UN ALGORITHME DE TIRAGE EQUILIBRE

G. ROY et A. VANHEUVERZWYN

MEDIAMETRIE - Direction Recherche et Méthodes

Cette communication a été présentée au Deuxième Colloque Francophone sur les Sondages (Bruxelles, 22-23 juin 2000) et sera publiée dans les Actes de ce Colloque, sous copyright.

1. Origine du problème

Lorsqu'on désire appliquer la technique du Bootstrap, plusieurs optiques sont envisageables comme respecter la structure *i.i.d.* telle que l'a décrite Efron à l'origine ou reproduire au mieux la méthode d'échantillonnage initiale. Dans le cadre des enquêtes par quotas marginaux, la deuxième optique nous conduit à la recherche d'un algorithme permettant de répondre, de manière exacte, aux contraintes de quotas.

Cet article propose trois démarches permettant de répondre à la problématique. La recherche de différentes démarches a été motivée par la nécessité de minimiser le temps de calcul de l'algorithme. En effet, l'intérêt du bootstrap est de pouvoir répliquer l'échantillon un grand nombre de fois (au moins 1000). L'algorithme proposé, pour pouvoir être utilisable, doit donc être le plus rapide possible et pouvoir s'adapter au nombre de critères de quotas.

2. Problématique

On dispose, dans une base de sondage, de Q variables qualitatives utilisées comme critères de quotas. On désire tirer un échantillon de taille fixe n, selon un procédé aussi proche que possible du sondage aléatoire simple mais respectant des contraintes de quotas marginaux sur ces Q critères. Il s'agit, par cette méthode, de reconstituer la démarche de l'enquêteur lors d'un sondage par quotas marginaux.

Prenons le cas idéal pour lequel tous les croisements de modalités des différents critères sont possibles. Dans ce cas, il suffit de tirer aléatoirement les unités une par une, de décrémenter les compteurs sur les quotas marginaux et d'annuler, dès qu'un quota est saturé, les probabilités de tirage des unités possédant la modalité en question. L'idée est donc de tirer des unités aléatoirement avec remise jusqu'à ce

que l'échantillon soit calé sur les quotas marginaux imposés. Or, dans la pratique, cette configuration est peu probable : certains croisements de modalités sont impossibles.

Prenons l'exemple d'une base de sondage de 18425 individus dans laquelle on dispose des critères d'âge et d'activité utilisés comme critères de quotas. Le croisement des deux critères nous donne les effectifs suivants pour chaque cellule :

	15-24 ans	25-34 ans	35-49 ans	50-64 ans	65 ans et plus
CSP +	43	765	1959	1192	26
CSP -	196	2172	2543	968	11
Retraités	0	0	6	707	3293
Autres inactifs	2774	432	476	557	305

Dans ce cas, on ne peut croiser les modalités « de 15 à 24 ans » et « de 25 à 34 ans » avec la modalité « retraités ». On ne trouve, dans notre échantillon, des retraités que parmi les individus des tranches d'âge « de 35 à 49 ans », « de 50 à 64 ans » et « 65 ans et plus ». Si on utilise l'algorithme simple exposé précédemment, il se peut qu'on sature les quotas sur ces dernières tranches d'âge avant d'avoir rempli le quota sur la modalité « retraités ». Ayant annulé les probabilités de tirage de ces individus, on ne pourra alors plus tirer de retraité et le quota sur cette modalité ne pourra être saturé.

Dans ce cas précis, il s'agit de ne pas saturer les classes d'âge dont le croisement est possible avec les retraités avant d'avoir rempli la classe des retraités. Notre problématique est donc d'automatiser la gestion des quotas d'une façon ne nous acculant pas à la recherche du « mouton à cinq pattes ». Pour cela, trois démarches sont envisagées.

Par la suite, nous appellerons « modalités problématiques » les modalités dont le croisement avec au moins une modalité d'un autre critère est impossible.

3. Les différentes démarches

Les trois approches envisagées, détaillées ci-dessous, diffèrent de par le niveau de restriction des contraintes imposées aux unités pour pouvoir être sélectionnées. En effet, intuitivement, le niveau de restriction semble être corrélé, négativement, avec la vitesse d'exécution du programme.

a) Une première idée, la plus simple à mettre en œuvre, consiste à s'astreindre, au début de la procédure, au tirage d'unités possédant une modalité problématique. Ce premier temps ayant évacué la question des modalités problématiques, on annule ensuite la probabilité de tirage des unités, déjà tirées ou non, présentant l'une ou l'autre de ces modalités problématiques. On

remplit notre échantillon en tirant aléatoirement avec remise parmi les unités restantes. Pour respecter les contraintes imposées par les quotas marginaux, dès qu'on atteint le nombre requis pour une modalité, on annule les probabilités de tirage des unités possédant cette modalité. Et ainsi de suite jusqu'à saturation des quotas sur toutes les modalités.

b) La seconde optique envisagée consiste à contrôler les écarts de remplissage entre modalités de quotas, c'est-à-dire que les quotas devront se remplir à un rythme similaire. Arrivé à un seuil de remplissage global fixé préalablement, on se préoccupera prioritairement des modalités problématiques comme dans la méthode a).

c) La dernière démarche représente un moyen terme entre les précédentes. On s'autorisera à accepter un pourcentage x d'unités non problématiques avant de remplir les modalités problématiques. Une fois le nombre désiré d'unités à modalités problématiques atteint, on poursuit de la même manière qu'en a).

Ces trois démarches sont équivalentes dans la mesure où elles permettent de respecter les contraintes de quotas. La première, la plus intuitive, est sans doute la plus proche de la démarche réelle de l'enquêteur, mais elle semble être la plus restrictive et donc la plus lente. Le critère de choix entre les trois méthodes sera la rapidité d'exécution du programme.

Les démarches proposées commencent toutes trois par une hiérarchisation des critères de quotas selon le nombre de modalités problématiques qu'ils possèdent afin de minimiser le nombre de tests dans la procédure. Si on reprend l'exemple du paragraphe **2**, on a trois modalités problématiques : la modalité « retraités » du critère d'activité et les modalités « de 15 à 24 ans » et « de 25 à 34 ans » du critère d'âge. Si on s'affranchit du problème des « retraités », on aura, par là-même, évacué celui des classes d'âge incompatibles avec la modalité « retraités ». De même, on pourrait résoudre le problème en s'occupant des classes d'âge problématiques. Fort de ce constat, c'est par le critère possédant le moins de modalités problématiques que le problème sera évacué, le nombre de tests dans la procédure étant ainsi minimisé.

4. Méthode (a) : algorithme

La première étape de l'algorithme consiste à détecter les modalités problématiques des différents critères présents dans la base de sondage. Les critères de quotas sont ensuite hiérarchisés selon le nombre de modalités problématiques qu'ils possèdent, l'idée étant de minimiser le nombre de tests dans la procédure. On notera $\mathbf{Q_m}$ le critère possédant le moins de modalités problématiques.

Dans la seconde étape de l'algorithme, les unités sont sélectionnées par la méthode suivante.

On répète jusqu'à atteindre la taille d'échantillon fixée préalablement :

① On tire une unité aléatoirement.

② **Si** les quotas sur les modalités problématiques du critère $\mathbf{Q_m}$ ne sont pas saturés et que cette unité ne présente aucune modalité problématique pour le critère $\mathbf{Q_m}$, on refuse l'unité et on retourne à l'étape ①.

Sinon :

2.1. On sélectionne l'unité.

2.2. On décrémente les quotas sur les modalités qu'elle présente.

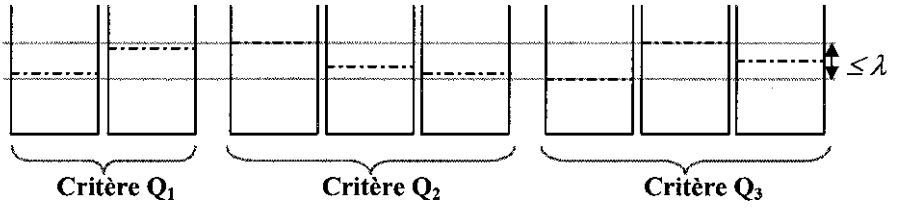
2.3.1. **Si** le quota sur une des modalités de l'unité sélectionnée est saturé, on annule les probabilités de sélection de toutes les unités présentant la modalité en question puis on repart à l'étape ①.

2.3.2. **Sinon**, on repart à l'étape ①.

Cette démarche est donc très contraignante dans la mesure où l'on s'astreint, dès le début de la procédure, à saturer les quotas des modalités problématiques. Mais c'est aussi la méthode la plus proche de la démarche réelle de l'enquêteur. En effet, afin de gérer le respect quotidien des quotas marginaux, le chef d'équipe impose à ses enquêteurs des priorités selon la difficulté d'atteinte de la cible et les contraintes de croisements. Ainsi, on préfèrera joindre en priorité des personnes retraitées et âgées de 50 ans et plus, car, d'une part, elles sont plus méfiantes à l'égard des appels téléphoniques et d'autre part, elles font partie des cibles problématiques.

5. Méthode (b) : Gestion des écarts de remplissage

Le processus de gestion des écarts de remplissage - méthode (b) - consiste à limiter à λ% l'écart maximal de remplissage des différentes modalités. Le schéma ci-dessous représente les taux de remplissage (en pourcentage) des différentes modalités de quotas.



λ sera déterminé empiriquement comme compromis entre la finesse de l'écart et la rapidité d'exécution du programme.

Toutefois, ce procédé ne permet pas de gérer les contraintes engendrées par l'existence de modalités problématiques. Il convient donc de déterminer un seuil global correspondant au pourcentage d'unités qu'il reste à tirer pour remplir l'échantillon ; à partir de ce seuil, on s'attachera prioritairement aux modalités problématiques. Ce seuil sera noté s.

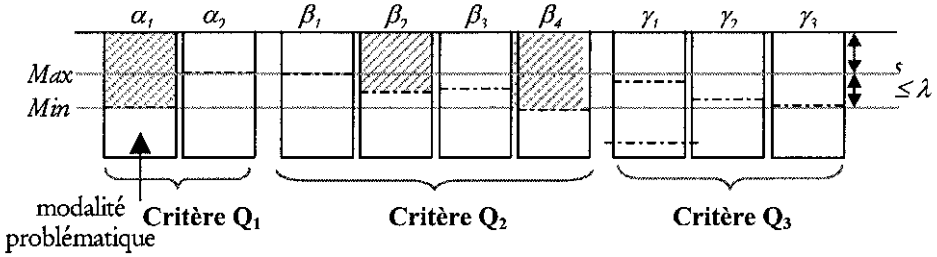
On note α_i les modalités du critère $\mathbf{Q_1}$, β_j les modalités du critère $\mathbf{Q_2}$ et γ_k les modalités du critère $\mathbf{Q_3}$. Comme pour les autres démarches, les critères de quotas sont hiérarchisés selon le nombre de modalités problématiques qu'ils possèdent. On supposera, dans l'ensemble de cette partie, que c'est le critère $\mathbf{Q_1}$ qui possède le moins de modalités problématiques.

5.1. Cas d'une seule modalité problématique

On suppose que la modalité α_1 est la seule modalité problématique du critère $\mathbf{Q_1}$.

Prenons le cas extrême où c'est la modalité problématique de notre critère qui est remplie « au plus bas » (*Min*) et où les autres modalités de ce même critère sont remplies « au plus haut » (*Max*).

On suppose que la modalité α_1 peut être croisée avec les modalités β_2 et β_4 pour le critère $\mathbf{Q_2}$ et γ_3 pour le critère $\mathbf{Q_3}$ et que les croisements de β_2 et β_4 avec γ_3 sont possibles. (Ces modalités α_1, β_2, β_4 et γ_3 sont hachurées dans le schéma ci-dessous).



Dans le pire des cas, il restera à remplir $(s + \lambda)$ % de la modalité problématique, où s% des modalités correspondent à des croisements possibles avec la modalité problématique.

$$(s+\lambda)n_{\alpha_1} = s.Min\left((n_{\beta_2} + n_{\beta_4}), n_{\gamma_3}\right)$$

Donc $\Updownarrow$

$$s = \frac{\lambda.n_{\alpha_1}}{Min\left((n_{\beta_2} + n_{\beta_4}), n_{\gamma_3}\right) - n_{\alpha_1}}$$

5.2. Cas de deux modalités problématiques ou plus

Soit :

I l'ensemble des modalités problématiques du critère $\mathbf{Q_1}$,

J_I l'ensemble des modalités du critère $\mathbf{Q_2}$ dont les croisements avec les modalités de I sont possibles,

K_I l'ensemble des modalités du critère $\mathbf{Q_3}$ dont les croisements avec les modalités de I sont possibles ; les croisements de J_I et K_I sont supposés possibles.

On peut alors généraliser la formule exposée dans le **5.1.** à savoir :

$$(s+\lambda)\sum_{i\in I} n_{\alpha_i} = s.Min\left(\sum_{j\in J_I} n_{\beta_j}, \sum_{k\in K_I} n_{\gamma_k}\right)$$

$$\Updownarrow$$

$$s = \frac{\lambda.\sum_{i\in I} n_{\alpha_i}}{Min\left(\sum_{j\in J_I} n_{\beta_j}, \sum_{k\in K_I} n_{\gamma_k}\right) - \sum_{i\in I} n_{\alpha_i}}$$

6. Méthode (c) : Recherche du x optimal

La recherche du x optimal correspond à la démarche (c) (cf. **3**). Le x optimal est le nombre maximal d'unités à sélectionner ne possédant pas de modalités problématiques, avant de s'attaquer au remplissage de ces modalités problématiques.

6.1. Exemple sur deux dimensions

Prenons le cas simple où on a deux critères qualitatifs. La modalité α_1 du critère $\mathbf{Q_1}$ n'admet pas de croisement avec les modalités β_3 et β_5 du critère $\mathbf{Q_2}$.

		Critère Q_2					
		β_1	β_2	β_3	β_4	β_5	
Critère Q_1	α_1	$n_{1,1}$	$n_{1,2}$		$n_{1,4}$		$n_{\alpha 1}$
	α_2	$n_{2,1}$	$n_{2,2}$	$n_{2,3}$	$n_{2,4}$	$n_{2,5}$	$n_{\alpha 2}$
	α_3	$n_{3,1}$	$n_{3,2}$	$n_{3,3}$	$n_{3,4}$	$n_{3,5}$	$n_{\alpha 3}$
		$n_{\beta 1}$	$n_{\beta 2}$	$n_{\beta 3}$	$n_{\beta 4}$	$n_{\beta 5}$	n

Idée : Il faut qu'on puisse encore trouver $n_{\alpha 1}$ unités parmi les unités des modalités β_1, β_2 et β_4.

On peut donc s'autoriser à remplir ces modalités β_1, β_2 et β_4 de façon que les trois effectifs correspondant respectent : $x \leq n_{\beta_1} + n_{\beta_2} + n_{\beta_4} - n_{\alpha_1}$

Remarque : Le remplissage de β_3 et β_5 n'influe en rien celui de α_1.

Dans le cas où on ne tirerait tout d'abord que des unités ne possédant pas la modalité α_1 du critère $\mathbf{Q_1}$, il se pourrait qu'on remplisse les modalités β_1, β_2 et β_4 du critère $\mathbf{Q_2}$. On aurait donc ensuite la modalité α_1 à remplir en ne disposant plus que d'unités à modalités β_3 et β_5 du critère $\mathbf{Q_2}$, ce qui est impossible. Il faut donc se laisser une marge de manœuvre nous assurant de ne pas être acculés à cette situation. Le nombre acceptable d'unités non problématiques doit donc être inférieur ou égal à la quantité $n_{\beta_1} + n_{\beta_2} + n_{\beta_4} - n_{\alpha_1}$.

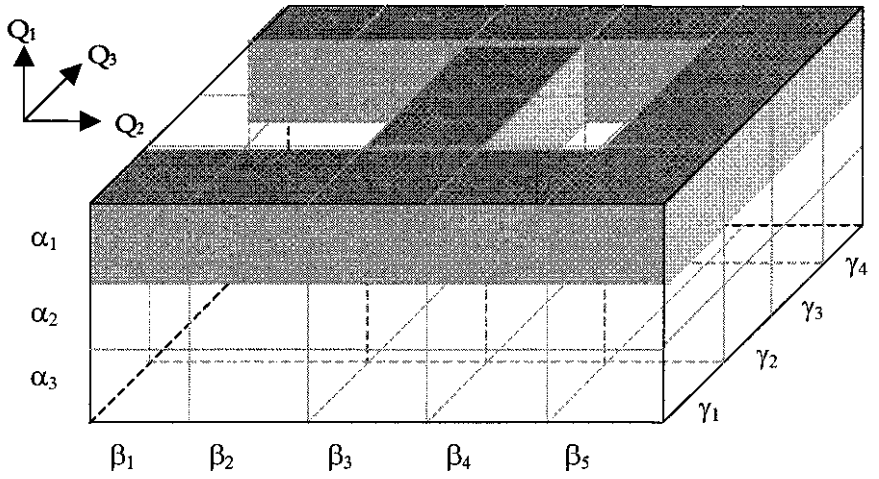
Le x optimal sera donc ici : $x_{opt} = n_{\beta_1} + n_{\beta_2} + n_{\beta_4} - n_{\alpha_1}$

6.2. Passage à trois dimensions

On ajoute un troisième critère à, par exemple, quatre modalités γ_1, γ_2, γ_3 et γ_4, où les croisements de γ_1 et γ_4 avec α_1 sont impossibles.

La modalité α_1 est alors incompatible avec les modalités β_3, β_5 et γ_1, γ_4, comme l'illustre le schéma ci-dessous.

Figure 1 : Modélisation du croisement de trois critères où α_1 est la seule modalité problématique du critère Q_1



Dans la même logique que précédemment, le x optimal vaut :

$$x_{opt} = n_{\beta_1\gamma_2} + n_{\beta_1\gamma_3} + n_{\beta_2\gamma_2} + n_{\beta_2\gamma_3} + n_{\beta_4\gamma_2} + n_{\beta_4\gamma_3} - n_{\alpha_1}$$

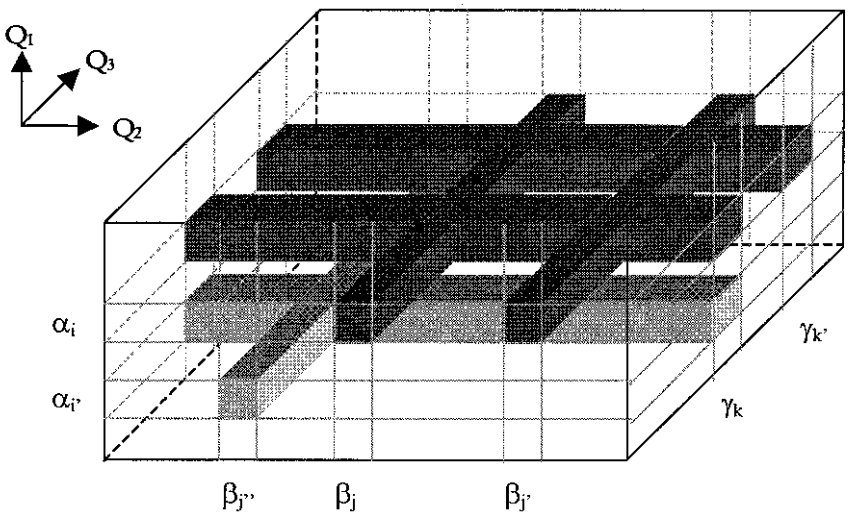
On peut, sans complication, généraliser le cas précédent à n_1, n_2, et n_3 modalités pour les critères $\mathbf{Q_1}$, $\mathbf{Q_2}$ et $\mathbf{Q_3}$. Dans ce cas, le x optimal vaut :

$$x_{opt} = \sum_{J \neq j,j'} \sum_{K \neq k,k'} n_{\beta_j\gamma_K} - n_{\alpha_i}$$

6.3. Généralisation à deux modalités problématiques et plus

On se place ici dans le cas où on dispose de trois critères possédant n_q modalités chacune (q = 1 à 3). Au moins deux modalités sont problématiques dans chaque critère.

Figure 2 : Modélisation du croisement de trois critères (chaque critère présente plus d'une modalité problématique)



On peut alors écrire $x_{opt_{Q_1}} = \sum_{J \neq j,j',j''} \sum_{K \neq k,k'} n_{\beta_J \gamma_K} - \left(n_{\alpha_i} + n_{\alpha_{i'}}\right)$ dans le cas où on prend le critère $\mathbf{Q_1}$ comme référence. Mais on pourrait de même considérer $\mathbf{Q_3}$ comme référence et alors, symétriquement, $x_{opt_{Q_3}} = \sum_{J \neq j,j',j''} \sum_{I \neq i,i'} n_{\alpha_I \beta_J} - \left(n_{\gamma_k} + n_{\gamma_{k'}}\right)$.

On montre facilement, mais au prix d'une formalisation un peu lourde (cf. annexe), que $x_{opt_{Q_1}} = x_{opt_{Q_3}} = x_{opt}$. De même, lors d'une généralisation à X critères dont p possèdent au moins une modalité problématique, on obtient l'unicité du x optimal.

Le choix du critère auquel on s'attachera en priorité est donc totalement arbitraire et on choisira de préférence celui possédant le moins de modalités problématiques afin de limiter le nombre de tests dans la procédure.

7. Conclusion

Les trois méthodes présentées ont été programmées et testées. Leurs vitesses d'exécution sont équivalentes et elles permettent de répondre de manière exacte à la problématique. C'est finalement la méthode (a) qui a été appliquée à la problématique initiale. Cet algorithme peut être utilisé aussi bien dans le cadre de l'échantillonnage que dans celui du rééchantillonnage. La technique du Bootstrap permet d'obtenir la meilleure approximation non paramétrique possible de la

statistique étudiée ; c'est notamment dans le cas de statistiques non linéaires que cette méthode prend tout son intérêt. Étant donné la place des enquêtes par quotas dans les instituts de sondage privés, l'utilisation du Bootstrap dans son approche « model-dependent » pourrait se développer dans les prochaines années.

Bibliographie

BICKEL P.J., FREEDMAN D.A. (1981). Some asymptotic theory for the bootstrap. *The Annals of Statistics*, Vol. 9, n°6, pp 1196-1217.

BICKEL P.J., FREEDMAN D.A. (1984). Asymptotic normality and the bootstrap in stratified sampling. *The Annals of Statistics*, Vol. 12, pp 470-482.

DEVILLE J.C. (1991). Une théorie des enquêtes par quotas. *Techniques d'enquête*, Vol. 17, n°2, pp 177-195. Statistique Canada.

EFRON B. (1977). Bootstrap methods : another look at the Jackknife. *The Annals of Statistics*, Vol. 7, n°1, pp 1-26.

KISH L., FRANKEL M.R. (1974). Inference from complex samples. *Journal of the Royal Statistical Society*, B, pp 1-37.

LECOUTRE J.P., TASSI P. (1987). *Statistique non paramétrique et robustesse*. Coll. Economie et Statistiques avancées. Economica.

SHAO J., TU D. (1996). *The Jacknife and Bootstrap*. Coll. Springer Series in Statistics. Springer-Verlag.

TILLÉ Y. (1998). *Théorie des sondages*. Polycopié ENSAI.

Annexe

Il s'agit ici de démontrer l'unicité du x optimal, c'est-à-dire l'égalité des $x_{opt_{Q_1}}$, $x_{opt_{Q_2}}$ et $x_{opt_{Q_3}}$. Pour cela, on reprend le contexte exposé dans le paragraphe **6.3**.

Développons l'expression de $x_{opt_{Q_1}}$:

$$
\begin{aligned}
x_{opt_{Q_1}} &= \sum_{J\neq j,j',j''} \sum_{K\neq k,k'} n_{\beta_J\gamma_K} - \left(n_{\alpha_i} + n_{\alpha_{i'}}\right) \\
&= \sum_{I} \sum_{J\neq j,j',j''} \sum_{K\neq k,k'} n_{\alpha_I\beta_J\gamma_K} - \sum_{I=i,i'} \sum_{J} \sum_{K} n_{\alpha_I\beta_J\gamma_K} \\
&= \sum_{I\neq i,i'} \sum_{J\neq j,j',j''} \sum_{K\neq k,k'} n_{\alpha_I\beta_J\gamma_K} - \sum_{I=i,i'} \sum_{J=j,j',j''} \sum_{K=k,k'} n_{\alpha_I\beta_J\gamma_K} \\
&= \sum_{K\neq k,k'} \sum_{J\neq j,j',j''} \sum_{I\neq i,i'} n_{\alpha_I\beta_J\gamma_K} - \sum_{K=k,k'} \sum_{J=j,j',j''} \sum_{I=i,i'} n_{\alpha_I\beta_J\gamma_K} \\
&= \sum_{K} \sum_{J\neq j,j',j''} \sum_{I\neq i,i'} n_{\alpha_I\beta_J\gamma_K} - \sum_{K=k,k'} \sum_{J} \sum_{I} n_{\alpha_I\beta_J\gamma_K} \\
&= \sum_{J\neq j,j',j''} \sum_{I\neq i,i'} n_{\alpha_I\beta_J} - \left(n_{\gamma_k} + n_{\gamma_{k'}}\right) \\
&= x_{opt_{Q_3}}
\end{aligned}
$$

On montre, de la même manière, que $x_{opt_{Q_1}} = x_{opt_{Q_2}}$

On peut donc écrire : $\boxed{x_{opt_{Q_1}} = x_{opt_{Q_2}} = x_{opt_{Q_3}} = x_{opt}}$