Les méthodes de calculs d'allocations optimales pour les enquêtes entreprises : l'exemple de l'ESA-EAP

S. Quenum – C. Imberti

Séminaire de méthodologie statistique

DG White - 20/11/19

Insee

Mesurer pour comprendre

01 Stratification et allocation : deux facteurs de gain de précision

L'échantillonnage des enquêtes entreprises

Une constante, le sondage aléatoire simple (SAS) stratifié :

- Une strate exhaustive : selon la taille (CA, effectif, contribution...)
- Une partie sondée stratifiée, avec un sondage aléatoire simple dans chaque strate

Critères de stratification : souvent l'activité (plus ou moins agrégée) croisée avec une tranche de taille (effectif ou CA), voire la région.

Unité échantillonnée : entreprise (ie groupe), unité légale, établissement (parfois un salarié au 2nd degré de tirage)

Une coordination négative entre enquêtes différentes et positive avec la même enquête N-1

L'apport du SAS stratifié

Une réduction de la variance d'échantillonnage

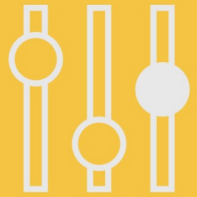
La variance est presque toujours plus faible que pour un SAS :

- ne dépend que de la dispersion intra-strate
- car sinon les chiffres d'affaires sont très variables (x1000)

Gain supplémentaire de précision en jouant sur les allocations :

- allocations plus optimales que l'allocation proportionnelle pour minimiser la variance de l'estimateur d'une variable d'intérêt principale (en général le CA)

02 L’allocation optimale de Neyman



L'allocation optimale de Neyman

Minimise la variance de la variable d'intérêt Y pour n fixé

- améliore la précision globale de l'estimateur de Y

$$V_{OPTI}(\hat{\bar{Y}}_{ST}) \leq V_{PROP}(\bar{y}) \leq V_{SAS}(\bar{y})$$

- il faut disposer d'un proxy des dispersions intra-strates de la variable Y

$$\begin{cases} \underset{n_1,\ldots,n_H}{Min}\ V(\hat{\bar{Y}}) \\ \text{s/c} \quad \sum_{h=1}^{H} n_h = n \end{cases}$$

En généralisant à une contrainte de coût total fixé :

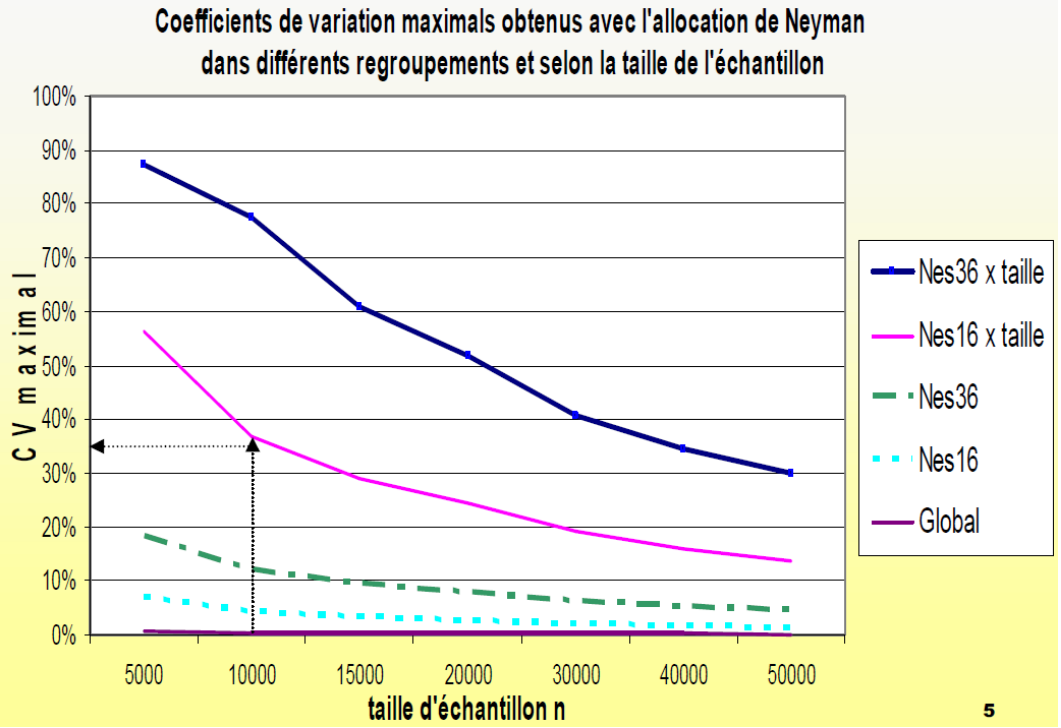
$$C_0 + \sum_{h=1}^{H} C_h n_h = C$$

Les limites de l'allocation de Neyman

Pas de gain de précision partout

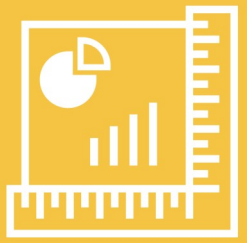
- ne garantit pas une bonne précision pour des estimations sur des sous-groupes de la population (« regroupements de publication ») (Koubi et Mathern, JMS 2009)

- peut dégrader la précision d'autres variables non corrélées positivement



Source : DADS, masse salariale 2004

03 L’allocation optimale de Neyman sous contraintes (de précisions locales)



Neyman sous contraintes

Garantir une précision minimale dans les domaines de diffusion

- il faut déterminer le CV local seuil
- il faut viser une précision locale sans détériorer la précision globale
- les domaines de diffusion sont des regroupements des strates d'optimisation
- il faut toujours disposer d'un proxy des Sh

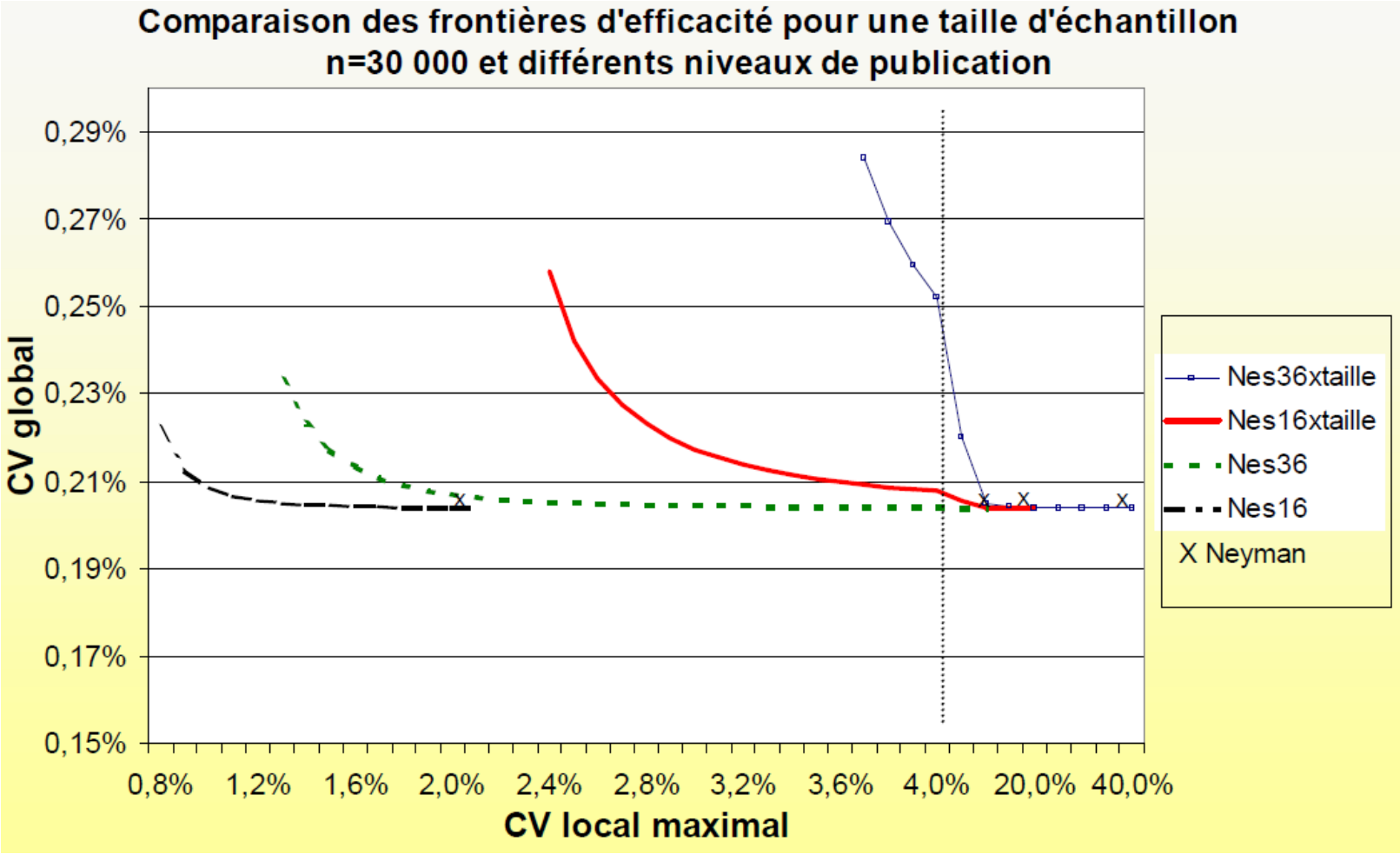
$$\begin{cases} \underset{n_1,\ldots,n_H}{Min}\ V(\hat{\bar{Y}}) \\ \text{s/c} \sum_{h=1}^{H} n_h = n \\ \text{s/c}\ n_h \leq N_h \\ \text{s/c}\ \underset{p \in pub}{Max}\, CV_p \leq CV_{seuil} \end{cases}$$

Visualisation du résultat de Neyman sous contraintes

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

Source : DADS, masse salariale 2004

Choix du CV local max : précision locale versus précision globale



Source : DADS, masse salariale 2004

L’allocation mixte

Compromis pour préserver la précision partout

Une moyenne pondérée entre différentes allocations :

- l'optimum de l'allocation de Neyman est reputé plat : s'en éloigner un peu ne détériore pas trop la précision
- mix d’allocations de Neyman pour des domaines de diffusion différents
- mix d’allocations de Neyman et proportionnelle (Merly-Alpa/Rebecq 2016) pour préserver la précision sur les autres variables

04 L'exemple de l'échantillonnage de l'ESA-EAP en groupes

Les deux enquêtes ESA et EAP

- Un tirage utilisant l'allocation de Neyman sous contraintes

Deux enquêtes entreprises faisant partie du dispositif Esane permettent de produire des statistiques structurelles :

- Enquête Sectorielle Annuelle (ESA) :

Champ : activités de commerce, de construction, de services et de transport

116 000 unités légales interrogées en France métropolitaine

- Enquête Annuelle de Production (EAP)

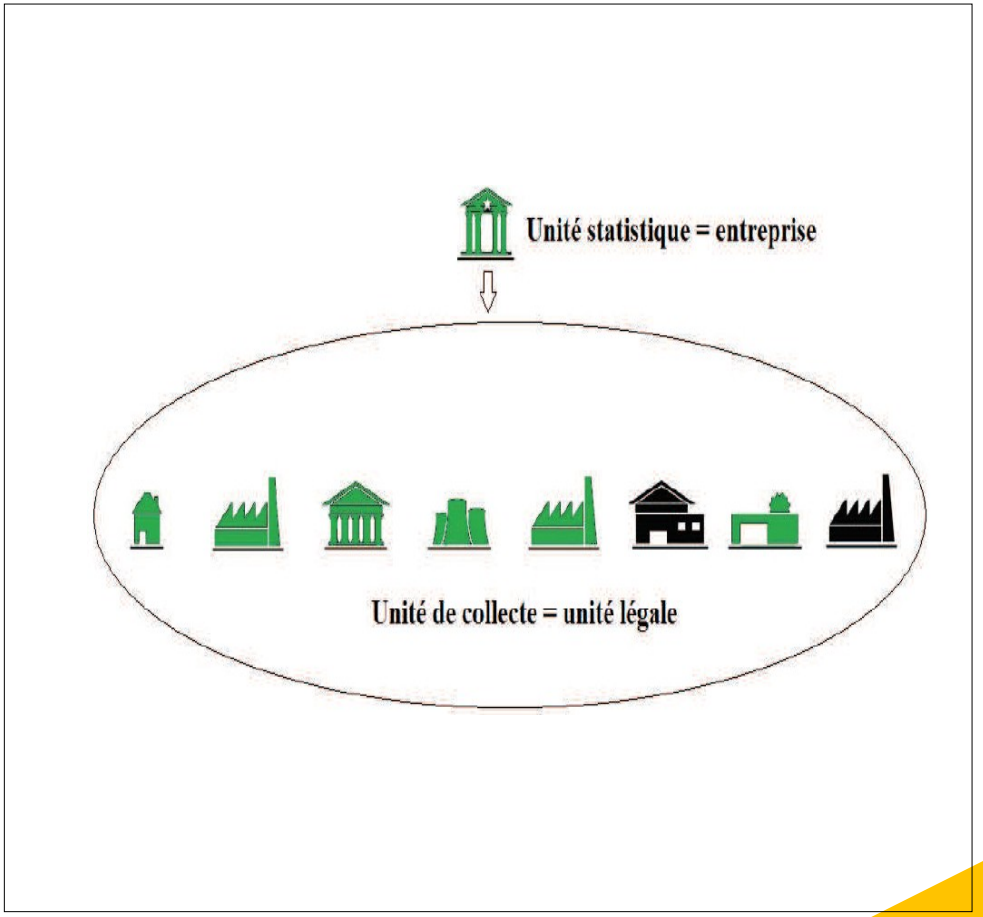
Champ : secteur industriel

35 000 unités légales interrogées en France métropolitaine

Un échantillon d'entreprises

vu comme un tirage en grappes d'unités légales (UL)

- unité interrogée = unité légale
- l'unité statistique = entreprise
- Sélection des UL "majeures" de l'entreprise échantillonnée



Les principales contraintes
- avec le nouveau concept entreprises

Réoptimiser le plan de sondage :

- pour une diffusion des résultats en entreprises...
- ...tout en pouvant exploiter les résultats en unités légales...
- ...la diffusion se faisant sur deux domaines de publication...
- ...en conservant le même nombre d'unités légales qu'actuellement à enquêter pour chacune des deux enquêtes...
- ...et que ce nombre reste stable d'une année sur l'autre !

La stratification

Séparation des champs des 2 enquêtes au niveau entreprises

La stratification est définie par le croisement entre :

- secteur d'activité de l'entreprise (au niveau APE)
- effectif de l'entreprise (9 tranches d'effectifs – teff)
- et une strate exhaustive pour chaque enquête

Les deux domaines de publication sont les suivants :

- APE (activité sur 5 positions)
- Groupe (activité sur 3 positions) teff regroupées

Allocation de Neyman sous contraintes

Allocations d'entreprises sous contraintes :

- de précisions locales (CV loc) pour le CA niveau entreprise
- de maîtrise du nombre d'UL à enquêter (par enquête) puisque l'unité enquêtée reste l'unité légale
- grace à l'introduction d'une fonction coût = nombre moyen d'UL par entreprise dans la strate h

$$\begin{cases} \min\limits_{n_1,\dots,n_H} \mathbb{V}_p[\hat{t}_{y\pi}] \\ s.c. \sum\limits_{h=1}^{H} C_h n_h = N_{UL} \\ s.c.\ n_h \leq N_h \\ s.c. \max\limits_{d\in D} CV_d \leq CV_{loc} \end{cases}$$

Avec $C_h = \bar{N}_{UL,h}$

Un nombre d'UL variable en théorie, mais peu en pratique

La contrainte de coût dans Neyman conduit en moyenne au nombre d'unités légales voulu NUL.

Mais le nombre d'unités légales qui seront réellement enquêtées est aléatoire et dépend de l'échantillon d'entreprises sélectionné

	EAP	*ESA*	*Total*
n_{ent}	27 000	82 700	109 700
$\mathbb{E}_p\left[\hat{N}_{UL}\right] = N_{UL}$	35 000	116 000	151 000
$IC_{95\%}(N_{UL})$	[34 970 ; 35 030]	[115 840 ; 116 160]	[150 830 ; 151 170]

Table – Resultats relatifs au nombre d'entreprises à tirer (n_{ent}) et au nombre d'unités légales à enquêter (N_{UL}).

Choix du CV local max : précision locale versus précision globale

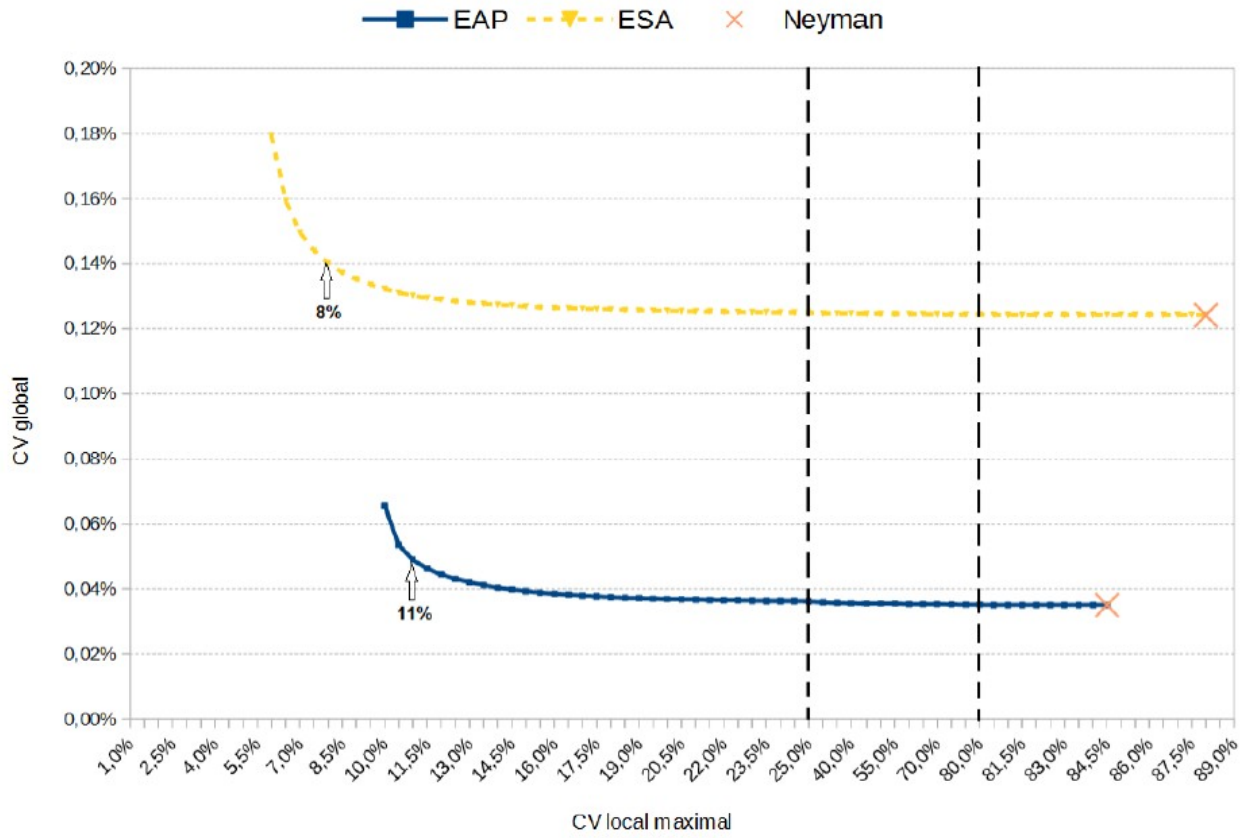


Figure – Frontière d'efficacité pour le domaine de diffusion Groupe × teff.

Choix du CV local max : précision locale versus précision globale

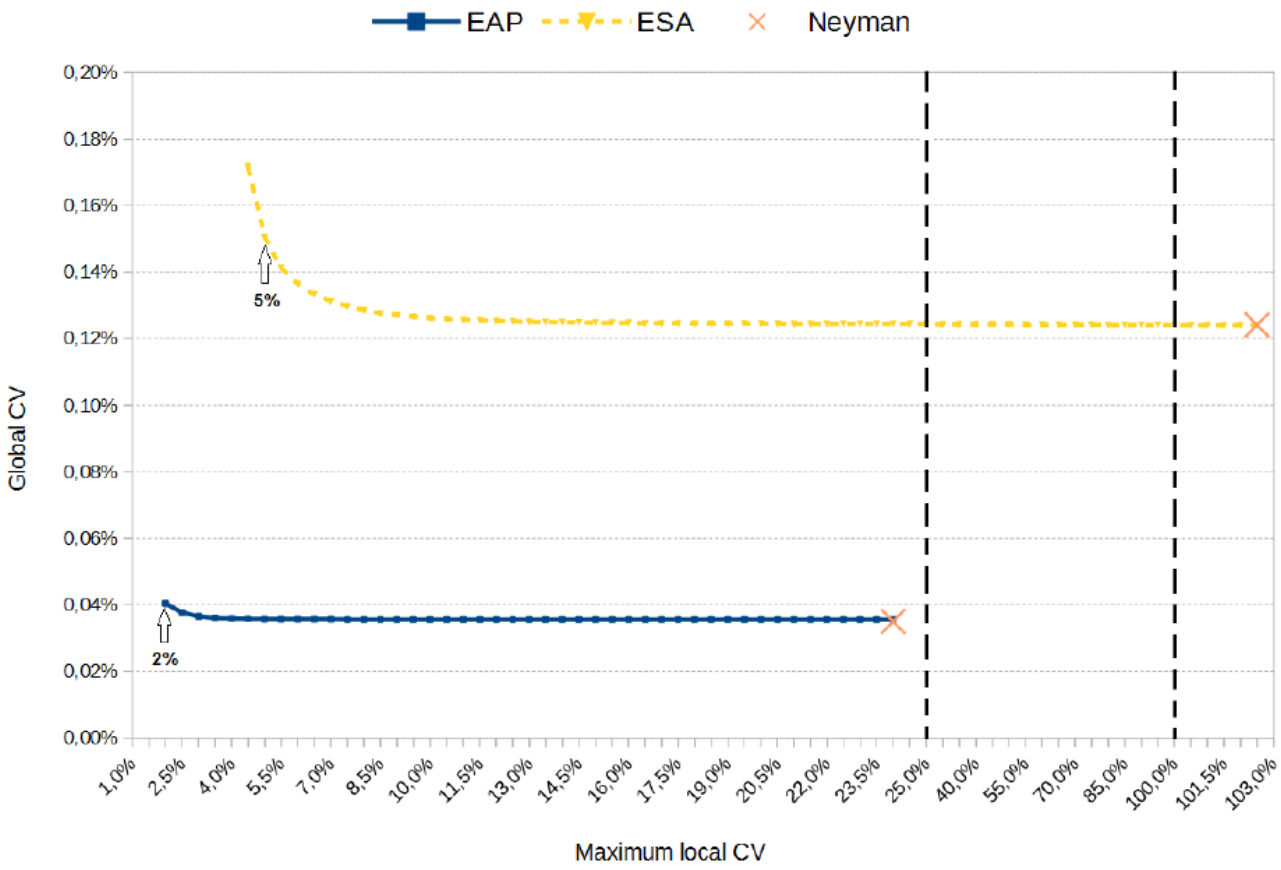


FIGURE – Frontière d'efficacité pour le domaine de diffusion APE.

L'allocation mixte: moyenne de 2 Neyman sous contraintes

$$n_{\text{mixte}} = \frac{1}{2} n_1 + \frac{1}{2} n_2$$

	Domaines de diffusion					
	APE (1)			Groupe × teff (2)		
Niveaux	$n_{ent,1}$	$n_{ent,2}$	$n_{ent,mix}$	$n_{ent,1}$	$n_{ent,2}$	$n_{ent,mix}$
100% Max	5%	74.4%	**23.1%**	89.3%	11%	**43.1%**
90%	5%	9%	**6.3%**	20.8%	11%	**12.5%**
75% Q3	5%	4.9%	4.4%	9.2%	8%	8.9%
50% Médiane	2%	2%	2%	4.2%	4.6%	4.2%
25% Q1	0.9%	0.8%	0.8%	0.1%	0.2%	0.2%
10%	0.2%	0.1%	0.2%	0%	0%	0%
0% Min	0%	0%	0%	0%	0%	0%

Table – Distribution des CV locaux des estimations de CA au niveau entreprise selon l'allocation et le domaine considérés (sans la strate exhaustive des plus de 200 salariés pour le domaine Groupe × teff).

Conclusion

Avec le passage aux groupes, le tirage de l'ESA-EAP est une forme de sondage stratifié avec grappes. Toutefois, le besoin reste d'avoir la meilleure précision globale possible sous contraintes locales. L'allocation optimale de Neyman avec contrainte de coût de collecte permet de répondre à cette problématique.

Pour en savoir plus:

R. Le Gleut-T. Merly-Alpa : L'impact du profilage sur la refonte du plan de sondage des Enquêtes Sectorielles Annuelles, JMS 2018

Merci de votre attention. Si vous avez des questions...

Sylvain Quenum – Caroline Imberti
Section méthodes de sondages pour les enquêtes auprès des entreprises
DMCSI/DMS/Division sondages
01 87 69 55 72
sylvain.quenum@insee.fr
caroline.imberti@insee.fr

Annexe : formules

$$V(\hat{\bar{Y}}) = (1 - \frac{n}{N})\frac{S^2}{n}$$

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

$$\forall h \in \{1, \ldots, H\} \quad n_h = n \times \frac{N_h}{N}$$

$$f_h = \frac{n_h}{N_h} = \frac{n}{N} = f$$

$$V(\hat{\bar{Y}}^{prop}_{SAS-str}) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 (1-f_h)\frac{S_h^2}{n_h} = \frac{1-f}{n}\sum_{h=1}^{H}\frac{N_h}{N}S_h^2 \simeq (1-f)\frac{S^2_{intra}}{n}$$

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

$$n_h = \frac{N_h \cdot S_h}{\sqrt{c_h}} \cdot \frac{C}{\sum_{h=1}^{H} N_h \cdot S_h \cdot \sqrt{c_h}}$$

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

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

$$n_h = N_h$$

$$n_\alpha = \alpha \cdot n_{\text{Prop}} + (1-\alpha) \cdot n_{\text{Neyman}}$$

$$n_h = (C - C_{sat})\frac{(N_h S_{yh})/\sqrt{C_h}}{\sum_{j \in H'} N_j S_{yj}\sqrt{C_j}}$$

$$\hat{N}_{UL} = \sum_{h=1}^{H}\sum_{k \in S_h} N_{UL,k}$$