.03  4055.38    1579073.1  643 108 0.17
stratum 4 | take-some 0.98  1 |   29550.98  7142.89    3828258.5  593 155 0.26
stratum 5 |  take-all 1.00  1 | 486367.49 18735.89 164741159.9  221 221 1.00
Total                                                          2000 536 0.27
```

En présence de mortalité, n passe de 404 à 536. C'est une augmentation importante.

Exemple 3: Enquête mensuelle sur le commerce de détail, mortalité + non-réponse

On peut également postuler une non réponse igorable à l'intérieur de chaque strate

```
plan4<-strata.LH(.... , rh=c(0.9,0.9,.95,1), model="loglinear",
model.control=(list(beta = 0.9, sig2 = 0.015,ph= c(.95,.95,.975,1)))
```

```
Strata information:
           |      type   ph   rh |        bh     E(Y)      Var(Y)   Nh  nh   fh
stratum 1 | take-none 1.00    - |   1992.40   594.60     64167.2   65   0 0.00
stratum 2 | take-some 0.95 0.90 |   7414.94  2006.28    606796.7  452  50 0.11
stratum 3 | take-some 0.95 0.90 |  13861.43  3916.98   1449136.2  618 106 0.17
stratum 4 | take-some 0.98 0.95 |  28525.19  6937.68   3670632.3  632 172 0.27
stratum 5 |  take-all 1.00 1.00 | 486367.49 18305.04 159759452.3  233 233 1.00
Total                                                          2000 561 0.28
```

En tenant compte de la non-réponse, n passe de 536 à 561. Notons que les tailles des strates ont changé. Un simple ajustement pour la non réponse a posteriori donne

$$n = \frac{52}{0.9} + \frac{108}{0.9} + \frac{155}{0.95} + 221 = 563$$

DISCUSSION

- La façon dont les tailles d'échantillon sont traitées (réelles ou entières) dans les algorithmes de stratification a un certain impact sur les résultats
- Pour obtenir une stratification optimale il faut les considérer comme des valeurs entières
- Le package `stratification` permet de réaliser ces analyses et de planifier des enquêtes qui utilisent l'échantillonnage stratifié.

Techniques d'enquête, juin 2011
Vol. 37, N° 1, pp. 59-72
Statistique Canada, N° 12-001-X au catalogue

Élaboration de plans stratifiés en R à l'aide du programme *stratification*

Sophie Baillargeon et Louis-Paul Rivest [1]