OPTIMUM DE TYPE NEYMAN POUR L'ECHANTILLONNAGE EQUILIBRE SUR DES MARGES

Daniel BONNERY, Guillaume CHAUVET, Jean-Claude DEVILLE
CREST-ENSAI

25/03/2009

Introduction

Lors de la conception d'un plan de sondage utilisant un découpage en strates de la population, une des questions à résoudre est celle de l'allocation du nombre de questionnaires (i.e. d'individus enquêtés) dans chacune des strates, étant donné un nombre n total de questionnaires à ne pas dépasser.

Pour un sondage stratifié avec sondage aléatoire simple dans chaque strate, on définit l'allocation optimale de Neyman comme l'allocation qui permet d'obtenir une variance minimale pour l'estimateur de Horvitz-Thompson d'un total d'une variable d'intérêt donnée sous contrainte de taille globale d'échantillon inférieure ou égale à n.

La donnée d'un nombre de questionnaires à allouer à chacune des strates est équivalente à la donnée d'un taux de sondage dans chaque strate. La recherche de l'allocation optimale peut donc être vue comme une recherche de probabilités de sondages optimales constantes sur chaque strate et de somme égale à n.

Pour résumer, rechercher l'allocation optimale de Neyman revient à minimiser la formule de variance d'un sondage stratifié avec sondage aléatoire simple au sein de chaque strate, qui est une fonction des probabilités d'inclusion, sous des contraintes linéaires sur les probabilités d'inclusion.

Un sondage stratifié peut être vu comme un sondage équilibré sur une variable catégorielle, avec des probabilités d'inclusion égales sur chaque strate. On peut se poser la question de savoir quel est le meilleur choix de probabilités d'inclusions lorsqu'on s'apprête à effectuer un sondage équilibré sur une ou plusieurs variables auxiliaires, étant donné des contraintes linéaires sur les probabilités d'inclusion.

L'objet de l'article est de rechercher les probabilités de sondage optimales pour un sondage équilibré pour un jeu de variables autre que l'ensemble des variables indicatrices d'appartenance à chaque strate.

Deux cas pratiques ont notamment été envisagés : on définit des sous population par croisement de deux variables catégorielles. Dans le premier cas, on souhaite que les estimateurs de nombre d'individus dans les sous populations soit exact ; dans le second, le tirage est équilibré sur le nombre d'individus enquêtés dans chaque sous population (il peut par exemple s'agir de fixer le nombre d'individus à enquêter par tranche d'age, et de fixer séparément le nombre d'individus à enquêter par sexe sans fixer au préalable le nombre d'individus par croisement des

deux variables).

Il s'agit donc de poser un problème d'optimisation, de discuter de l'existence de solutions au problème, et de proposer une méthode numérique de recherche de solution adaptée à chaque cas particulier.

1 Le problème d'optimisation

1.1 Présentation du problème d'optimisation

On se donne une population $U = [\![1, N]\!]$ $N \in \mathbb{N}^*$ individus, k désignant un individu. Si on munit $\mathscr{P}(U)$, l'ensemble des échantillons possibles (sans remise) issus de U, d'une probabilité $\begin{array}{rcl} p & : \mathscr{P}(U) & \to [0,1] \\ & s & \mapsto p(s) \end{array}$, on peut définir l'estimateur de Horvitz Thompson d'un total Y d'une variable d'intérêt y définie sur U par $\hat{Y} = \hat{Y}(s) = \sum_{k \in s} \frac{y_k}{\pi_k}$, $\pi_k = \sum_{s|k \in s} p(s)$ étant la probabilité de tirer un échantillon contenant l'individu k. Dans la suite, π désignera un vecteur de probabilités d'inclusion. On imposera que les probabilités d'inclusion soient non nulles : $\pi \in]0,1]^N$.

On se donne aussi : $z_\pi^1, ..., z_\pi^m$ m variables d'équilibrage dépendant de π définies sur U. On note $z_{\pi,k}^i$ la valeur prise par z_π^i pour l'individu $k \in U$.

La méthode du cube est un algorithme permettant de tirer un échantillon selon un plan de sondage p tel que les estimateurs d'Horvitz Thompson de variables auxiliaires données soient exacts (on parle de tirage équilibré sur les variables auxiliaires), et qui respecte des probabilités d'inclusion données.

On dispose d'une formule d'approximation de la variance pour l'estimateur de Horvitz Thompson obtenu suite à un tirage équilibré sur les variables $z_\pi^1 \dots z_\pi^m$, **exact** et à entropie maximale, respectant un vecteur de probabilités d'inclusion π donné ([5, 6]) :

$$V : \pi \mapsto \sum_{k=1}^{N} (\frac{1}{\pi_k} - 1)(y_k - \hat{y}_k(\pi))^2$$

$\hat{y}_k(\pi)$ désignant l'estimateur de y_k par la régression de y sur $z^1(\pi) \dots z^m(\pi)$ pondérée par les poids $(\frac{1}{\pi_k} - 1)$

La pratique a montré que cette approximation était proche de la vraie variance lorsque le tirage équilibré n'était pas exact.

$V(\pi)$ est une formule d'approximation de la variance. Elle présente un avantage par rapport à la formule de Horvitz-Thompson, qui donne une expression exacte de la variance de l'estimateur de Horvitz-Thompson en fonction des probabilités d'inclusion double : $Var(\hat{Y}) = \sum_{(kl) \in U^2} \frac{y_k y_l (\pi_{kl} - \pi_k \pi_l)}{\pi_k \pi_l}$ (on rappelle que $\pi_{kl} = \sum_{s \in \mathscr{P}(U) | s \supset \{kl\}} p(s)$ est la probabilité que les individus k et l appartiennent simultanément à l'échantillon tiré). En effet, la formule d'approximation ne nécessite pas le calcul des probabilités d'inclusion double.

Au moment du tirage équilibré sur les variables auxiliaires, le responsable d'enquête a la liberté de choisir les probabilités d'inclusion de chaque individu (le vecteur π). Dans l'esprit du problème d'optimisation posé par Neyman, il s'agit de déterminer les meilleures probabilités d'inclusion sous certaines contraintes linéaires afin de réduire la variance.

Plus précisément, on cherche à déterminer $\pi \in]0,1]^N$ tel que $V(\pi)$ soit minimum, sous des contraintes linéaires, résumées par l'expression $A\pi = B$ (à ne pas confondre avec les contraintes d'équilibrage : il s'agit de contraintes sur le vecteur des probabilités d'inclusion). avec A, B de

même nombre de lignes (le nombre de contraintes linéaires) et de nombre de colonnes respectivement N et 1. A est pris de plein rang.

Par la suite, les contraintes linéaires posées seront :

– $\sum_{k=1}^{N} \pi_k = n$

– π_k constant sur chaque ensemble d'une partition donnée de U.

1.2 Résolution du problème d'optimisation

La fonction V définie plus haut peut être non convexe, et peut admettre plusieurs minima locaux. Lorsque certaines conditions sur les variables y et z_π^i sont respectées, on arrive à prouver l'existence d'un minimum global. Toutefois, les méthodes numériques envisagées pour la recherche du minimum n'assurent pas la convergence vers le minimum global. Elles permettent toutefois de diminuer la variance par rapport à un premier jeu de probabilités d'inclusion.

Dans la suite, on présente les méthodes envisagées pour les deux cas pratiques évoqués en introduction.

2 Applications

2.1 Neyman et le sondage stratifié ([3, 2]).

Avec ce premier exemple, il s'agit de donner (rappeler) une première méthode de résolution du problème d'optimisation, qui inspirera une méthode de résolution pour un problème plus général (l'équilibrage sur plusieurs marges).

On se donne une partition de U en I ensembles : $U_1 \dots U_I$ de taille $N_1 \dots N_I$. On pose pour $i \in \{1 \dots I\}$, $z_k^i = 1$ si $k \in U_i$, 0 sinon. On note $\bar{Y}_i = \frac{\sum_{k \in U_i} y_k}{N_i}$.

Si on applique le programme de minimisation pour les variables z^i et si on impose de plus π_k constant sur chaque U_i, alors le programme d'optimisation nous donne approximativement les probabilités d'inclusion qui nous donnent l'allocation de Neyman.

En effet, le tirage équilibré sur les marges réalisé par la méthode du cube est alors un sondage stratifié avec sondage aléatoire simple par strate. Les probabilités d'inclusion optimales, qui minimisent la variance de l'estimateur de Horvitz Thompson de Y sont celles qui donnent l'allocation de Neyman.

Les probabilités d'inclusion qui minimisent l'approximation de la variance sont donc approximativement les probabilités optimales qui donnent l'allocation de Neyman.

En notant $n_i = N_i \pi_k$ pour $k \in U_i$ l'approximation de la variance s'écrit :

$$\begin{aligned} V(\pi) &= \sum_{k=1}^{N} (\tfrac{1}{\pi_k} - 1)(y_k - \hat{y}_k(\pi))^2 \\ &= \sum_{i=1}^{I} (\tfrac{1}{n_i} - \tfrac{1}{N_i}) N_i^2 \frac{\sum_{k \in U_i} (y_k - \bar{Y}_i)^2}{N_i - 1} \frac{N_i - 1}{N_i} \end{aligned}$$

A comparer avec la vraie variance d'un sondage stratifié :

$$\sum_{i=1}^{I} (\frac{1}{n_i} - \frac{1}{N_i}) N_i^2 \frac{\sum_{k \in U_i} (y_k - \bar{y}_i)^2}{N_i - 1}$$

En effet dans ce cas (ce qui apparaîtra comme une particularité de l'optimisation dans le cas de l'équilibrage sur une variable catégorielle), $\hat{y}_k(\pi) = \bar{Y}_i$ si $k \in U_i$: dans une ANOVA à un facteur, l'estimation $\hat{y}_k(\pi)$ de y_k est la moyenne de y sur la strate d'appartenance de k, indépendamment des poids de la régression, donc **indépendammment de** π.

Remarque : comme dans le cas de l'allocation de Neyman, le programme d'optimisation ne nous donne pas forcément des n_i entiers : l'équilibrage n'est possible que si l'on arrondit les n_i.

On obtient donc pour $k \in U_i$, en notant $S_i = \frac{\sum_{l \in U_i}(y_l - \bar{Y}_i)^2}{N_i - 1}$:

$$\begin{aligned} \pi_k &= n \frac{N_i \sqrt{\frac{\sum_{l \in U_i}(y_l - \bar{Y}_i)^2}{N_i}}}{\sum_j N_j \sqrt{\frac{\sum_{l \in U_j}(y_l - \bar{Y}_i)^2}{N_j}}} \\ &= n \frac{\sqrt{\frac{N_i - 1}{N_i}} N_i S_i}{\sum_j \sqrt{\frac{N_j - 1}{N_j}} N_j S_j} \end{aligned}$$

On arrondit ensuite π_k de telle façon que $N_i \pi_k$ soit entier en respectant $\sum \pi_k = n$ (à comparer avec la véritable allocation de Neyman qui donne avant troncature : $\pi_k = n \frac{N_i S_i}{\sum_j N_j S_j}$)

Remarque : on obtient exactement les mêmes probabilités d'inclusion et la même procédure de tirage (sondage aléatoire stratifié) si on pose pour $i \in \{1 \dots m = I\}$: $z_k^i = 1$ si $k \in U_i$, 0 sinon, et si on impose $A\pi = B$: les π_k sont constants sur chaque U_i.

Ce premier exemple peut apparaître comme un cas particulier des deux exemples présentés par la suite, où on introduit plusieurs variables catégorielles comme variables d'équilibrage (sans perte de généralité et pour simplifier les notations, on se limite à 2 variables catégorielles d'équilibrage).

2.2 Choix des meilleures probabilités pour un sondage équilibré sur 2 marges

2.2.1 De quoi s'agît il ?

Il s'agit de réaliser un sondage équilibré sur les marges de deux variables catégorielles. Le responsable de l'enquête veut par exemple obtenir des estimateurs de Horvitz Thompson exacts des effectifs dans chaque tranche d'âge, et des estimateurs de Horvitz Thompson exacts du nombre d'hommes et du nombre de femmes, sans pour autant chercher à estimer exactement les effectifs du nombre d'hommes ou de femmes d'une tranche d'âge donnée.

2.2.2 Exposé du problème

Soit $I, J \in \mathbb{N}$, $(U_{ij})_{(i,j) \in [\![1,I]\!] \times [\![1,J]\!]}$ une partition de U. On pose $U_{i.} = \cup_j U_{ij}$, $U_{.j} = \cup_i U_{ij}$
On note $N_{i.} = \#(U_{i.})$, $N_{.j} = \#(U_{.j})$ et $N_{ij} = \#(U_{ij})$
On se donne les variables d'équilibrage $(z_{i.})_{i \in [\![1,I]\!]}, (z_{.j})_{j \in [\![1,J-1]\!]}$ [1] :

$$z^{i.} : k \mapsto z_k^{i.} = \begin{array}{ll} 1 & \text{si } k \in U_{i.} \\ 0 & \text{sinon} \end{array}$$

$$z^{.j} : k \mapsto z_k^{.j} = \begin{array}{ll} 1 & \text{si } k \in U_{.j} \\ 0 & \text{sinon} \end{array}$$

On souhaite déterminer les probabilités d'inclusion constantes sur tout U_{ij} qui minimiseront la variance d'un sondage équilibré sur les $z_{i.}, z_{.j}$ pour $i \in [\![1, I]\!], j \in [\![1, J-1]\!]$ à entropie maximum sous contrainte de taille d'échantillon égale à n.

1. j varie de 1 à $J-1$ pour ne pas avoir une redondance de l'information :$z_{.J} = \sum z_{i.} - \sum_{j<J} z_{.j}$

Dorénavant, $\hat{y}_k(\pi)$ dépend vraiment de π, ce qui complexifie le problème de minimisation : il n'est plus possible d'obtenir une expression formelle simple du minimum de

$$V(\pi) : \pi \mapsto \sum_{k=1}^{N} (\frac{1}{\pi_k} - 1)(y_k - \hat{y}_k(\pi))^2$$

Une première approche consiste à considérer que $\sum_{k \in U_{ij}} (y_k - \hat{y}_k)^2$ ne dépend pas de (π), et de résoudre le problème en utilisant le Lagrangien. On retrouve cette approche dans les travaux de Tillé et Favre ([2]).

Toutefois, cette approche doit être justifiée : il convient de s'assurer que le gain de variance obtenu n'est pas annulé par la variation de $\sum_{k \in U_{ij}} (y_k - \hat{y}_k)^2$ en fonction de la variation de π, ce qui est démontré par la suite.

Pour se rapprocher le plus possible d'un éventuel minimum local de V, on peut réitérer la méthode qui consiste à séparer la variable de la fonction V : on écrit $V(\pi) = v(\pi, \hat{y}(\pi))$, on minimise ensuite $v(\pi, \tilde{y}$ par rapport à π en gardant fixé $\tilde{y}$, puis on minimise v en gardant fixé π, le point clef étant que $argmin_{\tilde{y}} v(\pi, \tilde{y}) = \hat{y}(\pi)$.

Par la suite, on discute des conditions d'existence d'un minimum global de la fonction V sur $]0, 1]^N$, ce qui assure la convergence de $V(\pi^r)$ lorsque (π^r) désigne la suite de vecteurs de probabilités obtenue à partir de la méthode d'itération. On démontre que la méthode d'itération permet de diminuer l'approximation de la variance de l'estimateur d'Horvitz Thompson pour un tirage équilibré, et on démontre enfin qu'en cas de convergence de (π^r), un minimum local de la fonction V est atteint.

2.2.3 Existence d'un minimum global de la fonction V

Pour ce problème, à condition que la dispersion de la variable y par cellule U_{ij} soit non nulle, il existe une probabilité d'inclusion π minimisant V (il existe un minimum global contraint pour la fonction V sur $]0, 1]^N$).

En effet, dans ce cas, on a l'inégalité :

$$\begin{array}{rcl} V(\pi) & = & \sum_{ij} (\frac{1}{\pi_{ij}} - 1) \sum_{k \in U_{ij}} (y_k - \hat{y}_k(\pi))^2 \\ & \geq & \sum_{ij} (\frac{1}{\pi_{ij}} - 1) \sum_{k \in U_{ij}} (y_k - \bar{y}_{ij})^2 \end{array}$$

Soit π^0 vérifiant les contraintes de départ. On a :

$$\begin{array}{rcl} V(\pi) \leq V(\pi^0) & \Rightarrow & \sum_{ij} (\frac{1}{\pi_{ij}} - 1) \sum_{k \in U_{ij}} (y_k - \bar{y}_{ij})^2 < V(\pi^0) \\ & \Rightarrow & \pi_{ij} \geq \frac{\sum_{k \in U_{ij}} (y_k - \bar{y}_{ij})^2}{V(\pi^0) + \sum_{k \in U_{ij}} (y_k - \bar{y}_{ij})^2} \\ & \Rightarrow & \pi_k \geq m = min_{ij} \{ \frac{\sum_{k \in U_{ij}} (y_k - \bar{y}_{ij})^2}{V(\pi^0) + \sum_{k \in U_{ij}} (y_k - \bar{y}_{ij})^2} \} \end{array}$$

Cette inégalité permet de réduire le domaine de recherche d'un minimum au sous ensemble fermé $[m, 1]^N \supset V^{-1}([0, V(\pi^0)])$ de $]0, 1]^N$.

La fonction V étant continue sur $]0, 1]^N$, il existe un unique minimum global sur le sous ensemble compact $[m, 1]^N$, qui est aussi un minimum global sur $]0, 1]^N$ d'après l'inégalité ci dessus.

2.2.4 Détail de la méthode : recherche d'un point fixe

On pose

$$\begin{array}{rcccl} v & : &]0, 1]^N \times \mathbb{R}^N & \rightarrow & \mathbb{R}^+ \\ & : & (\pi, \tilde{y}) & \mapsto & \sum_{k=1}^{N} (\frac{1}{\pi_k} - 1)(y_k - \tilde{y}_k)^2 \end{array}$$

On note $\hat{y}(\pi)$ la régression de y sur les variables d'équilibrage avec les poids $(\frac{1}{\pi_k} - 1)$

On a :

$$V(\pi) = v(\pi, \hat{y}(\pi))$$

On définit par récurrence la suite $(\pi^r)_{r\in\mathbb{N}}$:
Soit $\pi^0 \in]0,1]^N$ tel que $A\pi^0 = B$.
Si $argmin_{\{x\in]0,1]^N | Ax=B\}}\{v(x, \hat{y}(\pi^r))\}$ existe et est réduit à un élément, alors on pose :

$$\pi^{r+1} = argmin_{\{x\in]0,1]^N | Ax=B\}}\{v(x, \hat{y}(\pi^r))\}$$

sinon abandon.

On remarque que la condition d'existence de minimum global (la dispersion de y est non nulle sur chaque U_{ij}) est suffisante pour que la suite soit définie $\forall r \in \mathbb{N}$.
En effet, si $\forall i, j \sum_{l\in U_{ij}}(y_l - \bar{Y}_{ij})^2 > 0$, alors quelque soit $\tilde{y} \in \mathbb{R}^N$, $\sum_{l\in U_{ij}}(y_l - \tilde{y}_l)^2 > 0$, et alors $argmin_{\{x\in]0,1]^N | Ax=B\}}\{v(x, \tilde{y})\}$ est défini car x défini par $x_k = \frac{N_{i'j'}\sqrt{\frac{\sum_{l\in U_{i'j'}}(y_l-\tilde{y}_l)^2}{N'_{i'j'}}}}{\sum_{ij} N_{ij}\sqrt{\frac{\sum_{l\in U_{ij}}(y_l-\tilde{y}_l)^2}{N_{ij}}}}$ si $k \in U_{i'j'}$ convient.

De plus, la suite des $V(\pi^r) = v(\pi^r, \hat{y}(\pi^r))$ est décroissante :

$$\begin{array}{rcll} V(\pi^{r+1}) & = & v(\pi^{r+1}, \hat{y}(\pi^{r+1})) & \\ & \leq & v(\pi^{r+1}, \hat{y}(\pi^r)) & \text{car } \hat{y}(\pi^{r+1}) = argmin_{\tilde{y}} v(\pi^{r+1}, \tilde{y}) \\ & \leq & v(\pi^r, \hat{y}(\pi^r)) & \text{car } \pi^{r+1} = argmin_{\tilde{\pi}} v(\tilde{\pi}, \hat{y}(\pi^r)) \\ & = & V(\pi^r) & \end{array}$$

La suite des $V(\pi^r)$ étant décroissante et positive, elle est convergente.

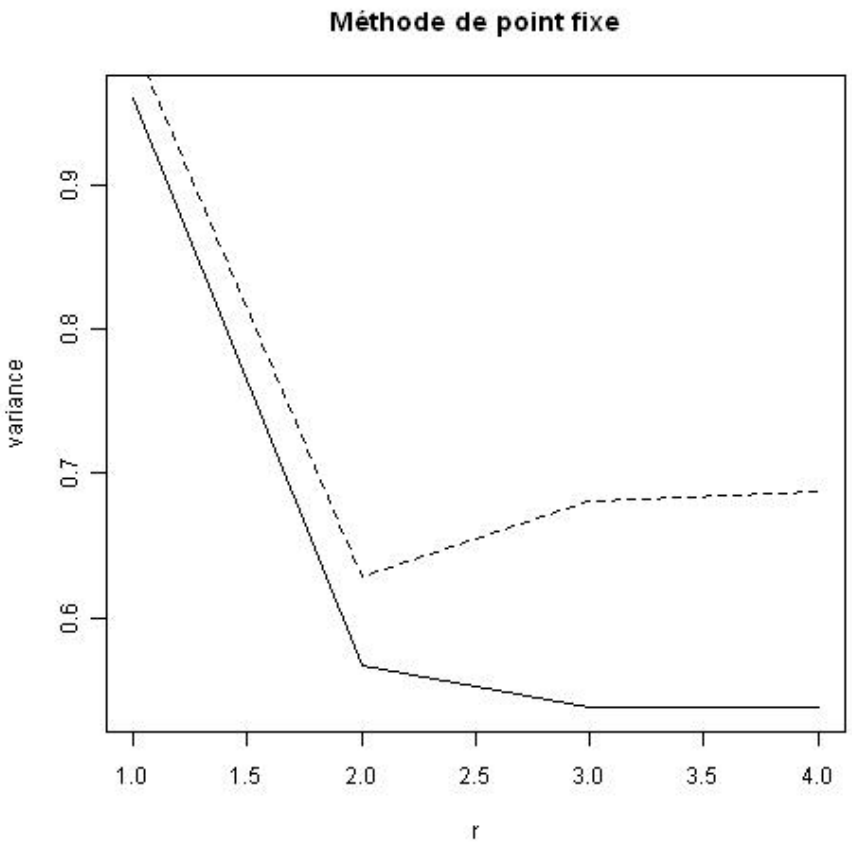
Cette procédure peut donc permettre d'améliorer la variance d'un estimateur issu d'un tirage équilibré, en fournissant un jeu de probabilités d'inclusion meilleur que celui de départ.

De plus, en cas de convergence des π^r (ce qu'on observe en pratique), on atteint un minimum local contraint de la fonction V (Démonstration en annexe 5).

2.2.5 Simulations

Des simulations ont été effectuées. On génère aléatoirement 2 variables catégorielles sur une population de 600 individus et on génère une variable y selon un modèle ANOVA à 2 facteurs et intéraction, les paramêtres du modèle étant générés aléatoirement.

Le graphique ci dessous donne l'évolution de la variance calculée par simulation au fur et à mesure des itérations.



Les différences entre les deux courbes sont dues aux erreurs d'estimation de la variance empirique, à l'approximation effectuée dans la formule d'approximation de la variance, et à la variance de la phase de vol dans le cas de problèmes non exacts.

Ce problème pouvant s'avérer non exact, on pourrait imaginer le cas ou le gain apporté par le choix de probabilités pour lesquelles le problème est exact serait plus grand que l'amélioration due à l'optimisation de la variance due à la phase de vol (qui seule est prise en compte par la formule de variance approchée).

2.3 Quotas probabilistes

2.3.1 De quoi s'agît il ?

Supposons qu'un responsable doive conduire une enquête en s'imposant d'interroger un certain nombre de personnes, en respectant des contraintes marginales sur le sexe et sur l'age (des nombre donnés d'hommes, de femmes, de mineurs, et de majeurs doivent être interrogés), sans se donner de contraintes sur nombre de personnes à interroger par croisement des deux variables. D'ou le rapprochement avec les enquêtes par quotas.

Il s'agit toutefois de rester dans un cadre ou tout individu a une probabilité non nulle maîtrisée et connue d'être tirée. D'où le terme probabiliste.

On se place donc dans le cas ou le responsable envisage d'effectuer un tirage équilibré sur le nombre d'hommes, de femmes, de mineurs et de majeurs interrogés à partir d'une base de sondage donnée. Le responsable a toutefois la possibilité de fixer lui même les contraintes marginales : il peut donc choisir celles qui optimisent la variance de l'estimateur de Horvitz Thompson d'une variable donnée.

2.3.2 Formalisation du problème

Soit $I, J \in \mathbb{N}$ $(U_{ij})_{(i,j)\in[\![1,I]\!]\times[\![1,J]\!]}$ une partition de U. On pose $U_{i.} = \cup_j U_{ij}$, $U_{.j} = \cup_i U_{ij}$
On note $N_{i.} = \#(U_{i.})$, $N_{.j} = \#(U_{.j})$ et $N_{ij} = \#(U_{ij})$
On se donne les variables d'équilibrage $(z_{i.})_{i\in[\![1,I]\!]}, (z_{.j})_{j\in[\![1,J-1]\!]}$:

$$z_\pi^{i.} : k \mapsto z_{\pi,k}^{i.} = \begin{array}{cc} \pi_k & \text{si } k \in U_{i.} \\ 0 & \text{sinon} \end{array}$$

$$z_\pi^{.j} : k \mapsto z_{\pi,k}^{.j} = \begin{array}{cc} \pi_k & \text{si } k \in U_{.j} \\ 0 & \text{sinon} \end{array}$$

On souhaite déterminer les probabilités d'inclusion constantes sur tout U_{ij} qui minimiseront la variance d'un sondage équilibré sur les $z_\pi^{i.}, z_\pi^{.j}$ à entropie maximum sous contrainte de taille d'échantillon égale à n ($A\pi = B \Leftrightarrow \pi_k$ constant sur U_{ij} et $\sum \pi_k = n$)

Dans ce cas, la méthode de point fixe ne peut pas être appliquée : en effet dans ce cas, on peut avoir $v(\pi^{r+1}, \hat{y}^{r+1}) > v(\pi^r, \hat{y}^r)$ car $\hat{y}^r \notin vect(z^1_{\pi^{r+1}} \dots z^m_{\pi^{r+1}})$ et par suite, augmentation de la variance au fur et à mesure des itérations (cf. annexe 4). Des simulations ont notamment permis d'observer ce phénomène.

2.3.3 Existence d'un minimum global

Pour ce problème, à condition que la dispersion de la variable y par cellule U_{ij} soit non nulle, il existe une probabilité d'inclusion π minimisant la variance (il existe un minimum global contraint pour la fonction V sur $]0,1]^N$). (La démonstration du 2.2.3 est aussi valable). Toutefois, la méthode envisagée ne permet d'atteindre qu'un minimum local.

2.3.4 Méthode de descente

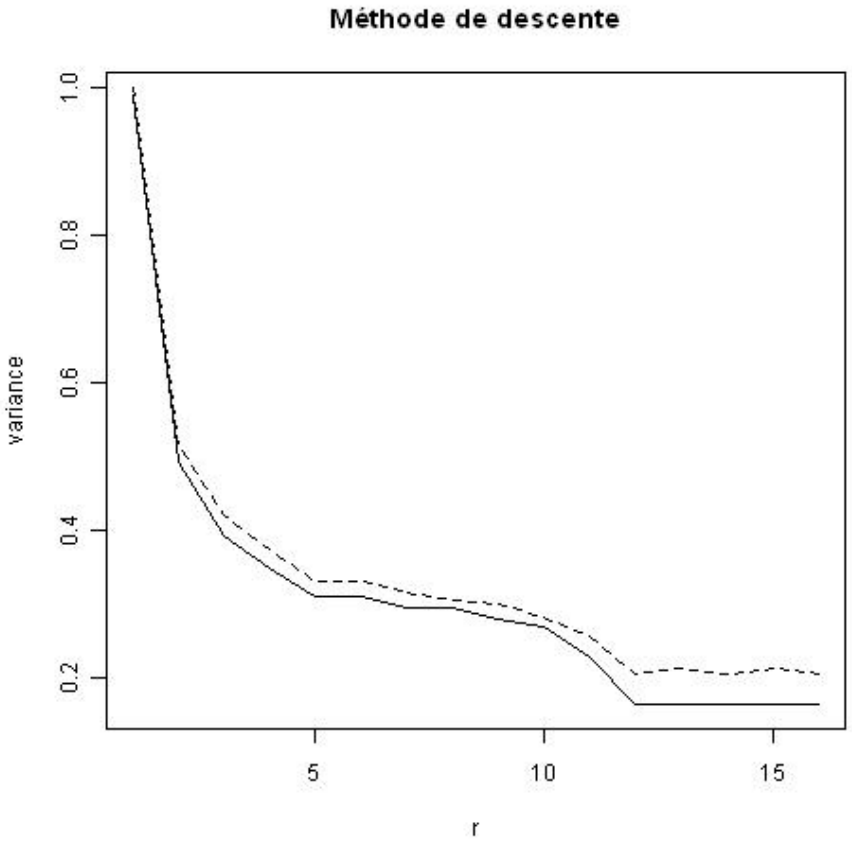
Dans le cas ou les variables d'équilibrage dépendent d'une probabilité d'inclusion π, on préfèrera utiliser pour minimiser la fonction V une méthode de descente. En effet, la méthode appliquée dans le cas de l'équilibrage sur deux variables ne marge ne convient pas dans ce cas, du fait que les variables d'équilibrage dépendent de π (cf. annexe 5).

Il s'agit d'une méthode itérative basée sur le calul de la dérivée de V par rapport à π. A chaque étape, on se déplace dans le sens opposé au projeté du gradient de V, d'une distance assez faible pour que la variance diminue, jusqu'à ce que le gradient de V soit quasi nul. Cette méthode a l'avantage de toujours converger et de toujours atteindre un minimum local dès lors qu'il existe un minimum global.

Les détails du calcul du gradient sont donnés en annexe 4.

2.3.5 Simulations

La méthode de descente telle qu'elle est décrite converge moins vite que la méthode du point fixe (il faut en général une trentaine d'itérations pour un précision de 10^{-5}). Le problème des quotas probabilistes présente l'avantage d'être toujours exact. De ce fait, la formule de variance approchée est beaucoup plus proche de la variance obtenue par simulation, ce qu'illustre le graphique suivant :



Deux variables catégorielles ont été générées aléatoirement sur une population de 600 individus. Une variable y a aussi été générée selon un modèle ANOVA avec intéractions utilisant ces deux facteurs. Les paramêtres du modèle ont été générés aléatoirement.

La méthode de descente a convergé au bout de 16 itérations pour $n = 60$. On dispose alors d'une suite de 15 vecteurs de probabilités d'inclusion. La suite des variances approchées associées à ces vecteurs est une suite décroissante représentée par la courbe pleine.

La courbe en pointillés représente la variance obtenue par simulations de l'estimateur de Horvitz Thompson otbtenu suite à un tirage équilibré sur les deux facteurs et de taille d'échantillon égale à $n = 60$. Pour chaque point, 20000 simulations ont été effectuées.

Il s'agit par ce graphique de constater que les variations de la variance approximée et de la vraie variance en fonction des probabilités d'inclusion sont les mêmes, ce qui justifie la méthode : les gains de variance approchée apportés par la méthode sont plus grands que les écarts entre variance approchée et variance empirique.

2.3.6 En pratique

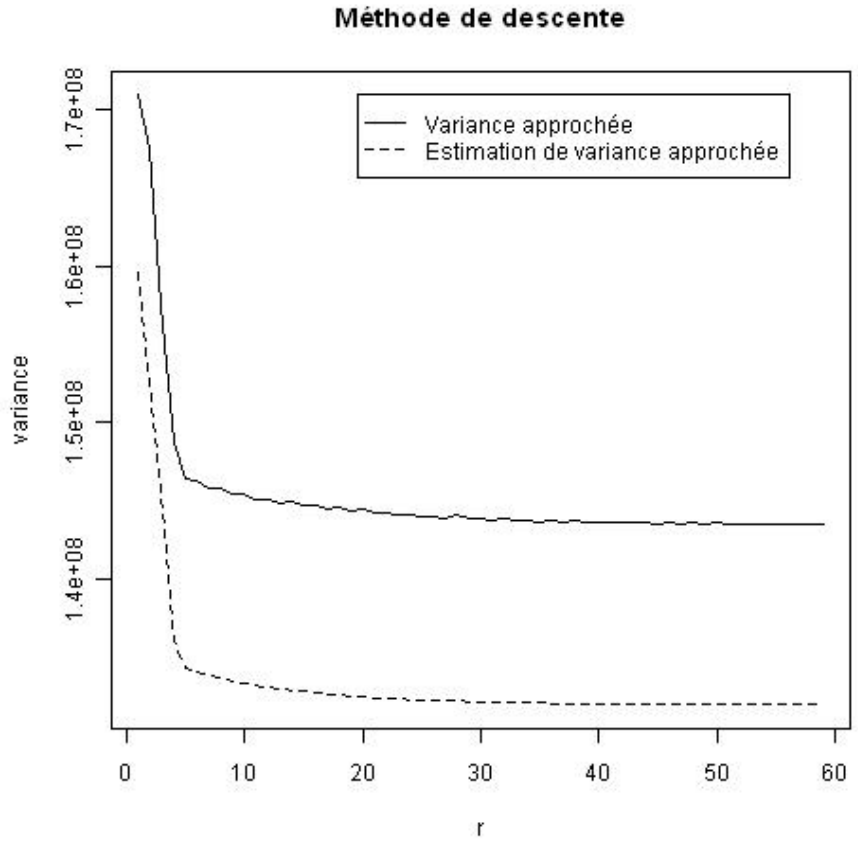
On remarque tout d'abord que la méthode de recherche de probabilités optimales nécessite uniquement la connaissance de $\sum_{k\in U_{ij}} y_k$ et $\sum_{k\in U_{ij}} y_k^2$ et N_{ij} pour tout i, j, c'est à dire la moyenne, la dispersion de y sur tout U_{ij} ainsi que la taille de tout U_{ij}

En pratique on dispose de valeurs approchées de y, $z^{i\cdot}$, $z^{\cdot j}$ sur un échantillon s de la population (par exemple un échantillon obtenu l'année précédente, comme souvent on en utilise dans le cas de Neyman pour estimer la variance par strate). Cet échantillon est accompagné de poids de sondages $(w_k)_{k\in s}$ non nécessairement constants par croisement $s \cap U_{ij}$.

Cet échantillon nous permet d'obtenir $\hat{Y}_{ij}$, $\widehat{Y_{ij}^2}$ $\hat{N}_{ij}$ des estimateurs de $Y_{ij} = \sum_{k\in U_{ij}} y_k$, $Y_{ij}^2 = \sum_{k\in U_{ij}} y_k^2$, N_{ij}.

On peut alors lancer le programme d'optimisation à partir des valeurs estimées.

Le graphique suivant rend compte des variations de la variance approchée (en trait continu) au fur et à mesure des itérations qui visent à diminuer l'estimation de la variance approchée (en pointillés).



On a estimé $\hat{Y}_{ij}$, $\widehat{Y^2}_{ij}$ $\hat{N}_{ij}$ à partir d'un sondage aléatoire simple de 60 individus parmi 600. On lance alors le programme d'optimisation à partir des valeurs estimées. On obtient une suite π^r tel que $V(\pi^r, \hat{Y}_{ij}, \widehat{Y^2}_{ij}, \hat{N}_{ij})$ décroit (courbe en pointillés). On observe ensuite les mêmes variations pour la vraie variance $V(\pi^r, Y_{ij}, Y_{ij}^2, N_{ij})$ du moins lorsque les variations sont assez grandes (lorsque $\|\pi^{r+1} - \pi^r\|$ est assez grand devant $\|(N_{ij}, Y_{ij}, Y_{ij}^2) - (\hat{N}_{ij}, \hat{Y}_{ij}, \widehat{Y^2}_{ij})\|$)

Conclusion

Ce travail a consisté à appliquer des méthodes d'analyse numérique pour minimiser une formule d'approximation de la variance d'un sondage équilibré. Dans les cas présentés, équilibrage sur des variables catégorielles ou quotas probabilistes, il est concevable que le responsable d'enquête dispose de suffisament d'informations pour pouvoir estimer l'approximation de la variance et lancer le programme d'optimisation. Selon le point de départ des itérations, le gain de variance peut être très important.

Références

[1] Ciarlet. *Introduction à l'analyse numérique matricielle et à l'optimisation*. 1982.

[2] Tillé Favre. Optimal allocation in balanced sampling. 2005.

[3] Neyman. On the two differnet aspects of representative method : the method of stratified sampling and the method of purposive selection. *Journal of royal statistical society*, 1934.

[4] Tillé. *Théorie des sondages*. 2001.

[5] Deville Tillé. Efficient balanced sampling : the cube method. *Biometrika*, 2004.

[6] Deville Tillé. Variance approximation under balanced sampling. *Journal of statistical planning and Inference*, 2005.

Annexes - Convergence des méthodes itératives

Annexe 1. Ecriture matricielle de la fonction V

On introduit ici des notations matricielles pour expliquer simplement pourquoi la résolution formelle du problème en introduisant le lagrangien n'est pas toujours possible car cela revient à calculer formellement les coefficients de l'inverse d'une matrice. On note z_π la matrice dont la $i^{\text{ème}}$ colonnes est z^i_π. Les autres notations utilisées par la suite sont celles introduites dans l'article.

$V(\pi)$ est la norme du vecteur des résidus de la régression de y sur les variables z^i pondérée par P_π la matrice de $\mathscr{M}_N(\mathbb{R}^{*+})$ de terme diagonal $p_k = \frac{1}{\pi_k} - 1$

Soit :

$$V(\pi) = \|\sqrt{P_\pi}\hat{\varepsilon}(P_\pi)\|^2$$

avec

$$\hat{\varepsilon}(P_\pi) = (I - z_\pi({}^t z_\pi P_\pi z_\pi)^{-1}\,{}^t z_\pi P_\pi)y$$

$z_\pi \in \mathscr{M}_{N,m}$ étant la matrice des variables $z^i(\pi)$

On cherche donc :

$$argmin_{\pi\in]0,1]^N}\{{}^t yP_\pi y - {}^t yP_\pi z_\pi({}^t z_\pi P_\pi z_\pi)^{-1}\,{}^t z_\pi P_\pi y\} \text{ s.c. } A\pi = B$$

avec A, B de même nombre de lignes (le nombre de contraintes linéaires) et de nombre de colonnes respectivement N et 1.

La fonction : $\pi \mapsto {}^t yP_\pi y - {}^t yP_\pi z_\pi({}^t z_\pi P_\pi z_\pi)^{-1}\,{}^t z_\pi P_\pi y$ est continue sur $]0,1]^N$ et minorée par 0 (car norme au carré d'un résidu).

On peut trouver en fonction des contraintes sur π des conditions sur y qui permettent de prouver l'existence d'un minimum global sur $]0,1]^N$.

Annexe 2. Recherche d'un minimum à partir du Lagrangien

Malheureusement, l'expression du minimum ne peut s'obtenir formellement sans inverser $({}^t z_\pi P_\pi z_\pi)$ qui intervient dans le Lagrangien :

$$\begin{array}{rcccl} \mathcal{L} & : &]0,1]^N \times \mathbb{R} & \to & \mathbb{R} \\ & : & \pi, \lambda & \mapsto & {}^t yP_\pi y - {}^t yPz_\pi({}^t z_\pi P_\pi z_\pi)^{-1}\,{}^t z_\pi P_\pi y - \lambda(A\pi - B) \end{array}$$

Exception faite de cas triviaux ou la matrice à inverser est diagonale car les variables d'équilibrage sont perpendiculaires entre elles, on ne peut calculer formellement les coefficients de l'inverse de la matrice $({}^t z_\pi P_\pi z_\pi)$, donc on ne peut calculer la dérivée du Lagrangien, ni la dérivée seconde dans le but d'obtenir des conditions nécessaires sur π pour atteindre un minimum.

En effet, la dérivée de $({}^t z_\pi P_\pi z_\pi)^{-1}$ est :

$$\frac{d({}^t z_\pi P_\pi z_\pi)^{-1}}{d\pi} = ({}^t z_\pi P_\pi z_\pi)^{-1}\big(\frac{{}^t dz_\pi}{d\pi}P_\pi z_\pi + {}^t z_\pi \frac{dP_\pi}{d\pi} z_\pi + {}^t z_\pi P_\pi \frac{dz_\pi}{d\pi}\big)({}^t z_\pi P_\pi z_\pi)^{-1}$$

Le jacobien de P_π, est une matrice diagonale, dont le $k^{\text{ème}}$ élément diagonal est $[J_{P_\pi}]_{kk} = \frac{-1}{\pi_k^2}$

Annexe 3. Méthode de Newton

On peut utiliser la méthode de Newton pour annuler la dérivée du Lagrangien, mais le calcul de la pente fait intervenir la dérivée seconde de $({}^t z_\pi P_\pi z_\pi)^{-1}$, qui fait elle même intervenir plusieurs fois l'inverse de $({}^t z_\pi P_\pi z_\pi)$. On peut craindre que les approximations numériques dans les calculs d'inverses de matrice rendent la démarche chaotique.

La méthode de Newton est donc écartée.

Annexe 4. Méthode de descente

Une approche plus raisonnable consiste à utiliser une méthode de descente ([1]).

On suppose que les contraintes ne sont pas redondantes, i.e. que A est de plein rang.

On sait calculer le gradient de :

$$\begin{array}{rcccl} V & : &]0,1]^N & \rightarrow & \mathbb{R}^+ \\ & & \pi & \mapsto & {}^t y P_\pi y - {}^t y P_\pi z_\pi ({}^t z_\pi P_\pi z_\pi)^{-1}\ {}^t z_\pi P_\pi y \end{array}$$

On a :

$$\begin{array}{rcccl} \frac{dV}{d\pi} & : &]0,1]^N & \rightarrow & \mathbb{R}^+ \\ & & \pi & \mapsto & {}^t y \frac{dP_\pi}{d\pi}(\pi) y - \frac{d({}^t y P_\pi z_\pi ({}^t z_\pi P_\pi z_\pi)^{-1}\ {}^t z_\pi P_\pi y)}{d\pi}(\pi) \end{array}$$

et

$$\begin{array}{l} d({}^t y P_\pi z_\pi ({}^t z_\pi P_\pi z_\pi)^{-1}\ {}^t z_\pi P_\pi y) = \\ \qquad {}^t y dP_\pi z_\pi ({}^t z_\pi P_\pi z_\pi)^{-1}\ {}^t z_\pi P_\pi y \\ \qquad + {}^t y P_\pi dz_\pi ({}^t z_\pi P_\pi z_\pi)^{-1}\ {}^t z_\pi P_\pi y \\ \qquad + {}^t y P_\pi z_\pi ({}^t z_\pi P_\pi z_\pi)^{-1} ({}^t dz_\pi P_\pi z_\pi + {}^t z_\pi dP_\pi z_\pi + {}^t z_\pi P_\pi dz_\pi)({}^t z_\pi P_\pi z_\pi)^{-1}\ {}^t z_\pi P_\pi y \\ \qquad + {}^t y P_\pi z_\pi ({}^t z_\pi P_\pi z_\pi)^{-1}\ {}^t dz_\pi P_\pi y \\ \qquad + {}^t y P_\pi z_\pi ({}^t z_\pi P_\pi z_\pi)^{-1}\ {}^t z_\pi dP_\pi y \end{array}$$

on pose

$$grad(V)(\pi) = \sum_k \big(\frac{dV}{d\pi}(\pi)\big)(e_k).e_k$$

où e_k est le vecteur ayant pour valeur 1 pour la $k^{\text{ème}}$coordonnée, 0 pour les autres.

On a $grad(V) = (A({}^t AA)^{-1}\ {}^t A) grad(V) + (I - (A({}^t AA)^{-1}\ {}^t A)) grad(V)$

Pour rester dans l'espace des contraintes, on "descendra" en restant dans $A^\perp$. Il suffit pour obtenir la direction de descente $\Delta(\pi)$ à partir du point π de projeter $-grad(V)(\pi)$ sur $A^\perp = Ker(A)$

On pose $\Delta(\pi) = -(I - (A({}^t AA)^{-1}\ {}^t A)) grad(V)(\pi)$

On définit la méthode d'itération suivante :

Soit π^0 tel que $A\pi^0 = B$

On définit la suite π^r par récurrence :

$$h^r = min\{k \in \mathbb{N} | \pi^r + \frac{\Delta(\pi^r)}{k} \in]0,1]^N, V(\pi^r + \frac{\Delta(\pi^r)}{k}) < V(\pi^r)\}$$

$$\pi^{r+1} = \pi_r + \frac{\Delta(\pi^r)}{h^r}$$

Annexe 5. Méthode du point fixe

On pose

$$\begin{array}{rcccl} v & : &]0,1]^N \times \mathbb{R}^N & \rightarrow & \mathbb{R}^+ \\ & : & (\pi, \tilde{y}) & \mapsto & \sum_{k=1}^N (\frac{1}{\pi_k} - 1)(y_k - \tilde{y}_k)^2 \end{array}$$

On note $\hat{y}(\pi) = z_\pi ({}^t z_\pi P_\pi z_\pi)^{-1} ({}^t z_\pi P_\pi y)$ le projeté orthogonal de y sur $Vect(z^1(\pi) \dots z^m(\pi))$ suivant le produit scalaire défini par P_π : $< a|b >= \sum_k (\frac{1}{\pi_k} - 1) a_k b_k$

On a :

- $\begin{array}{rcl} \hat{y}(\pi) & = & argmin_{\tilde{y} \in vect(z_1 \dots z_m)} \sum_{k=1}^N (\frac{1}{\pi_k} - 1)(y_k - \tilde{y}_k)^2 \\ & = & argmin_{\tilde{y} \in vect(z_1 \dots z_m)} v(\pi, \tilde{y}) \end{array}$
- $V(\pi) = v(\pi, \hat{y}(\pi))$

Soit $\pi^0 \in]0,1]^N$ tel que $A\pi^0_k = B$. On définit par récurrence la suite $(\pi^r)_{r \in \mathbb{N}}$:

- $\tilde{y}^0 = \hat{y}(\pi^0)$
- si $argmin_{\{x \in]0,1]^N | Ax = B\}} \{v(x, \tilde{y}^r)\}$ existe et est réduit à un élément, alors on pose : $\pi^{r+1} = argmin_{\{x \in]0,1]^N | Ax = B\}} \{v(x, \tilde{y}^r)\}$ sinon abandon.
- $\tilde{y}^{r+1} = \hat{y}(\pi^r)$

Dans le cas ou z_π ne dépend pas de π, si $\forall r$, $argmin_{\{x \in]0,1]^N | Ax = B\}} \{v(x, \tilde{y}^r)\}$ existe, alors la suite des $v(\hat{y}(\pi^r), \pi^r)$

est décroissante (car $v(\hat{y}(\pi^{r+1}), \pi^{r+1}) < v(\hat{y}(\pi^r), \pi^{r+1}) < v(\hat{y}(\pi^r), \pi^r)$) et positive donc convergente.

Cette procédure peut donc permettre d'améliorer la variance d'un estimateur issu d'un tirage équilibré, en fournissant un jeu de probabilités d'inclusion meilleur que celui de départ.

Mais

- on n'est pas sûr de pouvoir passer à l'étape suivante (problème d'existence de l'argmin).
- on n'a pas démontré qu'il y avait convergence des π^r quand $r \rightarrow +\infty$
- même en cas de convergence, on n'est même pas sûr d'avoir atteint un minimum local contraint de la fonction V.

Toutefois, en cas de convergence des π^r vers un certain π^*, on a :

$$\pi^* = argmin_{\{x \in]0,1]^N | Ax = B\}} \{v(x, \hat{y}(\pi^*))\}$$

$$\hat{y}(\pi^*) = argmin_{\{\tilde{y} \in vect(z_\pi)\}} \{v(\pi^*, \tilde{y})\}$$

π^* est un point fixe de : $x \mapsto argmin_{\{x \in]0,1]^N | Ax = B\}} \{v(x, \hat{y}(x))\}$

- $x \mapsto v(x, \hat{y}(\pi^*))$ atteignant un minimum local sous contrainte que $Ax = B$ en $x = \pi^*$, on en déduit que pour $d\pi$ tel que $Ad\pi = 0$, on a :

$$\Big(\frac{dv(x, \tilde{y})}{dx} (\pi^*, \hat{y}(\pi^*)) \Big) (d\pi) = 0$$

- $\tilde{y} \mapsto v(\pi, \tilde{y})$ atteignant un minimum local sur $vect(z_\pi)$ en $\tilde{y} = \hat{y}(\pi^*)$, on en déduit que pour $d\tilde{y}$ tel que $d\tilde{y} \in vect(z_\pi)$, on a :

$$\Big(\frac{dv(x, \tilde{y})}{d\tilde{y}} (\pi^*, \hat{y}(\pi^*)) \Big) (d\tilde{y}) = 0$$

Or

$$\Big(\frac{d\hat{y}(\pi)}{d\pi}(\pi)\Big) = \Big(\frac{d\big(z_\pi({}^t z_\pi P_\pi z_\pi)^{-1}\ {}^t z_\pi P_\pi y\big)}{d\pi}(\pi)\Big)$$

Dans le cas ou z_π ne dépend pas de π (i.e. $z = z_\pi$), on a :

$$\begin{array}{rcl}
\big(\frac{d\hat{y}(\pi)}{d\pi}(\pi)\big)(d\pi) & = & \frac{dz({}^t z P_\pi z)^{-1}\ {}^t z P_\pi y}{d\pi} d\pi \\
& = & -z({}^t z P_\pi z)^{-1}({}^t z \frac{dP_\pi}{d\pi} d\pi z)({}^t z P_\pi z)^{-1}\ {}^t z P_\pi y \\
& & +z({}^t z P_\pi z)^{-1}\ {}^t z \frac{dP_\pi}{d\pi} d\pi y \\
& = & z({}^t z P_\pi z)^{-1}({}^t z \frac{dP_\pi}{d\pi} d\pi)\big(- z({}^t z P_\pi z)^{-1}\ {}^t z P_\pi + I_N\big) y \\
& \in & vect(z)
\end{array}$$

Dans le cas ou z_π dépend de π, on a :

$$\begin{array}{rcl}
\big(\frac{d\hat{y}(\pi)}{d\pi}(\pi)\big)(d\pi) & = & \frac{d\Big(z_\pi({}^t z_\pi P_\pi z_\pi)^{-1}\ {}^t z_\pi P_\pi y\Big)}{d\pi} d\pi \\
& = & \big(\frac{dz_\pi}{d\pi}(d\pi)\big)({}^t z_\pi P_\pi z_\pi)^{-1}\ {}^t z_\pi P_\pi y \\
& & +z_\pi \frac{d\Big(({}^t z_\pi P_\pi z_\pi)^{-1}\ {}^t z_\pi P_\pi\Big)}{d\pi}(d\pi) y
\end{array}$$

On en déduit qu'on atteint un extremum local si et seulement si

$$\forall d\pi \in A^\perp, \big(\frac{dz_\pi}{d\pi}(\pi^*)\big)(d\pi)({}^t z_{\pi^*} P_{\pi^*} z_{\pi^*})^{-1}\ {}^t z_{\pi^*} P_{\pi^*} y \in vect(z_{\pi^*})$$

Dans le cas ou z_π ne dépend pas de π ($z = z_\pi$), on a : $d\tilde{y} = \frac{d\hat{y}(\pi)}{d\pi} d\pi$ vérifie $d\tilde{y} \in vect(z_\pi)$
Ce qui entraine :

$$\begin{array}{rcccc}
\big(\frac{dV}{d\pi}(\pi)\big)(d\pi) & = & \big(\frac{dv(\pi,\hat{y}(\pi))}{d\pi}(\pi)\big)(d\pi) & & \\
& = & \big(\frac{\partial v}{\partial x}(\pi,\hat{y}(\pi))\big)(d\pi) & + & \frac{\partial v}{\partial \tilde{y}}(\pi,\hat{y}(\pi)) \frac{d\hat{y}(\pi)}{d\pi} d\pi \\
& = & 0 & + & \frac{\partial v}{\partial \tilde{y}}(\pi,\hat{y}(\pi)) \frac{d\hat{y}(\pi)}{d\pi} d\pi \\
& = & 0 & + & 0
\end{array}$$

En cas de convergence de π^r vers π^*, et dans le cas ou les variables d'équilibrage ne dépendent pas de π, on peut conclure que V admet un minimum local en π^*.