Optimisation de l'échantillon de l'enquête sur la structure des salaires

Pascal ARDILLY (UMS)

Malik KOUBI (DERA)

INSEE

Objectif

Evaluer l'effet des caractéristiques des employeurs et des employés sur le niveau des salaires et sur le mode de rémunération des salariés

Processus d'optimisation

Plan stratifié, à 2 degrés, avec SAS à chaque degré

Strates préalables au 1er degré :
Taille X activité X grande région

1er degré : établissements

Strates préalables au 2ème degré : cadres / non cadres

2ème degré : salariés

Processus d'optimisation

Variable d'intérêt Y principale :
salaire net annuel total

On cherche les tailles d'échantillon
à ***chaque degré qui MINIMISENT***

la variance

Notations

M_h = nombre d'établissements dans la strate h

m_h = nombre d'établissements tirés en strate h

N_i = nombre total de salariés dans l'établissement i

n_i = nombre total de salariés tirés dans l'établissement i

Si on "oublie" la stratification au second degré

$$\hat{Y} = \sum \hat{Y}_h = \sum_h \frac{M_h}{m_h} \sum_{Sh} N_i \times \overline{y}_i \Rightarrow \quad V(\hat{Y}) = \sum_h V(\hat{Y}_h)$$

avec $$V(\hat{Y}_h) = M_h^2 (1 - \frac{m_h}{M_h}) \frac{S_{T,h}^2}{m_h} + \frac{M_h}{m_h} \sum_i N_i^2 (1 - \frac{n_i}{N_i}) \frac{S_i^2}{n_i}$$

$S_{T,h}^2$: dispersion des masses salariales entre les établissements de la strate h.

S_i^2 : dispersion des salaires dans l'établissement i.

Autre variable d'intérêt : le salaire horaire

Si Y = Salaire horaire = Salaire net total / durée total du travail, on travaille sur les résidus :

Résidu = Salaire net total - salaire horaire X durée du travail

OPTIMISATION SANS CONTRAINTE BUDGETAIRE

On impose 2 simplifications :

1) $\forall h, \forall \mathrm{i},\ n_i = \overline{n_h}$ ⇨ on notera $n_h = \overline{n_h} \times m_h$.

2) On distingue **cadres** et **non cadres** (strates préalables au second degré) , et on pose : $n_{h,cadr} = \lambda_h \times n_h$ et $n_{h,non-cadr} = (1-\lambda_h) \times n_h$,

avec $$\lambda_h = \frac{N_{h,cadr}.S_{h,cadr}}{N_{h,cadr}.S_{h,cadr} + N_{h,non-cadr}.S_{h,non-cadr}}$$

- C’est plus simple (moins d’inconnues) ;
- Ca donne plus de souplesse dans la gestion des contraintes de seuil ;
- Ca évite le problème du calcul des n_i pour les créations d’établissements;
- De toutes façons, la stratification par taille doit limiter la variabilité des n_i ;

En prenant en compte la stratification cadres / non cadres :

$$V(\hat{Y}) = \sum_h \left(\frac{a_h}{m_h} + \frac{b_h}{n_{h,cadr}} + \frac{c_h}{n_{h,non-cadr}}\right) + Cste$$

Avec les simplifications, on obtient

$$V(\hat{Y}) = \sum_h \left(\frac{a_h}{m_h} + \frac{d_h}{n_h}\right) + Cste$$

où **les coefficients a_h et d_h sont calculés à partir de la base de sondage (DADS 2000)**

Juste par curiosité …

$$\begin{cases} a_h = M_h^2 . S_{T,h}^2 - M_h \left(\sum_{i=1}^{M_h} N_{i,cadres} . S_{i,cadres}^2 + \sum_{i=1}^{M_h} N_{i,non-cadres} . S_{i,non-cadres}^2 \right) \\ d_h = \dfrac{b_h}{\lambda_h} + \dfrac{c_h}{1-\lambda_h} \\ b_h = M_h \left(\sum_{i=1}^{M_h} N_{i,cadres}^2 . S_{i,cadres}^2 \right) \\ c_h = M_h \left(\sum_{i=1}^{M_h} N_{i,non-cadres}^2 . S_{i,non-cadres}^2 \right) \end{cases}$$

Allocation optimale à m et n fixés

On note :

Nombre total d'établissements tirés = m

Nombre total de salariés tirés = n

On cherche les m_h et n_h optimaux pour le programme :

$$\min \sum_h \left(\frac{a_h}{m_h} + \frac{d_h}{n_h}\right)$$

$$\text{sous contraintes } \sum m_h = m \quad \sum n_h = n$$

Allocation optimale à m et n fixés

On trouve

$$m_h^{opt} = m.\frac{\sqrt{a_h}}{\sum\sqrt{a_h}} \quad \text{et} \quad n_h^{opt} = n.\frac{\sqrt{d_h}}{\sum\sqrt{d_h}}$$

Ensuite $$n_{h,cadres}^{opt} = n_h^{opt}.\lambda_h$$

et $$n_{h,non-cadr}^{opt} = n_h^{opt}.(1-\lambda_h)$$

Expression de la variance optimale

$$V^{opt}(m,n) = \frac{C_m}{m} + \frac{C_n}{n} + Cste$$

avec

$$C_m = (\sum \sqrt{a_h})^2 \quad C_n = (\sum \sqrt{d_h})^2$$

calculés à partir de la base de sondage

OPTIMISATION AVEC CONTRAINTE BUDGETAIRE

p_m = prix d'un questionnaire « établissement »

p_n = prix d'un questionnaire « salarié »

$$p_m.m + p_n.n = C$$

On pose $r = \frac{p_m}{p_n}$ et $c = \frac{C}{p_n}$. On cherche :

$$\min \frac{C_m}{m} + \frac{C_n}{n}$$

sous contrainte : $r.m + n = c$

OPTIMISATION AVEC CONTRAINTE BUDGETAIRE

On obtient finalement :

$$m^{opt} = \frac{c}{r + \sqrt{\frac{C_n}{C_m}}\sqrt{r}} \quad \text{et} \quad n^{opt} = \frac{c}{1 + \sqrt{\frac{C_m}{C_n}}\sqrt{r}}$$

En particulier on a : $\left(\frac{n}{m}\right)^{opt} = \sqrt{\frac{p_m.C_n}{p_n.C_m}}$.

Optimisation de l'échantillon de l'enquête sur la structure des salaires

(aspects pratiques)

Partie 1 : Détermination de la taille de l'échantillon

Influence de la taille (m,n) de l'échantillon sur la variance

Optimisation de la taille sous contrainte budgétaire connaissant le prix relatif des questionnaires

Contribution des niveaux salarié et établissement à la variance

La contribution du niveau établissement relativement au niveau salarié (soit Cm/Cn) est ordonnée de la manière suivante

VOLUME >> SALAIRE ANNUEL >> SALAIRE HORAIRE

Variable d'intérêt	Coefficient C_m	Coefficient C_n
Nombre d'heures	2,06511 E+20	1,98363 E+20
Nombre de jours	7,64593 E+18	6,51069 E+18
Salaire horaire	1,07674 E+22	4,23739 E+22
Salaire journalier	1,31981 E+22	5,00739 E+22
Salaire net annuel	3,96275 E+22	6,78497 E+22

Une interprétation possible

Le salaire horaire dépend de normes légales et/ou conventionnelles

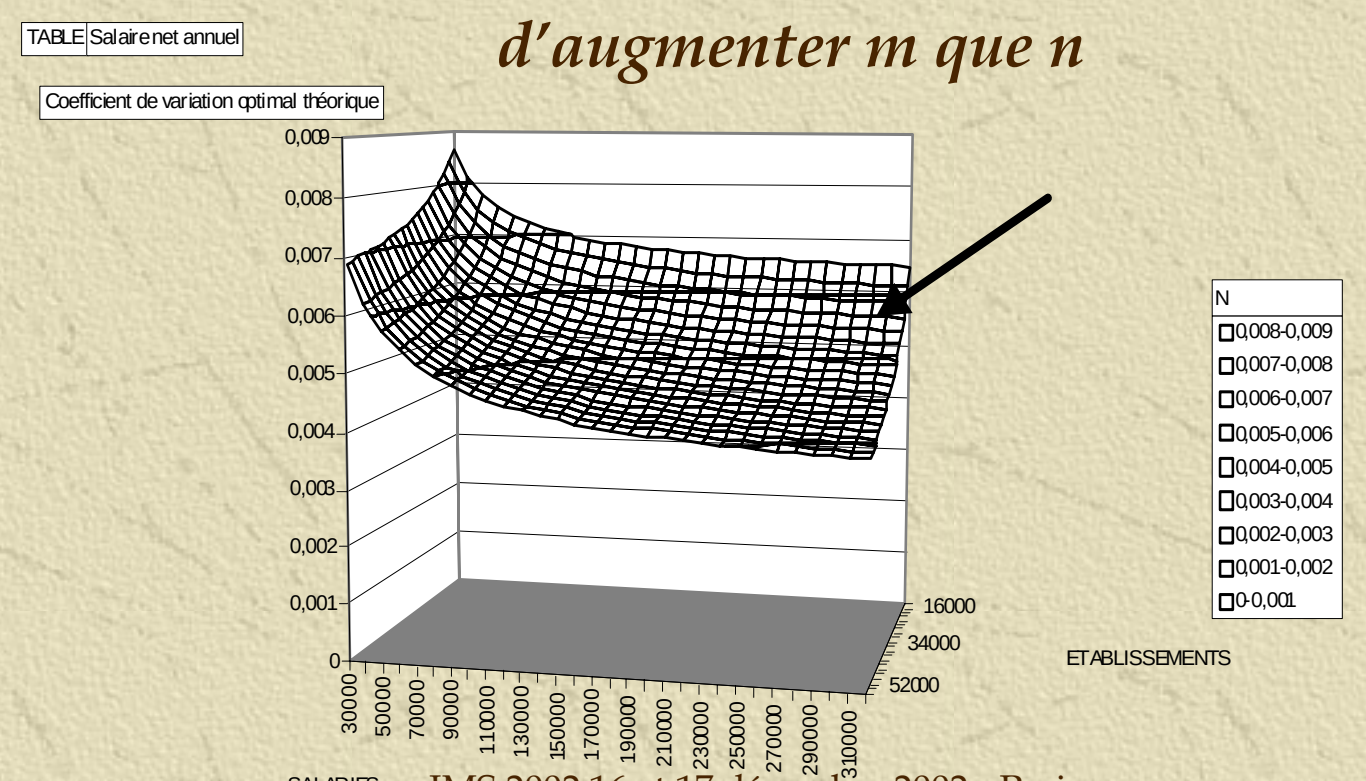
L'établissement agit donc plutôt sur le volume de travail de chaque salarié

Contribution des niveaux salarié et établissement à la variance

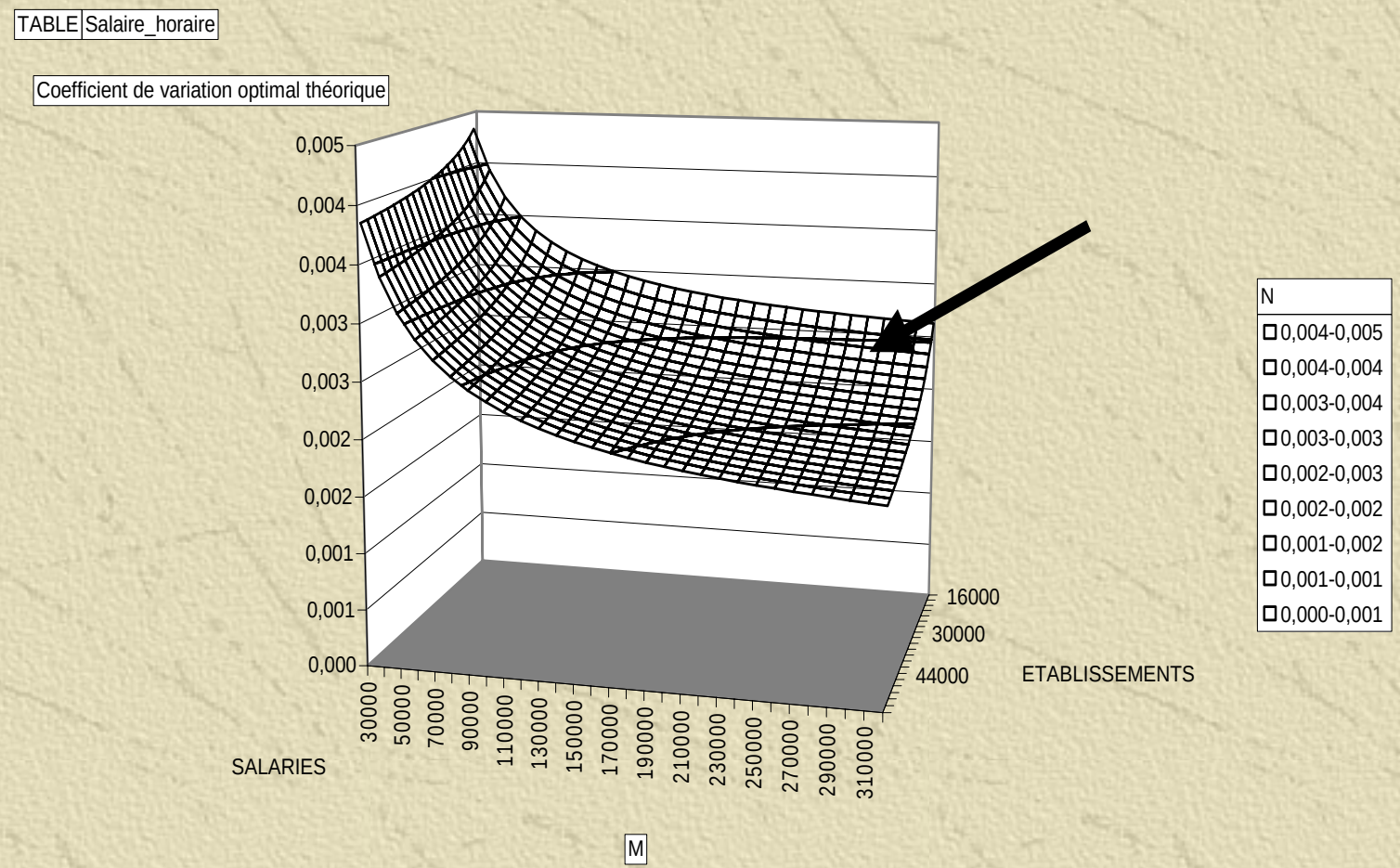
$$V^{opt}(m,n) = \frac{C_m}{m} + \frac{C_n}{n} + Cste$$

Coefficient de variation de l'estimateur en fonction de la taille de l'échantillon

Avec la taille d'échantillon actuelle, il est plus efficace d'augmenter m que n

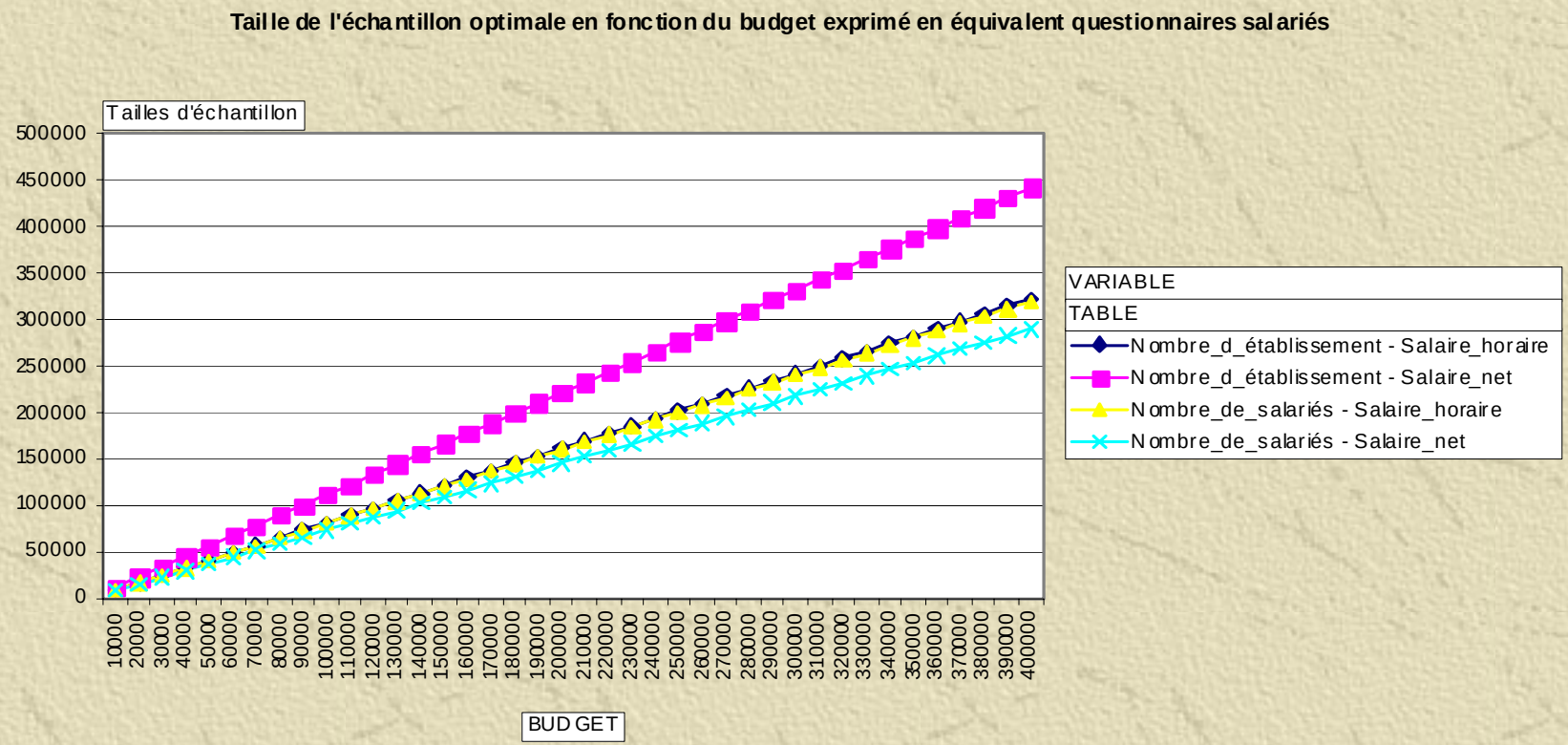


C'est également vrai mais de façon moins marquée pour le salaire horaire



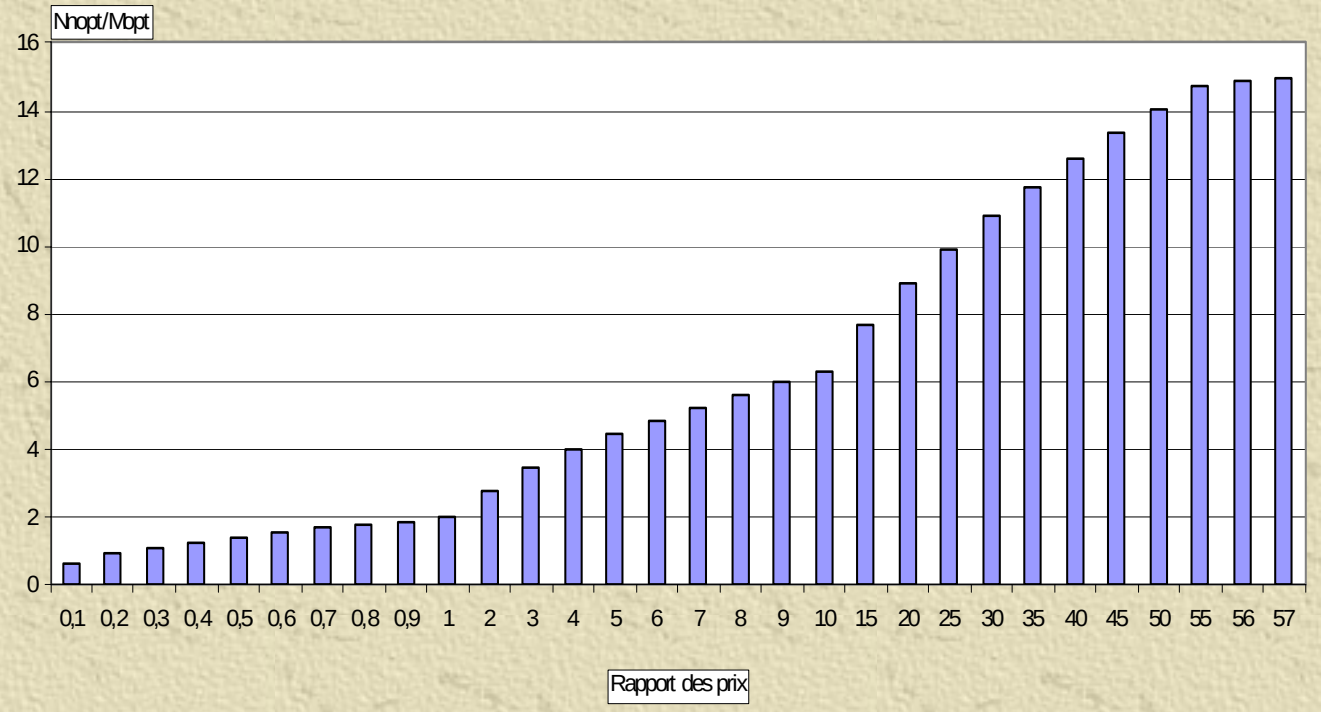
Optimisation sous contrainte budgétaire

Si on fixe le rapport des prix des questionnaires (r=0,25) et le budget (en « équivalent questionnaires salariés »)



La structure de l'échantillon optimal est cependant sensible au rapport des prix des questionnaires.

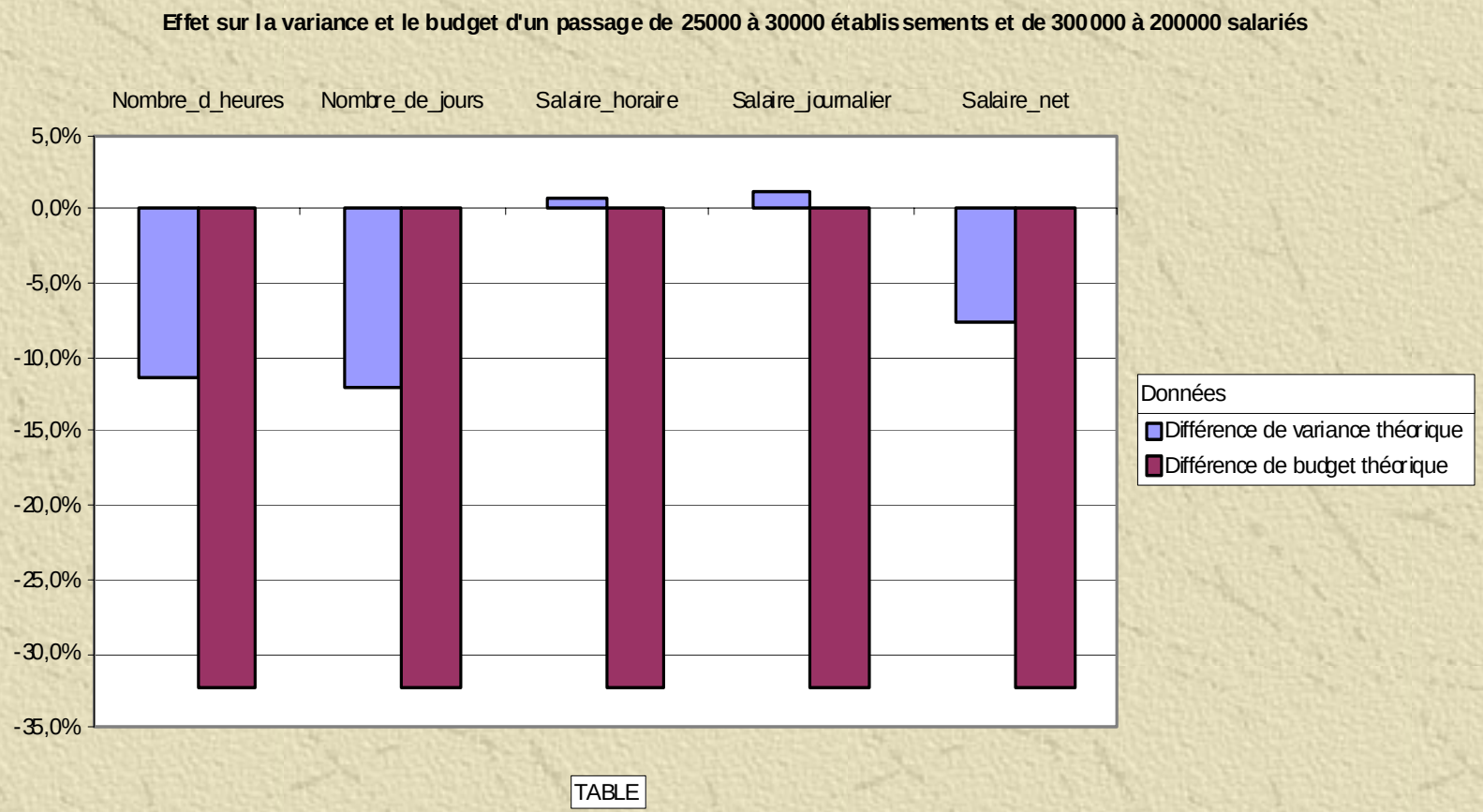
Rapport du nombre de salariés au nombre d'établissements, dans l'échantillon optimal, en fonction du prix relatif des questionnaires



Effets d'un changement de taille de l'échantillon

On se place dans l'hypothèse où r=0,25 et on considère :

- *une augmentation de m : 25 000 -> 30 000*
- *une diminution de n : 300 000 -> 200 000*



➔ *La taille actuelle n'est pas Pareto-optimale : on peut réduire le coût sans perdre en précision sur les principales variables, notamment en tirant (un peu) plus d'établissements et (beaucoup) moins de salariés*

➔ *Il faut cependant tenir compte d'autres considérations extérieures à cette étude : multiplicité des utilisations de l'enquête, charges des entreprises, etc…*

On abandonne désormais l'optimalité sous contrainte de budget et on impose m=20 000 et n=200 000

Partie 2 : Détermination de l'allocation à taille fixée m=20 000 et n=200 000

Prise en compte de contraintes supplémentaires

Description de l'allocation obtenue

L'allocation théorique ne tient pas compte de certaines contraintes implicites

Les m_h et n_h optimaux vérifient en réalité le programme suivant :

$$\min \sum_h \left(\frac{a_h}{m_h} + \frac{d_h}{n_h}\right)$$

$$\text{sous contraintes } \sum m_h = m$$
$$\sum n_h = n$$

La solution théorique à ce stade donne des m_h^{opt} proportionnels à $\sqrt{a_h}$

contraintes supplémentaires $m_h \le M_h$, $n_h \le N_h$

- *Pas de saturation au degré « salarié »*

- *Certaines strates sont saturées au degré « établissement » -> réallocation des effectifs excédentaires sur les strates non saturées*

Le principe de la réallocation est fondé sur les deux remarques suivantes

➔ *Le procédure d'optimisation est identique sur tout sous ensemble de strates (en modifiant comme il se doit la taille (m,n))*

➔ Il existe un *ordre de saturation des contraintes bien déterminé.*

Ces 2 remarques conduisent à un algorithme de ré allocation assez simple.

L'algorithme comprend 2 étapes :

** Calcul sans tenir compte des contraintes supplémentaires*

** Ré allocation* de *l'excédent sur les strates restantes, proportionnellement à* $\sqrt{a_h}$

Juste par curiosité …

On peut donner un moyen analytique de calculer le nombre de contraintes saturées dans la « vraie » solution :

c'est le plus petit entier k tel que l'optimisation sur les strates $(k+1, k+2, ..., H)$, donne une solution non saturée.

Plus précisément, le programme…

$$\min \sum_{h=k+1}^{H} \left(\frac{a_h}{m_h} + \frac{d_h}{n_h}\right)$$

…sous contraintes…

$$\sum_{h=k+1}^{H} m_h = m - \sum_{h=1}^{k} M_h$$

$$\sum_{h=k+1}^{H} n_h = n - \sum_{h=1}^{k} N_h$$

…et sans contrainte supplémentaire…

…n'est pas saturé,

Alors que le même programme avec k'=k-1 l'est. Cette condition définit l'entier k (nombre de contraintes saturées dans la vraie solution).

Description de l'allocation

Distribution des taux de sondage aux différents degrés

➔ *Les établissements sont plus sondés et de manière plus hétérogène que les salariés*

➔ *Les cadres sont plus sondés que les non-cadres*

➔ *La réallocation modifie peu la distribution des taux de sondage*

Variable d'intérêt :	Salaire_horaire (allocation n=200 000 et m=20 000)

Distribution des taux de sondage	**Selon le degré envisagé**				
Centiles de la distribution	**Cadres**	**Etab(réalloué)**	**Etablissements**	**Non-cadres**	**Salariés**
1	0,15%	0,25%	0,21%	0,28%	0,31%
5	0,42%	0,28%	0,24%	0,42%	0,44%
10	0,56%	0,30%	0,26%	0,48%	0,53%
25	1,58%	0,34%	0,29%	0,54%	0,63%
50	2,35%	0,42%	0,36%	0,69%	0,78%
75	3,98%	0,99%	0,84%	0,97%	1,28%
90	5,26%	2,15%	1,83%	1,34%	2,13%
95	6,34%	4,44%	3,77%	1,46%	2,42%
99	8,99%	20,91%	17,77%	2,06%	4,89%

Conclusion

➔ **Une optimisation avec de fortes implications pratiques, sur une enquête d'envergure**

➔ **Des méthodes classiques, mais une démarche complète de recherche d'un optimum**

➔ **Difficulté pratique pour établir une contrainte de budget pertinente**

➔ **Répercussions sur l'optimum (sensibilité de la taille optimale au rapport des prix des questionnaires)**

➔ **Des contraintes de nature stratégique conduisent à s'éloigner - très fortement - de l'optimum**

➔ **Les résultats de cette étude ont néanmoins été utilisés qualitativement pour modifier l'échantillonnage de l'ESS dans une perspective d'économie budgétaire**