RÉSOLUTION D'UNE DES LIMITES DE L'ALLOCATION DE NEYMAN

Malik KOUBI (*), Sandrine MATHERN (**)

(*) Insee, (**) Drees

Introduction

Dans un plan de sondage stratifié, l'allocation de Neyman optimise la précision pour l'estimateur de la moyenne d'une variable d'intérêt au niveau de l'ensemble de la population. Cette allocation peut toutefois souffrir d'imprécision lorsqu'on veut obtenir des estimations sur des sous-groupes de la population.

L'amélioration de la méthode de Neyman proposée ici permet d'assurer une précision fixée à l'avance dans ces sous-groupes appelés « regroupements de publication ». L'hypothèse faite est que chaque regroupement de publication est une réunion de strates d'échantillonnage.

La méthode consiste à résoudre un programme de Neyman modifié, dans lequel sont ajoutées des contraintes de précision au niveau des regroupements de publication. La solution du programme se présente sous la forme d'une fonction par morceau, dont il faut déterminer les intervalles de définition et l'expression.

Une illustration de la méthode est proposée. Elle consiste à calculer de manière exacte la frontière d'efficacité, expression analytique du dilemme entre précision globale et précision locale. Cette frontière permet d'évaluer la perte en précision globale lorsque l'on tient à fixer une précision minimale dans des regroupements de publication. Elle constitue de ce fait une aide à la décision précieuse pour l'échantillonnage, car elle synthétise les arbitrages devant lesquels le statisticien est placé.

1. Limites de l'allocation de Neyman et amélioration proposée

1.1. Cadre de référence et notations

On souhaite estimer la moyenne $\widehat{\overline{Y}}$ d'une variable d'intérêt Y.
Considérons une population de référence, notée U, de taille N, dont les unités sont notées i. On stratifie cette population, selon des variables fortement corrélées à Y, en H strates de taille $\mathrm{N_h}$. La variance de la variable Y dans la strate *h* est notée $\mathrm{S_h^2}$.

On sélectionne un échantillon de n_h unités dans chacune des strates de manière aléatoire et indépendante, afin de constituer un échantillon de taille globale n.

La précision de notre estimateur de Y est mesurée par sa variance $V(\widehat{\overline{Y}})$ ou par son coefficient de variation $CV(\widehat{\overline{Y}})$.

1.2. Allocation de Neyman et ses limites

L'allocation optimale de Neyman
Dans un sondage stratifié avec tirage aléatoire simple dans chaque strate, l'allocation de Neyman est celle qui minimise la variance de l'estimateur de la moyenne de Y pour une taille d'échantillon fixée.

Formellement, on résout :

$$\begin{cases} \underset{n_1,\ldots,n_H}{Min}\ V(\widehat{\bar{Y}}) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 * \left(1 - \frac{n_h}{N_h}\right) * \frac{S_h^2}{n_h} \\ \text{s/c} \quad \sum_{h=1}^{H} n_h = n \end{cases}$$

L'allocation de Neyman est de la forme :

$$n_h = \frac{N_h * S_h}{\sum_{h=1}^{H} N_h * S_h} * n$$

Le nombre d'unités tirées dans chaque strate *h* est proportionnel à la grandeur de la strate *h* en termes d'effectifs et à la dispersion de la variable d'intérêt dans cette strate.

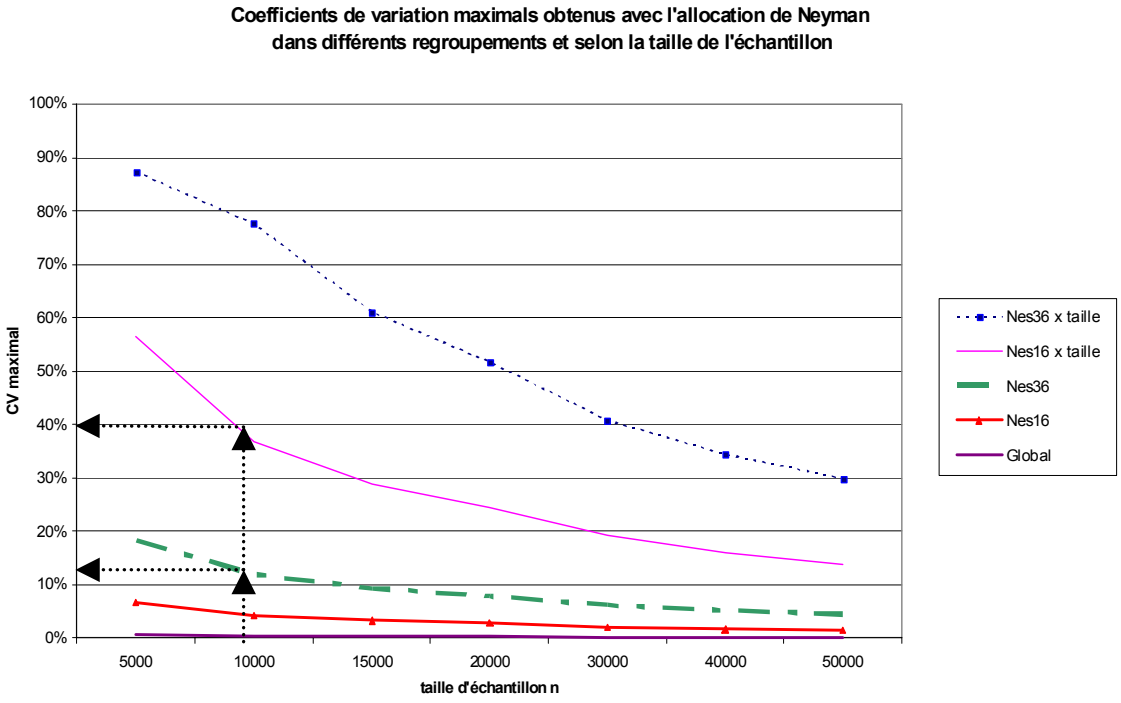
Deux problèmes potentiels liés à l'allocation de Neyman

En pratique, on peut trouver un (ou plusieurs) n_h optimal(aux) supérieur(s) au nombre d'individus N_h présents dans la strate *h*. Ce premier problème peut être résolu en imposant dès le départ de la minimisation $n_h = N_h$ pour les strates concernées.

Le deuxième problème est plus conséquent : l'allocation de Neyman optimise la variance d'un estimateur portant sur *l'ensemble de la population* mais peut introduire des différences notables entre les variances *au niveau des regroupements de publication* : on ne maîtrise pas la précision dans ces regroupements.

Le graphique ci-contre illustre cette deuxième limite

Le graphique ci-contre donne le plus mauvais coefficient de variation obtenu avec l'allocation de Neyman dans différents niveaux de publication (publication par secteur d'activité en 16 postes, en 36 postes, par secteur d'activité x taille d'établissement) et selon une taille d'échantillon fixée. On se base ici, comme dans le reste de l'article, sur les données issues des DADS avec comme variable d'intérêt la masse salariale versée par les établissements français en 2004.



Note de lecture : pour un échantillon de taille n = 10 000, le coefficient de variation maximal de notre estimateur dans les regroupements par secteur d'activité (nes16) x taille d'établissement est de 36,8 % (ce qui est considérable !).

L'allocation de Neyman donne un estimateur très précis au niveau global, le coefficient de variation reste inférieur à 1 % quelque soit l'échantillon de taille supérieure à 5 000. Par contre, si l'on souhaite publier les résultats en nes16, le plus mauvais CV d'un estimateur est de 12 % pour un échantillon de taille n = 10 000. Pour des résultats dans un regroupement encore plus fin, comme en nes16 x taille d'établissement et pour une taille d'échantillon de 10 000, le plus mauvais CV atteint 36 %. De manière générale, même avec une taille d'échantillon « raisonnable » telle que n = 30 000, la précision dans certains regroupements de publication n'est pas toujours satisfaisante à des niveaux fins d'agrégations.

1.3. Formalisation du problème à résoudre

L'idée est donc de contrôler la précision dans ces regroupements de publications en imposant des contraintes supplémentaires dans le programme d'optimisation.
On distingue :

- **les strates d'échantillonnage *h***, desquelles sont tirées les unités interrogées ;
- **les regroupements de publication *p***, auxquels sont diffusés les résultats d'une enquête et au niveau desquels on veut contrôler la précision.

Par ailleurs, nous faisons **une hypothèse importante (H)** : on suppose que chaque regroupement de publication est la réunion d'une ou plusieurs strates d'échantillonnage (on exclut les domaines quelconques).

Pour une taille d'échantillon *n* fixée, on se propose de résoudre le programme suivant :

$$\begin{cases} \underset{n_1,\ldots,n_H}{Min}\ V(\hat{\bar{Y}}) & (1) \\ \text{s/c } \sum_{h=1}^{H} n_h = n & (2) \\ \text{s/c } n_h \le N_h & (3) \\ \text{s/c } \underset{p \in pub}{Max}\, CV_p \le CV_{seuil} & (4) \end{cases}$$

L'optimisation (1) et la contrainte (2) correspondent au problème de Neyman. Les contraintes (3) et (4) sont nouvelles par rapport à l'approche classique :

- les contraintes (3) permettent de s'assurer que le nombre d'unités tirées dans chaque strate est bien inférieur au nombre d'unités présentes dans celle-ci ;

- les contraintes (4) sont les contraintes de précision locale. Elles expriment que le plus grand des coefficients de variation dans les regroupements de publication doit être inférieur à un seuil de précision que l'on s'est fixé à l'avance. C'est la véritable nouveauté qu'apporte cette étude.

Nous nous proposons donc d'en trouver une solution exacte et d'illustrer cette solution à l'aide du concept de frontière d'efficacité entre les deux objectifs de précision globale (au niveau de la population) et de précision locale (au niveau des regroupements de publication).

2. L'ingrédient principal de la résolution : les fonctions par morceaux

Le programme que l'on s'attache à résoudre peut être simplifié. Sous l'hypothèse (H) précédente, on montrera en partie 2.5 que la contrainte de précision locale dans les regroupements de publication se traduit par un ensemble de contraintes sur les strates d'échantillonnage : $CV_p \leq CV_{seuil} \Leftrightarrow \{n_h \geq n_{h\min}(CV_{seuil})\}_{h\in p}$, avec $n_{h\min}(CV_{seuil})$ connu pour tout *h*. Fixer un seuil de précision dans un regroupement de publication revient à tirer un nombre minimal d'unités dans chaque strate d'échantillonnage qui la compose.

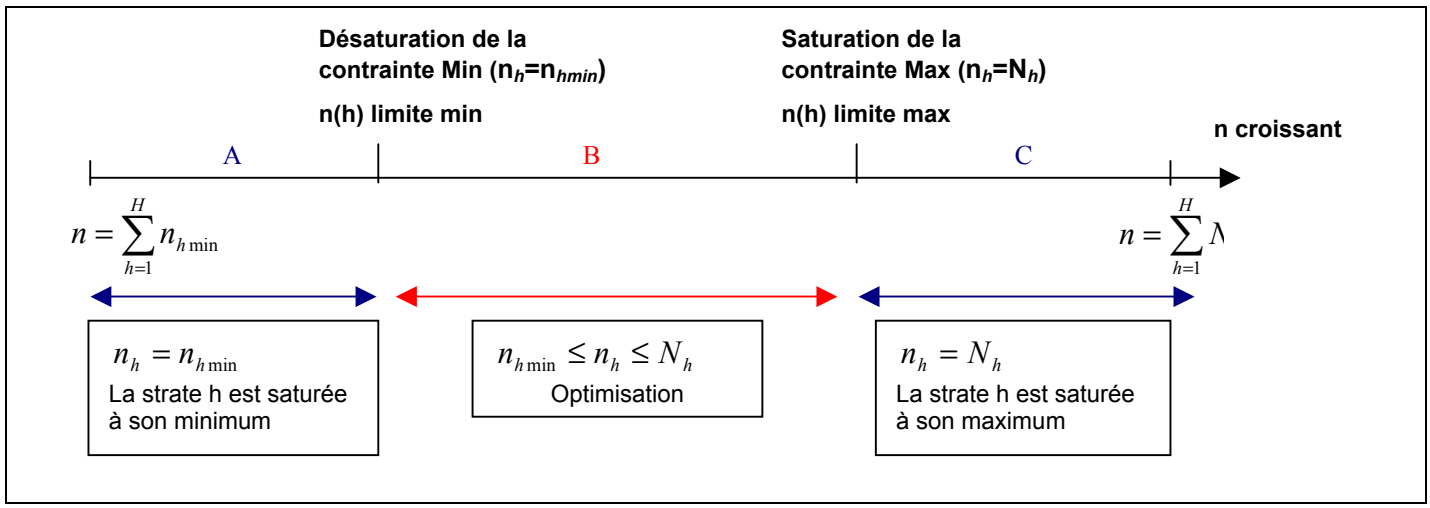
2.1. Une allocation de Neyman par morceaux

Considérons le programme suivant, auquel nous nous ramènerons par la suite. Cette partie s'attache à montrer que la solution à cette fonction est une fonction n_h par morceaux, dont les intervalles de définition dépendent du paramètre *n*, taille globale de l'échantillon, et dont l'expression est simple sur chacun de ces intervalles.

$$\begin{cases} \underset{n_1,\dots n_H}{Min}\, V(\hat{\bar{Y}}) = \sum_h (\frac{N_h}{N})^2 . \underbrace{(1-\frac{n_h}{N_h}).\frac{S_h^2}{n_h}}_{\text{var } h} & (1) \\ \text{s/c} \sum_{h=1}^{H} n_h = n & (2) \\ \text{s/c}\,(n_{h\min}(CV_{seuil}) \leq n_h \leq N_h)_{h=1\dots H} & (3+4) \end{cases}$$

(4) (3)

Notons tout d'abord que les fonctions solutions $n_h(n)$ sont croissantes. Elles sont strictement croissantes à l'intérieur des limites de saturation des contraintes (3+4)[1] et constantes en dehors de ces limites. Typiquement, la fonction n_h a la forme indiquée par la figure suivante.



[1] Une contrainte de type (3+4) du programme sature lorsque l'une des deux inégalités devient une égalité : $n_h = n_{h\min}(CV_{seuil})$ ou $n_h = N_h$.

Examinons ce qui se passe lorsque n augmente :

Tout d'abord, les contraintes (4) ne peuvent être satisfaites que si $n \geq \sum_{h=1}^{H} n_{h\min}$, valeur minimale à envisager. Pour des petites valeurs de *n* au dessous d'une certaine limite n(h) limite min, la contrainte min est saturée et $n_h = n_{hmin}$. A partir d'une certaine valeur de *n*, notée n(h) limite max, c'est la contrainte max qui est saturée et $n_h = N_h$. Entre les deux valeurs, n_h est une fonction strictement croissante de *n*.

2.2. L'ordre de saturation des contraintes

Au total, les 2H valeurs limites ainsi définies définissent des intervalles sur lesquels aucune saturation (ou désaturation) n'intervient. Sur un tel intervalle, en dehors des strates dont une des contraintes est saturée, l'optimisation est une résolution de Neyman classique avec une taille d'échantillon diminuée de la valeur des strates saturées. Pour une valeur de n donnée, un certain nombre de contraintes sont saturées. L'allocation doit en tenir compte :

$$n_h = \frac{N_h * S_h}{\sum_{h \in Hnonsat} N_h * S_h} * (n - n_{sat})$$

pour les strates *h* non saturantes
n_{sat} désignant les effectifs de l'ensemble des strates saturées ;

$$n_h = n_{h\min}$$
$$n_h = N_h$$

pour les strates *h* saturantes
à leur valeur minimale ou maximale.

Cette formule permet par ailleurs de déterminer l'ordre de saturation des contraintes. En effet, la saturation aux limites de l'intervalle se traduit par l'égalité : $n_h = n_{h\min}$ pour les contraintes min et par $n_h = N_h$ pour les contraintes max. La valeur limite n qui réalise l'égalité vaut donc :

$$n(h)_{\lim ite} = n_{sat} + \frac{a_h}{N_{h_0} * S_{h_0}} * \sum_{h \in Hnonsat} N_h * S_h$$

$a_h = N_h$ ou $n_{h\min}$ selon le cas envisagé.

Le seul paramètre qui dépend de *h* détermine donc l'ordre de saturation des contraintes, qui est donné par :

$$Ordre_{h,\max} = \frac{N_h}{N_h * S_h} \quad \text{et} \quad Ordre_{h,\min} = \frac{n_{h\min}}{N_h * S_h}$$

Ordre $_{h,max}$ donne l'ordre de saturation des contraintes Max ($n_h = N_h$), et *Ordre* $_{h,min}$ donne l'ordre de désaturation des contraintes Min ($n_h = n_{hmin}$).

Lorsque la valeur de n augmente, la première contrainte qui sera saturée est celle pour laquelle *Ordre* $_{h,max}$ est la plus petite et la dernière celle pour laquelle *Ordre* $_{h,max}$ est la plus grande. Il en va de même pour la désaturation des contraintes donnée par *Ordre* $_{h,min}$ (car ces deux quantités sont fonction croissantes de n).

On détermine ainsi la taille d'échantillon limite *n(h)* $_{limite}$, pour laquelle la contrainte de la strate *h* sature (ou désature selon le cas envisagé). Or, le seul terme qui dépend de *h* dans cette égalité est la variable *Ordre*. Il suffit donc de connaître sa valeur pour chacune des strates d'échantillonnage, pour en déduire immédiatement dans quel ordre les contraintes saturent lorsque la taille globale de l'échantillon *n* augmente.

2.3. Détermination des intervalles de définition de la fonction par morceaux n_h

Exemple illustratif avec 3 strates

On considère trois strates avec les caractéristiques suivantes :

Strate	n_{hmin}	N_h	S_h
1	2	8	1/4
2	9	30	1/10
3	32	48	1/12

Notre programme d'optimisation sous contrainte de précision locale est de la forme :

$$\begin{cases} \underset{n_1,n_2,n_3}{Min}\ V(\hat{\bar{Y}}) = \sum_{h=1}^{3} (\frac{N_h}{N})^2 .(1-\frac{n_h}{N_h}).\frac{S_h^2}{n_h} \\ \text{s/c } \sum_{h=1}^{3} n_h = n \\ \text{s/c } (2 \le n_1 \le 8) \text{ et } (9 \le n_2 \le 30) \text{ et } (32 \le n_3 \le 48) \end{cases}$$

On définit l'ordre de saturation des six contraintes, en calculant les valeurs de la variable *Ordre*.

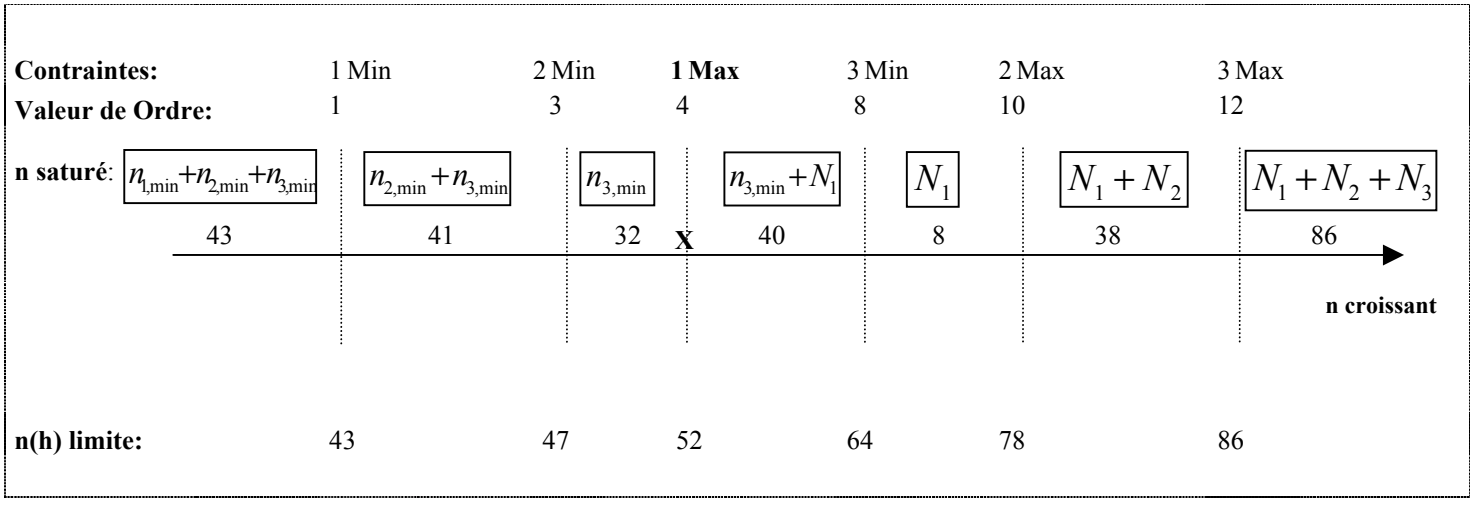
Contrainte	Valeur de la variable *Ordre*	Ordre de saturation
1 Min	1	*1*
1 Max	4	*3*
2 Min	3	*2*
2 Max	10	*5*
3 Min	8	*4*
3 Max	12	*6*

Exemple détaillé pour la contrainte 1 :

$Ordre_{1\ min} = n_{1\ min} / (N_1 * S_1)$
$= 2 / (8*1/4) = 1$

$Ordre_{1\ max} = N_1 / (N_1 * S_1) = 1/S_1$
$= 4$

Visualisation de l'ordre de saturation des contraintes



En connaissant l'ordre de saturation des contraintes, on peut calculer pas à pas la valeur de n_{sat}, la taille de l'échantillon saturé.

Pour une taille d'échantillon globale n assez grande permettant de respecter les contraintes de précision locale, on a tout d'abord l'ensemble des contraintes saturées à leur minimum ($n_{sat} = n_{1\min} + n_{2\min} + n_{3\min} = 43$). Puis lorsque n augmente, la contrainte 1 Min désature la première, n_{sat}

vaut alors $n_{sat} = n_{2\min} + n_{3\min} = 41$. La contrainte 2 Min désature en deuxième, n_{sat} vaut alors $n_{sat} = n_{3\min} = 32$, et ainsi de suite.

Ensuite, connaissant la valeur de la variable *Ordre* et celle de n_{sat} on peut déterminer la taille d'échantillon *n(h) limite*, pour laquelle les différentes contraintes désaturent et saturent.

On se place par exemple lors de la saturation de la contrainte 1 Max (par la gauche) : la seule strate saturée à ce moment est la strate 3 : $n_3 = n_{3min} = n_{sat} = 32$. Par conséquent, on peut déterminer la taille d'échantillon *n(1) limite max*, pour laquelle la contrainte 1 sature à son maximum :

$$n(1)_{\lim ite\ \max} = n_{sat} + \frac{N_1}{N_1 * S_1} * (N_1 * S_1 + N_2 * S_2) = 32 + 4 * [(8 * 1/4) + (30 * 1/10)] = 52$$

Finalement, en fonction de la taille d'échantillon *n* retenue pour notre échantillonnage final, on peut connaître l'ensemble des strates saturées, et appliquer l'optimisation de Neyman aux seules strates non saturées.

Si l'on souhaite tirer un échantillon de taille $n = 50$, la strate 3 sera saturée à son minimum et $n_3 = n_{3min} = n_{sat} = 32$. Il suffira de réaliser une allocation de Neyman sur les strates 1 et 2 non saturantes pour cette valeur de *n*, avec une taille d'échantillon de 50 - 32 = 18.

En conclusion, pour obtenir une allocation de Neyman dont les résultats sont meilleurs localement pour une taille d'échantillon *n*, il suffit de repérer les strates saturées et de faire une allocation de Neyman sur les strates non saturantes, avec une taille d'échantillon *n*, diminuée de la taille de l'échantillon saturé.

2.4. Sur chaque intervalle de définition, des formules simples

A ce stade, nous avons montré comment résoudre le programme du 2.1. Dans celui-ci, les contraintes portent sur les n_h et non sur les précisions CV_p, comme c'est le cas pour le problème initial. Il reste donc, pour compléter la résolution, à montrer comment traduire une contrainte portant sur la précision des regroupements de publication en une contrainte portant sur le nombre d'unités n_h à sélectionner dans chaque strate d'échantillonnage.

Tout d'abord, notons que sur un intervalle de définition donné, les relations qui existent entre les variables du problème ont des expressions simples.

<u>Relation liant CV et *n*</u>
Partons de la formule de la variance de l'estimateur de la moyenne.

$$V(\hat{\bar{Y}}) = \sum_{h \in Hnonsat} \left(\frac{N_h}{N}\right)^2 (1 - \frac{n_h}{N_h}) \frac{S_h^{\ 2}}{n_h}$$

$$N^2.\ V(\hat{\bar{Y}}) = \sum_{h \in Hnonsat} (1 - \frac{n_h}{N_h}) \frac{N_h^2 . S_h^{\ 2}}{n_h}$$

Or, sur l'intervalle on a par ailleurs comme au 2.2 pour les strates non saturées :

$$n_h = \frac{N_h . S_h}{\sum_{h \in Hnonsat} N_h . S_h} (n - n_{sat})\ (*)$$

En injectant cette formule dans l'équation précédente, on obtient :

$$N^2.\ V(\hat{\bar{Y}}) = \frac{\left(\sum_{h \in Hnonsat} N_h . S_h\right)^2}{n - n_{sat}} - \sum_{h \in Hnonsat} N_h . S_h^2$$

Soit en faisant apparaître le coefficient de variation plutôt que la variance :

$$N^2.\bar{y}^2.CV^2 = \frac{\left(\sum_{h\in Hnonsat} N_h.S_h\right)^2}{n - n_{sat}} - \sum_{h\in Hnonsat} N_h.S_h^2$$

où $\bar{y}$ désigne la valeur moyenne de la variable d'intérêt. En dehors des variables CV et *n*, tous les autres paramètres sont constants sur un intervalle de définition donné. La relation est donc une fonction simple et explicitement inversible. La fonction inverse est la suivante :

$$\boxed{n = n_{sat} + \frac{\left(\sum_{h\in Hnonsat} N_h.S_h\right)^2}{N^2.\bar{y}^2.CV^2 + \sum_{h\in Hnonsat} N_h.S_h^2}}$$

De plus, toujours en vertu de cette relation univoque liant CV et *n*, les intervalles de définition peuvent être définis par des valeurs de CV plutôt que par des valeurs de *n*. Au total, la fonction *n* est une fonction par morceaux de CV, avec les limites d'intervalles définies comme valeurs CV et les paramètres intervenant dans l'expression sont constants et calculables sur chaque intervalle.

Relation liant n_h et CV

En tenant compte de la relation (*), on obtient une fonction n_h dépendant de CV. Cette fonction est notée $Neyman\_c(CV)$ pour rappeler qu'elle résulte d'une allocation de Neyman avec contraintes.

$$\boxed{n_h = \frac{N_h.S_h}{\sum_{h\in Hnonsat} N_h.S_h}(n(CV) - n_{sat}) = Neyman\_c(CV)}$$

2.5. Traduction des contraintes de précision en contraintes sur les n_h

L'analyse précédente peut également être faite au niveau d'un regroupement de publication plutôt qu'au niveau de l'ensemble du champ. C'est là qu'intervient l'hypothèse (H) faite dans l'ensemble de l'étude, qui pose que les regroupements de publication sont des ensembles disjoints de strates d'échantillonnage. De cette manière, ce qui se passe dans un regroupement de publication n'interfère pas avec ce qui se passe dans un autre regroupement de publication.

Sur chaque strate *p*, la méthode de résolution exposée s'applique au programme suivant :

$$\begin{cases} \underset{n_{h,h\in p}}{Min} \sum_{h\in p}\left(\frac{N_h}{N_p}\right)^2 (1-\frac{n_h}{N_h})\frac{S_h^{\ 2}}{n_h} \\ s/c \sum_{h\in p} n_h = n_p \\ s/c\ \left(n_h \le N_h\right)_{h\in p} \end{cases}$$

Ainsi, pour une strate de publication *p*, on peut facilement établir la relation liant n_p et CV_p et celles liant n_h et CV_p. Toutes ces relations s'obtiennent comme des fonctions par morceaux, dont les intervalles peuvent être définis comme des valeurs de CV_p et dont la forme dépend de paramètres fixes sur chaque intervalle et également calculables.

$$\boxed{n_h = Neyman\_c, p(CVp)}$$

De ce fait, en fixant la valeur CV_{seuil} à respecter dans chaque regroupement de publication, les valeurs n_{hmin} qui doivent figurer dans les contraintes du problème à la place des contraintes portant sur CV_p, sont simplement les $n_h(CV_p)$ obtenus selon cette méthode. Pour chaque $h \in p$,

$$\boxed{n_{h\min} = Neyman\_c, p(CV_{seuil})}$$

2.6. Synthèse du schéma de résolution global

Au total, l'algorithme de résolution suit le schéma suivant. On se fixe une précision CV_{seuil} à respecter dans chaque regroupement de publication. Notons que pour simplifier la valeur est la même pour tous les regroupements de publication mais que la méthode permettrait également de fixer des valeurs-seuil différentes d'un regroupement à l'autre.

On suit alors les étapes suivantes :

1) Détermination des n_{hmin} par résolution, au niveau des regroupements de publication, du programme de résolution du 2.5 ne contenant que des contraintes max.

$$n_{h\min} = Neyman\_c, p(CV_{seuil})$$

2) Résolution du programme du 2.1 comprenant les contraintes max et les contraintes min ainsi calculées :

$$n_h = Neyman\_c(CV_{seuil})$$

Le schéma suivant résume le processus de résolution

Programme sans contrainte min au niveau de la strate p

$$\underset{n_{h,h\in p}}{Min}\ V(\hat{\bar{Y}}_p) = \sum_{h\in p}\left(\frac{N_h}{N_p}\right)^2 (1-\frac{n_h}{N_h})\frac{S_h^{\ 2}}{n_h}$$

$$s/c \sum_{h\in p} n_h = n_p$$

$$s/c\ \left(n_h \le N_h\right)_{h\in p}$$

Méthode des fonctions par morceaux

$$fonctions \quad n_h = Neyman\_c, p(CVp)$$

Pour un CV_{seuil} fixé

$$n_{h\min}(CV_{seuil}) = Neyman\_c, p(CV_{seuil})$$

Injection des valeurs des contraintes

$$\underset{n_1,\ldots n_H}{Min}\, V(\hat{\bar{Y}}) = \sum_h (\frac{N_h}{N})^2 . \underbrace{(1-\frac{n_h}{N_h}).\frac{S_h^2}{n_h}}_{\text{var}\, h}$$

$$\text{s/c}\ \sum_{h=1}^{H} n_h = n$$

$$\text{s/c}\,(n_{h\min}(CV_{seuil}) \le n_h \le N_h)_{h=1\ldots H}$$

Méthode des fonctions par morceaux

$$n_h = Neyman\_c(CV_{seuil})$$

3. Cas pratiques de la méthode d'optimisation

3.1. Comparaison de l'allocation de Neyman et de l'allocation sous contrainte de précision locale

Nous évaluons ici l'efficacité de l'allocation avec contrainte de précision dans les regroupements de publication en comparant sa précision avec celle de l'allocation classique de Neyman, pour trois niveaux de publication différents.

Cas d'application : dans chacun des trois exemples qui vont suivre, on estime la masse salariale versée par les établissements français à partir d'un échantillon de 10 000 établissements sélectionnés dans les DADS 2004.

Les établissements ont été stratifiés par secteur d'activité en 36 postes (Nes36) et par taille en six modalités [**(1)** : 10-19 salariés, **(2)** : 20-49 salariés, **(3)** : 50-99 salariés, **(4)** : 100-249 salariés, **(5)** : 250-499 salariés, **(6)** : 500 salariés et plus)].

On sélectionne ensuite un échantillon d'établissements selon la méthode de Neyman puis un autre selon notre méthode sous contrainte de précision locale et on compare les deux allocations en termes de précision.

Taille d'échantillon n = 10 000 et publication par secteur d'activité (Nes16)

Pour notre allocation contrainte, on se fixe un coefficient de variation maximal de 2 % dans les secteurs d'activité.

Secteur d'activité	CV Neyman	CV Allocation contrainte au seuil de 2%
Global	**0,44%**	**0,46%**
EB	**2,6%**	**2,0%**
EC	1,9%	1,9%
ED	1,4%	1,5%
EE	1,5%	1,7%
EF	1,3%	1,4%
EG	**2,6%**	**2,0%**
EH	1,8%	1,9%
EJ	1,2%	1,4%
EK	1,4%	1,5%
EL	1,6%	1,8%
EM	**4,2%**	**2,0%**
EN	1,0%	1,1%
EP	**2,7%**	**2,0%**
EQ	1,6%	1,8%
ER	**2,2%**	**2,0%**

Pour un échantillon de taille n = 10 000 établissement, l'allocation de Neyman donne de bons résultats tant au niveau global qu'au niveau des regroupements de publication par secteur d'activité. L'allocation contrainte permet néanmoins de diviser par deux le coefficient de variation de certains petits secteurs, comme celui de l'immobilier (EM). L'augmentation du coefficient de variation global qui en résulte est faible (+0,02 point).

L'efficacité de l'allocation contrainte augmente lorsque la taille de l'échantillon est faible, car l'allocation de Neyman donne alors de moins bons résultats au niveau local.

Taille d'échantillon n = 10 000 et publication par secteur d'activité (Nes36)

Pour notre allocation contrainte, on se fixe un coefficient de variation maximal de 3 % dans les secteurs d'activité.

Secteur d'activité	CV Neyman	CV allocation contrainte au seuil de 3%	n publication Neyman	n publication allocation contrainte au *seuil* de 3%	Variation n pub
GLOBAL	**0,44%**	**0,49%**	**10 000**	**10 000**	/
B0	2,6%	3,0%	365	298	-67
C1	**6,5%**	**2,9%**	**49**	**187**	**138**
C2	3,6%	2,9%	136	184	47
C3	2,7%	2,9%	163	152	-12
C4	4,1%	3,0%	87	150	63
D0	1,4%	1,7%	147	129	-18
E1	2,0%	2,5%	88	77	-11
E2	2,3%	3,0%	260	173	-87
E3	3,0%	2,8%	238	259	21
F1	3,8%	3,0%	88	132	44
F2	**6,0%**	**3,0%**	**46**	**161**	**115**
F3	4,2%	3,0%	95	176	81
F4	2,3%	2,8%	260	200	-61
F5	2,5%	2,8%	257	210	-47
F6	3,0%	3,0%	154	154	0
G1	4,2%	2,9%	22	28	6
G2	3,0%	3,0%	177	177	0
H0	1,8%	2,3%	486	323	-163
J1	3,2%	2,9%	144	177	34
J2	1,8%	2,3%	685	469	***-216***
J3	1,9%	2,4%	455	303	-152
K0	1,4%	1,7%	584	446	-137
L0	1,6%	2,0%	742	552	***-190***
M0	4,2%	2,9%	152	278	126
N1	1,0%	1,3%	322	293	-29
N2	1,5%	1,9%	1 321	927	***-394***
N3	3,1%	2,9%	706	789	83
N4	2,3%	2,9%	87	71	-16
P1	3,1%	2,9%	219	244	26
P2	4,8%	3,0%	335	715	381
P3	**11,9%**	**3,0%**	**21**	**254**	**232**
Q1	4,3%	3,0%	159	311	152
Q2	1,8%	2,2%	577	384	***-193***
R1	2,1%	2,9%	199	132	-67
R2	**5,3%**	**3,0%**	**174**	**484**	**311**

A un niveau de publication plus fin, l'allocation contrainte apporte un vrai gain en terme de précision notamment pour les petits secteurs tels que l'habillement et cuir C1, l'industrie textile F2, les services personnels et domestiques P3 et les activités associatives R2. Avec une allocation classique de Neyman, leur coefficient de variation peut aller jusqu'à près de 12 % (secteur P3) ; celui-ci peut être réduit à 3 %, grâce à une meilleure redistribution de l'allocation. On tire ainsi 254 établissements dans P3 contre seulement 21 avec l'allocation classique de Neyman et trois fois plus d'établissements dans les secteurs C1, F2 et R2. Ce gain d'établissements dans ces secteurs d'activité se fait au détriment des secteurs les plus précis L0, J2, N2, Q2, qui voient leur nombre d'établissements interrogés diminuer, mais tout en leur préservant une bonne précision au niveau local.

La précision globale reste quant à elle, peu affectée par cette redistribution de l'allocation de Neyman : le coefficient de variation est respectivement de 0,44 % pour l'allocation selon Neyman et de 0,49 % pour l'allocation contrainte par un seuil de précision local de 3 %.

Taille d'échantillon n = 10 000 et publication par secteur d'activité (Nes16) x taille d'établissement

Pour notre allocation contrainte, on se fixe un coefficient de variation maximal de 5 % dans regroupements par secteurs d'activité x taille d'établissement.

Secteur d'activité x taille	CV Neyman	CV allocation contrainte au seuil de 5 %	$n_{publication}$ Neyman	$n_{publication}$ allocation contrainte au seuil de 5 %	Variation CV ou n_{pub}
GAIN DE PRECISION en CV_p > 5 points					
Global	**0,44%**	**0,59%**	**10 000**	**10 000**	**-0,15pts**
EB x 1	**11,5%**	**4,9%**	46	234	6,6pts
ED x 1	**36,8%**	**5,0%**	7	179	31,8pts
ED x 2	**20,1%**	**5,0%**	8	95	15,1pts
ED x 3	**16,8%**	**4,9%**	4	32	11,9pts
ED x 4	**10,7%**	**4,9%**	17	57	5,8pts
EG x 1	**15,1%**	**5,0%**	16	134	10,1pts
EM x 1	**12,2%**	**5,0%**	42	234	7,2pts
EM x 3	**12,2%**	**5,0%**	18	83	7,2pts
EM x 5	**13,2%**	**4,7%**	10	31	8,5pts
ER x 1	**10,0%**	**5,0%**	68	265	5pts
PERTE DE PRECISION en CV_p > 2 points					
EB x 6	**0,0%**	**3,5%**	78	71	-3,5pts
EC x 6	**1,6%**	**5,0%**	102	59	-3,4pts
EE x 6	**0,4%**	**4,4%**	171	101	-4pts
EF x 6	**0,5%**	**4,2%**	212	146	-3,7pts
EG x 6	**1,8%**	**4,8%**	42	29	-3pts
EM x 6	**0,0%**	**3,7%**	22	21	-3,7pts
EN x 4	**2,7%**	**5,0%**	381	125	-2,3pts
EN x 5	**2,8%**	**5,0%**	249	107	-2,2pts
EN x 6	**0,3%**	**3,0%**	483	298	-2,7pts
ER x 6	**1,4%**	**4,7%**	101	66	-3,3pts
VARIATIONS de $n_{publication}$ > 100					
EB x 1	11,5%	4,9%	46	234	**188**
ED x 1	36,8%	5,0%	7	179	**172**
EG x 1	15,1%	5,0%	16	134	**118**
EL x 1	7,9%	4,9%	105	260	**155**
EL x 2	7,7%	4,9%	161	358	**197**
EM x 1	12,2%	5,0%	42	234	**191**
ER x 1	10,0%	5,0%	68	265	**197**
EF x 4	2,9%	4,6%	186	80	-106
EJ x 1	2,9%	4,4%	292	126	-166
EJ x 2	2,4%	4,3%	351	115	-236
EJ x 3	3,0%	4,6%	196	85	-112
EJ x 4	2,9%	4,6%	231	100	-131
EN x 2	3,2%	4,5%	527	285	-242
EN x 3	3,4%	4,8%	274	148	-126
EN x 4	2,7%	5,0%	381	125	-256
EN x 5	2,8%	5,0%	249	107	-142
EN x 6	0,3%	3,0%	483	298	-185

Lorsque l'on passe à une publication par secteur d'activité x taille d'établissement, les gains de précision avec l'allocation contrainte sont toujours plus accentués (*1ère partie du tableau*) mais la perte de précision globale est également plus importante. Il y a un vrai arbitrage entre précision locale et globale.
Tous les niveaux de publication peuvent être ramenés à un seuil maximal de variation de 5 %. En contrepartie, certains secteurs d'activités voient leur précision diminuer (*2ème partie du tableau*) jusqu'à un maximum de 4 points par rapport à l'allocation de Neyman.
Enfin, la *3ème partie du tableau* met en évidence les plus grandes redistributions d'unités à sélectionner entre les strates d'échantillonnage et donc dans leurs regroupements de publication associés.

3.2. Concept et illustration de la frontière d'efficacité

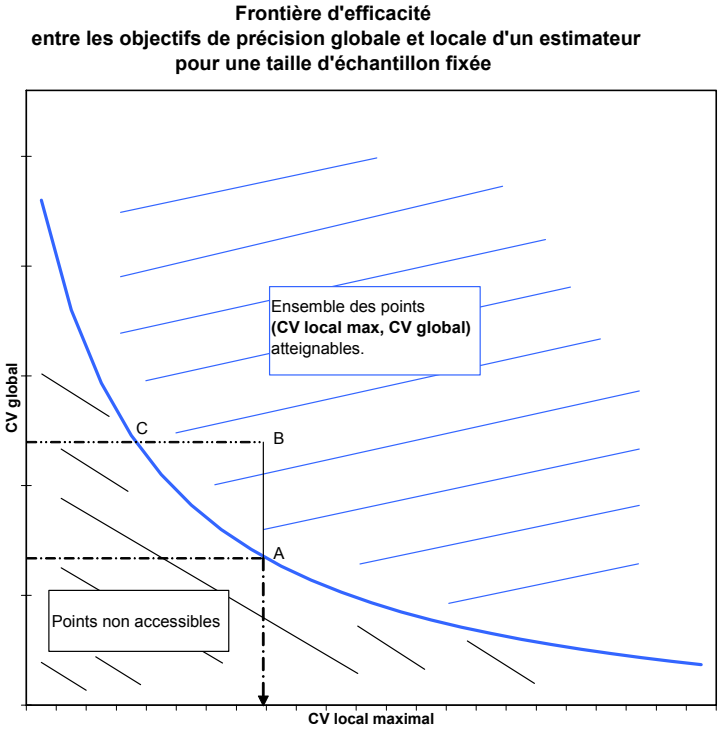
Le dilemme entre précision globale et locale d'un estimateur que rencontre le statisticien est illustré ici par le graphique de la frontière d'efficacité entre ces deux objectifs. Cette frontière permet d'évaluer la perte en précision globale lorsque l'on tient à fixer une précision minimale dans des regroupements de publication.

On rappelle que la précision locale est mesurée par le plus mauvais coefficient de variation des estimateurs dans les regroupements de publication $CV_{\text{local maximal}} = \underset{p \in pub}{Max}\, CV_p$.

Frontière d'efficacité

On appelle frontière d'efficacité, l'ensemble des allocations (n_1,…, n_H) à taille d'échantillon *n* fixée tel qu'on ne peut augmenter la précision locale dans les regroupements de publication, sans détériorer la précision globale d'un estimateur (et inversement). Cette frontière d'efficacité est représentée dans le repère (CV_{global}, $CV_{\text{local maximal}}$).

Schéma :



Par abus de langage, nous parlerons de la frontière d'efficacité comme l'ensemble des couples (CV_{global}, $CV_{\text{local maximal}}$) réalisables pour une taille d'échantillon fixée, de la courbe ci-dessus. Ce sont les points optimaux.

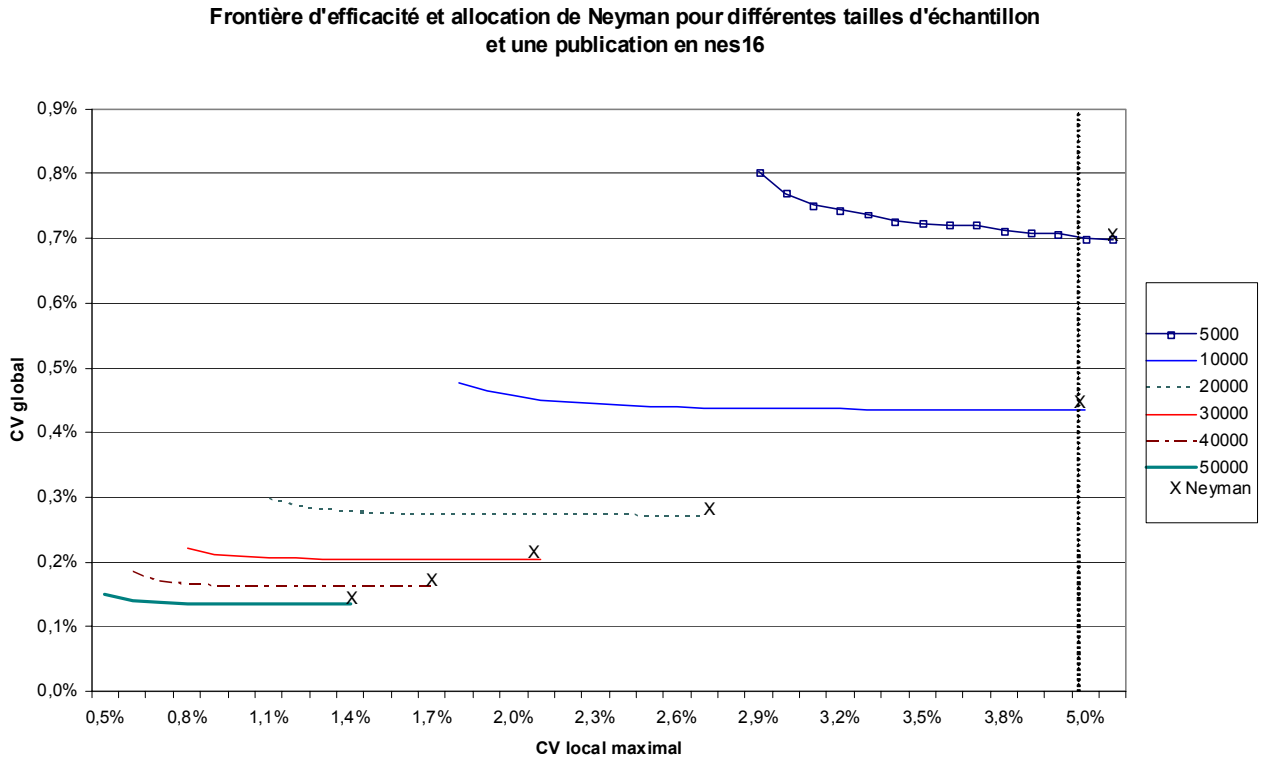
Ceux au-dessus de la courbe, comme le point B, sont possibles mais des situations meilleures peuvent être atteintes : pour le même niveau de précision globale que B, la précision locale maximale peut être réduite à celle de C ; pour le même CV local maximal que B, on peut atteindre une meilleure précision globale (point A).

Cette courbe constitue une aide à la décision précieuse pour l'échantillonnage car elle synthétise les différents arbitrages devant lesquels le statisticien est placé.

En pratique, la courbe est obtenue en faisant varier le seuil de précision souhaité dans les regroupements de publication et en enregistrant pour chaque valeur de cette variable, le résultat de l'optimisation CV_{global}.

On illustre ce concept lors d'une estimation de la masse salariale versée par les établissements français à partir d'un échantillon d'établissements sélectionnés dans les DADS 2004 selon notre méthode d'optimisation sous contrainte.

Exemple de frontière d'efficacité pour une publication par secteur d'activité (Nes16)



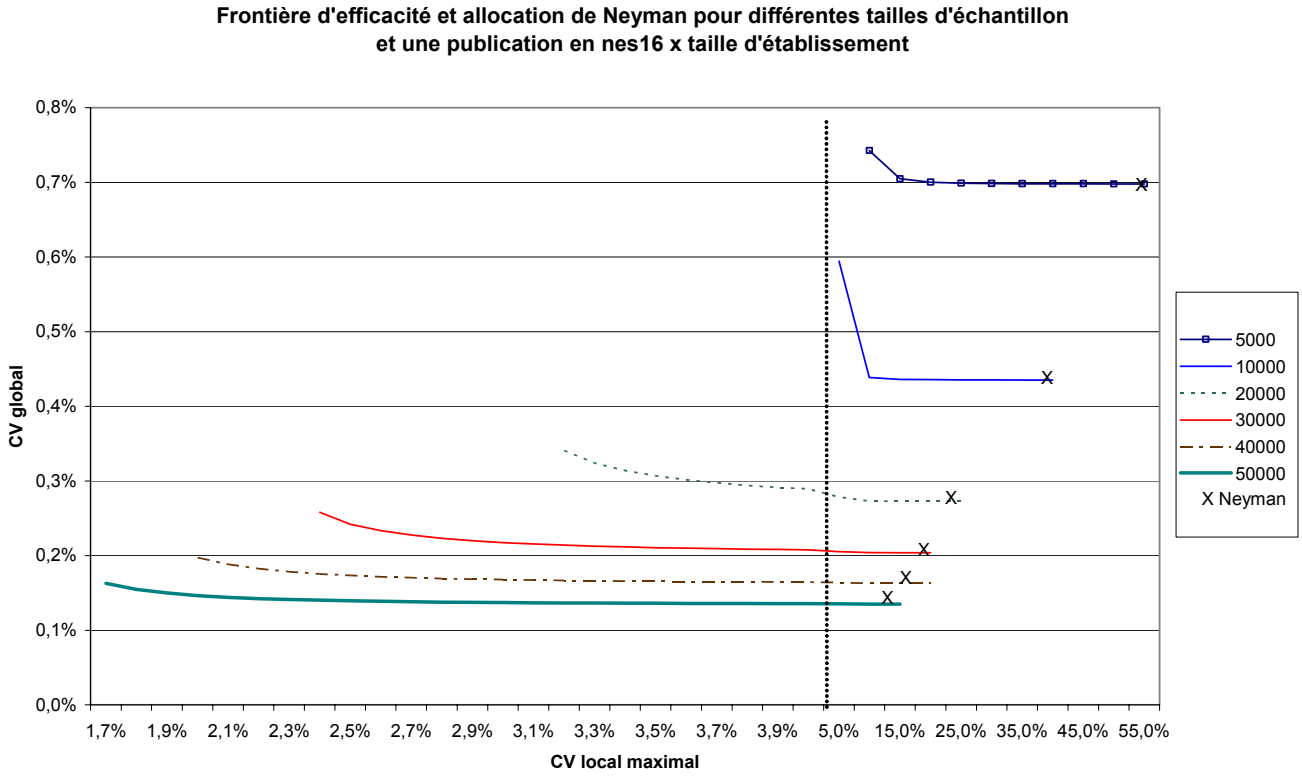
Ce graphique montre que les précisions globales ainsi que les précisions locales d'un estimateur sont fonctions croissantes de la taille de l'échantillon tiré. La précision globale d'un estimateur est très satisfaisante puisqu'elle reste sous le seuil de 1 % quelle que soit la taille d'échantillon de 5 000 à 50 000 établissements ; la précision locale par contre est plus dispersée, puisque le $CV_{local\ maximal}$ varie de plus de 5 % pour un échantillon de taille n = 5 000 à 0,5 % pour n=50 000.

La précision locale obtenue avec l'allocation de Neyman (repérée par x) peut être sensiblement améliorée au prix d'une dégradation minimale de la précision globale de notre estimateur : la précision globale est en effet décroissante avec la précision locale, mais se stabilise rapidement quelque soit le $CV_{local\ maximal}$ (courbes plates). Le minimum de la variance globale est un minimum plat.

Exemple : pour un échantillon de 20 000 établissements, le CV local maximal peut être diminué de 2,7 % (Neyman) à 1,1 % (allocation sous contrainte), en impactant la précision globale de seulement 0,03 point (précision globale de 0,27 % à 0,30 %).

Mieux vaut alors s'assurer la meilleure précision locale, à savoir choisir l'échantillon tel que le pire des CV locaux soit de 1,1 %, au détriment de l'allocation de Neyman à partir de laquelle il n'y a plus de contrainte de précision locale. Des précisions locales inférieures au seuil de 1,1 % ne sont pas atteignables, car la contrainte $n \geq \sum_h n_{h\min}$ n'est alors plus vérifiée.

Regroupements de publication : par secteur d'activité (Nes16) x taille d'établissement

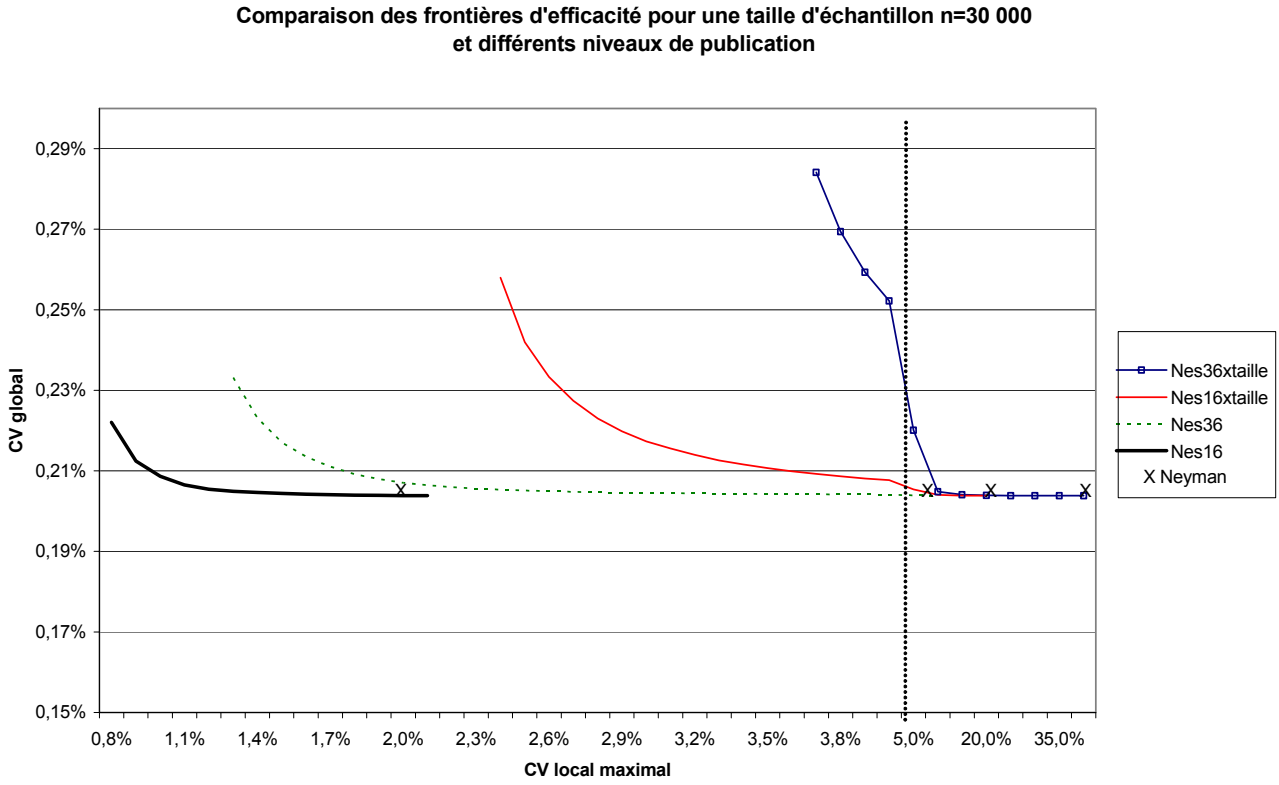


Lorsque l'on publie des résultats dans des niveaux de nomenclature plus fins (nes16 x taille d'établissement), les contraintes de précisions locales deviennent plus exigeantes : les résultats locaux de l'allocation de Neyman peuvent être améliorés mais avec un impact un peu plus fort sur la précision globale.

Exemple : pour n = 20 000, le $CV_{local\ maximal}$ peut être ramené de 24,4 % (allocation de Neyman) à 3,2 % (allocation sous contrainte) avec une perte de précision globale de 0,07 point (passage du CV_{global} de 0,27 % à 0,34 %).

Néanmoins, les gains de précisions suite à l'allocation sous contrainte sont conséquents et ceci en partie à cause de quelques mauvaises précisions locales données par l'allocation de Neyman, dans les petits regroupements de publication.

Résultats pour n = 30 000 et différents niveaux de publication



La méthode d'optimisation sous contrainte de précision locale apporte un fort gain de précision locale par rapport à l'allocation de Neyman dans le cas de regroupements de publication fins. Mais cette amélioration de la précision locale dans les strates fines demande également une détérioration plus conséquente de la précision globale.

Exemple : pour n = 30 000, la précision en nes16 x taille d'établissement peut être ramenée de 19,1 % à 2,4 % grâce à l'allocation sous contrainte. La précision globale s'en voit dégrader de 0,06 point (CV_{global} respectivement de 0,20 % avec l'allocation de Neyman et de 0,26 % avec l'allocation sous contrainte). La plus mauvaise précision en nes36 de 6,1 % en CV peut également être améliorée jusqu'au seuil de 2,4 % avec une perte de précision globale de moins de 0,01 point.

3.3. Application

Cette méthode d'optimisation sous contrainte de précision locale a été mise en œuvre en 2006 dans le cadre de la refonte des enquêtes sur l'activité et les conditions d'emploi de la main-d'œuvre (Acemo) du ministère du Travail. Elle a notamment été appliquée à l'enquête trimestrielle Acemo.

Bibliographie

[1] Ardilly P., *« Les techniques de sondage »*, Éditions technip, Paris, 2006.
[2] Bankier M.D., *« Power Allocations : Determining sample Sizes for Sub-national Areas »*, The American Statistician, 1988, vol.42 p.174-177.
[3] Tillé Y., « *Théorie des sondages : Échantillonnage et estimation en populations finies : cours et exercices »,* Dunod, Paris, 2001.
[4] Koubi M., Mathern S., « *La nouvelle méthode d'échantillonnage de l'enquêtes trimestrielle Acemo depuis 2006 »*, Document d'études Dares N°146, 2009.