Résolution d'une limite

de

l'allocation de Neyman

Sandrine Mathern - Malik Koubi

Plan de la présentation

- Limites de l'allocation de Neyman et amélioration proposée
- Résolution du nouveau programme d'optimisation
- Illustration de l'arbitrage entre précision globale et précision locale

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

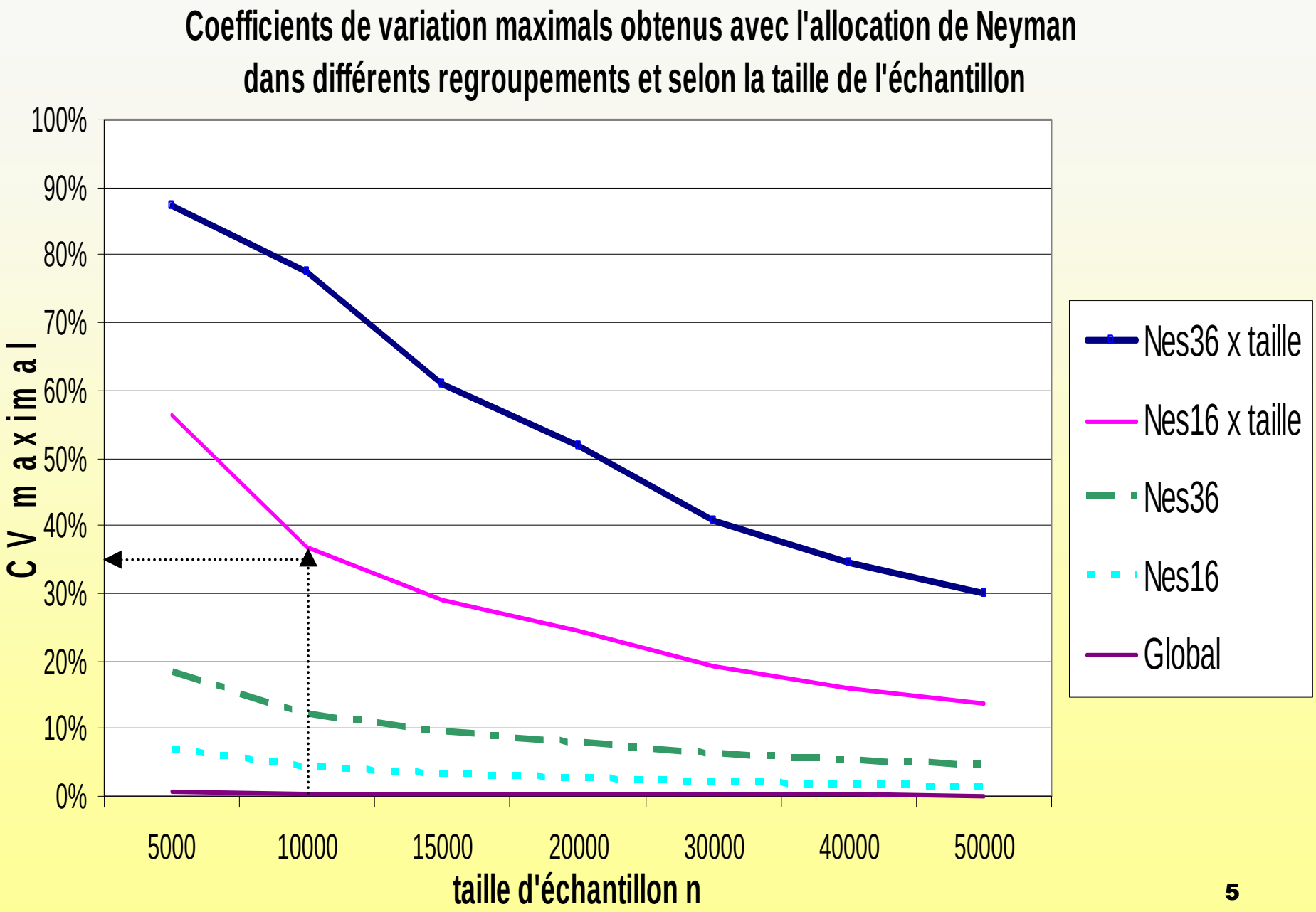
- Sondage stratifié et allocation optimale de Neyman:

  à taille d'échantillon fixée, assure une précision maximale pour l'estimateur de la moyenne **au niveau de l'ensemble de la population**.

$$\begin{cases} \underset{n_1,\ldots,n_h}{Min}\, V(\bar{\bar{Y}}) \\ slc \sum_{h=1}^{H} n_h = n \end{cases} \longrightarrow \boxed{n_h = \frac{N_h * S_h}{\sum_h N_h * S_h} * n}$$

- Problème:

  l'allocation peut être imprécise pour des estimations sur **des sous-groupes de la population** (« regroupements de publication »).


Coefficients de variation maximals obtenus avec l'allocation de Neyman
dans différents regroupements et selon la taille de l'échantillon
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
C V m a x i m a l
5000
10000
15000
20000
30000
40000
50000
taille d'échantillon n
Nes36 x taille
Nes16 x taille
Nes36
Nes16
Global

- Solution proposée:

réallouer l'allocation de Neyman afin de **respecter un seuil minimal de précision** dans tous les « regroupements de publication p »

$$\underset{p \in pub}{Max}\, CV_p \leq CV_{seuil}$$

... sans trop détériorer la précision globale.

- Hypothèse fondamentale et notations:

h : indice pour les strates d'échantillonnage

p : indice pour les regroupements de publication

Hypothèse : chaque regroupement de publication est une réunion de strates d'échantillonnage.

Autrement dit, les regroupements ***p*** forment une partition des strates ***h***.

Par abus de langage:

- H désignera à la fois l'ensemble et le nombre des strates d'échantillonnage
- on utilisera la notation : $h \in p$

Programme à résoudre

$$\begin{cases} \underset{n_1,\ldots,n_h}{Min}\ V(\hat{\bar{Y}}) & (1) \\ \text{slc} \quad \sum_h n_h = n & (2) \\ \text{slc} \quad n_h \leq N_h & (3) \\ \text{slc} \quad \underset{p}{Max}\ CV_p \leq CV_{seuil} & (4) \end{cases}$$

(1) et (2) optimisation classique de Neyman

(3) nombre d'unités tirées dans chaque strate est inférieure au nombre d'unités présent dans celle-ci

(4) traduction des contraintes de précision locale. Véritable ajout de cette méthode.

Exemple de réallocation

Secteur d'activité	CV Neyman « classique »	CV Alloc. contrainte CV_{seuil}= 3%	Effectifs Neyman	Effectifs Allocation contrainte CV_{seuil}= 3%	Ecart des deux méthodes
Ensemble	**0,44 %**	**0,49 %**	**10 000**	**10 000**	/
Habillement	**6,5 %**	**2,9 %**	49	187	138
Ind. Textile	**6,0 %**	**3,0 %**	46	161	115
Commerce de gros	1,8 %	2,3 %	685	469	*-216*
Finances	1,6 %	2,0 %	742	552	*-190*
Conseils	1,5 %	1,9 %	1 321	927	*-394*
Service domestique	**11,9 %**	**3,0 %**	21	254	232
Act. sociale	1,8 %	2,2 %	577	384	*-193*
Association	**5,3 %**	**3,0 %**	174	484	311

2. Résolution du nouveau programme d'optimisation à l'aide de fonctions par morceaux

Étape 1: simplification du programme

contraintes uniquement sur les strates d'échantillonnage

$$\begin{cases} \underset{n_1,\ldots,n_h}{Min}\ V(\hat{\bar{Y}}) & (1) \\ \sum_h n_h = n & (2) \\ n_h \le N_h & (3) \\ \underset{p}{Max}\ CV_p \le CV_{seuil} & (4) \end{cases}$$

$$\begin{cases} \underset{n_1,\ldots,n_h}{Min}\ V(\hat{\bar{Y}}) \\ \sum_h n_h = n \\ (n_{h\min}(CV_{seuil}) \le n_h \le N_h)_{h=1\ldots H} \end{cases}$$

Étape 2: la méthode des fonctions par morceaux

C'est l'ingrédient principal de la résolution.

Elle permet de calculer les solutions n_h du programme suivant :

$$\begin{cases} \underset{n_1,\ldots,n_h}{Min}\ V(\hat{\bar{Y}}) \\ \sum_h n_h = n \\ (n_{h\min} \le n_h \le N_h)_{h=1\ldots H} \end{cases}$$

La méthode des fonctions par morceaux

Plus précisément, supposons les n_{hmin} fixés.

- **SI**, pour une valeur de *n*, nous connaissions la **LISTE** des contraintes « saturées »,
  alors on se ramènerait à un problème de Neyman « classique »

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

Or, **l'ORDRE DE SATURATION** quand *n* augmente des 2H contraintes peut être établi

- Ceci permet de déterminer la liste des contraintes saturées pour chaque valeur de *n*

- Au total, on obtient les solutions n_h comme des fonctions par morceaux définies chacune sur 2H intervalles.

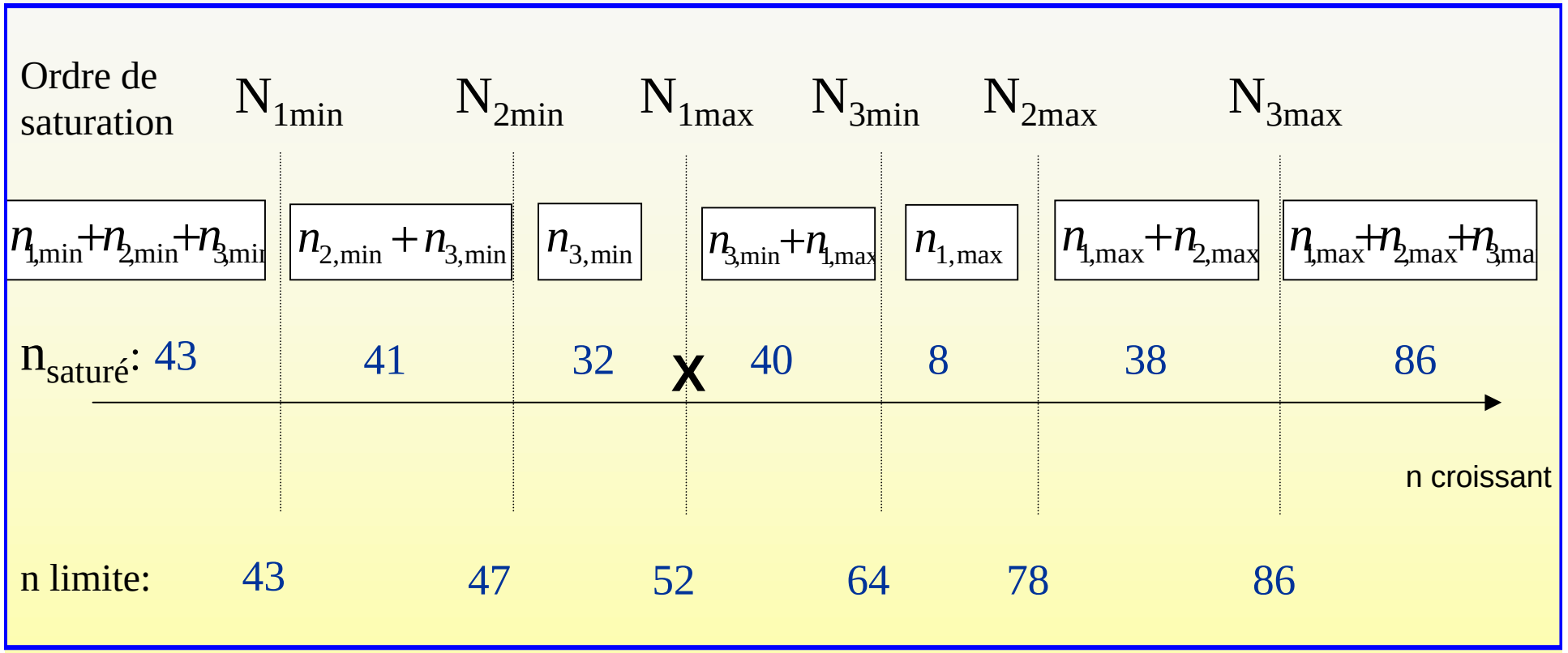
Exemple fictif avec 3 strates:

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

$$(2 \le n_1 \le 8)$$
$$(9 \le n_2 \le 30)$$
$$(32 \le n_3 \le 48)$$

Contrainte	Valeur de la variable *ordre*	Ordre de saturation
N_1 min	1	*1*
N_1 max	4	*3*
N_2 min	3	*2*
N_2 max	10	*5*
N_3 min	8	*4*
N_3 max	12	*6*

Rappel: $(2 \le n_1 \le 8)$ $(9 \le n_2 \le 30)$ $(32 \le n_3 \le 48)$



Les solutions sont des fonctions par morceaux calculables :

- les intervalles de définition dépendent des contraintes n_{hmin} (c)
- sur chaque intervalle, la dépendance en ***n***, a une forme simple :

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

De plus, sur chaque intervalle, on a des relations simples également calculables entre différentes quantités.

En particulier entre CV et n :

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

Par substitution, la relation $n_h = N\_c(n)$ peut s'écrire :

$$n_h = Neyman\_c(CV)$$

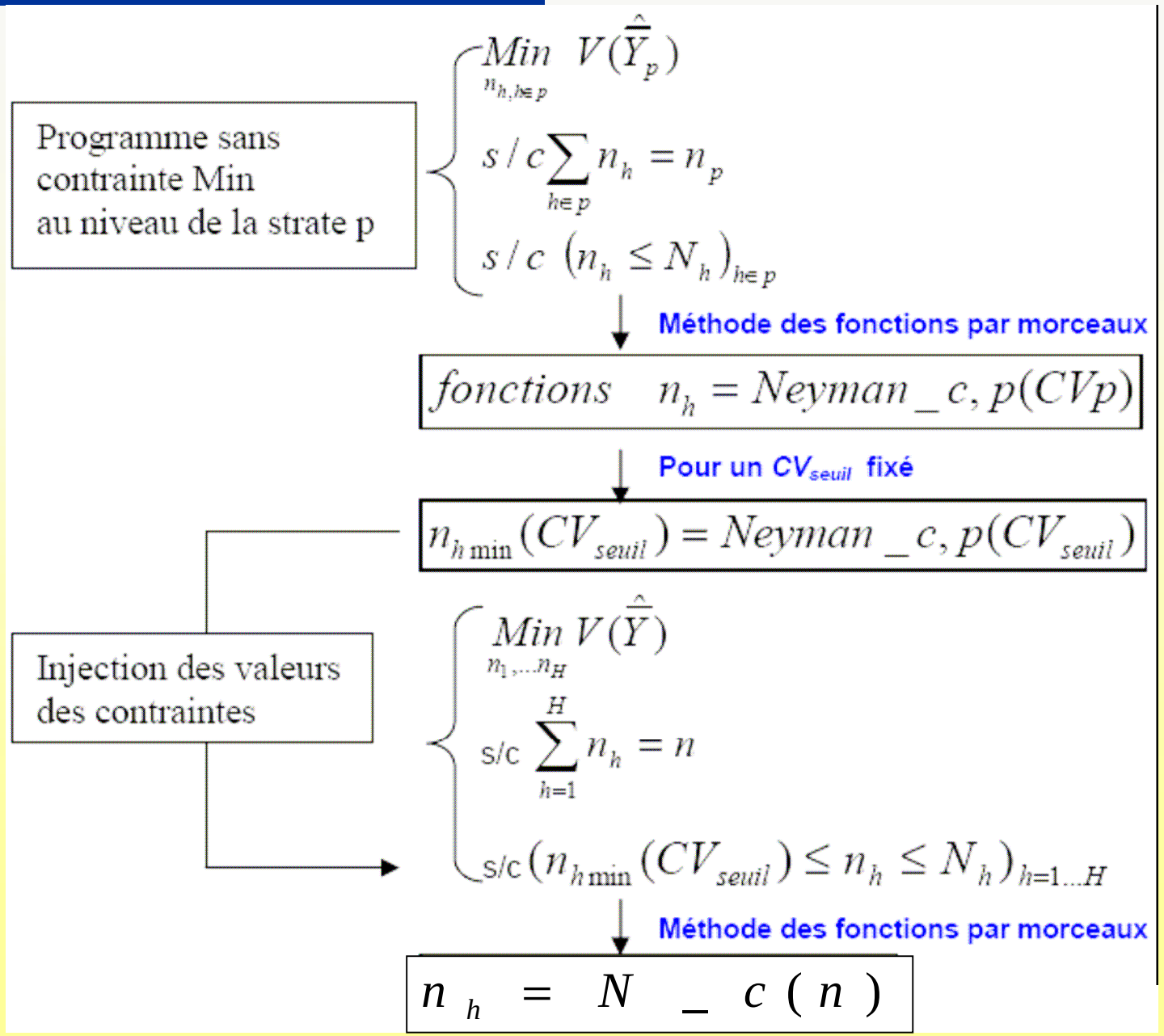
Dernière observation :

l'analyse peut être menée au niveau de chaque regroupement de publication

$$n_h = N\_c, p(n_p)$$

$$n_h = Neyman\_c, p(CV_p)$$

Schéma global de résolution



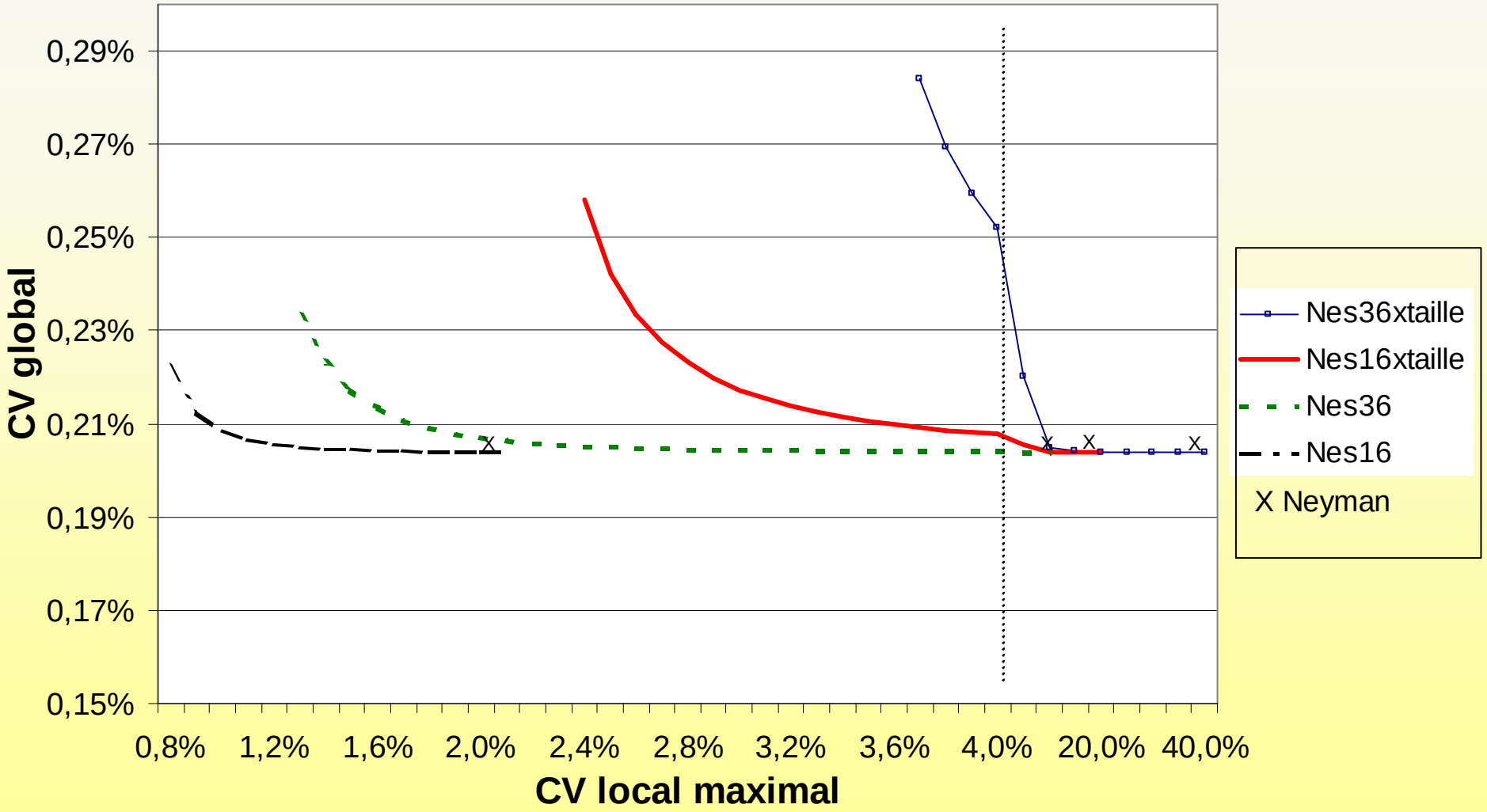
3. Illustration de l’arbitrage entre précision globale et précision locale

Concept de frontière d'efficacité

Ensemble des couples (CV_{global}, $CV_{local\ maximal}$) réalisables pour une taille d'échantillon donnée, tel qu'on ne peut améliorer l'un des deux termes sans détériorer l'autre.

Aide à la décision précieuse pour l'échantillonnage d'une enquête.

Comparaison des frontières d'efficacité pour une taille d'échantillon n=30 000 et différents niveaux de publication



Compléments

- L'ordre de saturation est donné par:

$$Ordre_{h,\max} = \frac{N_h}{N_h * S_h} \qquad Ordre_{h,\min} = \frac{n_{h\min}}{N_h * S_h}$$

$Ordre_{h,\max}$: ordre de saturation des contraintes Max ($n_h = N_h$)

$Ordre_{h,\min}$: ordre de désaturation des contraintes Min ($n_h = n_{h\min}$).

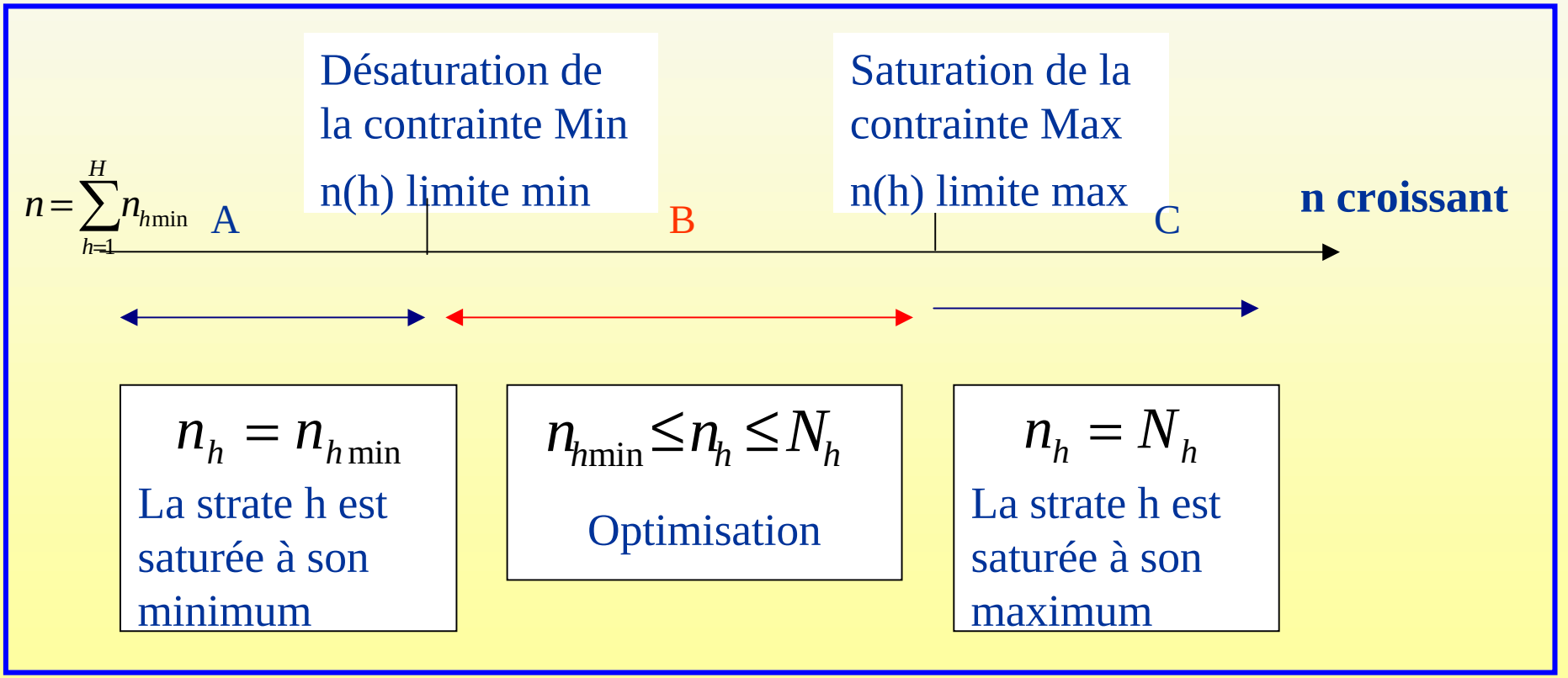
- On détermine la taille d'échantillon pour laquelle chaque strate sature.

$$n(h_0) = n_{sat} + \frac{N_{m_0}}{N_{h_0} * S_{h_0}} * \sum_{h \in Hnonsat} N_h * S_h$$

- Pour notre taille d'échantillon fixée, on connaît l'ensemble des strates saturantes et non saturantes.

Ce qui se passe pour une strate h quand n augmente

$$(n_{h\min}(CV_{seuil}) \le n_h \le N_h)$$



Détermination de l'ordre de saturation

les n_h optimaux pour les strates non saturantes sont de la forme :

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

, h strate non saturante.

La strate h_0 sature lorsque : $n_{h_0} = \frac{N_{h_0} * S_{h_0}}{\sum_{h \in Hnonsat} N_h * S_h} * (n - n_{sat}) = N_{m_0}$

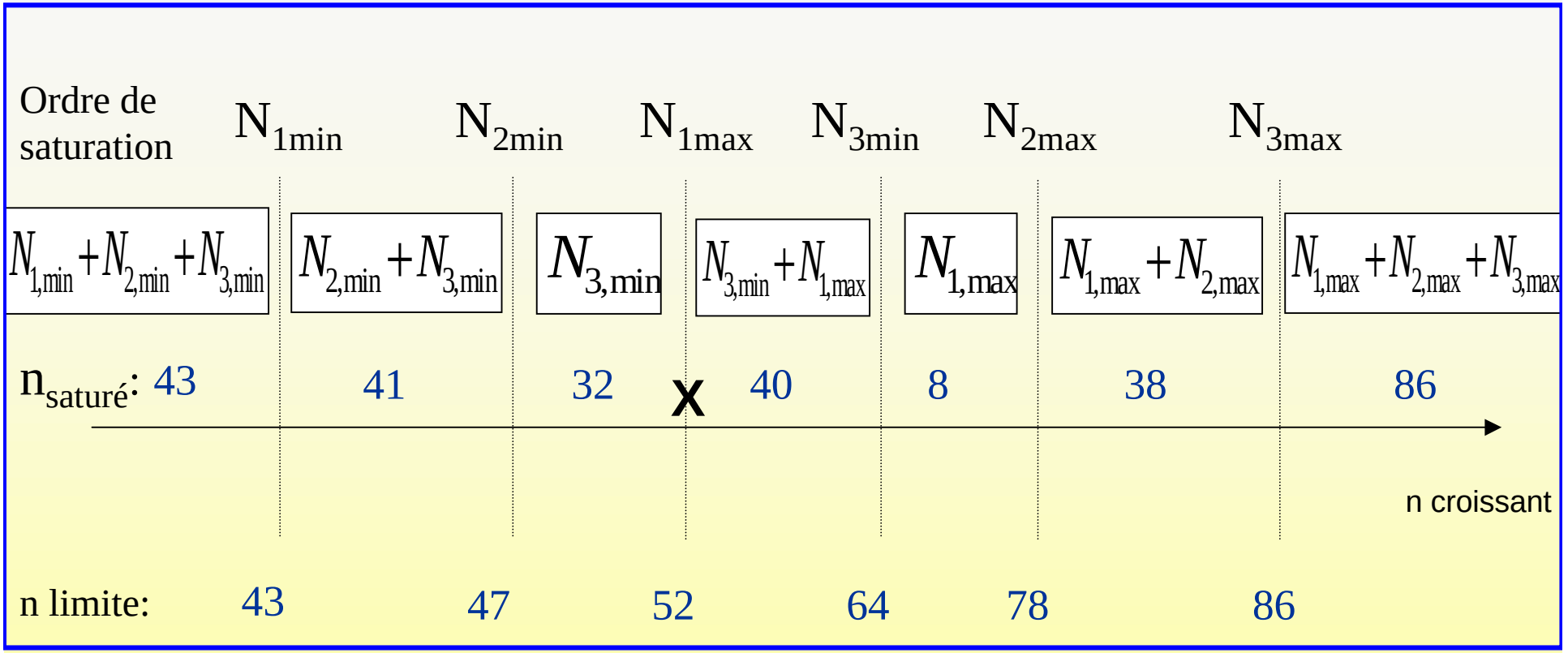
$N_{m_0} = N_{h_0 \min} ou\ N_{h_0}$

soit

$$n(h_0) = n_{sat} + \frac{N_{m_0}}{N_{h_0} * S_{h_0}} * \sum_{h \in Hnonsat} N_h * S_h$$

« Ordre »

Rappel: $(2 \le n_1 \le 8)$ $(9 \le n_2 \le 30)$ $(32 \le n_3 \le 48)$



Lors de la saturation de la contrainte N1 max : $n_3 = 32$, et $n_{saturé} = 32$.

Donc $$n(h)_{limite} = 32 + \frac{8}{2} * (2+3) = 52$$

Il suffit donc de faire une allocation de Neyman sur les strates non saturantes, càd les strates 1 et 2 pour $n_{nonsaturé} = 52 - 32 = 20$.