STA108 – Algorithmes de tirage

Cours n°05 du 16/10/2020

Sylvie Rousseau

le cnam

Sommaire

- **Nombres au hasard**
- Algorithmes de tirage
  - Plans simples sans remise
  - Plans à probabilités inégales

Nombres aléatoires

- Une suite de réalisations indépendantes de la loi uniforme U[0,1]
- Utiles pour réaliser des tirages ou simuler un phénomène aléatoire

Comment obtenir des nombres aléatoires ou pseudo-aléatoires ?

- Procédés physiques :
  - Jeu de la roulette ou tirage de type loterie :
    - *Par exemple, pour simuler une variable aléatoire discrète uniforme dans {0,1,...9}, on tire des boules avec remise dans une urne de 10 boules numérotées de 0 à 9*
  - Éclairage à intervalles irréguliers d'un disque divisé en 10 secteurs isométriques et numérotés de 0 à 9 (table de Kendall et Babington Smith)
  - ...
- Procédés déterministes
  - Décimales de π
  - Décimales de tables de logarithmes (table de Fisher et Yates)
  - Méthode de Von Neumann
  - Méthode de congruence
  - ...

Méthode du carré médian
(Von Neumann, 1946)

- On part d'un nombre entier
- On l'élève au carré
- On extrait les chiffres du centre comme nombres aléatoires

Exemple :

Etape j	x_j	x_j^2
1	*7534*	*56 761 156*
2	*7611*	*57 927 321*
3	*9273*	*85 988 529*
4	*9885*	*97 713 225*
....		

d'où la suite 761192739885

- Inconvénients majeurs :
  - Dépendance au nombre de départ
  - Régularités nombreuses (permanence de 0 ou de séries particulières).

le cnam

5

Méthode de congruence simple

- Repose sur des suites récurrentes
  - Choix arbitraire d'un entier x_0 appelé « germe » ou « seed » ou encore « graine »
  - Génération d'une séquence d'entiers $(x_1, \ldots, x_i, \ldots, x_n)$ selon

    $$x_{i+1} = ax_i + b \ [m] \quad \text{pour} \quad i \in \{0, \ldots, n-1\}$$

    où a, b et m sont des entiers appelés respectivement multiplicateur, incrément et modulo
    - On vérifie $0 \leq x_i \leq m$ pour $i \in \{1, \ldots, n\}$
- Intérêt : les nombres $(u_1, \ldots, u_i, \ldots, u_n)$ où $u_i = x_i / m$ forment un échantillon de la loi uniforme sur [0,1] si les entiers a, b et m sont « bien » choisis
  - Le procédé étant déterministe, ces nombres sont dits pseudo-aléatoires

Méthode de congruence simple

- Exemple *: $x_0 = 1$; $a = 6$; $b = 0$; $m = 25$*

  $x_0 = 1$

  $x_1 = 6\ [25] = 6$

  $x_2 = 36\ [25] = 11$

  $x_3 = 66\ [25] = 16$

  $x_4 = 96\ [25] = 21$

  $x_5 = 126\ [25] = 1 = x0$

  Ce cycle a pour longueur 5

- Remarques :
  - La séquence contient au plus *m* termes distincts
  - La suite est donc périodique de période *p* avec $m \geq p$
  - Si $p = m$, la période est dite pleine

Choix des entiers *a*, *b* et *m*

- Ils sont déterminés de telle sorte que la séquence ait les meilleures propriétés possibles
- En particulier, *m* est pris aussi grand que possible pour assurer une grande variété de valeurs
  - Hull et Dobell (1962) ont montré que les séquences de période pleine sont obtenues si et seulement si :
    - *b* et *m* sont premiers entre eux
    - *(a-1)* est un multiple de chaque nombre premier qui divise *m*
    - si *m* est un multiple de 4 alors *(a-1)* aussi

Méthode de Lehmer

- Un algorithme très usité est la méthode congruentielle de Lehmer (1948) qui pose b = 0 *:* $x_{i+1} = ax_i \ [m]$
  - m aussi grand que possible : m = 2^{x-1} -1 ou x = nombre de bits de l'ordinateur

Congruence avec retard

- Pour limiter certaines régularités, on peut calculer :
  - $x_{i+1} = ax_{i-r} + b$ [m] où r est un paramètre de retard
  - les r premiers termes selon une relation qui peut être différente, par exemple : $x_{i+1} = a'x_{i-r} + b'$ [m]
- Variantes
  - On peut introduire plusieurs termes de retard :

$$x_{i+1} = a\,(x_{i-r} + x_{i-s}) + b \text{ [m]}$$

  - Suites de Fibonacci : $u_{n+2} = u_{n+1} + u_n$ *[m]*

Validation et simulation d'autres lois

- Il existe des tests qui vérifient l'indépendance et l'uniformité des séquences de nombres pseudo-aléatoires ainsi obtenues
- Une séquence (v_1, …, v_i, ..., v_n) de la loi uniforme sur [a,b] s'obtient à partir des nombres (u_1, …, u_i, ..., u_n) de la loi uniforme sur [0,1] grâce à la relation suivante :

$$u_i = \frac{1}{b-a}(v_i - a)$$

- Les autres lois peuvent être simulées à partir grâce aux méthodes de la fonction de répartition inverse ou d'acceptation-rejet

Sommaire

- Nombres au hasard
- **Algorithmes de tirage**
  - Plans simples sans remise
  - Plans à probabilités inégales

Qualités recherchées d'un algorithme

- Exact : les probabilités d'inclusion d'ordre 1 sont respectées
- Général : adapté à toutes les probabilités d'inclusion d'ordre 1
- De taille fixe : la taille de l'échantillon vaut toujours n
- Sans remise : une unité ne peut pas être sélectionnée plusieurs fois
- Tel que les probabilités d'inclusion d'ordre 2 sont calculables sur toute la base de sondage et sont strictement positives
- Vérifiant les conditions de Sen,Yates et Grundy
- Rapide, évitant notamment d'énumérer tous les échantillons possibles
- Fonctionnant en une seule lecture de fichier
- Utilisable aussi quand la taille de la population n'est pas connue
- …

Plans simples sans remise

- Tirage successif des unités
- Tirage bernoullien (tirage de Bernoulli)
- Méthode du tri aléatoire
- Méthode de mise à jour de l'échantillon
- Méthode de sélection-rejet

Tirage successif des unités

- On tire une unité au hasard parmi *N*

  ⇨ *p = 1/N* d'obtenir l'individu $k_{(1)}$

- On ôte cet individu de la base de sondage
- Et on fait un nouveau choix

  ⇨ *p = 1/ (N-1)* d'obtenir $k_{(2)}$

- Etc. jusqu'à constituer un échantillon de taille n

Tirage successif des unités

- Avantage :
  - On a bien défini un plan simple
    - La probabilité d'obtenir les *n* individus ($k_{(1)}$, $k_{(2)}$, ... $k_{(n)}$) dans cet ordre vaut :

      $$1/N \times 1/(N-1) \times .... \times 1/(N-n+1) = (N-n)!/N!$$

    - Comme il y a *n!* permutations du *n*-uplet ($k_{(1)}$, $k_{(2)}$, ... $k_{(n)}$), la probabilité de sélectionner un échantillon de *n* unités vaut :

      $$n!(N-n)!/N! = 1/C_N^n$$

- Inconvénient :
  - Temps d'exécution (*n* lectures de fichiers et opérations de tri)

Tirage bernoullien

- On génère *N* réalisations d'une v.a de loi *U[0,1]* : $(u_1, ..., u_N)$
- On fixe *p* dans [0,1]
- Si $u_k < p$ alors l'unité *k* est échantillonnée
- Exemple :

k	p	u_k	I_k
1	0,4	0,323	1
2		0,5	0
3		0,361	1
4		0,78	0
5		0,012	1
6		0,252	1
7		0,578	0
8		0,997	0
9		0,112	1
10		0,645	0

$n_s = 5$

Tirage bernoullien

- Avantages :
  - Plan simple
  - Facile à programmer
  - Rapide
  - Tirage i.i.d. des unités
  - *N* n'a pas besoin d'être connu a priori
- Inconvénients:
  - La taille de l'échantillon obtenu est aléatoire
  - *s*=ø est possible

↳ Éviter ce plan de sondage surtout si *n* petit sauf si on ne peut faire autrement

Tirage bernoullien

- Remarques
  - $\pi_k = p$ = constante
  - $\pi_{kl} = p^2$ par indépendance des tirages
  - Si $p = n / N$ alors $E(n_s) = n$

Méthode du tri aléatoire

- On génère *N* réalisations d'une v.a de loi *U[0,1]* sur toute la population $(u_1, \ldots, u_N)$
- On trie la base de sondage par valeurs croissantes (ou décroissantes) des $(u_i)_{i = 1 \text{ à } N}$
- On retient les *n* 1ers (ou derniers) individus du fichier

Méthode du tri aléatoire

- Avantages :
  - Plan simple
  - Inutile de connaître *N* a priori
  - Facilité de mise en oeuvre
- Inconvénient :
  - Long si *N* est grand

Méthode de mise à jour de l'échantillon *(Mac Leod, Bellhouse, 1983)*

- On sélectionne les *n* 1ers individus de la base de sondage
- On répète pour $k = n+1$ à N :
  - simulation de *u* selon *U[0,1]*
  - si $u < n / k$, alors
    - sélectionner l'unité *k*
    - ôter une des unités déjà tirées, à probabilités égales *(1/n)*

Méthode de mise à jour de l'échantillon

- Exemple

k	π	I_k *(initial)*	I_k	u_k	n / k
1		1	0		
2		1	1		
3		1	0		
4		1	0		
5	0,4		1	0,354	0,800
6			1	0,025	0,667
7			0	0,987	0,571
8			1	0,153	0,500
9			0	0,485	0,444
10			0	0,800	0,400

4

Méthode de mise à jour de l'échantillon

- Avantages:
  - Plan simple
  - Inutile de connaître *N* a priori
  - Intéressant si *N* est grand
- Inconvénient :
  - Il faut réserver un vecteur de taille *n*

Méthode de sélection-rejet
(Fan, Muller, Rezucha, et Bellington, 1962, 1975)

- Au départ :
  - $k = 1$: nombre d'unités de la population déjà examinées (yc. celle en cours)
  - $j = 0$: nombre d'unités de la population déjà sélectionnées
- Puis tant que $j < n$ alors :
  - on génère u selon $U[0,1]$
  - si $u < (n-j) / (N-k+1)$ = nb d'unités restant à tirer / nb d'unités restant à examiner

    alors :
    - sélectionner k
    - $j = j + 1$
  - $k = k + 1$

Méthode de sélection-rejet

- Exemple

k	π	u_k	j	$(n - j) / (N - k + 1)$	I_k
1	0,4	0,375	0	0,400	1
2		0,624	1	0,333	0
3		0,045	1	0,375	1
4		0,517	2	0,286	0
5		0,632	2	0,333	0
6		0,246	2	0,400	1
7		0,927	3	0,250	0
8		0,325	3	0,333	1
9		0,645	4	0,000	0
10		0,178	4	0,000	0
Total	**4**				**4**

Méthode de sélection-rejet

- *Avantages*:
  - Plan simple
  - Une seule lecture de fichier suffit

- *Inconvénient* :
  - Il faut connaître *N* a priori

Plans à probabilités inégales

- Tirage poissonnien (tirage de Poisson)
- Méthode de Sunter
- Tirage systématique (algorithme de Madow)

Tirage poissonnien

- On dispose de π_k dans *[0,1]* pour toutes les unités *k* de la population
- On génère *N* réalisations d'une v.a de loi *U[0,1]: (u_1, ..., u_N)*
- Si $u_k < \pi_k$ alors l'unité *k* est échantillonnée
- Exemple :

k	π_k	u_k	I_k
1	0,234	0,997	0
2	0,078	0,645	0
3	0,139	0,112	1
4	0,518	0,361	1
5	0,267	0,578	0
6	0,791	0,500	1
7	0,389	0,780	0
8	0,894	0,323	1
9	0,333	0,452	0
10	0,357	0,012	1
Total	**4**		**5**

Tirage poissonnien

- Avantages :
  - Facilité de programmation
  - Tirage i.i.d. des unités
  - *N* n'a pas besoin d'être connu a priori
- Inconvénients :
  - La taille de l'échantillon obtenu est aléatoire
  - *s*=ø est possible

↳ éviter ce plan de sondage surtout si *n* petit sauf si on ne peut faire autrement

Tirage poissonnien

- Remarques
  - $\pi_{kl} = \pi_k \pi_l$ par indépendance des tirages
  - L'expression de la variance est assez simple
  - Sondage d'entropie maximale (le plus aléatoire possible)
  - Généralisation du tirage bernoullien

Méthode de Sunter
(généralisation de la méthode de sélection-rejet, 1977, 1986)

- Au départ :
  - *k = 1:* nombre d'unités de la population déjà examinées
  - *j = 0 :* nombre d'unités de la population déjà sélectionnées
  - *z = 0 :* cumul des probabilités d'inclusion

$$z_k = V_{k-1} = \pi_1 + \pi_2 + \ldots + \pi_{k-1}$$

- Puis tant que $j < n$ alors
  - on génère *u* selon *U[0,1]*
  - si $u < \pi_k \times (n-j) / (n-z)$ alors :
    - sélectionner *k*
    - $j = j + 1$
  - $k = k + 1$
  - $z = z + \pi_k$

Méthode de Sunter

- Exemple

k	X_k	π_k	V_k	u_k	j	$(n - j) / (n - V_{k-1}) \times \pi_k$	I_k
1	10	0,8	0,8	0,375	0	0,800	1
2	10	0,8	1,6	0,624	1	0,750	1
3	8	0,64	2,24	0,045	2	0,533	1
4	6	0,48	2,72	0,517	3	0,273	0
5	6	0,48	3,2	0,632	3	0,375	0
6	4	0,32	3,52	0,246	3	0,400	1
7	2	0,16	3,68	0,927	4	0,000	0
8	2	0,16	3,84	0,325	4	0,000	0
9	1	0,08	3,92	0,645	4	0,000	0
10	1	0,08	4	0,178	4	0,000	0
Total	**50**	**4**					**4**

Méthode de Sunter

- Avantage :
  - Une seule lecture de fichier suffit
- Inconvénient :
  - Il est possible que $\pi_k \times (n-j) / (n-z)$ dépasse 1

    => Cas rare qui conduit à *n-1* unités et non *n*

Tirage systématique (Madow, 1949)

- On cumule pour tous les individus les probabilités d'inclusion : $V_k = \pi_1+\pi_2 + ...+ \pi_k$
- On génère une seule réalisation u de la loi $U[0,1]$
- On sélectionne k tel que $V_{k-1} \leq u < V_k$
- Puis l tel que $V_{l-1} \leq u+1 < V_l$
- Etc ...

Tirage systématique

- Exemple : si *u = 0,67*

k	π_k	V_k	I_k
1	0,894	0,894	1
2	0,791	1,685	1
3	0,518	2,203	0
4	0,389	2,592	0
5	0,357	2,949	1
6	0,333	3,282	0
7	0,267	3,549	0
8	0,234	3,783	1
9	0,139	3,922	0
10	0,078	4	0
Total	**4**		**4**

Si u =0,265
Alors s={1,2,4,6}

Tirage systématique

- Avantages :
  - Taille fixe égale à *n*
  - Méthode générale
  - Méthode exacte
  - Méthode sans remise

- Inconvénients :
  - Beaucoup de probabilités d'inclusion d'ordre 2 sont nulles
  - L'échantillon obtenu dépend de l'ordre du fichier

Tirage équilibré

« Construire un échantillon équilibré, c'est un peu le rêve de tous les praticiens, économistes ou sociologues, qui doivent travailler sur une population à partir d'un échantillon : pour un ensemble de variables connues, on souhaite retrouver ce que l'on sait de la population. Pour ces variables, l'échantillon doit être « représentatif » pour être « bon ». Techniquement, c'est d'ailleurs la définition que donne J.HAJEK de la représentativité. L'ennui, c'est que l'échantillon se doit aussi d'être tiré au hasard pour que, d'après J.NEYMANN, on puisse utiliser ses propriétés statistiques pour dire qu'il est exempt de biais et pouvoir évaluer sa précision »

J-C. DEVILLE

Tirage équilibré

- **Un échantillon est dit équilibré** sur un jeu de variables auxiliaires *x* si

$$\hat{T}_{x\pi} = \sum_{k \in s} \frac{X_k}{\pi_k} = T_x = \sum_{k \in U} X_k$$

Le total T_x est donc parfaitement estimé (variance d'échantillonnage nulle)

- Un plan de sondage est dit équilibré sur les variables *X* si seuls les échantillons équilibrés sur *X* ont une probabilité non nulle d'être sélectionnés

Suppose une information auxiliaire (X connue au niveau individuel sur toute la population)

Tirage équilibré

- Intérêts : choix aléatoire d'un échantillon qui permet de :
  - Retrouver ce qu'on sait de la population, pour un ensemble de variables connues x (« variables de contrôle » ou « variables d'équilibrage »)
  - Gagner en précision sur l'estimation du(es) paramètre(s) d'intérêt (la variance d'échantillonnage dépend seulement de la variabilité non expliquée par les variables de contrôle)
    - Gain d'autant plus grand que le lien entre la(es) variable(s) d'intérêt et les variables d'équilibrage est fort

Tirage équilibré

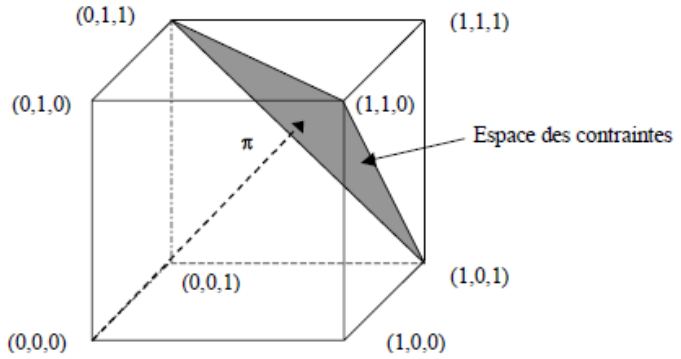
- Exemples :
  - Un échantillon équilibré sur la variable constante égale à 1 restitue exactement la taille N de la population
  - Équilibrer sur la variable des probabilités d'inclusion permet d'obtenir un échantillon de taille fixe n
  - Tirage probabiliste respectant des quotas

Tirage équilibré

- Exemples
  - Un échantillon équilibré sur la variable constante égale à 1 restitue exactement la taille N de la population
  - Équilibrer sur la variable des probabilités d'inclusion permet d'obtenir un échantillon de taille fixe n
  - Tirage probabiliste respectant des quotas
- Parfois impossible
  - Exemple : estimer le total de Y avec un sondage aléatoire simple sans remise de taille 2 dans une population de 5 unités où $Y_1 = Y_2 = Y_3 = 1$ et $Y_4 = Y_5 = 0$

Méthode du cube
Deville, Tillé, 2004

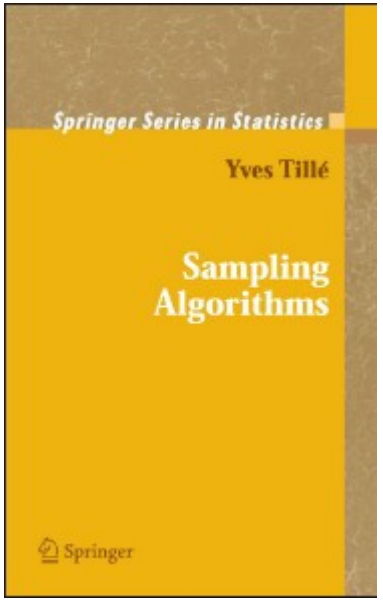
- Représentation dans un hypercube de tous les échantillons possibles d'une base de sondage de N individus. Chaque sommet désigne l'un des 2^N échantillons sans remise possibles, y compris l'ensemble vide.
- Deux phases
  - Envol
  - Atterrissage (équilibrage approché)
- Exemple d'équilibrage toujours exact sondage aléatoire simple sans remise de 2 unités parmi 3

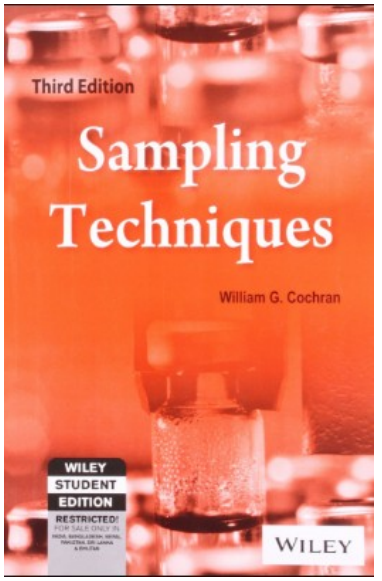


cnam

Beaucoup d'autres méthodes

- De nombreuses méthodes pour des plans à probabilités inégales en particulier
  - Mais aucune ne satisfait tous les critères de qualité requis
- Sampling Algorithms, Yves Tillé, Springer, 2006
- Sampling Techniques, 3ed.,William G.Cochran, Wiley, 2007





le cnam