La macro SAS CUBE d'échantillonnage équilibré

Documentation de l'utilisateur

Sylvie ROUSSEAU[1]
Frédéric TARDIEU[2]

08 avril 2004

[1] Division « Echantillonnage et traitement statistique des données », UMS, INSEE
[2] Service statistique, direction régionale de Poitou-Charentes, INSEE

Les auteurs remercient par avance les utilisateurs de la macro qui les informeront des erreurs qui peuvent subsister ainsi que ceux qui voudront bien leur faire part d'éventuelles suggestions pour l'améliorer.

Sommaire

Préface

Propos recueillis auprès de Jean-Claude Deville,
Directeur du laboratoire de statistique d'enquête,
CREST, INSEE

« 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 le dit 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. Un échantillon équilibré par CUBE est donc un bon échantillon, aléatoire et représentatif.
On doit cependant comprendre pourquoi, mathématiquement, un tel échantillon est bon. La réponse est simple : il est bon parce qu'il est précis. Plus exactement, la variance des estimateurs naturels (Horvitz-Thompson) qui lui sont associés ne dépend que de la variabilité non expliquée par les variables que l'on contrôle. Dans certains cas, le gain peut s'avérer assez extraordinaire.
Les applications nécessitent donc l'usage de bases de sondage contenant des variables auxiliaires assez nombreuses et connues pour toutes les unités tirables. Par ailleurs, il vaut mieux que la collecte ne laisse pas prise à la non-réponse, car celle-ci détruirait de façon incontrôlée les savants équilibres concoctés pour l'échantillonnage. Bien entendu, le nombre de contraintes doit rester relativement petit par rapport à la taille espérée de l'échantillon.
Les applications les plus naturelles ont donc trait à l'échantillonnage dans des populations d' « objets » tels que des communes, des aires géographiques, des entités administratives. A l'INSEE par exemple, une utilisation majeure réside dans le tirage des unités primaires des enquêtes auprès des ménages (agglomérations urbaines ou cantons ruraux), ou, potentiellement, des aires de l'enquête sur l'emploi.

On remarquera que, même pour un échantillon à probabilités égales (ce que CUBE fait aussi !), la faisabilité d'un échantillonnage équilibré probabiliste n'est pas évidente du tout (sans CUBE, bien entendu). Par conséquent, on peut utiliser CUBE pour obtenir facilement un échantillon à probabilités égales de communes (ou d'IRIS 2000) dans le cadre du Recensement Rénové. Or miraculeusement, lorsqu'on a tiré un échantillon équilibré à probabilités égales, ce qui reste de la population demeure équilibré, de sorte qu'on peut tirer dans le reliquat un nouvel échantillon équilibré. CUBE permet donc de partager une population (par exemple les communes rurales d'un département) en groupes de rotation, chacun d'eux constituant un échantillon aléatoire équilibré.
De façon un peu plus complexe mais tout aussi efficace, CUBE permet de tirer des échantillons à probabilités inégales DISJOINTS (ou dont on contrôle les intersections). Exemple : pour le contrôle de

qualité du recensement, tirer dans un lot de traitement un échantillon de districts pour vérifier l'analyse ménage-famille (probabilités dépendant du nombre de gros ménages) et un autre échantillon destiné au contrôle de la codification des professions (probabilités dépendant du nombre d'actifs).
Autre application assez intéressante : les enquêtes par quotas qui, par définition, sont des enquêtes sur échantillons équilibrés, mais que l'on réalise généralement en laissant les enquêteurs faire des « choix raisonnés ». CUBE est un algorithme permettant de constituer des échantillons probabilistes par quotas.
Mais, de façon un peu surprenante, les applications potentielles les plus prometteuses se situent dans le domaine de la correction de la non-réponse par imputation. Cette technique consiste à remplacer une donnée manquante par une donnée plausible, par exemple en allant chercher la valeur de la variable manquante auprès d'un individu répondant (hot-deck). On constitue ainsi un échantillon de répondants (les donneurs) auprès desquels on va recueillir les valeurs à imputer. Le problème réside dans le fait que cette technique, contrairement aux méthodes d'imputation déterministe et aux méthodes de repondération, génère une variance, parfois importante, non liée à l'échantillonnage. Elle est donc parfaitement parasitaire. Le recours à l'échantillonnage équilibré permet de réduire ou d'annuler cette variance superflue. Considérons, par exemple, le cas élémentaire dans lequel la réponse manque au hasard avec la même probabilité pour chaque unité. Si l'échantillon est de taille n et qu'il y a m répondants, on repondèrera les données par n/m si on peut choisir cette technique. Si on doit imputer, un échantillonnage aléatoire simple de n-m unités parmi les m répondants semble être la solution. Cette pratique augmente pourtant la variance des estimateurs (car chaque répondant va compter aléatoirement pour 1 ou 2 ; on peut montrer que la variance d'une moyenne peut ainsi augmenter de plus de dix pour cent sans pathologie). Tirer un échantillon de donneurs équilibré sur la moyenne des répondants (et éventuellement sur quelques autres caractéristiques de cette variable, -écart-type, médiane-) va restituer un échantillon imputé qui fournit la même moyenne qu'avec la technique de repondération, et donc annule la variance parasite due aux aléas de l'imputation. Cette idée se généralise à des cas très nombreux et complexes et constitue une voie d'avenir pour l'amélioration des techniques d'imputation ».

I. Introduction

La macro CUBE est un algorithme d'échantillonnage qui réalise des tirages équilibrés : elle assure le choix aléatoire d'un échantillon apte à restituer les vraies structures de la base de sondage pour des informations auxiliaires données. En outre, cette méthode améliore également, parfois de manière notable, la précision des résultats obtenus par l'enquête.
Ses applications sont potentiellement nombreuses, pourvu que l'on dispose d'une base de sondage munie d'information auxiliaire, qualitative ou quantitative, connue au niveau individuel. L'INSEE l'a mise en œuvre pour construire l'échantillon maître, réserve intermédiaire et localisée de logements qui alimente la plupart des enquêtes ménages. Dans le cadre du nouveau recensement de la population, l'institut l'a également employée pour désigner les groupes de rotation, de communes ou d'adresses, affectés à une année donnée du cycle quinquennal. Dans les deux cas, la sélection s'appuie sur des données socio-démographiques de référence disponibles sur l'ensemble du territoire.

Les fondements théoriques de l'algorithme ont été conçus par Jean-Claude Deville (Insee) et Yves Tillé (Université de Neuchâtel).
La macro CUBE a été développée à l'INSEE en SAS version 8 et sous l'environnement WINDOWS. Mise gratuitement à disposition sur le site Internet de l'INSEE, elle émane de contributeurs successifs : initialement écrite par des élèves de l'ENSAI, Frédéric Tardieu (service statistique de l'INSEE Bretagne) a par la suite rédigé le socle principal du programme actuel, finalisé par Bernard Weytens au centre national informatique de Lille grâce à des améliorations proposées par l'équipe méthodologique du nouveau recensement de la population de l'INSEE Rhône-Alpes.

I. 1. Définition d'un tirage équilibré

Le concept d'échantillon équilibré est simple et intuitif : un échantillon est dit équilibré sur une ou plusieurs variables disponibles dans la base de sondage, lorsque pour chacune d'entre elles, l'estimateur de Horvitz-Thompson[3] du total coïncide exactement avec le vrai total issu de la base de sondage.

Un échantillon *S* équilibré sur la variable de contrôle X respecte donc la contrainte suivante :

$$\sum_{k \in S} \frac{X_k}{\pi_k} = \sum_{k=1}^{N} X_k$$

où pour tout individu *k* de la base de sondage (*k = 1* à *N),* π_k désigne sa probabilité d'être sélectionné dans *S* et X_k la valeur qui lui est associée pour la variable X.

Par exemple, un sondage aléatoire simple (sans remise, de taille fixe) est équilibré sur la variable constante égale à « 1 ». L'échantillon obtenu restitue ainsi la taille exacte de la population dont il est issu. Pour s'en convaincre, il suffit de vérifier que la taille de la population s'obtient aussi bien en sommant la variable constante égale à « 1 » sur la base de sondage qu'en calculant sur l'échantillon l'estimateur de Horvitz-Thompson associé au total de cette variable. De même, un plan stratifié où l'on prélève un échantillon de taille fixe dans chaque strate, selon un tirage simple sans remise, est équilibré sur chacune des variables indicatrices caractérisant l'appartenance à chaque strate. Un tel échantillonnage permet de restituer l'effectif réel de chaque strate.

Un algorithme de tirage équilibré respecte donc à la fois les probabilités d'inclusion individuelles fixées à l'avance et autant de contraintes ou équations dites « d'équilibrage » que de variables auxiliaires de contrôle. Il s'agit donc de résoudre un système à *p* équations, où *p* est le nombre de contraintes, et à *N* inconnues définies par les indicatrices d'appartenance à l'échantillon des *N* individus de la base de sondage, tout en respectant un jeu de probabilités d'inclusion.

[3] L'estimateur d'Horvitz-Thompson du total d'une variable se calcule en multipliant chaque valeur individuelle observée sur l'échantillon par un coefficient d'extrapolation à la population entière (ou poids de sondage) égal à l'inverse de sa probabilité d'inclusion.

I. 2. Avantages d'un tirage équilibré

Les échantillons équilibrés offrent deux avantages majeurs. Par définition, ils « représentent »[4] bien la population au regard de l'information auxiliaire choisie : ils en assurent des estimations exactes, donc non soumises à la variance d'échantillonnage.

En outre, ils peuvent améliorer de manière notable la précision des estimateurs de paramètres issus de l'enquête, surtout si les variables d'équilibrage sont choisies avec soin. En effet, s'il existe un lien entre l'information auxiliaire et la variable d'intérêt, on peut assez naturellement imaginer que l'estimation du total de la variable d'intérêt sera, elle aussi, bien retranscrite par l'échantillon. On montre que la variance de l'estimateur de Horvitz-Thompson associé à la variable d'intérêt ne dépend que de la part de sa variabilité non expliquée par les variables de contrôle. En conséquence, l'estimation sera d'autant plus précise que le lien entre la variable d'intérêt et les variables d'équilibrage est fort. Dans certains cas, ce gain peut s'avérer substantiel.

Il est à noter que le même principe prévaut pour justifier le bien-fondé des redressements, par exemple des techniques de calage. Mais à la différence d'un ajustement effectué après collecte, les tirages équilibrés permettent d'améliorer la précision des estimateurs dès la phase d'échantillonnage. De ce fait, ils supposent néanmoins une information auxiliaire plus riche car connue a priori au niveau individuel sur l'ensemble de la base de sondage et non pas au niveau agrégé et à celui des seules unités échantillonnées comme cela suffit pour un calage.

I. 3. Difficultés de mise en œuvre d'un tirage équilibré

Le concept d'échantillon équilibré, si naturel, n'est pas nouveau, mais toute la difficulté réside dans la mise en œuvre d'algorithmes de tirage à la fois respectueux des probabilités d'inclusion, sans remise, rapides et généralisables à tout plan de sondage. Or, pour réaliser un tirage équilibré, il faut remplir deux conditions potentiellement irréalisables. D'abord, il n'existe pas toujours d'échantillon vérifiant exactement les équations d'équilibrage, et quand bien même, un tel échantillon existerait, l'équilibrage pourrait s'avérer tellement contraint qu'il pourrait conduire à un choix déterministe. Or la sélection doit demeurer aléatoire pour que les propriétés statistiques de biais et de variance d'échantillonnage conservent leur sens et pour respecter les probabilités d'inclusion.

Par exemple, dans une population composée de 100 individus dont la moitié est de sexe féminin, aucun échantillon aléatoire simple sans remise de taille impaire, 1 par exemple, ne peut être exactement équilibré sur le sexe - et restituer le nombre exact de femmes - sinon il faudrait sélectionner un effectif non entier de femmes.

De nombreuses raisons contribuent à rendre l'équilibrage impossible : les problèmes d'arrondis ou l'abus de variables de contrôle, obligent souvent à se contenter d'un équilibrage « approché ». La conséquence de cette approximation devient négligeable pour de grands échantillons.

Pour appréhender la notion d'équilibrage, avant que CUBE n'existe, des méthodes de sélection dites réjectives avaient été développées. Une des approches consiste à considérer d'abord un très grand nombre d'échantillons pour ensuite ne conserver que ceux qui respectent les contraintes (de façon approximative le cas échéant) et au final n'en choisir qu'un seul, de façon aléatoire. On comprend aisément qu'un tel procédé peut nécessiter un temps de calcul considérable. La bibliographie proposée au paragraphe VI indique quelques références sur ces méthodes.

[4] La notion intuitive de représentativité définie par Hájek caractérise un échantillon qui fournit des résultats « raisonnablement » proches des paramètres à estimer. Les échantillons équilibrés se conforment par construction à cette définition puisque les estimations associées aux variables de contrôle coïncident exactement avec les vraies valeurs, propres à la population.

I. 4. Objectif de la macro CUBE

La macro CUBE permet de désigner un échantillon équilibré, de manière aléatoire et « optimale ». Le critère d'optimalité, spécifié par l'utilisateur, concerne le traitement de l'équilibrage approché : il vise à minimiser l'éloignement à l'équilibre lorsque les contraintes ne peuvent pas être exactement respectées. Dans le cas où seuls certains échantillons sont exactement équilibrés, pour respecter le caractère aléatoire du tirage, la macro ne peut pas garantir à l'utilisateur d'obtenir l'un de ces échantillons.

Pour inférer à la population d'intérêt, les unités échantillonnées par CUBE sont pondérées par l'inverse des probabilités d'inclusion initiales fixées par l'utilisateur.

L'algorithme permet le choix d'échantillons équilibrés dans des bases de sondage comptant jusqu'à 100000 individus. Des bases plus volumineuses peuvent être considérées, moyennant les étapes intermédiaires indiquées au paragraphe II-4.

La durée de fonctionnement dépend essentiellement de la taille de la base de sondage et du nombre de contraintes imposées. On trouvera au paragraphe II-4 des indications pour évaluer a priori le temps requis en fonction de ces différents paramètres. Par ailleurs, la macro informe l'utilisateur de sa durée de fonctionnement, a posteriori.

I. 5. Contexte d'utilisation de la macro CUBE

La macro CUBE s'avèrera d'autant plus efficace que les objectifs de l'enquête auront été clairement établis au préalable car ils conditionnent le choix des critères appropriés au tirage. Algorithme d'échantillonnage, CUBE ne dispense effectivement pas le méthodologue de toute la phase de réflexion préalable au choix du plan de sondage le plus adapté à sa problématique. En particulier, procéder à un tirage équilibré avec la macro CUBE nécessite de disposer en amont à la fois des probabilités d'inclusion, propres au plan choisi, et d'informations auxiliaires appropriées connues à un niveau individuel.

On notera cependant que même en l'absence d'informations auxiliaires, il existe toujours - dans le cadre des sondages probabilistes - deux variables d'équilibrage essentielles : la constante égale à « 1 » et les probabilités d'inclusion elles-mêmes. Respecter ces contraintes permet respectivement d'estimer de manière exacte l'effectif de la population et d'obtenir un échantillon de taille fixe. Dans un plan où tous les individus possèdent la même probabilité d'inclusion, ces deux contraintes sont redondantes, comme l'illustrent les paragraphes II-3 et V-2 ci-dessous. En effet, si deux variables sont colinéaires, équilibrer sur l'une revient à équilibrer sur l'autre.

II. Présentation de l'algorithme

La méthode du Cube tire son nom d'une représentation géométrique des différents échantillons possibles d'une base de sondage de taille N dans un hypercube de dimension N. Chacun des sommets de l'hypercube désigne l'un des 2^N échantillons sans remise possibles, y compris l'ensemble vide. Les N coordonnées d'un sommet sont définies par les valeurs des indicatrices d'appartenance à l'échantillon de chacun des N individus. Par exemple, le sommet de coordonnées (1,0,…,0,1) désigne l'échantillon de taille 2 composé du premier et du dernier individu de la base de sondage.

L'algorithme génère une marche aléatoire dans cet hypercube qui part des probabilités d'inclusion individuelles fixées a priori, et qui arrive à un sommet symbolisant l'échantillon final. Il se décompose en deux phases successives dites de vol et d'atterrissage. La deuxième étape ne se déclenche que dans le cas où le vol n'aurait pas abouti à un sommet respectant exactement les contraintes et permet alors de désigner l'échantillon final. En pratique, cette seconde phase se déclenche dans la plupart des cas, pour résoudre des problèmes d'arrondis notamment.

Dans la suite, on appelle « matrice des contraintes » la matrice *A* à *p* lignes et *N* colonnes où la $k^{ème}$ colonne donne les valeurs prises par le $k^{ème}$ individu de la base de sondage sur les *p* variables de contrôle $\left(X_1,...,X_j,...,X_p\right)$ divisées par la probabilité d'inclusion π_k de cet individu :

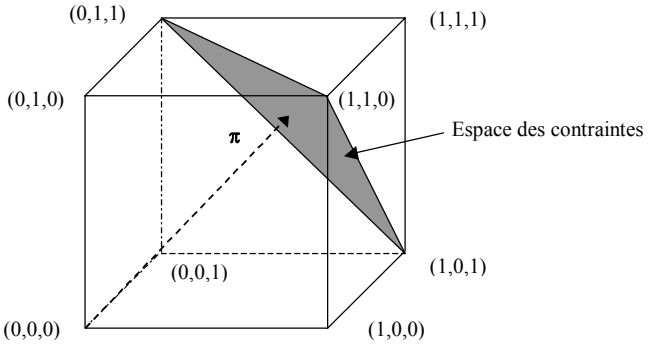
$$A = \begin{pmatrix} \frac{X_{1,1}}{\pi_1} & \cdots & \frac{X_{k,1}}{\pi_k} & \cdots & \frac{X_{N,1}}{\pi_N} \\ \cdots & & \cdots & & \cdots \\ \frac{X_{1,j}}{\pi_1} & \cdots & \frac{X_{k,j}}{\pi_k} & \cdots & \frac{X_{N,i}}{\pi_N} \\ \cdots & & \cdots & & \cdots \\ \frac{X_{1,p}}{\pi_1} & \cdots & \frac{X_{k,p}}{\pi_k} & \cdots & \frac{X_{N,p}}{\pi_N} \end{pmatrix}$$

Par ailleurs, on appelle « espace des contraintes » l'espace affine de R^N passant par le point π dont les coordonnées sont les probabilités d'inclusion individuelles initiales et dirigé d'après une base du noyau de la matrice des contraintes (il s'agit du translaté du noyau de *A* par π , c'est-à-dire *Ker A +* π*).*

Les échantillons équilibrés sont donc les sommets de l'hypercube qui appartiennent à l'espace des contraintes.

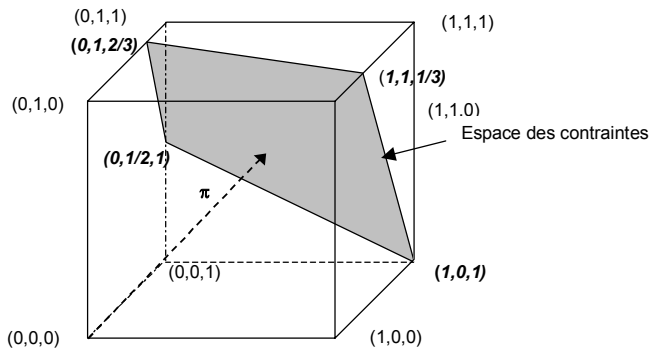
Les représentations ci-dessous illustrent les différents concepts utilisés par l'algorithme dans le cas d'une population de 3 unités, munies de probabilités d'inclusion toutes égales à 2/3. Chaque schéma caractérise une contrainte particulière.

<u>Exemple (i) d'équilibrage toujours exact</u> : sondage aléatoire simple sans remise équilibré sur la constante « 1 »



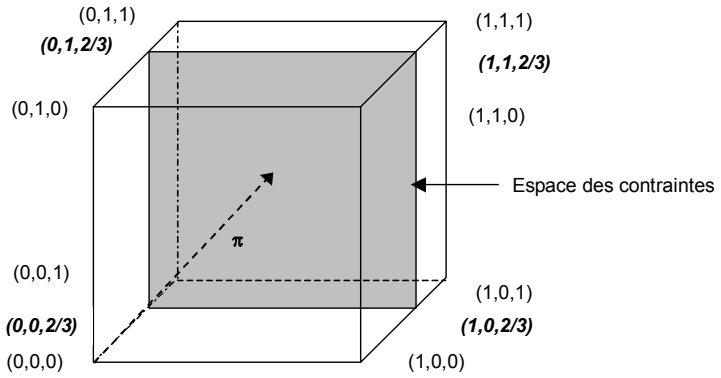
Il existe trois échantillons équilibrés, composés de 2 unités distinctes. L'équilibrage est toujours exact et la phase de vol suffit à désigner l'un de ses échantillons.

<u>Exemple (ii) d'équilibrage parfois exact</u> : sondage aléatoire sans remise équilibré sur le numéro d'ordre des individus dans la population



L'équation d'équilibrage n'est vérifiée ici que pour le seul sommet (1,0,1) composé de la première et de la dernière unité. Selon le chemin suivi par la phase de vol, celle-ci peut le sélectionner, mais aussi aboutir sur l'un des 3 autres points à l'intersection de l'espace des contraintes et du cube. La phase d'atterrissage se déclenche alors. Elle peut conduire, en fonction du critère d'optimalité choisi, à sélectionner l'un des 5 autres échantillons approximativement équilibrés qui sont apparentés aux sommets (0,1,0), (0,1,1), (0,0,1), (1,1,0) et (1,1,1). Le critère d'optimalité choisi pour la phase d'atterrissage permet de retenir plus probablement les échantillons les plus « proches » de l'équilibre. Cependant, pour préserver la nature aléatoire du tirage, la macro ne peut pas garantir à l'utilisateur qu'il obtienne l'unique échantillon équilibré.

Exemple (iii) d'équilibrage exact impossible : sondage aléatoire sans remise équilibré sur la variable indicatrice valant 1 seulement pour le dernier individu de la population



L'intersection entre l'espace des contraintes et les sommets du cube est vide donc aucun échantillon ne peut exactement satisfaire à l'équation d'équilibrage : la phase d'atterrissage est indispensable pour désigner un échantillon vérifiant approximativement la contrainte imposée.

II. 1. La phase de vol

Durant la phase de vol, toutes les contraintes sont exactement respectées. Cette phase s'arrête soit parce qu'elle est parvenue à un sommet parfaitement équilibré, soit parce que, avec le chemin suivi, il est devenu impossible de respecter exactement toutes les contraintes. Par exemple, lorsque les contraintes se réduisent à des conditions de stratification et de taille fixe et qu'il n'y pas de problèmes d'arrondis a priori, la macro fournit dès cette étape l'échantillon équilibré voulu. L'exemple (i) supra illustre cette situation.

Initialisée par les probabilités d'inclusion individuelles, la phase de vol fonctionne par itérations successives où chaque pas décide du sort d'au moins un individu. Le principe de chaque étape consiste à choisir de manière aléatoire une direction dans l'espace des contraintes et à la suivre jusqu'à ce qu'elle conduise sur une face du cube. Les différentes positions qu'il est possible d'atteindre lors d'une étape donnée sont en moyenne centrées sur le point auquel menait l'étape précédente, de sorte que les probabilités d'inclusion initiales soient respectées.

A chaque étape, la macro CUBE actualise un vecteur appelé *« PISTAR »* dont les composantes sont les coordonnées de la position qui vient d'être atteinte. Petit à petit, au hasard des directions suivies, la phase de vol chemine ainsi dans l'espace des contraintes et dans le cube jusqu'à s'arrêter sur un sommet si toutes les contraintes sont vérifiées ou sur une face le cas échéant. En fin de vol, l'état du vecteur *« PISTAR »* renseigne donc sur le sort réservé à chaque individu : ses composantes valent 0 pour les unités de la base de sondage définitivement écartées, 1 pour celles définitivement échantillonnées, ou un nombre strictement compris entre 0 et 1 pour les autres.

La phase d'atterrissage va régler le sort encore incertain de ces derniers individus dont le nombre (noté *Nrestants)* est au plus égal à la dimension de l'espace des contraintes (c'est-à-dire au nombre de contraintes non colinéaires).

A l'issue de la phase de vol, si les contraintes ont été rigoureusement respectées, l'utilisateur dispose donc d'un échantillon équilibré constitué des unités pour lesquelles *« PISTAR »* vaut 1. Sinon, la phase d'atterrissage s'amorce, mais déjà, avec un nombre raisonnable de contraintes, le sort de la majorité d'individus est définitivement réglé.

II. 2. La phase d'atterrissage

Lorsque la phase de vol n'a pas abouti sur l'un des sommets du cube, l'algorithme amorce la phase d'atterrissage (exemples supra (iii) et parfois (ii)). L'objectif est de parvenir à un sommet du cube « proche » de l'équilibre souhaité et du point déjà atteint. L'atterrissage concerne donc seulement les *Nrestants* individus au sort encore incertain et conduit à un échantillon final qui comprend bien entendu les individus déjà retenus en cours de vol.

Cette phase suppose d'affaiblir les contraintes, puisqu'elles se sont avérées trop rigides pour pouvoir être toutes systématiquement respectées. Pour cela, la version actuelle de la macro propose à l'utilisateur de choisir parmi trois critères, en fonction de la problématique rencontrée. La phase d'atterrissage livre in fine un échantillon aléatoire, issu d'un choix optimal au sens du critère choisi.

a. L'option d'atterrissage régie par un ordre de priorité sur les contraintes (option A)

Cette option consiste à relâcher successivement les contraintes de façon à pouvoir recommencer la phase de vol jusqu'à ce qu'elle parvienne à un échantillon équilibré. Ainsi, peu à peu, les contraintes sont levées, en commençant par les moins importantes[5]. A l'extrême, il ne reste plus qu'un seul individu sur lequel statuer ; l'algorithme le retient alors de manière randomisée avec une probabilité égale à sa composante dans le vecteur *« PISTAR »*. ***Pour employer cet atterrissage à bon escient, l'utilisateur doit lister les contraintes par ordre de priorité décroissante dans le paramètre CONTR= de la macro.***

Par exemple, pour s'assurer d'un échantillon de taille fixe avec cette option, il faut mentionner les probabilités d'inclusion individuelles comme première variable de contrôle. Le nombre d'unités sélectionnées est alors égal à la somme des probabilités d'inclusion, habituellement entière. Sinon l'algorithme propose, de manière randomisée, des échantillons dont la taille vaut l'un ou l'autre des deux entiers les plus proches.

Il est à noter que le respect d'une contrainte donnée dans l'équilibrage final ne dépend pas seulement de sa place dans le paramètre CONTR=, mais aussi de sa liaison éventuelle avec les variables qui lui sont préférées et de l'aléa des chemins suivis par l'algorithme. Par exemple, satisfaire 3 conditions dont la dernière est redondante avec la première conduira naturellement à mieux respecter la première (et aussi dernière) contrainte au détriment de la seconde. La comparaison des écarts relatifs dans l'équilibrage final ne peut donc de fait reprendre l'ordre des contraintes. L'exemple 1 du paragraphe V-2 ci-dessous présente un cas de ce type.

L'INSEE a notamment employé cette option pour désigner l'échantillon maître et les groupes de rotation du nouveau recensement de la population (de communes de moins 10 000 habitants ou d'adresses dans les communes plus grandes) sollicités par roulement au cours du cycle quinquennal.

b. L'option d'atterrissage la plus générale, retenue par défaut (option B)

Cette option commence d'abord par envisager tous les échantillons sans remise possibles et compatibles avec les résultats de la phase de vol, en nombre $2^{Nrestants}$ (avec l'ensemble vide). L'algorithme associe un coût à chacun des échantillons énumérés, coût d'autant plus important que l'échantillon s'éloigne de l'équilibre souhaité. On reviendra ci-dessous sur le choix du critère de coût. L'objectif consiste alors à trouver le plan de sondage, conforme à l'issue de la phase de vol, et dont le coût attendu est minimum (c'est à dire celui qui privilégie les échantillons compatibles les plus proches des contraintes). Pour rechercher ce plan optimum, on utilise un programme d'optimisation linéaire sur

[5] L'option d'atterrissage A détecte d'abord, parmi les *Nrestants* individus sur lesquels elle doit statuer, les éventuelles relations de colinéarité entre les contraintes. Elle élimine les contraintes colinéaires aux autres en commençant par la contrainte citée en dernier. Une fois résumées de cette sorte, les contraintes restantes sont alors relâchées une à une.

des jeux de probabilités de tirage des échantillons. Assuré de l'existence de solution, il ne reste alors qu'à tirer un des échantillons possibles conformément au jeu de probabilités optimum.

Pour mesurer l'éloignement d'un échantillon à l'équilibre, deux critères de coût sont disponibles dans le paramètre « COUT » de la macro :

- Le choix par défaut (CV) raisonne sur l'écart entre l'estimation associée à une contrainte et son vrai total rapporté à ce vrai total. Plus précisément, le coût de chaque échantillon *s* réalisable vaut :

$$coût_{CV}(s) = \sum_j \left(\frac{\hat{Z}_{j,HT}(s) - Z_j}{Z_j} \right)^2$$

où Z_j et $\hat{Z}_{j,HT}(s)$ désignent respectivement le vrai total de la $j^{ème}$ contrainte[6] et son estimateur d'Horvitz-Thompson.
A partir des coûts de chaque échantillon réalisable, l'algorithme calcule le coût global d'un plan de sondage selon :

$$\sum_s p(s) coût_{CV}(s)$$

où $p(s)$ désigne la probabilité de tirage de l'échantillon *s* (conditionnelle à l'issue de la phase de vol).
La phase d'atterrissage recherche le plan de sondage qui minimise ce coût global, représentant l'espérance de la somme sur toutes les contraintes des carrés des écarts relatifs.

- Une autre notion de coût (DIST) considère le carré de la distance euclidienne entre un échantillon (apparenté à un sommet de l'hypercube) et sa projection sur l'espace des contraintes. Le coût d'un échantillon *s* réalisable vaut ici :

$$coût_{DIST}(s) = (s - \pi)' M (s - \pi)$$

où

$$\begin{cases} s & \text{vecteur composé de 0 et de 1 symbolisant l'échantillon} \\ \pi & \text{vecteur composé des probabilités d'inclusion initiales} \\ M = A'(AA')^{-1}A & \text{matrice de projection sur l'espace des contraintes} \\ A & \text{matrice des contraintes} \end{cases}$$

On en déduit le coût global d'un plan de sondage :

$$\sum_s p(s) coût_{DIST}(s)$$

où $p(s)$ désigne la probabilité de tirage de l'échantillon *s* (conditionnelle à l'issue de la phase de vol).
La phase d'atterrissage se résout en minimisant ce coût global, représentant l'espérance du carré de la distance d'un échantillon à sa projection sur l'espace des contraintes.

Retenue par défaut, cette option d'atterrissage est la plus générale des trois. Elle donne un équilibrage de bonne qualité globale, en ce sens où elle ne pénalise ni ne favorise aucune contrainte a priori. Toutefois, en contrepartie de sa généralité, elle peut s'avérer parfaite sur des contraintes jugées secondaires au détriment de contraintes plus prioritaires pour la problématique rencontrée. En outre, elle consomme davantage de temps de calcul. L'option peut être recommandée pour sélectionner de grands échantillons, lorsqu'on ne regarde pas à l'unité près en terme de taille d'échantillon.

c. L'option d'atterrissage de taille fixe, hors cas d'arrondis (option C)

Cette option consiste à favoriser la sélection d'échantillons de taille fixe, elle n'a donc de sens que si la variable probabilité d'inclusion figure parmi les variables de contrôle. Elle fonctionne comme l'option par défaut à ceci près qu'elle n'envisage que les échantillons compatibles respectant exactement

[6] Dans le cas où le vrai total *Z* est nul et que cet atterrissage s'avère nécessaire, alors un message d'erreur s'affiche et la macro s'arrête. Pour pallier cette limite, une possibilité est de décaler l'échelle de mesure de la variable z à la condition sine qua non d'être certain d'estimer exactement la taille de la population. Une autre approche consiste à utiliser le critère DIST.

l'équilibrage sur les probabilités d'inclusion, c'est-à-dire les échantillons dont la taille est égale à la somme des probabilités d'inclusion initiales de tous les individus de la base de sondage (ou ce qui revient au même à la somme des coordonnées du vecteur « PISTAR »).
Si cette somme est entière, l'option C conduit toujours à des échantillons finaux de taille fixe égale à cet entier. Les deux autres options[7], notamment l'option B, ne garantissent pas systématiquement une telle propriété. En contrepartie, cette restriction sur l'éventail des échantillons autorisés peut conduire à pénaliser d'autres contraintes (celles qui sont les moins liées aux probabilités d'inclusion).
Si la somme des probabilités d'inclusion en fin de vol n'est pas un entier, les deux entiers qui l'encadrent sont les tailles possibles de l'échantillon final ; l'un d'entre eux est retenu de manière randomisée.

II. 3. Quelques illustrations

a. Comment tirer un échantillon de taille fixe ?

Si on équilibre avec succès sur la variable des probabilités d'inclusion, on obtiendra un échantillonnage de taille fixe, gagnant ainsi en précision par rapport à un échantillonnage de taille variable toutes choses égales par ailleurs.

En effet, dans une base de sondage de taille *N* où chaque individu *k* possède la probabilité π_k d'être retenu dans l'échantillon *S* de taille fixe *n*, on cherche à vérifier l'équation d'équilibrage suivante :

$$\sum_{k\in S}\frac{\pi_k}{\pi_k}=\sum_{k=1}^{N}\pi_k \quad \text{c'est-à-dire} \quad \sum_{k\in S}1=n\,.$$

Ceci équivaut donc à disposer d'un échantillon de la taille voulue *n*.

Par exemple, dans le cas d'un sondage aléatoire simple sans remise de taille fixe *n* (comme illustré dans l'exemple 8 infra), on a $\pi_k = n/N$, ce qui assure l'équilibrage sur la variable π_k puisque :

$$\sum_{k\in S}1=\sum_{k=1}^{N}\frac{n}{N}=n\,.$$

Comme spécifié en II-2, les options d'atterrissage A et C permettent d'obtenir des échantillons de taille fixe, égale à la somme des probabilités d'inclusion sur la base de sondage (ou l'un des deux entiers encadrant cette somme si elle n'est pas entière).

b. Comment obtenir un échantillon qui estime exactement la taille de la population ?

Si on équilibre avec succès sur la constante « 1 », dont la somme vaut la taille de la population, l'estimateur d'Horvitz-Thompson donnera exactement cet effectif, qui sera par conséquent estimé avec une variance nulle.

Dans le cas d'une population de taille *N* où chaque individu *k* possède la probabilité π_k d'être retenu dans l'échantillon *S,* vérifier l'équation d'équilibrage sur la constante « 1 » consiste à imposer :

$$\sum_{k\in S}\frac{1}{\pi_k}=\sum_{k=1}^{N}1\ .$$

Cette contrainte équivaut à $\hat{N}_{HT}=N$ c'est-à-dire conduit à obtenir un échantillon tel que l'estimateur d'Horvitz-Thompson $\hat{N}_{HT}$ de la taille de la population (somme des poids de sondage) coïncide avec l'effectif réel.

[7] Dans le cas où la somme des probabilités d'inclusion est entière -égale au nombre d'unités à sélectionner-, l'option A assure un échantillon de la taille voulue à condition de spécifier les probabilités d'inclusion comme première variable de contrôle.

On remarque bien sûr que dans le cas d'un sondage aléatoire simple sans remise de taille fixe *n*, estimer exactement la taille de la population équivaut à obtenir un échantillon de taille fixe puisque les deux variables de contrôle, la constante « 1 » et les probabilités d'inclusion individuelles $\pi_k = n/N$ sont proportionnelles. L'exemple (i) ci-dessus illustre en particulier ce propos.

c. Comment obtenir un échantillon qui estime exactement une moyenne d'une variable auxiliaire donnée ?

Puisque l'estimation d'une moyenne s'obtient en divisant celle d'un total par la taille de la population, si l'on veut retrouver la moyenne exacte d'une variable de contrôle à partir de l'échantillon, il faut également s'assurer que l'échantillon estime parfaitement la taille de la population.

d. Comment procéder pour des plans stratifiés ?

Un tirage stratifié peut être obtenu par un équilibrage sur des variables ad-hoc, faisant intervenir les variables de contrôle et les indicatrices d'appartenance aux strates.

Par exemple, pour que les contraintes imposent une allocation fixe dans chaque strate, on équilibrera sur les produits des probabilités d'inclusion par les indicatrices de strate.

Considérons en effet une base de sondage de taille *N* partitionnée en *H* strates notées U_h pour *h=1* à *H*, où l'on prélève des échantillons S_h de taille fixe n_h indépendants les uns des autres. Equilibrer sur les *H* probabilités d'inclusion individuelles des strates, revient à vérifier :

$$\sum_{k\in S}\frac{\pi_k}{\pi_k}\delta_{k,h} = \sum_{k=1}^{N}\pi_k\delta_{k,h} \quad \text{où} \quad \delta_{k,h} = \begin{cases} 1 & \text{si } k \in U_h \\ 0 & \text{sinon} \end{cases}, \quad \text{c'est-à-dire :} \quad \sum_{k\in S_h} 1 = \sum_{k=1}^{N_h}\pi_k = n_h \ ,$$

ce qui impose d'obtenir *H* échantillons de taille n_h.

Au demeurant, pour que cet équilibrage soit possible, il faut que la somme des probabilités d'inclusion dans chaque strate soit un entier. Sinon, face à ce problème d'arrondi, la macro CUBE désigne, en phase d'atterrissage, des allocations par strate conformes en espérance à la somme des probabilités d'inclusion de cette strate sur la base de sondage. En particulier, en spécifiant d'abord la contrainte de taille fixe dans l'option A d'atterrissage, l'effectif global de l'échantillon choisi vaut, de manière randomisée, l'un des deux entiers entourant la somme des probabilités d'inclusion. Mais cette propriété ne peut s'appliquer aux effectifs échantillonnés dans chaque strate puisque les variables d'équilibrage associées ne peuvent évidemment pas être toutes prioritaires.

On trouvera au paragraphe V-6 une application directe de l'algorithme à ce type de plan de sondage.

e. Comment tirer des échantillons probabilistes vérifiant des quotas ?

L'algorithme permet d'obtenir des échantillons équilibrés dans différents domaines y compris lorsque ceux-ci s'entrecroisent sans former une partition emboîtée de la population. Il suffit pour cela d'utiliser des variables ad-hoc faisant intervenir, à titre de variables de contrôle, les indicatrices d'appartenance aux différents sous-groupes. Il est par ailleurs possible d'ajouter d'autres critères de contrôle quantitatifs, l'équilibrage sur ces dernières variables se superposant alors au respect des quotas.

Par exemple, pour satisfaire des quotas marginaux par tranches d'âge et sexe, on considèrera les deux critères d'âge et de sexe à partir desquels on construira autant de variables indicatrices que nécessaire, 5 si on distingue 3 classes d'âge en sus des 2 strates de sexe. Les équations d'équilibrage s'écrivent alors :

$$\begin{cases} \displaystyle\sum_{k\in S}\frac{\delta_{k,sexe_h}}{\pi_k} = \sum_{k=1}^{N}\delta_{k,sexe_h} = & \text{Nombre d'individus de la base de sondage du sexe } h,\ h=1,2 \\ \displaystyle\sum_{k\in S}\frac{\delta_{k,age_h}}{\pi_k} = \sum_{k=1}^{N}\delta_{k,age_h} = & \text{Nombre d'individus de la base de sondage dans la tranche d'âge } h,\ h=1,2,3 \end{cases}$$

avec

$$\delta_{k,sexe_h} = \begin{cases} 1 & \text{si l'individu } k \text{ est du sexe } h\,,\ h = 1,2 \\ 0 & \text{sinon} \end{cases}$$

$$\delta_{k,age_h} = \begin{cases} 1 & \text{si l'individu } k \text{ est dans la tranche d'âge } h\,,\ h = 1,2,3 \\ 0 & \text{sinon} \end{cases}$$

L'exemple 6 vu en V-7 applique l'algorithme dans un cas analogue.

II. 4. Les limites de la macro CUBE[8]

a. Dans quels cas est-on assuré d'obtenir un échantillon exactement équilibré ?

Avec l'algorithme du CUBE, l'utilisateur n'est assuré d'un échantillon exactement équilibré que lorsque l'intersection entre le cube et l'espace des contraintes ne rencontre que des sommets du cube (comme dans l'exemple (i) ci-dessus et plus généralement dans le cas d'un plan simple, stratifié ou non, de taille fixe et toujours entière). Sinon :

- ou bien il n'existe pas de solution exacte car aucun échantillon possible ne peut satisfaire les contraintes (exemple (iii) supra),

- ou bien il existe un certain nombre de solutions exactes, mais parce que l'algorithme du CUBE produit des échantillons aléatoires, il ne désignera pas nécessairement un échantillon exact in fine (exemple (ii) supra).

b. Les limites informatiques de la représentation des nombres

Du fait des calculs intermédiaires réalisés par l'algorithme (par exemple élévation au carré de certains termes issus des équations d'équilibrage), la prudence s'impose lorsque le nombre de chiffres significatifs manipulé devient important. Au-delà de huit chiffres significatifs pour la partie entière, des informations risquent d'être perdues. En se propageant dans le calcul des probabilités d'inclusion géré par CUBE, cette perte d'information dégradera l'ensemble de la phase d'échantillonnage. Pour pallier cet écueil, lorsqu'elle suspecte un problème d'arrondi, la macro CUBE produit un message d'avertissement (comme dans l'exemple 8 décrit en V-9). Un remède consiste alors à changer d'unité, par exemple en divisant la variable d'équilibrage concernée par une puissance de 10 afin de ramener sa valeur dans une fourchette raisonnable de 10 à 10 000. Expédiant certes commun, mais il permet à l'algorithme de se dérouler convenablement en ne faisant pas porter la dégradation sur la partie entière des contraintes.

c. Quel est le nombre limite de contraintes ?

Lorsque le nombre de contraintes augmente, les calculs deviennent de plus en plus volumineux et le recours à la phase d'atterrissage se systématise. Dans sa version actuelle et compte-tenu des capacités informatiques, l'algorithme ne limite pas le nombre de contraintes avec l'option A d'atterrissage et il permet de traiter jusqu'à 19 contraintes avec l'option B et 21 avec l'option C. Il peut ainsi traiter des bases de sondage comptant jusqu'à 100 000 individus dans un délai « raisonnable » comme indiqué ci-dessous.

d. Comment évaluer la durée de fonctionnement de la macro ?

La durée de fonctionnement dépend essentiellement de la taille de la base de sondage et du nombre de contraintes. A titre d'ordre de grandeur, pour un tirage dans une population de taille *N,* contraint par *p* conditions, on peut estimer le temps de vol à environ *T* secondes où

$$T = 1{,}37 \times 10^{-6} \times p^{0{,}39} \times N^{1{,}9}$$

[8] Les performances indiquées ont toutes été obtenues sur un micro-ordinateur équipé d'un processeur INTEL CELERON cadencé à 1,7 Ghz disposant de 256 Mo de mémoire RAM.

Ce temps *T* estime également assez bien la durée globale de fonctionnement avec l'option A d'atterrissage. Pour l'option B, il faut ajouter au temps de vol une durée d'atterrissage fonction du nombre de contraintes. L'option C quant à elle fonctionne en un temps intermédiaire. On pourra évaluer, indépendamment de la fonction de coût utilisée, le temps d'atterrissage de l'option B selon les indications suivantes :

Nombre de contraintes	Temps d'atterrissage estimé
Moins de 3	1,5 s
De 3 à 12	De 1,5 à 3 s
15	30 s
18	15 min
19	56 min

Ces règles se vérifient assez bien pour des bases de sondage comptant jusqu'à 100 000 individus. Elle donne par exemple, un résultat en quelques minutes pour une base de sondage de quelques milliers d'individus et quelques contraintes (2 minutes environ pour 10 000 individus et 6 conditions). Avec 9 contraintes et l'option A d'atterrissage, on peut escompter un résultat en moins de 3 heures pour une base de 100 000 unités.

e. Comment traiter le cas des grandes bases de sondage ?

En l'état actuel du programme et étant donnés les performances du logiciel et les temps de calcul, il n'est pas conseillé d'échantillonner dans des bases de sondage dépassant 100 000 individus. Toutefois, ceci peut s'avérer incontournable, il est alors conseillé de procéder en plusieurs étapes :

i. Partitionner la base et conduire une phase de vol sur chacune des parties ;

ii. Récupérer, pour chaque partie, les individus sur lesquels la macro CUBE n'a pas statué. Ils se repèrent grâce à l'état du vecteur des probabilités d'inclusion en fin de vol (0 < PISTAR < 1) dans la table DATAFIC_SORTIE décrite au paragraphe IV-2 infra ;

iii. Réunir ces individus et relancer la macro CUBE en utilisant comme probabilités d'inclusion les valeurs de la variable PISTAR et en multipliant les variables d'équilibrage utilisées en (i) par $\frac{PISTAR}{PI}$ où PI désigne la variable des probabilités d'inclusion initiales (présente dans DATAFIC) ;

iv. Pondérer correctement les individus échantillonnés à l'issue de l'étape précédente en leur attribuant comme poids de sondage l'inverse des probabilités d'inclusion initiales PI. Pour inférer à la population, il convient en effet de modifier les pondérations livrées par CUBE à l'étape (iii) car celles-ci renvoient à la variable PISTAR et non aux probabilités initiales ;

v. Rassembler tous les individus échantillonnés et convenablement pondérés (étapes (i) et (iv)).

III. Syntaxe et paramètres

La macro CUBE a été développée avec la version 8 de SAS, sous l'environnement WINDOWS. Elle utilise les modules SAS/IML et SAS/GRAPH.

Pour la mettre en œuvre, il faut au préalable allouer la librairie qui contient le catalogue où se trouve la macro CUBE et la déclarer en début du programme SAS selon les instructions suivantes :

```
LIBNAME lib "c:\……\" ;                    /* Chemin du catalogue */
OPTIONS MSTORED SASMSTORE = lib ;          /* Options SAS indiquant la
                                           présence de macros compilées
                                           dans le catalogue */
```

Par ailleurs, au début de son exécution, la macro assigne les options suivantes à la session SAS en cours : NOMPRINT, NOSYMBOLGEN, NONOTES, LABEL et COMPRESS=YES. Par conséquent si,

après avoir lancé CUBE, l'utilisateur veut revenir à des options différentes, il doit utiliser l'instruction OPTIONS pour y parvenir.

Si l'algorithme détecte des anomalies, alors il les affiche dans la fenêtre « LOG » de SAS.

III.1. Syntaxe de la macro CUBE

```
%CUBE       (DATAFIC     =     ,
             ID          =     ,
             PI          =     ,
             CONTR       =     ,
             ATTER       =     ,
             COUT        =     ,
             POND        =     ,
             SORTIE      =     ,
             COMMENT     =      ) ;
```

L'écriture des paramètres respecte les règles suivantes :

- L'ordre des paramètres n'a pas d'importance ;
- Ils sont séparés par des virgules (et non des points virgules) ;
- Les paramètres obligatoires doivent être impérativement renseignés par l'utilisateur et respecter les différentes valeurs possibles quand une liste est proposée. Leur absence ou une valeur distincte de celles qui sont prévues provoque l'arrêt de la macro et l'émission de messages d'erreurs ad-hoc. Les paramètres qui possèdent des valeurs par défaut peuvent être omis ;
- Ils se conforment aux règles d'usage de SAS, en particulier les noms de variables ou de tables peuvent compter jusqu'à 32 positions et les noms de librairies jusqu'à 8 positions ;
- Ils peuvent s'écrire indifféremment en majuscules ou minuscules ;
- Ils peuvent comporter des caractères « blancs » (espaces) mais pas de virgule ni être mis explicitement à valeur manquante.

III.2. Paramètres relatifs à la base de sondage

Paramètre	Rôle	Statut
DATAFIC =	Désigne la base de sondage	Obligatoire
ID =	Désigne l'identifiant des individus	Obligatoire
PI =	Spécifie les probabilités d'inclusion individuelles	Obligatoire
CONTR =	Spécifie la(les) variable(s) d'équilibrage	Obligatoire

DATAFIC=*nom de table SAS*

Désigne la table SAS dans laquelle la procédure tire l'échantillon. Cette base de sondage doit contenir les probabilités d'inclusion, les variables d'équilibrage et l'identifiant de chaque individu. Elle doit aussi remplir les règles de qualité usuellement recherchées (exhaustivité, sans défaut de couverture ni de double compte, etc.). Elle peut aussi bien se trouver dans une librairie personnelle que dans la librairie temporaire de SAS.

ID=*nom de variable*

Désigne l'identifiant des individus dans la base de sondage. De type caractère ou numérique, sa valeur caractérise de manière unique chaque individu.

PI=*nom de variable*

Spécifie les probabilités d'inclusion individuelles. Elles ne doivent pas présenter de valeur manquante et doivent être comprises entre 0 et 1 :

- Si des unités ont une probabilité de sélection nulle, celles-ci ne participent pas à l'algorithme. Elles ne peuvent donc jamais figurer dans l'échantillon final et n'interviennent pas dans le calcul des totaux de référence des variables de contrôle.
- De même, les éventuels individus qui possèdent une probabilité d'inclusion égale à 1 ne participent pas à l'algorithme. Ils sont intégrés, d'office et in fine, à l'échantillon final avec une pondération égale à 1.
- Tout autre valeur en dehors de l'intervalle [0,1] cause l'interruption de la macro et l'émission d'un message d'avertissement. C'est le cas en particulier dès qu'une probabilité d'inclusion dépasse 1 (situation parfois obtenue avec un tirage à probabilités proportionnelles à une variable de taille).

CONTR=*nom(s) de variable(s)*
Spécifie la(les) variable(s) d'équilibrage, séparées par au moins un blanc s'il y a au moins 2 variables concernées. Elles sont de type numérique. Les variables qualitatives peuvent être prises en compte par l'intermédiaire de variables indicatrices d'appartenance aux différentes modalités (de façon à donner un sens à la notion de total). Les variables d'équilibrage ne doivent présenter de valeur manquante. La présence d'au moins une valeur manquante pour au moins une variable provoque l'interruption de l'algorithme et l'émission d'un message d'avertissement à l'utilisateur. En outre, chaque variable d'équilibrage ne doit figurer qu'une seule fois sous un même nom, sinon l'algorithme s'arrête et signale cette redondance. ***Avec l'option A d'atterrissage, les variables d 'équilibrage doivent figurer par ordre d'importance décroissante car le programme relâche d'abord les dernières variables citées.***

III.3. Paramètres de la phase d'atterrissage

Paramètre	Rôle	Valeur par défaut
ATTER =	Spécifie l'option d'atterrissage	B
COUT =	Spécifie le critère de coût	CV

ATTER =*option*
Désigne l'option retenue pour la phase d'atterrissage. Les valeurs possibles sont les suivantes :

`ATTER = A`
Les contraintes sont abandonnées successivement en commençant par la dernière variable citée dans le paramètre CONTR.

`ATTER = B`
L'atterrissage se fait par minimisation de l'espérance du coût calculée sur tous les échantillons réalisables à la fin de la phase de vol. Le coût se mesure selon le critère précisé par le paramètre COUT =. C'est l'option par défaut.

`ATTER = C`
L'atterrissage s'obtient par minimisation de l'espérance du coût calculée sur tous les échantillons réalisables et dont la taille est égale à la somme des probabilités d'inclusion à la fin de la phase de vol (sous-ensemble des échantillons possibles issus de l'option B). Le coût se mesure selon le critère précisé par le paramètre COUT =.

COUT =*critère*
Spécifie le critère de coût associé aux échantillons proposés par la phase d'atterrissage pour les options ATTER= B et ATTER=C. L'option ATTER=A ignore ce paramètre alors sans objet. La macro propose le choix entre :

COUT = CV

Il s'agit du critère par défaut. La phase d'atterrissage se résout en minimisant sous toutes les contraintes l'espérance de la somme des carrés des écarts relatifs comme indiqué au paragraphe II-2.

COUT = DIST

La phase d'atterrissage se résout en minimisant sous toutes les contraintes l'espérance du carré de la distance entre l'échantillon et sa projection sur le sous espace des contraintes comme indiqué au paragraphe II-2.

III.4. Paramètres relatifs à l'échantillon tiré

Paramètre	Rôle	Valeur par défaut	Statut
POND =	Spécifie la variable de pondération finale	POND	
SORTIE =	Spécifie le nom de la table en sortie qui contient l'échantillon tiré		Obligatoire

SORTIE = *nom de table SAS*

Désigne le nom de la table de sortie qui contient les individus échantillonnés. Elle se trouve dans la librairie temporaire de SAS (et non dans une librairie personnelle). Elle comprend toutes les variables de la base de sondage, plus des informations propres à l'échantillonnage : pondération finale, jugement rendu par la phase de vol et mention de l'appartenance à l'échantillon final. Cette variable indicatrice de la sélection vaut 1 pour toutes les unités et porte le même nom que la table SORTIE=. La pondération des unités est renseignée dans la variable POND= tandis que la variable PISTAR contient l'état du vecteur des probabilités d'inclusion individuelles en fin de vol.

POND = *nom de variable*

Désigne la variable de pondération des unités échantillonnées par CUBE. Elle est égale à l'inverse des probabilités d'inclusion initiales des unités sélectionnées. Cette variable figure dans la table SORTIE= . Son nom par défaut est POND.

III.5. Paramètre relatif aux commentaires

Paramètre	Rôle	Valeur par défaut
COMMENT=	Spécifie les commentaires à afficher	OUI

COMMENT = *affichage*

Spécifie l'affichage des commentaires.

COMMENT = OUI

Affiche un bilan des différentes étapes de l'échantillonnage qui permet notamment d'apprécier la qualité de l'équilibrage.

COMMENT = GRAPHE

Affiche les commentaires de l'option OUI ainsi que deux graphiques qui permettent de juger des caractéristiques de l'échantillon retenu par la phase d'atterrissage lorsque celle-ci procède par optimisation d'une fonction de coût. Cette option n'a donc de sens que pour les options d'atterrissage ATTER=B ou ATTER=C (si ATTER=A, les commentaires de l'option OUI sont affichés).

COMMENT = NON

Supprime l'affichage des commentaires et des graphiques.

IV. Les sorties

L'algorithme s'achève en ayant désigné un échantillon, exactement ou approximativement équilibré, qu'il décrit dans une table SAS. La macro crée également une table SAS en sortie qui contient tous les individus et toutes les variables de la base de sondage complétées d'informations propres à l'échantillonnage. Cette table pourra par exemple s'employer pour résoudre la problématique des grandes bases de sondage comme indiqué au paragraphe II-4. De plus, à la demande de l'utilisateur, deux autres sorties peuvent retracer le déroulement de l'échantillonnage : il s'agit de commentaires affichés dans la fenêtre OUTPUT et/ou de graphiques créés dans des fenêtres GRAPHE de SAS.

Par ailleurs, la durée de fonctionnement de l'algorithme s'affiche dans la fenêtre LOG (comme les éventuels messages d'avertissement envoyés à l'utilisateur).

IV.1. Table SAS listant l'échantillon tiré

Les unités sélectionnées sont listées dans la table SAS répondant au nom que lui a donné l'utilisateur dans le paramètre SORTIE=. Présente dans la librairie temporaire de SAS, cette table reprend pour les individus échantillonnés toutes les informations présentes dans la base de sondage, complétées des variables suivantes :

- L'indicatrice d'appartenance à l'échantillon. Cette variable porte le même nom que la table SORTIE= et vaut donc 1 pour toutes les unités.

- La pondération des unités tirées. Cette variable POND= du nom choisi par l'utilisateur (ou de POND attribué par défaut) est égale à l'inverse des probabilités d'inclusion initiales pour les unités sélectionnées.

- La variable PISTAR contenant l'état du vecteur des probabilités d'inclusion individuelles en fin de vol. Elle vaut 1 pour les unités de la base de sondage définitivement échantillonnées par la phase de vol ou un nombre strictement compris entre 0 et 1 pour les autres.

IV.2. Table SAS complétant la base de sondage

Par défaut, la macro CUBE crée dans la librairie temporaire une table SAS qui ajoute, dans la base de sondage, les informations relatives au déroulement de l'échantillonnage. Son nom reprend ceux spécifiés par l'utilisateur dans les paramètres DATAFIC= et SORTIE= concaténés en DATAFIC_SORTIE[9]. Elle enrichit donc la base de sondage des mêmes variables que celles présentes dans la table décrivant l'échantillon :

- L'indicatrice d'appartenance à l'échantillon. Cette variable porte le même nom que la table SORTIE= et vaut donc 1 pour les unités tirées et 0 sinon.

- La pondération des unités tirées. Cette variable POND= du nom choisi par l'utilisateur (ou de POND attribué par défaut) est égale à l'inverse des probabilités d'inclusion initiales pour les unités sélectionnées et vaut 0 pour les autres.

- La variable PISTAR contenant l'état du vecteur des probabilités d'inclusion individuelles en fin de vol. Elle vaut 0 pour les unités de la base de sondage définitivement écartées par la phase de vol, 1 pour celles définitivement échantillonnées à cette étape ou un nombre strictement compris entre 0 et 1 pour les autres.

[9] Afin que cette table se trouve dans la librairie temporaire, on supprime de DATAFIC la référence éventuelle à une librairie.

IV.3. Commentaires de la fenêtre OUTPUT

Si l'utilisateur a spécifié le paramètre COMMENT= OUI ou COMMENT= GRAPHE, il obtient dans la fenêtre OUTPUT de SAS un bilan commenté de l'échantillonnage. Ces sorties standards résument les différentes étapes du tirage depuis les contraintes introduites et l'issue de la phase de vol jusqu'aux résultats obtenus sur l'échantillon.

Elles rappellent le nombre *p* de variables d'équilibrage et la dimension *q* de l'espace des contraintes où $p \geq q$. Une différence éventuelle entre ces deux nombres résulte de contraintes colinéaires, ce qui s'observe quand des combinaisons linéaires de *q* contraintes permettent d'en déduire *p-q* autres.

Un bilan de la phase de vol précise le nombre d'individus pour lesquels elle n'a pas définitivement statué, effectif toujours inférieur ou égal à la dimension *q* de l'espace des contraintes. Le nombre d'individus qui reste à échantillonner s'obtient en sommant les composantes du vecteur « PISTAR » des individus pour lesquels la phase de vol n'a pas statué. Dans le cas où cette somme ne serait pas entière, un message avertit de l'éventualité d'une des configurations suivantes :

- Dans la base de sondage déjà, la somme des probabilités d'inclusion n'est pas entière (concevable en théorie, mais rare en pratique) ;

- L'utilisateur n'a pas explicitement demandé un échantillon de taille fixe : la variable « probabilité d'inclusion » ne figure pas comme variable d'équilibrage (scénario également autorisé, mais rare en pratique) ;

- Il y a un véritable problème d'arrondi lié à la manipulation de valeurs trop grandes. Il convient alors d'effectuer des corrections adéquates (comme un changement d'unités par exemple) pour remédier à cette situation.

Le commentaire précise également s'il a fallu faire appel à la phase d'atterrissage, l'option et le critère de coût utilisés s'il y a lieu, ainsi que la taille de l'échantillon final.

Pour juger de la fidélité de cet échantillon à chaque condition d'équilibrage, les résultats fournissent également l'estimation obtenue sur l'échantillon et la comparent à la vraie valeur de la base de sondage. Ainsi, pour chaque information auxiliaire de total Z estimé par $\hat{Z}_{HT}$ avec l'échantillon *s,* l'écart relatif entre ces quantités : $\frac{\hat{Z}_{HT}(s) - Z}{Z} \times 100$ est indiqué[10].

Exemple : d'après l'exemple 2 commenté au paragraphe V-3. ci-dessous

```
            BILAN DE VOTRE ECHANTILLONNAGE
_____________________________________________________________

Résumé des spécifications relatives aux contraintes
-----------------------------------------------------
Vos contraintes ne sont pas colinéaires :
    le nombre de variables d'équilibrage est    :             4
    pour un espace des contraintes de dimension :             4


A l'issue de la phase de vol
-----------------------------------------------
Nombre d'individus sur lesquels il reste à statuer :          4
Nombre d'individus qu'il reste à échantillonner    :          2
```

[10] Dans le cas où le vrai total *Z* d'une variable de contrôle est nul et si seule une phase de vol a eu lieu ou un atterrissage A, alors l'indicateur d'écart vaut 0 si $\hat{Z}_{HT}(s)$ est nul (l'estimation est correcte), sinon il vaut ' . '.

```
Résultats obtenus
--------------------------------------------------
Lancement de la phase d'atterrissage :                      OUI
Option de la phase d'atterrissage    :                        B
Critère de coût de la phase d'atterrissage    :              CV

Nombre d'individus échantillonnés    :                      500

1ère colonne : valeur réelle du total
2ème colonne : estimateur de Horvitz-Thompson du total
3ème colonne : écart relatif en pourcentage


                                          Ecart
Valeur                Valeur              relatif
réelle                estimée             (en %)

PISAS           500 | PISAS           500 | PISAS       0.00
PERSONNES     22133 | PERSONNES     22160 | PERSONNES   0.12
HOMMES        10747 | HOMMES        10760 | HOMMES      0.12
NB018          4201 | NB018          4200 | NB018      -0.02
```

IV.4. Sorties graphiques

Si l'utilisateur a spécifié le paramètre COMMENT= GRAPHE, il obtient deux sorties graphiques dans des fenêtres GRAPHE de SAS, en sus des commentaires standards de la fenêtre OUTPUT. Ces graphiques permettent d'apprécier les caractéristiques de l'échantillon retenu par la phase d'atterrissage lorsque celle-ci procède par optimisation d'une fonction de coût. Le choix de cet affichage se justifie donc seulement pour les options d'atterrissage ATTER=B ou ATTER=C. Rappelons en effet que lorsque la phase de vol n'a pas pu statuer sur tous les individus, ces deux options envisagent un certain nombre d'échantillons réalisables et leur associent un coût, fonction du critère spécifié pour mesurer la distance à l'équilibre. ***Les différents concepts représentés (taille, coût, probabilité) s'entendent donc conditionnellement à l'issue de la phase de vol.*** Chaque échantillon possible est repéré par un numéro incrémenté par taille croissante, puis par coût croissant pour une taille donnée.

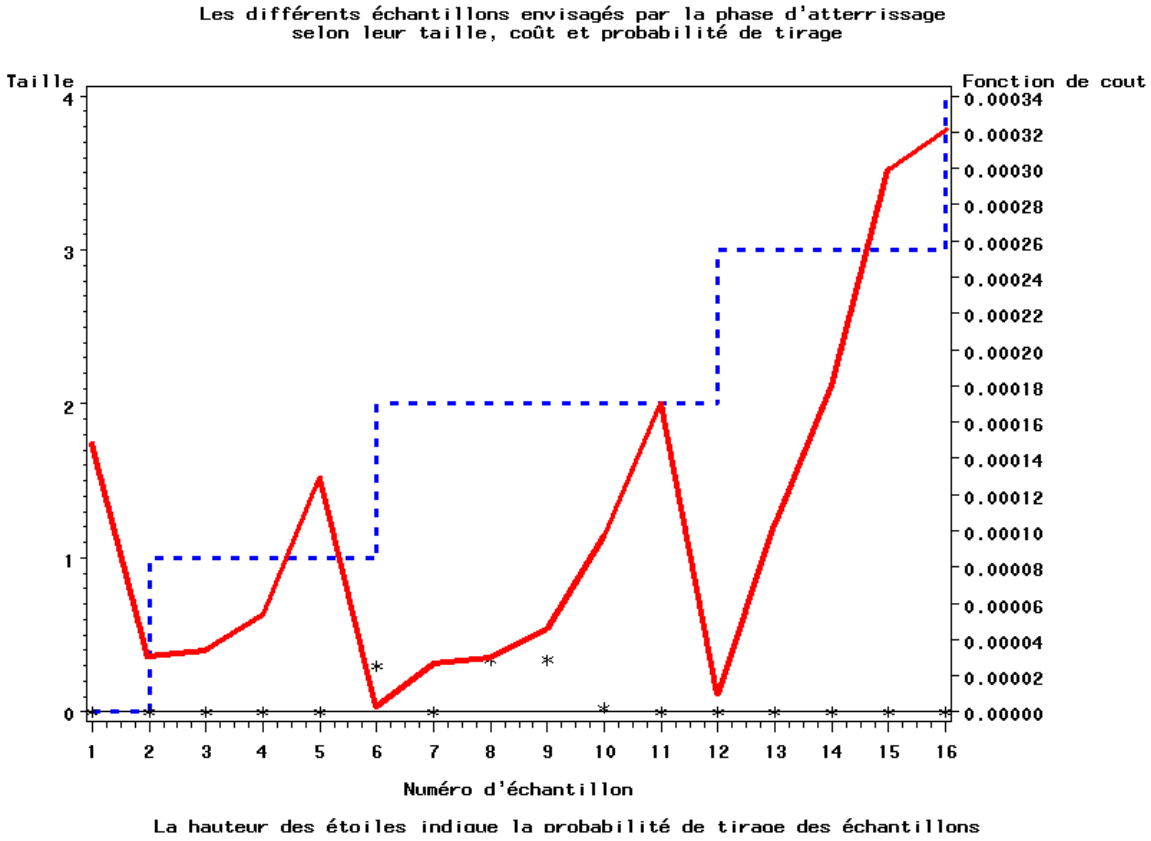
Un premier graphique intitulé *«Les différents échantillons envisagés par la phase d'atterrissage selon leur taille, coût et probabilité de tirage*» précise le coût de chacun des échantillons compatibles possibles ainsi que leur taille et probabilité de tirage. Sa lecture repose sur 3 axes : l'axe des abscisses liste les différents échantillons dans l'ordre croissant de la numérotation. L'axe des ordonnées à gauche précise à la fois la taille d'un échantillon et sa probabilité de tirage. L'axe des ordonnées à droite renseigne le coût d'un échantillon. Cette représentation permet de superposer les informations suivantes :

- un pointillé bleu trace une fonction en escalier continue à droite qui renvoie la taille d'un échantillon identifié par son numéro d'abscisse. La taille se lit sur l'axe des ordonnées à gauche. La courbe relie les différents points sans marquer de discontinuité.

- un tracé plein de couleur rouge donne les coûts des différents échantillons repérés par leurs abscisses respectives. Les coûts se lisent sur l'axe des ordonnées de droite et sont reliés entre eux sans marque de discontinuité. Du fait de la numérotation choisie, la courbe est croissante pour une taille donnée.

- des étoiles dont la hauteur indique la probabilité de tirage de l'échantillon considéré, lue sur l'axe des ordonnées à gauche. En particulier, les étoiles non confondues avec l'axe des abscisses repèrent les échantillons solutions du programme d'optimisation, donc ceux pour lesquels la probabilité de tirage est strictement positive. On distingue ainsi les échantillons plausibles de ceux qui s'avèrent irréalisables. Le second graphique détaille davantage cet aspect.

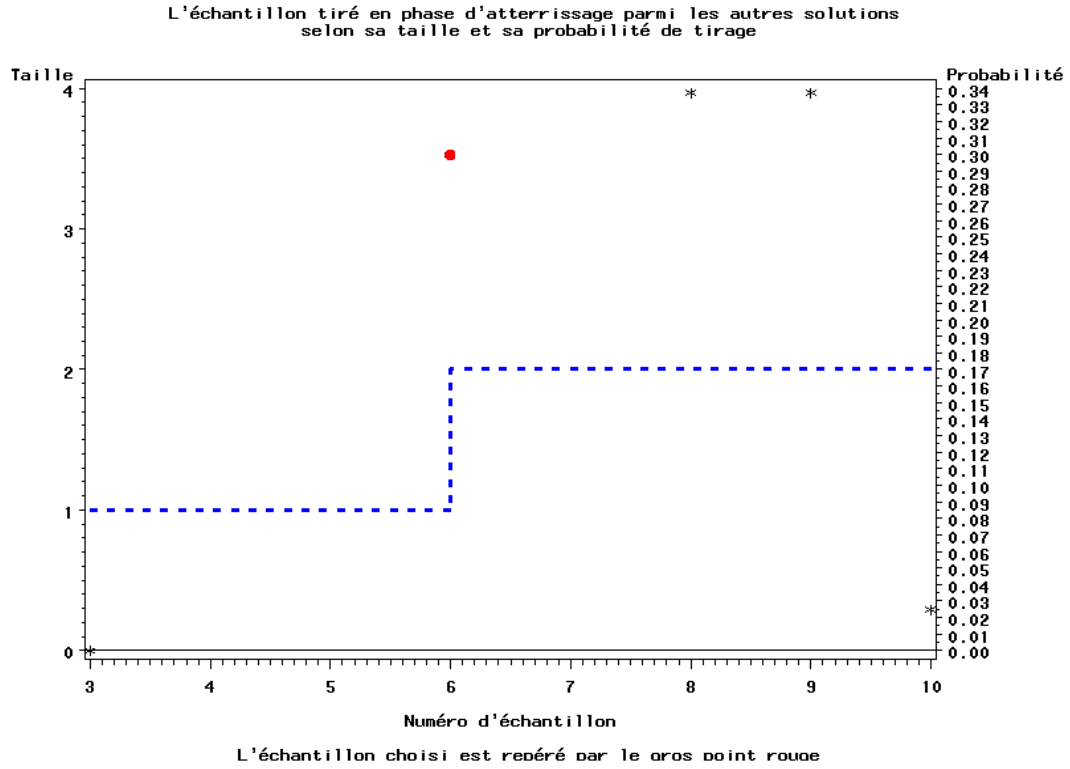
Exemple : d'après l'exemple 2 du paragraphe V-3 ci-dessous :



<u>Note de lecture</u> : l'échantillon n°6 est celui qui possède le plus petit coût parmi les échantillons tirables de taille 2. Les échantillons n°8, 9 et 10 comportent aussi 2 unités mais s'éloignent davantage des contraintes d'équilibrage (au sens du critère de coût respectivement voisin de 0,00004; 0,00005 et 0,0001). Le plan optimum n'accorde aucune possibilité de tirage aux autres échantillons compatibles de taille 2.

Le second graphique intitulé « *L'échantillon tiré en phase d'atterrissage parmi les autres solutions selon sa taille et probabilité de tirage* » met en évidence la probabilité de sélection des différents échantillons réalisables ainsi que leur taille. Il marque aussi l'échantillon retenu avec un gros point rouge distinctif. La représentation se restreint aux échantillons dont le numéro est compris entre ceux des premier et dernier échantillons de probabilité de tirage non nulle. L'axe des abscisses reprend la numérotation des échantillons ; l'axe des ordonnées à gauche précise la taille de chaque échantillon et celui de droite renseigne leur probabilité de tirage. L'échelle de chacun de ces axes est ajustée aux grandeurs possibles. On retrouve donc sur cette représentation le pointillé bleu de la courbe en escalier -continue à droite- qui se rapporte à la taille des différents échantillons. Et on lit en ordonnée à droite la probabilité de tirage d'un échantillon, symbolisée par des étoiles.

Exemple : d'après l'exemple 2 du paragraphe V-3 ci-dessous



Note de lecture : la phase d'atterrissage a ici retenu l'échantillon n°6 de taille 2 : l'échantillon final est donc composé de ces 2 unités et de celles sélectionnées au cours de la phase de vol. Sachant cette issue particulière du vol, il y avait 3 chances sur 10 d'obtenir l'échantillon n°6. L'échantillon n°3 de taille 1 est tirable mais cela se produit seulement dans de très rares cas.

V. Quelques exemples commentés

V.1. Description de la base de sondage employée

Les prochains exemples utilisent une base de sondage de 10 000 ménages issue du Recensement de la Population de 1999. L'unité statistique renvoie à un ménage, assimilé au logement qu'occupe à titre de résidence principale un ensemble de personnes, quels que soient les liens qui les unissent. Des données socio-démographiques décrivent ainsi cette population :

```
proc means data=bs n mean sum std ;
var un personnes hommes femmes nb018 nb1825 nb2540 nb4060 nb60;
run;
```

The MEANS Procedure

Variable	Label	N	Mean	Sum	Std Dev
un	constante 1	10000	1.0000000	10000.00	0
personnes	nombre de personnes du ménage	10000	2.2133000	22133.00	1.2492234
hommes	nombre d'hommes	10000	1.0747000	10747.00	0.8447433

femmes	nombre de femmes	10000	1.1386000	11386.00	0.7635760
nb018	nombre de personnes de moins de 18 ans	10000	0.4201000	4201.00	0.8595870
nb1825	nombre de personnes de 18 à 25 ans	10000	0.1582000	1582.00	0.4284753
nb2540	nombre de personnes de 25 à 40 ans	10000	0.4230000	4230.00	0.6983694
nb4060	nombre de personnes de 40 à 60 ans	10000	0.5987000	5987.00	0.7965687
nb60	nombre de personnes de plus de 60 ans	10000	0.6133000	6133.00	0.7730607

V.2. Exemple 1 SAS avec atterrissage A

Plan aléatoire simple sans remise de 500 ménages équilibré sur la taille de l'échantillon et celle de la population de ménages, les nombres totaux de personnes, d'hommes et d'individus de moins de 18 ans - Option d'atterrissage A

1- Introduction des probabilités d'inclusion individuelles dans la base de sondage

```
data bs;
set bs;
pisas=500/10000;
run;
```

2- *Lancement de la macro cube*

```
%CUBE (Datafic=bs,
       Id=ident,
       Pi=pisas,
       Pond=wa,
       Contr=pisas personnes hommes nb018 un,
       Sortie=echa,
       Atter=A,
       Comment = OUI);
```

3- ***Résultats*** *(en 1 min 58 s)*

```
            BILAN DE VOTRE ECHANTILLONNAGE
_____________________________________________________________

Résumé des spécifications relatives aux contraintes
-----------------------------------------------------
Certaines de vos contraintes sont colinéaires :
    le nombre de variables d'équilibrage est    :             5
    pour un espace des contraintes de dimension :             4


A l'issue de la phase de vol
-----------------------------------------------------
Nombre d'individus sur lesquels il reste à statuer :          4
Nombre d'individus qu'il reste à échantillonner    :          1


Résultats obtenus
-----------------------------------------------------
Lancement de la phase d'atterrissage :                      OUI
Option de la phase d'atterrissage    :                        A
```

```
Nombre d'individus échantillonnés    :                      500

1ère colonne : valeur réelle du total
2ème colonne : estimateur de Horvitz-Thompson du total
3ème colonne : écart relatif en pourcentage


                                          Ecart
Valeur                Valeur              relatif
réelle                estimée             (en %)

PISAS           500 | PISAS           500 | PISAS       0.00
PERSONNES     22133 | PERSONNES     22140 | PERSONNES   0.03
HOMMES        10747 | HOMMES        10740 | HOMMES     -0.07
NB018          4201 | NB018          4180 | NB018      -0.50
UN            10000 | UN            10000 | UN          0.00
```

Commentaires

Deux contraintes font double emploi ici : obtenir un échantillon de taille fixe équivaut à restituer l'effectif de la population (de ménages) pour un tirage sans remise à probabilités égales. Aussi ne reste-t-il que 4 conditions non colinéaires.
Comme dans la plupart des cas, la phase de vol s'est achevée ici en laissant une situation indécise pour un nombre d'individus égal à la dimension de l'espace des contraintes. Si elle n'a pas suffi pour désigner un échantillon équilibré sur toutes les contraintes, cette phase a déjà réglé le sort de 9 996 ménages, définitivement sélectionnés ou écartés. Elle laisse le soin à la phase d'atterrissage de choisir un ménage parmi les 4 candidats sur lesquels elle n'a pu statuer.
Avec l'option A, la phase d'atterrissage relâche les contraintes les unes après les autres en commençant par la dernière variable citée dans le paramètre CONTR= jusqu'à parvenir à un échantillon équilibré. Ici, on s'est donc successivement affranchi de l'équilibrage sur le nombre total de ménages, l'effectif total d'individus de moins de 18 ans, d'hommes et enfin de personnes. Mais comme l'utilisateur a souhaité en priorité un échantillon de taille fixe, contrainte respectée et redondante avec la dernière condition, celle-ci sera fort logiquement parfaitement estimée aux dépens des contraintes qui lui apparaissaient pourtant prioritaires.

V.3. Exemple 2 SAS avec atterrissage B

Plan aléatoire simple sans remise de 500 ménages équilibré sur la taille de l'échantillon, les nombres totaux de personnes, d'hommes et d'individus de moins de 18 ans - Option d'atterrissage B et critère de coût CV

Les probabilités d'inclusion valent toutes 500 / 10 000 (introduites dans l'exemple 1 supra).

1- Lancement de la macro cube

```
%CUBE (Datafic=bs,
       Id=ident,
       Pi=pisas,
       Pond=wb,
       Contr=pisas personnes hommes nb018,
       Sortie=echb,
       Cout=CV,
       Atter=B,
       Comment = GRAPHE);
```

2- Résultats (obtenus en 2 min 01 s)

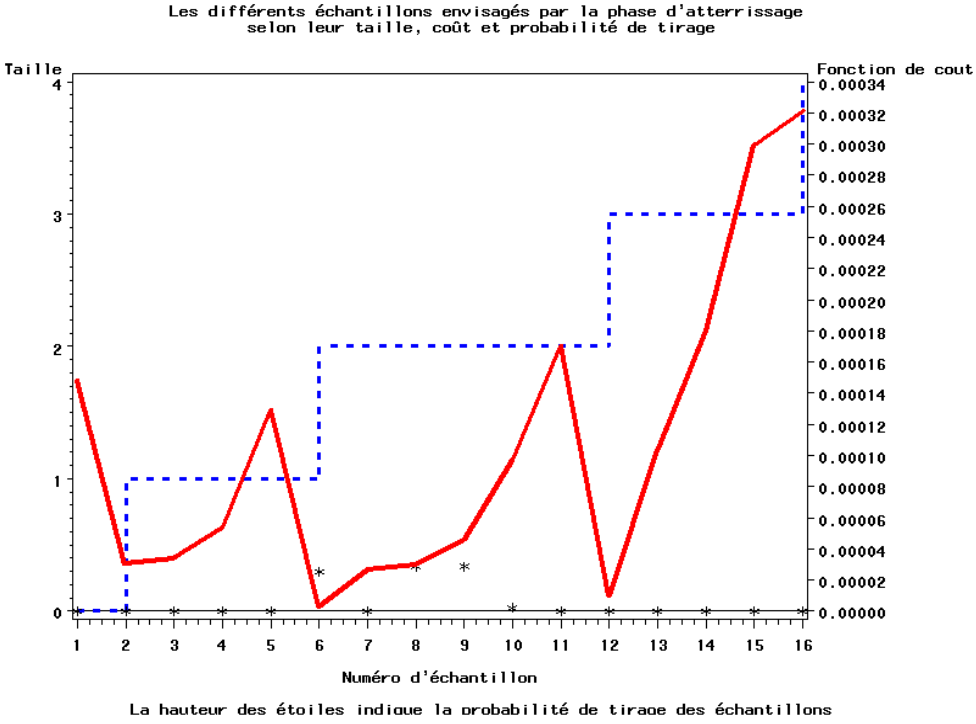
```
            BILAN DE VOTRE ECHANTILLONNAGE
_____________________________________________________________

Résumé des spécifications relatives aux contraintes
-----------------------------------------------------
Vos contraintes ne sont pas colinéaires :
    le nombre de variables d'équilibrage est    :             4
    pour un espace des contraintes de dimension :             4


A l'issue de la phase de vol
-----------------------------------------------------
Nombre d'individus sur lesquels il reste à statuer :          4
Nombre d'individus qu'il reste à échantillonner    :          2


Résultats obtenus
-----------------------------------------------------
Lancement de la phase d'atterrissage :                      OUI
Option de la phase d'atterrissage          :                 B
Critère de coût de la phase d'atterrissage :                CV

Nombre d'individus échantillonnés    :                     500

1ère colonne : valeur réelle du total
2ème colonne : estimateur de Horvitz-Thompson du total
3ème colonne : écart relatif en pourcentage


                                          Ecart
Valeur                Valeur              relatif
réelle                estimée             (en %)

PISAS           500 | PISAS           500 | PISAS       0.00
PERSONNES     22133 | PERSONNES     22160 | PERSONNES   0.12
HOMMES        10747 | HOMMES        10760 | HOMMES      0.12
NB018          4201 | NB018          4200 | NB018      -0.02
```

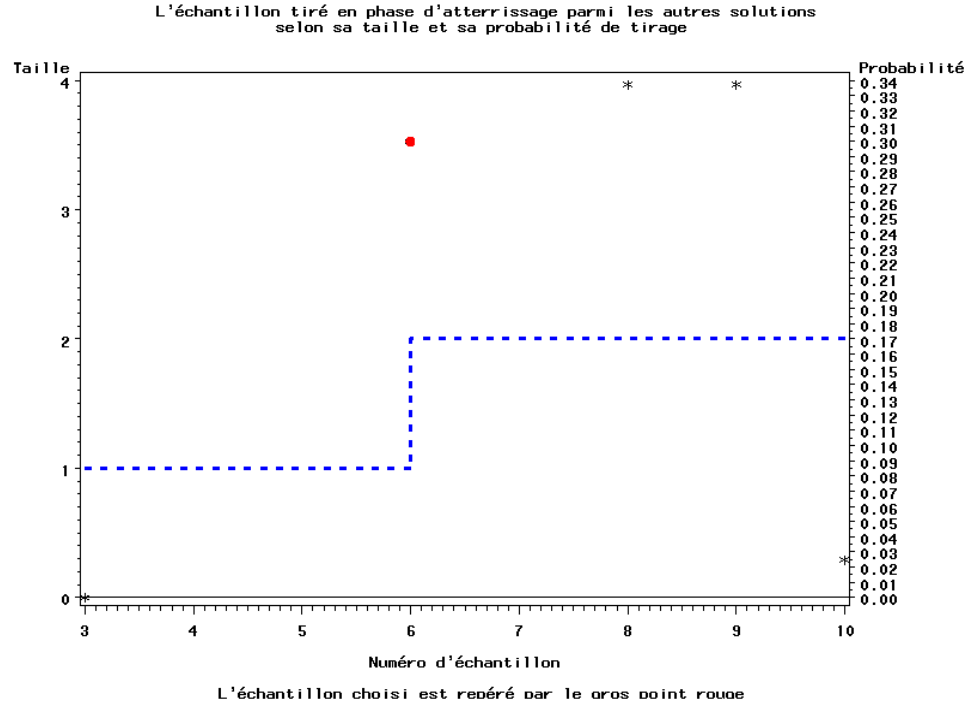
Commentaires

Le compte-rendu signale l'absence de redondance entre les 4 variables d'équilibrage. Comme dans la plupart des cas, la phase de vol s'est donc achevée en laissant une situation indécise pour un nombre d'unités égal à la dimension de l'espace des contraintes. La phase d'atterrissage s'est donc appliquée à ces 4 candidats, parmi lesquels il faut en retenir 2 afin de former un échantillon final de 500 ménages. Un tel scénario n'est pas garanti avec l'option B qui traite toutes les contraintes de manière similaire, sans accorder de préférence particulière à l'une d'entre elles. Le déroulement de l'atterrissage est commenté ci-dessous au vu des sorties graphiques. In fine, la macro a désigné un échantillon de 500 ménages qui se conforme donc à la taille demandée mais s'écarte légèrement des autres contraintes.

3- Sorties graphiques

Comme la phase de vol n'a pas statué sur 4 ménages, l'algorithme a envisagé les $2^4 = 16$ échantillons compatibles possibles. Il a calculé la taille et le coût de chacun de ses échantillons à partir du critère CV et a déduit le plan de sondage optimisant l'atterrissage. Les probabilités de tirage ainsi décernées autorisent seulement la sélection des échantillons n° 3, 6, 8, 9 et 10, de taille 1 ou 2. On constate ainsi sur cet exemple que la contrainte de taille fixe, traitée comme une autre par l'option B, ne peut de fait être toujours garantie (cas du n°3, au demeurant de probabilité si faible qu'elle atteint la limite de visibilité permise par l'échelle du graphique). Dans le cas présent, l'algorithme a désigné l'échantillon n°6, à la fois de probabilité de tirage élevée et de coût minimal.





V.4. Exemple 3 SAS avec atterrissage C

Plan aléatoire simple sans remise de 500 ménages équilibré sur la taille de l'échantillon, les nombres totaux de personnes, d'hommes et d'individus de moins de 18 ans - Option d'atterrissage C et critère de coût CV

Les probabilités d'inclusion valent toutes 500 / 10 000 (introduites dans l'exemple 1 supra).

1- Lancement de la macro cube

```
%CUBE (Datafic=bs,
       Id=ident,
       Pi=pisas,
       Pond=wc,
       Contr=pisas personnes hommes nb018,
       Sortie=outc,
       Cout=CV,
       Atter=C,
       Comment = GRAPHE);
```

2- Résultats (en 1min 59 s)

```
            BILAN DE VOTRE ECHANTILLONNAGE
_____________________________________________________________

Résumé des spécifications relatives aux contraintes
-----------------------------------------------------
Vos contraintes ne sont pas colinéaires :
    le nombre de variables d'équilibrage est    :              4
    pour un espace des contraintes de dimension :              4


A l'issue de la phase de vol
-----------------------------------------------------
Nombre d'individus sur lesquels il reste à statuer :           4
Nombre d'individus qu'il reste à échantillonner    :           2


Résultats obtenus
-----------------------------------------------------
Lancement de la phase d'atterrissage :                       OUI
Option de la phase d'atterrissage          :                   C
Critère de coût de la phase d'atterrissage :                  CV

Nombre d'individus échantillonnés    :                      500

1ère colonne : valeur réelle du total
2ème colonne : estimateur de Horvitz-Thompson du total
3ème colonne : écart relatif en pourcentage


                                          Ecart
Valeur                Valeur              relatif
réelle                estimée             (en %)

PISAS           500 | PISAS           500 | PISAS       0.00
PERSONNES     22133 | PERSONNES     22140 | PERSONNES   0.03
HOMMES        10747 | HOMMES        10740 | HOMMES     -0.07
NB018          4201 | NB018          4200 | NB018      -0.02
```

Commentaires

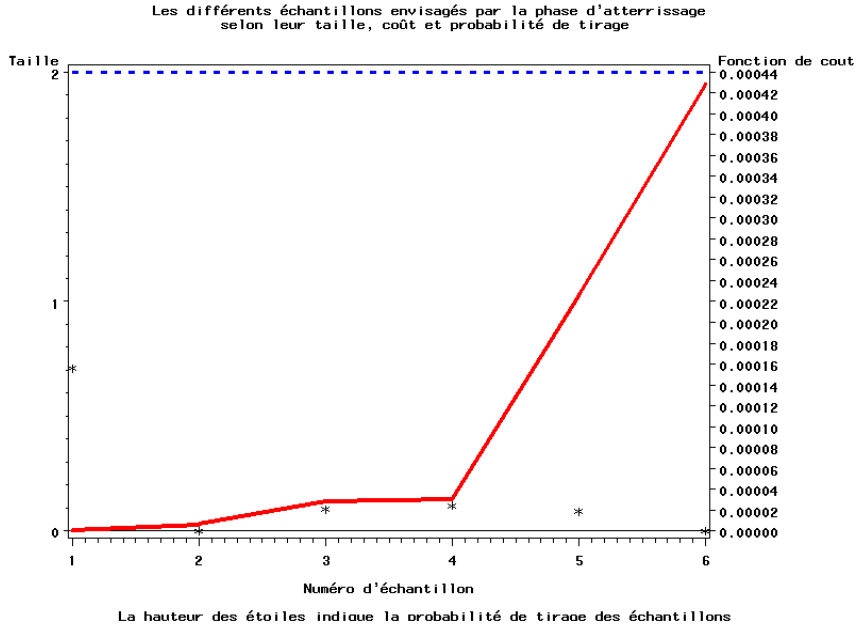
Les 4 variables d'équilibrage spécifiées n'étant pas colinéaires, elles forment donc un espace des contraintes de dimension 4. Comme avec la méthode B de l'exemple précédent, la phase de vol s'achève en laissant le sort de 4 ménages incertain, nombre à nouveau égal à la dimension de l'espace des contraintes. Comme l'option C sélectionne un échantillon de la taille voulue, la phase d'atterrissage doit choisir 2 ménages parmi les 4 candidats à la sélection. Son déroulement est commenté ci-dessous au vu des sorties graphiques. In fine, la macro a bien livré un échantillon de 500 ménages

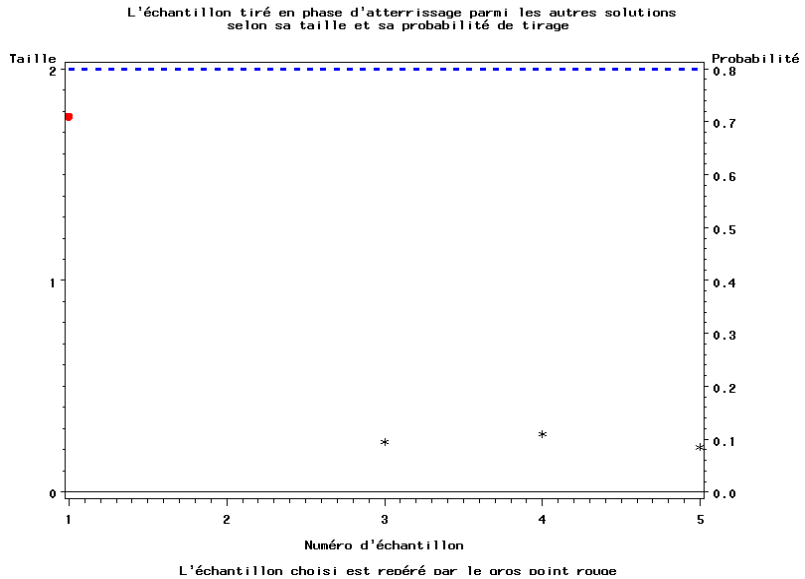
conformément à la taille voulue, mais par contre, les différents effectifs utilisés pour l'équilibrage sont légèrement sur ou sous estimés selon les cas.

3- Sorties graphiques

Avec l'option d'atterrissage C, comme la phase de vol n'a pas statué sur 4 ménages et qu'il en reste 2 à sélectionner pour respecter l'allocation voulue, l'algorithme se restreint donc aux échantillons de taille 2 parmi les 2^4 = 16 échantillons compatibles possibles. Il envisage ainsi $C_4^2 = 6$ échantillons distincts dont il calcule le coût à partir du critère CV spécifié.

Les graphiques illustrent le plan de sondage qui optimise l'atterrissage. Le premier précise le coût de des 6 échantillons possibles ainsi que leur taille et probabilité de tirage dans le plan optimum. Naturellement ici la taille vaut toujours 2 comme le montre le tracé en pointillé. On constate par ailleurs que les échantillons n°2 et 6 ne possèdent aucune chance d'être retenus. Le 1er échantillon, de loin le plus vraisemblable et aussi le plus proche des contraintes, a été retenu comme le montre le second graphique. Cependant, de façon moins probable par définition, le hasard aurait pu en désigner un autre. Privilégié par le programme d'optimisation qui lui a attribué la plus grande probabilité de sélection, l'échantillon n°1 offre de grandes chances d'aboutir à un équilibrage de bonne qualité.





V.5. Exemple 4 SAS de petite taille

Sondage aléatoire simple sans remise de 10 ménages équilibré sur la taille de l'échantillon le nombre total de personnes, d'hommes et d'individus de moins de 18 ans – Option d'atterrissage B et critère de coût DIST

1- Introduction des probabilités d'inclusion individuelles dans la base de sondage

```
data bs4;
set bs;
pi4=10/10000;
run;
```

2- Lancement de la macro cube

```
%CUBE (Datafic=bs4,
       Id=ident,
       Pi=pi4,
       Pond=w4,
       Contr=pi4 personnes hommes nb018,
       Sortie=ech4,
       Cout=DIST,
       Atter=B,
       Comment = Graphe);
```

3- Résultats (en 2 min 04 s)

```
            BILAN DE VOTRE ECHANTILLONNAGE
_____________________________________________________________

Résumé des spécifications relatives aux contraintes
-----------------------------------------------------
Vos contraintes ne sont pas colinéaires :
    le nombre de variables d'équilibrage est    :              4
    pour un espace des contraintes de dimension :              4


A l'issue de la phase de vol
-----------------------------------------------------
Nombre d'individus sur lesquels il reste à statuer :           4
Nombre d'individus qu'il reste à échantillonner    :           2


Résultats obtenus
-----------------------------------------------------
Lancement de la phase d'atterrissage :                       OUI
Option de la phase d'atterrissage          :                   B
Critère de coût de la phase d'atterrissage :                  CV

Nombre d'individus échantillonnés    :                       10

1ère colonne : valeur réelle du total
2ème colonne : estimateur de Horvitz-Thompson du total
3ème colonne : écart relatif en pourcentage


                                        Ecart
```

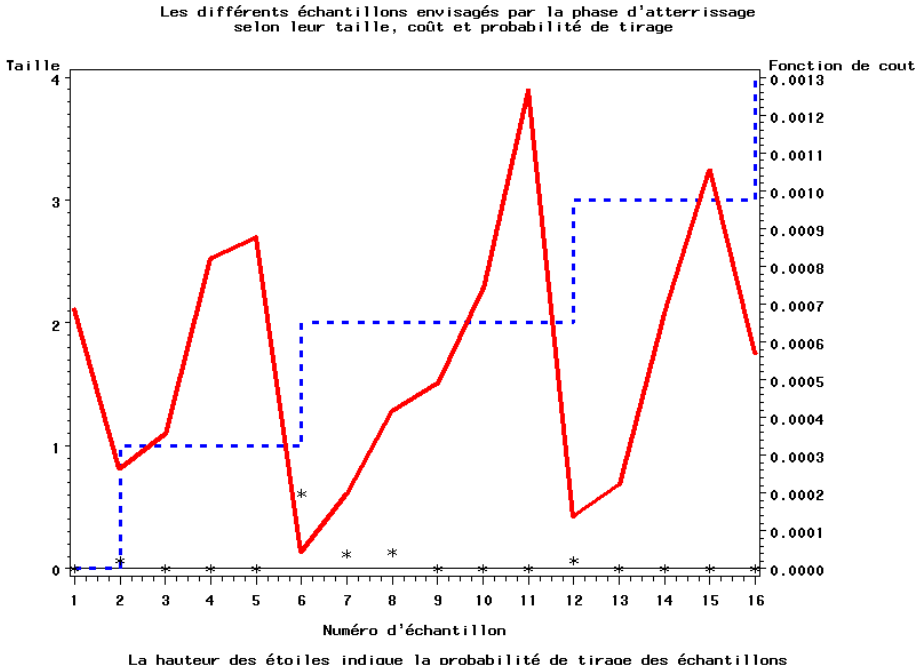
Valeur réelle		Valeur estimée		relatif (en %)	
PI4	10	PI4	10	PI4	0.00
PERSONNES	22133	PERSONNES	22000	PERSONNES	-0.60
HOMMES	10747	HOMMES	11000	HOMMES	2.35
NB018	4201	NB018	4000	NB018	-4.78

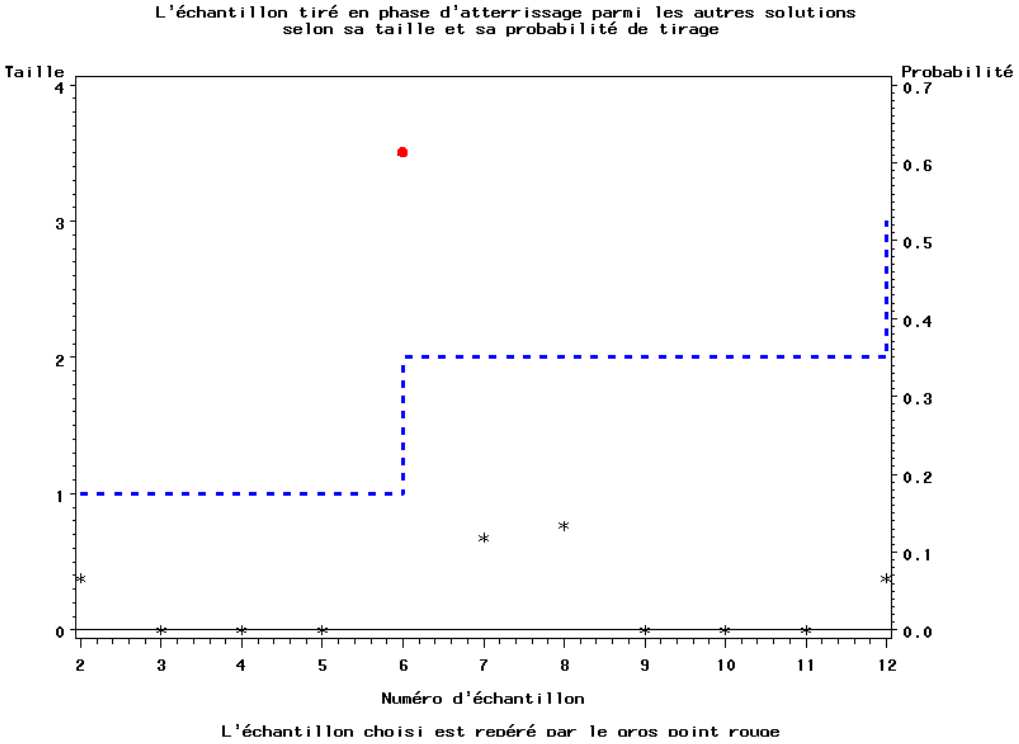
Commentaires

Cet exemple reprend les contraintes d'équilibrage envisagées dans les exemples 1 à 3 ci-dessus, mais cette fois-ci on veut sélectionner 10 individus seulement (et non 500) qui doivent être « représentatifs» de l'ensemble de la base de sondage en terme de taille de population, d'effectifs d'hommes et d'individus de moins de 18 ans. L'équilibrage est donc beaucoup plus difficile à atteindre.
A l'issue de la phase de vol, il reste 4 ménages sur lesquels statuer parmi lesquels il faudrait en sélectionner 2 pour vérifier la contrainte de taille fixe. Dans le cas présent, il s'avère que la phase d'atterrissage a effectivement retenu 2 ménages. De fait, l'échantillon final comporte donc 10 ménages comme souhaité. Les écarts relatifs sont naturellement sensiblement dégradés par rapport aux exemples précédents, mais leurs ordres de grandeurs restent relativement acceptables.

4- Sorties graphiques

L'option B envisage les 2^4 =16 échantillons compatibles avec l'issue du vol, calcule leur coût, taille puis probabilité de tirage dans le plan optimum. Celui-ci accorde seulement aux échantillons numérotés 2, 6, 7, 8 et 12 une réelle possibilité de tirage, en privilégiant le 6ème échantillon qui possède le coût le plus faible. Dans plus de 60% des cas, les deux ménages composant cet échantillon s'ajoutent à ceux déjà retenus par la phase de vol. Il se trouve que l'aléa du tirage l'a effectivement désigné ici.





.

V.6. Exemple 5 Plan PPT

Echantillonnage sans remise de 500 ménages à probabilités proportionnelles à la taille des ménages et équilibré sur la taille de l'échantillon (donc sur la population totale d'individus),le nombre total de ménages, les effectifs totaux des hommes, de personnes de moins de 18 ans, entre 18 et 25 ans, 25 et 40 ans, 40 et 60 ans (donc de plus de 60 ans aussi par déduction) - Option d'atterrissage B et critère de coût CV

1- Introduction des probabilités d'inclusion individuelles dans la base de sondage

```
data bs5;
set bs;
pippt=500*personnes/22133;
run;
```

2- Lancement de la macro cube

```
%CUBE (Datafic=bs5,
       Id=ident,
       Pi=pippt,
       Pond=w5,
       Contr=pippt un hommes nb018 nb1825 nb2540 nb4060,
       Sortie=ech5,
       Cout=CV,
       Atter=B,
       Comment = Graphe);
```

3 - Résultats (en 2 min 19 s)

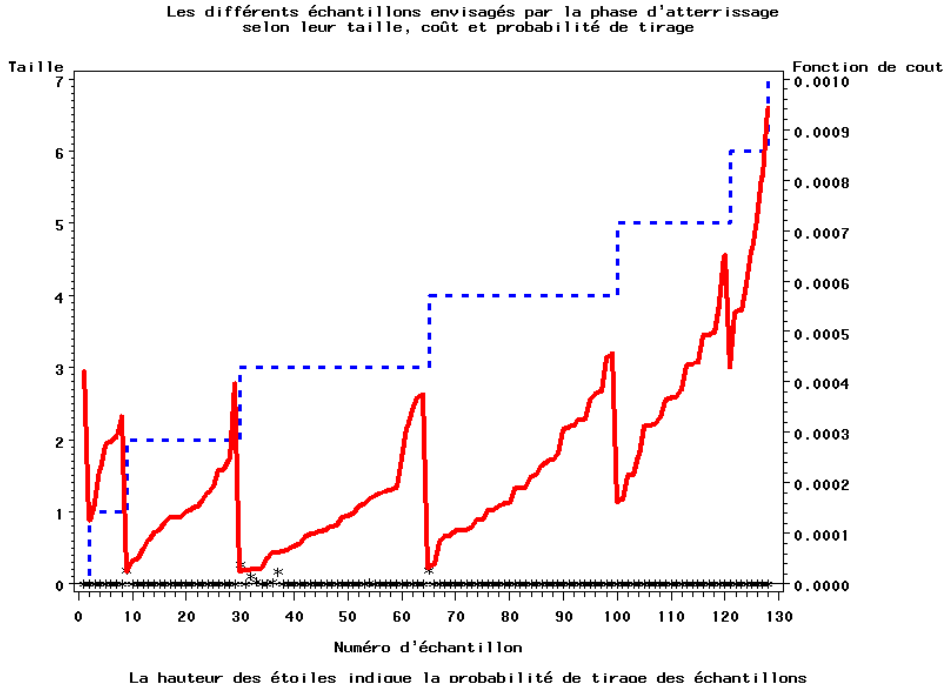
```
            BILAN DE VOTRE ECHANTILLONNAGE
_____________________________________________________________

Résumé des spécifications relatives aux contraintes
-----------------------------------------------------
Vos contraintes ne sont pas colinéaires :
    le nombre de variables d'équilibrage est    :             7
    pour un espace des contraintes de dimension :             7


A l'issue de la phase de vol
-----------------------------------------------------
Nombre d'individus sur lesquels il reste à statuer :          7
Nombre d'individus qu'il reste à échantillonner    :          3


Résultats obtenus
-----------------------------------------------------
Lancement de la phase d'atterrissage :                      OUI
Option de la phase d'atterrissage          :                  B
Critère de coût de la phase d'atterrissage :                 CV

Nombre d'individus échantillonnés    :                      500

1ère colonne : valeur réelle du total
2ème colonne : estimateur de Horvitz-Thompson du total
3ème colonne : écart relatif en pourcentage


                                    Ecart
Valeur             Valeur           relatif
réelle             estimée          (en %)

PIPPT        500 | PIPPT         500 | PIPPT    -0.00
UN         10000 | UN       9973.4109 | UN       -0.27
HOMMES     10747 | HOMMES   10764.051 | HOMMES    0.16
NB018       4201 | NB018    4212.1207 | NB018     0.26
NB1825      1582 | NB1825   1583.9499 | NB1825    0.12
NB2540      4230 | NB2540    4201.546 | NB2540   -0.67
NB4060      5987 | NB4060   5986.0982 | NB4060   -0.02
```

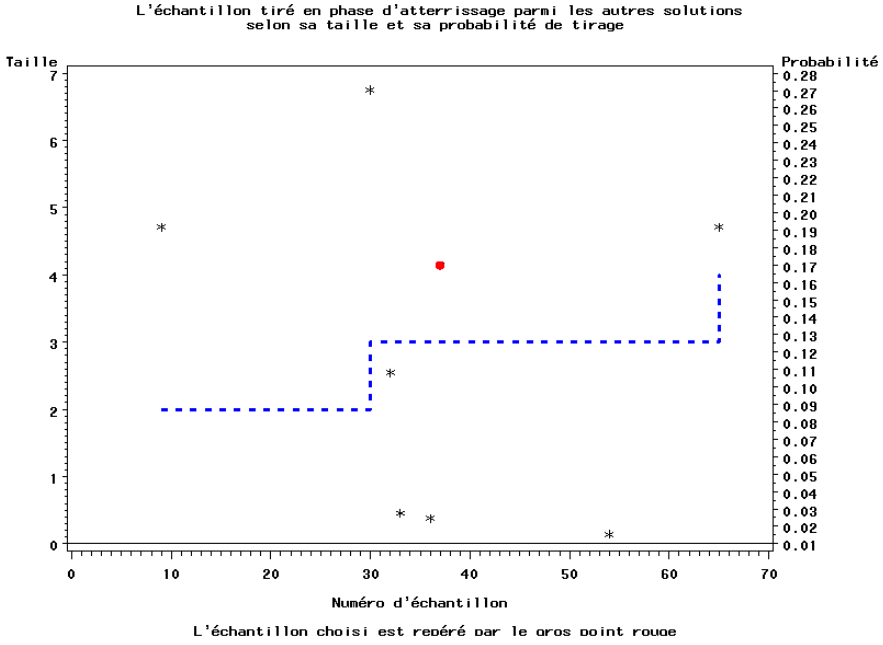
Commentaires

Il n'y a pas de colinéarité parmi les 7 variables d'équilibrage et à l'issue de la phase de vol, la dimension de l'espace des contraintes correspond, une nouvelle fois, au nombre de ménages sur lesquels il reste à statuer. La phase d'atterrissage devrait désigner 3 ménages au hasard parmi ces 7 candidats pour que soit respectée la contrainte de taille fixe, traitée sans priorité particulière par l'option B. Si l'utilisateur avait tenu expressément à obtenir un échantillon de taille fixe, il aurait eu intérêt à employer l'une des deux autres options. Sur cet exemple, on obtient bien un échantillon de 500 ménages, approximativement calibré sur les autres caractéristiques.

4- Sorties graphiques

La phase d'atterrissage considère les 2^7 = 128 échantillons compatibles avec la phase de vol, calcule leur taille, coût (critère CV) et probabilité de tirage à l'issue de la phase de vol. Au vu des graphiques, le plan optimum n'envisage qu'un petit nombre d'échantillons vraiment réalisables. Tous comportent 2 ou 3 ménages, ce qui ne permet pas toujours de respecter la contrainte de taille fixe, à une unité près. Dans le cas présent, la solution choisie correspond à l'échantillon n°37 qui ajoute 3 ménages à ceux désignés en cours de vol. In fine, l'allocation voulue est bien respectée.





V.7. Exemple 6 Plan stratifié

Tirage sans remise de 500 ménages répartis dans trois strates de densités d'habitat selon la tranche d'unité urbaine au recensement de 1999(*communes rurales, unités urbaines de moyenne ou grande taille)* avec des probabilités égales dans chacune des strates. Les contraintes d'équilibrage portent, dans chaque strate, sur la taille de l'échantillon de ménages et le nombre total d'individus de moins de 18 ans - Option d'atterrissage C et critère de coût CV

1- *Introduction des probabilités d'inclusion individuelles dans la base de sondage, calcul des indicatrices de strate et des contraintes ad-hoc*

```
data bs6;
set bs;
tu1=(tu99='0');
tu2=((tu99 >'0') and (tu99<'5'));
tu3=(tu99>='5');
if tu1=1 then pi=170/4622;
         else if tu2=1 then pi=165/2695;
                       else if tu3=1 then pi=165/2683;
run;

%macro disjonct (var,strate,nbstrate);
%do i=1 %to &nbstrate;
data bs6;
set bs6;
&var&strate&i =&var * &strate&i ;
run;
%end;
%mend;

%disjonct(pi,tu,3);
%disjonct(nb018,tu,3);
```

A l'issue de cette étape, les variables d'équilibrage ad-hoc ont été créées. Par exemple, pour que l'échantillon restitue le nombre de ménages dans chaque strate d'habitat, on calcule les variables nécessaires en multipliant les probabilités d'inclusion individuelles par les indicatrices d'appartenance aux différentes strates. On définit ainsi la variable *« pitu1 »* comme produit des probabilités d'inclusion fournies dans *« pi »* et de l'indicatrice d'appartenance à la strate des communes rurales *(« tu1 »)*. Un échantillon équilibré sur *« pitu1 »* permet de retrouver le nombre de ménages ruraux présents dans la base de sondage.

2- *Lancement de la macro cube*

```
%CUBE   (Datafic=bs6,
         Id=ident,
         Pi=pi,
         Pond=w6,
         Contr=pitu1 pitu2 pitu3 nb018tu1 nb018tu2 nb018tu3,
         Sortie=ech6,
         Cout=CV,
         Atter=C,
         Comment = Graphe);
```

3- ***Résultats*** *(en 2 min 12 s)*

```
                BILAN DE VOTRE ECHANTILLONNAGE
_____________________________________________________________

   Résumé des spécifications relatives aux contraintes
   ---------------------------------------------------
   Vos contraintes ne sont pas colinéaires :
       le nombre de variables d'équilibrage est    :            6
       pour un espace des contraintes de dimension :            6
```

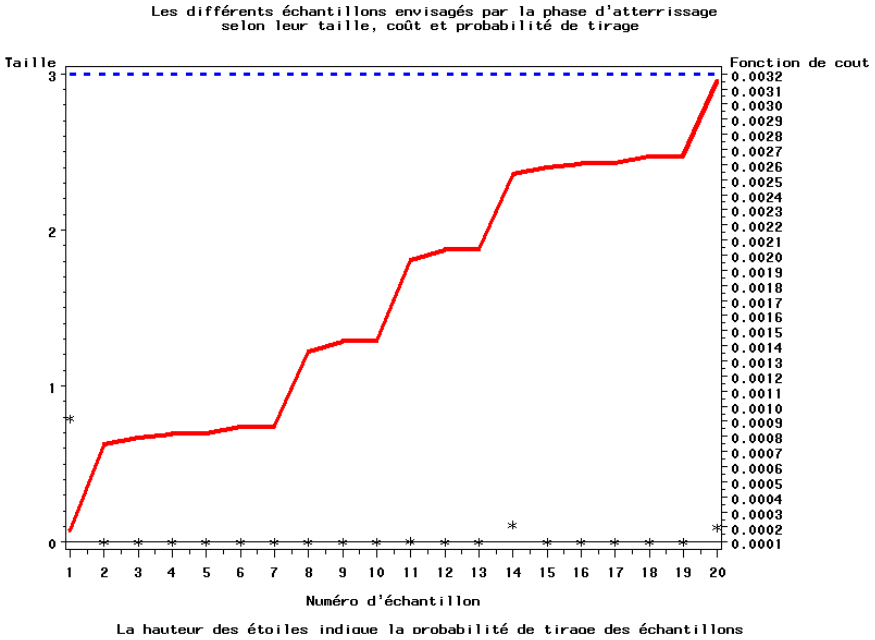
```
A l'issue de la phase de vol
--------------------------------------------------
Nombre d'individus sur lesquels il reste à statuer :          6
Nombre d'individus qu'il reste à échantillonner    :          3


Résultats obtenus
--------------------------------------------------
Lancement de la phase d'atterrissage :                      OUI
Option de la phase d'atterrissage          :                 C
Critère de coût de la phase d'atterrissage :                CV

Nombre d'individus échantillonnés    :                      500

1ère colonne : valeur réelle du total
2ème colonne : estimateur de Horvitz-Thompson du total
3ème colonne : écart relatif en pourcentage


                                        Ecart
Valeur               Valeur             relatif
réelle               estimée            (en %)

PITU1          170 | PITU1          170 | PITU1      0.00
PITU2          165 | PITU2          165 | PITU2     -0.00
PITU3          165 | PITU3          165 | PITU3     -0.00
NB018TU1      1962 | NB018TU1 1984.7412 | NB018TU1   1.16
NB018TU2      1055 | NB018TU2 1061.6667 | NB018TU2   0.63
NB018TU3      1184 | NB018TU3 1187.0242 | NB018TU3   0.26
```

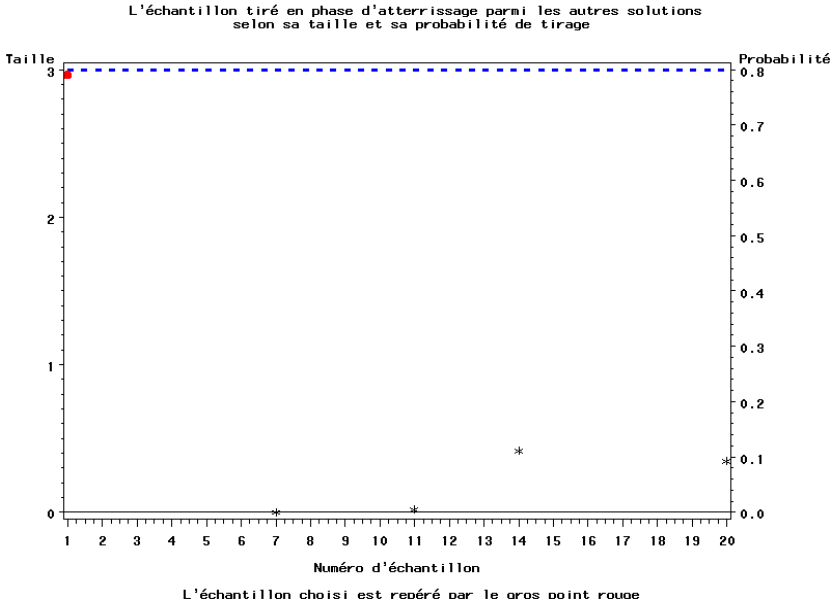
Commentaires

En l'absence de liaison de colinéarité directe entre elles, les 6 variables d'équilibrage forment bien un espace de la même dimension. Notons que les trois premières conditions impliquent que l'échantillon global compte 500 ménages, répartis entre les 3 strates de densité d'habitat. A l'issue de la phase de vol, il restait à statuer entre 6 ménages pour en désigner 3 au hasard. A nouveau, le nombre d'unités dont le sort est encore incertain vaut la dimension de l'espace des contraintes. In fine, la macro a sélectionné un échantillon respectant bien les allocations demandées dans chaque strate. Par contre, les estimations obtenues dans chaque strate pour le nombre total d'individus de moins de 18 ans déforment légèrement la réalité de la base de sondage.

4- Sorties graphiques

La phase d'atterrissage a envisagé les $C_6^3 = 20$ échantillons de taille 3 compatibles avec la phase de vol et calculé leur coût (critère CV) et probabilité de tirage à l'issue de la phase de vol. A nouveau, peu d'échantillons s'avèrent vraiment réalisables dans le plan optimum, tous comportent logiquement 3 ménages. Dans l'exemple, l'algorithme a désigné l'échantillon n°1 au hasard des directions suivies. Comme le programme d'optimisation le privilégie en lui attribuant la plus grande probabilité de sélection (environ 80% à l'issue de la phase de vol), cet échantillon a toutes les chances d'offrir un équilibrage de bonne qualité.





V.8. Exemple 7 Quotas marginaux

Tirage sans remise de 500 ménages à probabilités égales et équilibré selon une stratification qui superpose trois niveaux de densités d'habitat (*communes rurales, unités urbaines de moyenne ou grande taille)* et 4 tailles de ménages *(1, 2, 3 ou 4 personnes et plus)* avec des tailles fixées dans chacune des strates. Les contraintes d'équilibrage portent sur la taille de l'échantillon de ménages dans les 7 strates et le nombre total de personnes dans la population - Option d'atterrissage C et critère de coût CV.

1- *Introduction des probabilités d'inclusion individuelles dans la base de sondage, calcul des indicatrices de strate et des contraintes ad-hoc*

```
data bs7;
set bs;
tu1=(tu99='0');
tu2=((tu99 >'0') and (tu99<'5'));
tu3=(tu99>='5');
nper1=(nper='01');
nper2=(nper='02');
nper3=(nper='03');
nper4=(nper>='04');
pi=500/10000;
run;

%macro disjonct (var,strate,nbstrate);
%do i=1 %to &nbstrate;
data bs7;
set bs7;
&var&strate&i =&var * &strate&i ;
run;
%end;
%mend;

%disjonct(pi,tu,3);
%disjonct(pi,nper,4);
```

A l'issue de cette étape, les variables d'équilibrage ad-hoc ont été créées. Pour que l'échantillon restitue, comme on le souhaite ici, le nombre exact de ménages présents dans chaque strate, ces variables s'obtiennent en multipliant les probabilités d'inclusion individuelles par les indicatrices d'appartenance aux différentes strates. Par exemple, on obtient la variable *« pitu1 »* par le produit des probabilités d'inclusion *« pi »* et de l'indicatrice d'appartenance à la strate des communes rurales *(« tu1 »).* Un échantillon équilibré sur *« pitu1 »* permet de retrouver le nombre de ménages ruraux présents dans la base de sondage.

2- *Lancement de la macro cube*

```
%CUBE (Datafic=bs7,
       Id=ident,
       Pi=pi,
       Pond=w7,
       Contr= pitu1 pitu2 pitu3 pinper1 pinper2 pinper3,
       Sortie=ech7,
       Cout=CV,
       Atter=C,
       Comment = Graphe);
```

3- ***Résultats*** *(en 2 min 5 s)*

```
            BILAN DE VOTRE ECHANTILLONNAGE
_____________________________________________________________

Résumé des spécifications relatives aux contraintes
---------------------------------------------------
Vos contraintes ne sont pas colinéaires :
    le nombre de variables d'équilibrage est    :            6
    pour un espace des contraintes de dimension :            6
```

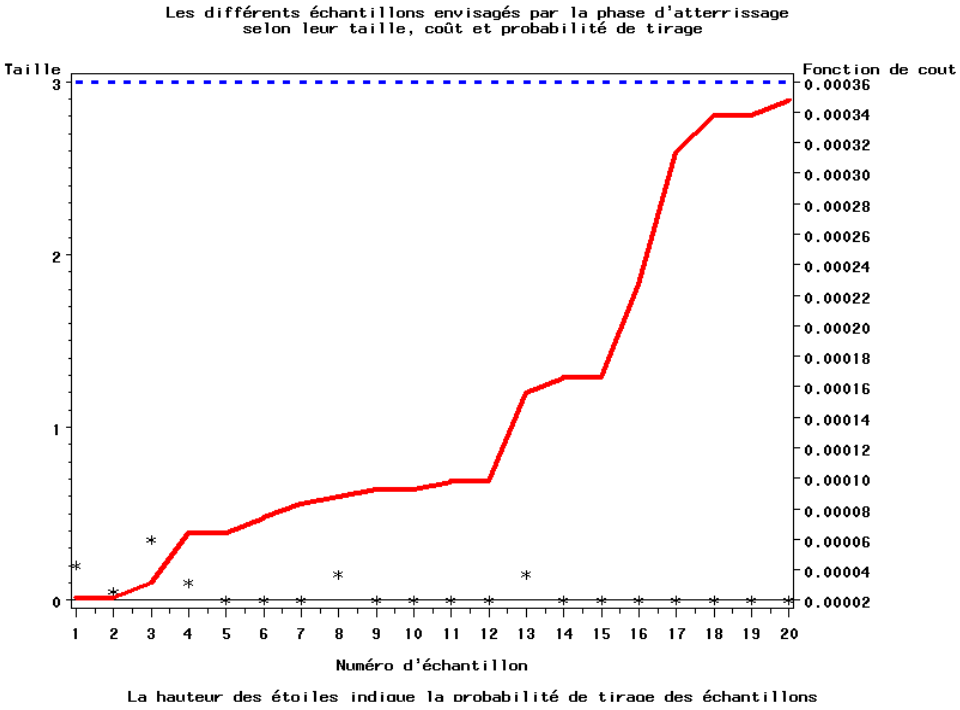
```
A l'issue de la phase de vol
--------------------------------------------------
Nombre d'individus sur lesquels il reste à statuer :          6
Nombre d'individus qu'il reste à échantillonner    :          3


Résultats obtenus
--------------------------------------------------
Lancement de la phase d'atterrissage :                      OUI
Option de la phase d'atterrissage          :                 C
Critère de coût de la phase d'atterrissage :                CV

Nombre d'individus échantillonnés    :                      500

1ère colonne : valeur réelle du total
2ème colonne : estimateur de Horvitz-Thompson du total
3ème colonne : écart relatif en pourcentage


                                      Ecart
Valeur              Valeur            relatif
réelle              estimée           (en %)

PITU1       231.1 | PITU1         231 | PITU1    -0.04
PITU2      134.75 | PITU2         134 | PITU2    -0.56
PITU3      134.15 | PITU3         135 | PITU3     0.63
PINPER1     170.5 | PINPER1       170 | PINPER1  -0.29
PINPER2    174.65 | PINPER2       175 | PINPER2   0.20
PINPER3     72.15 | PINPER3        72 | PINPER3  -0.21
```

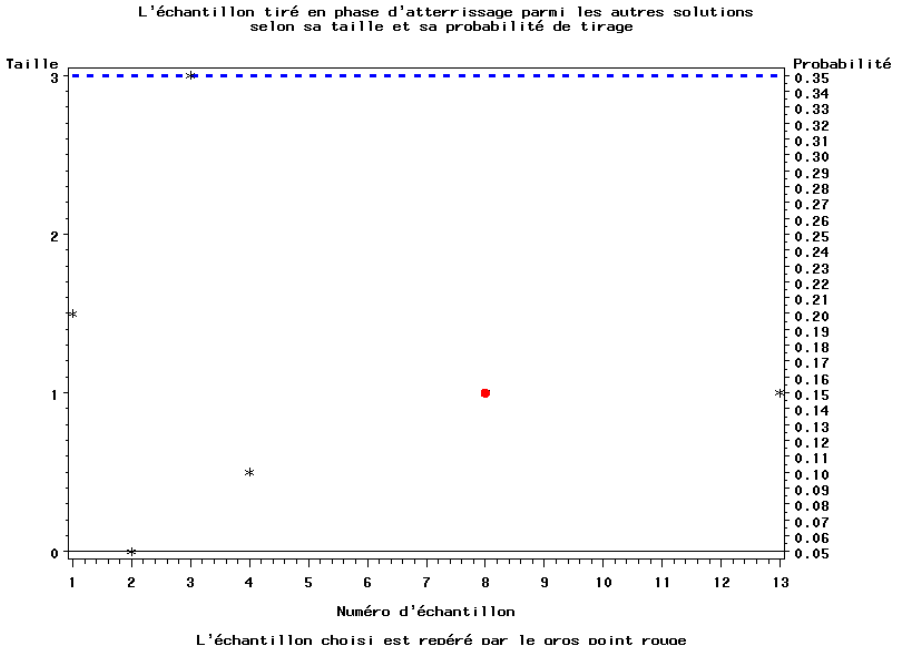
Commentaires

Comme signalé, les 6 contraintes ne sont pas colinéaires entre elles puisque le nombre total de ménages de grande taille n'a pas été introduit comme variable d'équilibrage (ce nombre se déduit des autres effectifs, attendu qu'il est égal au nombre total de ménages résidant dans les 3 strates de densité d'habitat moins le nombre total de ménages de taille inférieure ou égale à 3 personnes. De même, la contrainte relative à l'allocation globale de l'échantillon se déduit des trois premières conditions). De toute évidence, il n'était pas possible d'obtenir un échantillon parfaitement équilibré sur les effectifs par strate puisque aucune des allocations de strate n'est un entier. A l'issue de la phase de vol, il reste à statuer sur 6 ménages (dimension de l'espace des contraintes) pour en désigner 3 au hasard. De fait, la phase d'atterrissage a arrondi les allocations initiales non entières, et a permis de respecter ainsi, autant que possible, les quotas demandés.

4- Sorties graphiques

La phase d'atterrissage envisage $C_6^3 = 20$ échantillons de taille 3 compatibles avec la phase de vol, calcule leurs coûts (critère CV) et probabilités de tirage à l'issue de la phase de vol. A nouveau, peu d'échantillons s'avèrent réalisables dans le plan optimum et ils sont tous de taille 3. Dans l'exemple, l'algorithme a désigné l'échantillon 8 au hasard des directions suivies. Des chemins à plus forte probabilité auraient conduit aux échantillons n°3 ou n°1, plus proches des contraintes.





V.9. Exemple 8 SAS sans atterrissage

Sondage aléatoire simple sans remise de 500 ménages équilibré sur la taille de l'échantillon et le nombre total de ménages – Option d'atterrissage A

Les probabilités d'inclusion valent toutes 500 / 10 000 (introduites dans l'exemple 1 supra).

1- Lancement de la macro cube

```
%CUBE (Datafic=bs,
       Id=ident,
       Pi=pisas,
       Pond=wsas,
       Contr=pisas un,
       Sortie=echsas,
       Atter=A,
       Comment = OUI);
```

2- Résultats (obtenus en 1 min 50 s)

```
            BILAN DE VOTRE ECHANTILLONNAGE
_____________________________________________________________

Résumé des spécifications relatives aux contraintes
-----------------------------------------------------
Vos contraintes sont colinéaires :
    le nombre de variables d'équilibrage est    :              2
    pour un espace des contraintes de dimension :              1


A l'issue de la phase de vol
-----------------------------------------------------
Nombre d'individus sur lesquels il reste à statuer :           0
Nombre d'individus qu'il reste à échantillonner    :           0


Résultats obtenus
-----------------------------------------------------
Lancement de la phase d'atterrissage :                      NON

Nombre d'individus échantillonnés    :                      500

1ère colonne : valeur réelle du total
2ème colonne : estimateur de Horvitz-Thompson du total
3ème colonne : écart relatif en pourcentage


                                      Ecart
Valeur             Valeur             relatif
réelle             estimée            (en %)

PISAS        500 | PISAS         500 | PISAS     0.00
UN         10000 | UN          10000 | UN        0.00
```

Commentaires

Conformément à ce que l'on escomptait, un échantillon de 500 ménages a été désigné dès la phase de vol : celle-ci s'achève en ayant statué sur toutes unités de la base de sondage, ce qui rend, de fait, la phase d'atterrissage superflue. Au préalable, le commentaire a avisé l'utilisateur de la parfaite proportionnalité entre la constante « 1 » et les probabilités d'inclusion individuelles toutes égales à 500/10000. Ce cas illustre le développement traité au paragraphe II-3.

3- Caractéristiques de l'échantillon

```
proc means data=echsas sum;
var pisas un personnes  hommes femmes nb018 nb1825 nb2540 nb4060 nb60;
Weight wsas;
run;
```

The MEANS Procedure

Variable	Sum
pisas	500.0000000
un	10000.00
personnes	23700.00
hommes	11900.00
femmes	11800.00
nb018	5360.00
nb1825	1880.00
nb2540	4580.00
nb4060	6620.00
nb60	5260.00

V.10. Exemple 9 Un cas de dégénérescence

Illustration d'une dégénérescence de l'algorithme avec un tirage équilibré sur la taille de l'échantillon ainsi que sur le total d'une variable auxiliaire donnée « EFF ». Option d'atterrissage B et critère de coût CV.

Les commentaires affichés en OUTPUT donnent le compte-rendu suivant :

```
            BILAN DE VOTRE ECHANTILLONNAGE
_____________________________________________________________

Résumé des spécifications relatives aux contraintes
---------------------------------------------------
Vos contraintes ne sont pas colinéaires :
    le nombre de variables d'équilibrage est    :             2
    pour un espace des contraintes de dimension :             2


A l'issue de la phase de vol
---------------------------------------------------
Nombre d'individus sur lesquels il reste à statuer :          5
Nombre d'individus qu'il reste à échantillonner    :        2.4

                        Attention !....

Il reste à désigner un nombre non entier d'unités en fin de vol.
  Cela peut etre du à vos probabilités d'inclusion initiales,
  ou à votre choix de ne pas contraindre la taille d'échantillon,
  ou encore à la présence de valeurs individuelles très élevées
  pour certaines contraintes (au delà de 10 puissance 8)
```

```
Résultats obtenus
--------------------------------------------------
Lancement de la phase d'atterrissage :                      OUI
Option de la phase d'atterrissage    :                        B
Critère de coût de la phase d'atterrissage    :              CV

Nombre d'individus échantillonnés    :                       37

1ère colonne : valeur réelle du total
2ème colonne : estimateur de Horvitz-Thompson du total
3ème colonne : écart relatif en pourcentage

                                          Ecart
Valeur              Valeur                relatif
réelle              estimée               (en %)

PISORTIE      32.4 | PISORTIE          37 | PISORTIE   16.37
EFF   381954895289 | EFF    4845408249752 | EFFECTIF   26.88
```

Commentaires

Deux points éveillent l'attention ici : d'une part le nombre d'unités que la phase d'atterrissage doit sélectionner n'est pas entier ; d'autre part, le nombre d'individus restants sur lesquels elle opère dépasse la dimension de l'espace des contraintes. Si le premier point se conçoit - bien qu'il signale une configuration particulière comme l'une de celles présentées au paragraphe II-4 (en l'occurrence la somme des probabilités d'inclusion des individus de la base de sondage est non entière ce qui correspond à une taille irréalisable)-, le second avertit d'une réelle anomalie. De toute évidence, l'algorithme a dégénéré : il n'a pu traiter la variable EFF à cause d'un nombre trop important de chiffres significatifs. Malgré tout, si l'on veut conserver l'équilibrage sur cette variable, il est recommandé de la diviser, pour chacun des individus, par une puissance de 10 assez importante.

VI. Bibliographie

Ardilly, P. (1991) : Echantillonnage représentatif optimum à probabilités inégales*, Annales d'Economie et de Statistique,* **23,** 91-113.

Ardilly, P. (1994) : *Les techniques de sondage,* Paris, Editions Technip.

Bousabaa, A., Lieber, J., Sirolli, R. (1999) : La macro Cube. Technical report, INSEE, Rennes.

Caron, N. (2003) *:* Sondage équilibré, *La lettre du Système Statistique d'Entreprises*, n°53, INSEE

Cochran, W.-G.(1977) : *Sampling Techniques*, 3ème édition, New-York, Wiley.

Deville, J.-C., (1992) : Constrained samples, conditionnal inference, weighting : three aspects of the utilisation of auxiliary information, *Proceedings of the Workshop Auxiliary Information in Surveys, Örebro* (Suède).

Deville, J.-C., Grosbras, J.-M., Roth, N. (1988) : Efficient sampling algorithms and balanced samples*, COMPSTAT*, *Proceedings in computational statistics,* Physica Verlag, pp. 255-266.

Deville, J.-C., Tillé, Y. (2001) : Echantillonnage par la méthode du Cube, variance et estimation de variance » in *Enquêtes, modèles et applications,* 344-363, Paris, Dunod.

Deville, J.-C., Tillé, Y. (2002) : Variance approximation under balanced salmpling. *Soumis pour publication.*

Deville, J.-C., Tillé, Y. (2003) : Efficient Balanced Sampling : the Cube Method, à paraître dans *Biométrika.*

Dumais, J., Isnard, M. (2000) : Le sondage de logements dans les grandes communes dans le cadre du recensement rénové de la population, INSEE, *Actes des Journées de Méthodologie Statistique.*

Dussaix, A.-M., Grosbras, J.-M., (1993) : *Les sondages : principes et méthodes*, Que sais-je ? n° 701, Presses Universitaires de France.

Särndal, C.-E., Swensson B., Wretman J. (1992) : *Model Assisted Survey Sampling*, New-York, Springer.

Tillé, Y. (2001) : *Théorie des sondages : échantillonnage et estimation en populations finies*, Paris, Dunod.

Wilms, L. (2000) : Présentation de l'écnantillon maître en 1999 et application au tirage des unités primaires par la macro Cube, INSEE, *Actes des Journées de Méthodologie Statistique.*