UTILISATION D'INFORMATION AUXILIAIRE ET OPTIMISATION D'ÉCHANTILLON : LE CAS DE L'ENQUÊTE SUR LA STRUCTURE DES SALAIRES (ESS)

Pascal ARDILLY (*), Malik KOUBI (**)

(*) Insee, Unité Méthodes Statistiques
(**) Insee, Département de l'Emploi et des Revenus d'Activité

Introduction

Les *enquêtes sur la structure des salaires* (ESS) cherchent à évaluer, dans chaque pays européen, l'effet des caractéristiques des salariés et des employeurs sur le niveau des salaires et sur le mode de rémunération des salariés. L'information relative aux salariés échantillonnés (niveau et structure des salaires) est collectée auprès des établissements. L'échantillonnage de l'enquête se fait donc à deux degrés : on se donne une stratification des établissements et on cherche pour chaque strate le nombre d'établissements à tirer ainsi que le nombre de salariés à tirer par établissement. Certaines éditions de cette enquête ont en outre collecté de l'information sur les salariés **simultanément** auprès des établissements et d'un sous-échantillon de salariés.

D'une édition à l'autre, les tailles des échantillons de cette enquête (établissements comme salariés) ont beaucoup varié : ainsi en 1986, 23 500 établissements ont été interrogés sur 600 000 salariés ; en 1992, on a interrogé 26 000 établissements et 250 000 salariés via l'établissement ainsi que 25 000 salariés directement et, en 1994, 14 500 établissements et 160 000 salariés via l'établissement, ainsi que 20 000 salariés directement. La refonte du binôme d'enquêtes « coût » et « structure », entamée en 1999, prévoyait de retenir à l'avenir le même plan de sondage pour les deux enquêtes, qui comportent un tronc commun et doivent alterner tous les deux ans. L'échantillon de l'enquête sur le coût de la main-d'œuvre en 2000 (ECMO 2000) comportait 25 000 établissements et 300 000 salariés.

Le souci d'alléger la charge statistique pesant sur les entreprises d'une part, et les coûts associés à la collecte et au traitement du questionnaire adressé aux établissements ainsi que du questionnaire adressé directement aux salariés d'autre part, imposent de limiter la taille de l'échantillon sans sacrifier cependant la précision.

L'information auxiliaire sur les salaires contenue dans la base de sondage, constituée des fichiers établissements et salariés des déclarations annuelles de données sociales (DADS), permet de déterminer, sous contrainte budgétaire, la taille globale de l'échantillon d'établissements et de salariés qui minimise la variance de l'estimateur du total de la variable d'intérêt. Selon les objectifs que l'on assigne à l'enquête, la variable d'intérêt peut être le **salaire net annuel** ou le **salaire horaire moyen**. Les calculs seront donc menés pour ces deux variables et l'on s'efforcera de concilier les exigences de

précision relatives à chacune. Dans un premier temps, on détermine la plus petite variance qu'on peut obtenir lorsqu'on fixe à la fois la taille de l'échantillon d'établissements et la taille totale de l'échantillon de salariés. Dans un deuxième temps, en introduisant la contrainte budgétaire de l'enquête, on détermine les tailles totales respectives des échantillons d'établissements et de salariés minimisant la variance tout en respectant cette contrainte.

1. Notations et définition du champ

1.1 Notations et définitions

La lettre **h** désigne une strate d'établissements, M_h le nombre d'établissements de la strate, et N_h l'effectif salarié de la strate.

Les lettres **m** et **n** désignent respectivement le nombre total d'établissements et de salariés de l'échantillon, de même que $\mathbf{m_h}$ désigne le nombre d'établissements tirés dans la strate h et n_i désigne le nombre de salariés tirés dans l'établissement i. La variable d'intérêt est le salaire net annuel et on veut déterminer les tailles d'échantillon permettant d'estimer son total le plus précisément possible.

On note i l'identifiant de l'établissement, N_i, $N_{i,cadres}$ et $N_{i,non\text{-}cadres}$ désignent respectivement le nombre de salariés, de cadres et de non-cadres dans l'établissement i et T_i est la masse salariale de l'établissement i : ces informations « vraies » sont toutes issues de l'exploitation des DADS de l'année 2000.

Une optimisation «complète » consisterait à laisser n_i dépendre de i, mais conserver ce niveau de généralité conduirait à une gestion des calculs inextricable et d'un niveau de finesse illusoire : l'expression de variance contiendrait des millions d'inconnues (une par établissement existant…) et la démographie très évolutive des établissements ferait apparaître de nombreux cas d'établissements nouveaux sans allocation (l'optimisation se base en effet sur des données passées). Par ailleurs, la stratification des établissements tient compte, entre autres, d'un critère de taille : cette homogénéité tend à limiter la variabilité des allocations optimales par établissement. On imposera donc à l'allocation optimale du nombre de salariés tirés par établissement de ne dépendre que de la strate : $n_i = \tilde{n}_h$ pour tout établissement i dans la strate h. Si $\mathbf{n_h}$ est le nombre total de salariés tirés dans la strate h, on a par conséquent $n_h = m_h * \tilde{n}_h$.

1.2 Le champ statistique

Les règlements européens du 9 mars 1999 et du 8 septembre 2000 indiquent que doivent être couvertes toutes les activités des sections C à K de la NACE[1]. Les sections M, N et O sont facultatives pour 2002 et l'enquête ne les couvrira pas car l'étude de faisabilité n'est pas terminée.

Les établissements interrogés doivent :

- appartenir à une entreprise d'au moins dix salariés ;
- être localisés en France métropolitaine.

Néanmoins, toute l'étude qui suit a été effectuée à partir de l'ensemble des entreprises, sans limitation de taille : cette évolution du champ constitue en effet une perspective d'avenir à moyen terme.

L'univers des salariés est l'ensemble des salariés appartenant aux établissements définis ci-dessus.

[1] Voir annexe 2

2. Variance optimale à taille d'échantillon fixé

2.1 Expression de variance et programme à résoudre

Le plan de sondage effectivement utilisé est à deux degrés, stratifié préalablement au premier degré, avec sondage aléatoire simple à chaque degré. L'expression de la variance de l'estimateur d'un total (cas du salaire net) est assez facile à obtenir, et comprend deux termes : le premier (variance « inter ») traduit la variabilité liée au tirage des établissements et le second (variance «intra ») traduit la variabilité liée au tirage des salariés dans les établissements.
Le cas du salaire horaire est un peu plus complexe car le paramètre vrai et l'estimateur associé sont des ratios (salaire total au numérateur, durée totale du travail au dénominateur). La théorie de la linéarisation montre qu'il «suffit», de travailler sur des résidus : ainsi, pour calculer la variance de l'estimateur du ratio R - soit $\hat{R} = \hat{Y} / \hat{X}$ - on construira l'estimateur du total des résidus estimés définis par $Y_i - \hat{R}.X_i$ et on estimera sa variance. Sur le plan opérationnel, ce résultat est très pratique : tous les programmes écrits pour le salaire net peuvent donc être repris pour traiter le salaire horaire, à condition de travailler sur les résidus.

Si on choisit une allocation (m_h, n_h) dans la strate h, la vraie variance de l'estimateur du total de la variable d'intérêt a pour expression :

$$V = \sum_h \frac{a_h}{m_h} + \frac{d_h}{n_h} + Cste \qquad (1)$$

où Cste est une grandeur qui ne dépend pas des allocations m_h ni n_h. Pour le salaire net total, les a_h et les d_h ne dépendent que de la distribution vraie des salaires nets annuels, au travers de la dispersion du salaire total versé par les établissements à leurs salariés (masse salariale) et de la dispersion des salaires individuels perçus par les salariés de la strate h. Les expressions des a_h et des d_h sont données en annexe. Pour le salaire horaire, on manipule des dispersions de résidus (résidu du salaire sur le temps de travail).

En fait, ces coefficients tiennent compte de la distribution des salaires des cadres d'une part et des non-cadres d'autre part, au prix d'un (a priori) léger écart à l'allocation théorique optimale : en effet, on a établi les formules d'optimisation sur la base d'un nombre total de salariés par établissement, et on a ventilé a posteriori cet effectif entre cadres et non cadres, selon la philosophie de l'allocation de Neyman. On a pu ainsi réduire le nombre d'inconnues dans l'opération d'optimisation et gagner en souplesse, ne serait-ce que pour gérer plus facilement la répartition entre cadres et non cadres si une des deux populations s'avérait d'effectif trop faible dans un établissement donné.

Pour déterminer l'allocation optimale en supposant la taille globale de l'échantillon (m,n) fixée, il faut résoudre le programme suivant, la constante disparaissant :

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

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

2.2 Valeurs optimales théoriques des tailles d'échantillon

En utilisant la technique du lagrangien, le programme d'optimisation sous contraintes conduit à l'expression suivante des m_h et n_h optimaux :

Si $a_h > 0$, on trouve, $m_h^{opt} = m.\frac{\sqrt{a_h}}{\sum\sqrt{a_h}}$ $n_h^{opt} = n.\frac{\sqrt{d_h}}{\sum\sqrt{d_h}}$ (2)

Le cas où $a_h < 0$ pose problème, mais nous ne l'avons rencontré qu'une seule fois, dans la strate définie par (NACE=C, Tranche=4, Zeat=3)[2]. Cette strate comprend en tout 2 établissements et 241 salariés et figure parmi les strates exclues du traitement.

Expression de la variance optimale

A la constante près, de valeur négligeable, la valeur minimale de la variance est obtenue en remplaçant dans (1) les valeurs m_h et n_h par leurs valeurs optimales. Elle est de la forme :

$$V^{opt}(m,n) = \frac{C_m}{m} + \frac{C_n}{n} \text{ avec } C_m = (\sum\sqrt{a_h})^2 \text{ et } C_n = (\sum\sqrt{d_h})^2$$

Les deux constantes C_m et C_n sont des valeurs « vraies » calculées à partir de la situation traduite par les DADS, et qui dépendent de la distribution des salaires.

Le calcul des coefficients a_h et d_h relatifs au salaire net total et des grandeurs C_m et C_n a été effectué à partir des fichiers postes régionaux (l'unité d'observation dans ce fichier est le poste non annexe occupé par le salarié, dès lors qu'il vérifie certains critères de volume de travail). L'opération a nécessité le calcul, pour chaque établissement, en distinguant les cadres et les non-cadres, de l'effectif, de la masse salariale et de la dispersion des salaires.

Le coefficient de variation associé à l'estimateur du total $\hat{X}$ est $CV^{opt}(m,n) = \frac{\sqrt{V^{opt}(m,n)}}{\hat{X}}$.

2.3 Existence d'autres contraintes non explicitées

En réalité le programme, tel qu'il est résolu jusqu'ici, ne tient pas compte de l'existence d'autres contraintes que l'on n'explicite pas et qui doivent être vérifiées, comme le fait que la taille de l'échantillon doit être, dans chaque strate, inférieure à l'effectif de la strate. C'est pourquoi la solution trouvée est qualifiée de théorique. La complexité du calcul ne permet dans un premier temps que de procéder en «oubliant » ces contraintes supplémentaires. Celles-ci sont prises en compte dans un deuxième temps, en réallouant les effectifs correspondant aux contraintes saturées à l'aide d'un algorithme de ré allocation.

2.4 Résultats du calcul pour les deux variables d'intérêt

On a obtenu les valeurs suivantes de C_m et C_n selon les deux variables d'intérêt considérées. A titre indicatif, on a aussi calculé ces constantes pour d'autres variables d'intérêt : le nombre de jours, le nombre d'heures ou encore le salaire journalier.

[2] Voir annexe 2

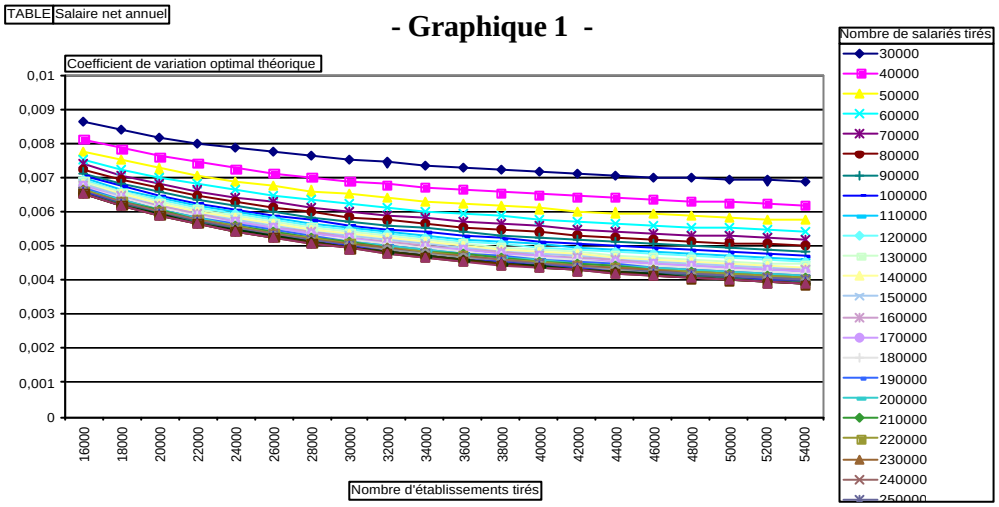
- Tableau 1 -

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

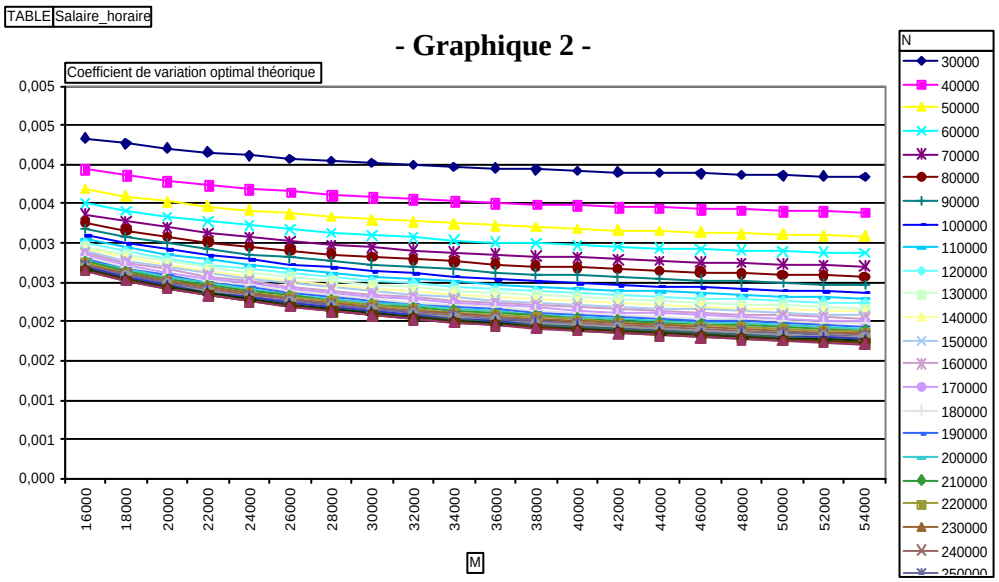
On a représenté sur les graphiques suivants, en fonction de la valeur du couple (m,n) et pour chacune des deux variables d'intérêt considérées, le coefficient de variation minimal que l'on peut obtenir théoriquement. Une première constatation est la faiblesse de ce coefficient, qui est de l'ordre de quelques « pour mille ». Ce résultat est naturel compte tenu des (très) grandes tailles d'échantillon en jeu.

Quelle que soit la variable d'intérêt retenue, il ne semble pas très rentable de tirer plus de 200 000 salariés : au-delà de ce seuil - déjà considérable -, les gains en précision sont très faibles. La sensibilité du coefficient de variation au nombre d'établissements est plus grande avec le salaire net qu'avec le salaire horaire. Dans les deux cas cependant, les gains en précision les plus importants sont naturellement acquis sur les 20 000 premiers questionnaires établissements, ce qui traduit le principe universel de variances évoluant comme l'inverse des tailles d'échantillon à chaque degré.

- Graphique 1 -



- Graphique 2 -



3. Optimisation de la taille de l'échantillon (m,n) sous contrainte budgétaire

3.1 Expression de la contrainte budgétaire

Si l'on tient compte du budget total de l'enquête et du prix relatif d'un questionnaire établissement par rapport à un questionnaire salarié, on peut déterminer l'échantillonnage optimal, c'est-à-dire le couple (m,n) optimal.

Soit C le budget total de l'enquête, p_m et p_n les coûts respectifs d'un questionnaire établissement et d'un questionnaire salarié divisés par le taux de réponse. Si m et n représentent les nombres de questionnaires répondants exploitables, la contrainte budgétaire s'écrit :

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

Dans cette expression, les paramètres $(p_m., p_n., C)$ sont supposés connus. Cette contrainte supplémentaire va nous permettre de déterminer les valeurs optimales de m et n. Les prix manipulés sont divisés par le taux de réponse afin de pouvoir travailler sur des tailles d'échantillon de répondants : le taux de réponse est le taux « établissement », la réponse d'un établissement étant de fait équivalente à la réponse des salariés qu'il emploie.

Il est pratique d'exprimer ces quantités en une unité sans dimension. Pour cela, prenons le prix du questionnaire salarié comme unité de coût et introduisons le prix relatif d'un questionnaire établissement par rapport à un questionnaire salarié $r = \frac{p_m}{p_n}$ ainsi que le rapport $c = \frac{C}{p_n}$, qui s'interprète comme le nombre total de questionnaires salariés équivalent au budget global disponible. La contrainte budgétaire s'écrit :

$$r.m + n = c$$

3.2 Optimisation de la taille d'échantillon en tenant compte de la contrainte budgétaire

Le programme à résoudre est le suivant :

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

sous la contrainte budgétaire $r.m + n = c$

En appliquant de nouveau la technique du lagrangien, on en déduit les m et n optimaux vérifiant la contrainte budgétaire :

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

conduisant à $V^{opt} = \frac{C_m}{m^{opt}} + \frac{C_n}{n^{opt}}$

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

Tout à fait logiquement, le rapport optimal entre le nombre de salariés et le nombre d'établissements de l'échantillon croît avec le rapport C_n/C_m des contributions des deux degrés à la variance, et décroît en fonction du prix relatif d'un questionnaire salarié par rapport à un questionnaire établissement. Il ne dépend pas du budget total de l'enquête. Ainsi, l'allocation optimale requiert d'autant plus de salariés que le niveau salarié contribue à la variance et d'autant moins que le prix d'un questionnaire salarié est élevé.

La dispersion des variables liées à la durée du travail provient bien davantage du niveau établissement que du niveau salarié : pour ces grandeurs, il y a un fort effet établissement, autrement dit la dispersion se situe sensiblement plus en « inter » qu'en « intra ». Par ailleurs, le salaire horaire est une grandeur relativement peu dispersée. Ce phénomène a tendance à augmenter les dispersions de type « inter » des salaires nets annuels - donc le paramètre C_m. Pour le salaire net annuel, cela conduit à une situation optimale où on a besoin de davantage d'établissements, relativement au nombre de salariés, que pour le salaire horaire.

3.3 Taille d'échantillon optimale pour le salaire net annuel et le salaire annualisé

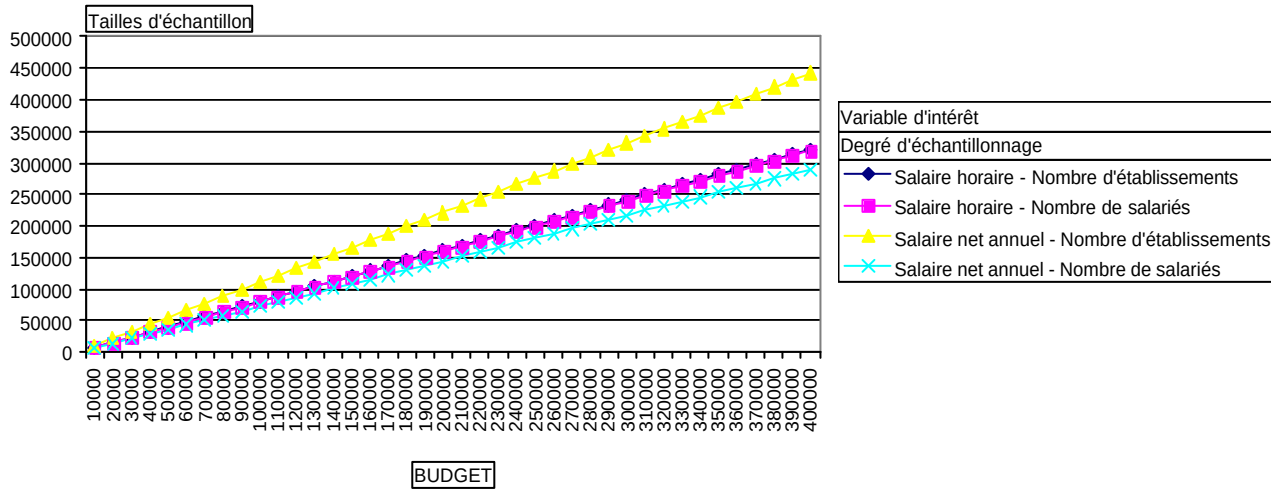
3.3.1 Résultat pour r = 0,25

Dans les graphiques suivants, on représente la taille optimale de l'échantillon minimisant la variance tout en respectant la contrainte budgétaire. Le rapport du prix du questionnaire établissement sur le prix du questionnaire salarié étant difficile à calculer avec précision a priori, on s'est appuyé sur une estimation sommaire en se fondant sur le rapport du nombre de caractères de chacun de ces questionnaires. On considère donc que le coût de traitement d'un questionnaire est à peu près proportionnel au nombre de caractères qu'il contient. Ce rapport a été estimé à r=0,25. L'abscisse des graphiques représente la variable c, précédemment définie comme le rapport du budget total sur le prix du questionnaire salarié.

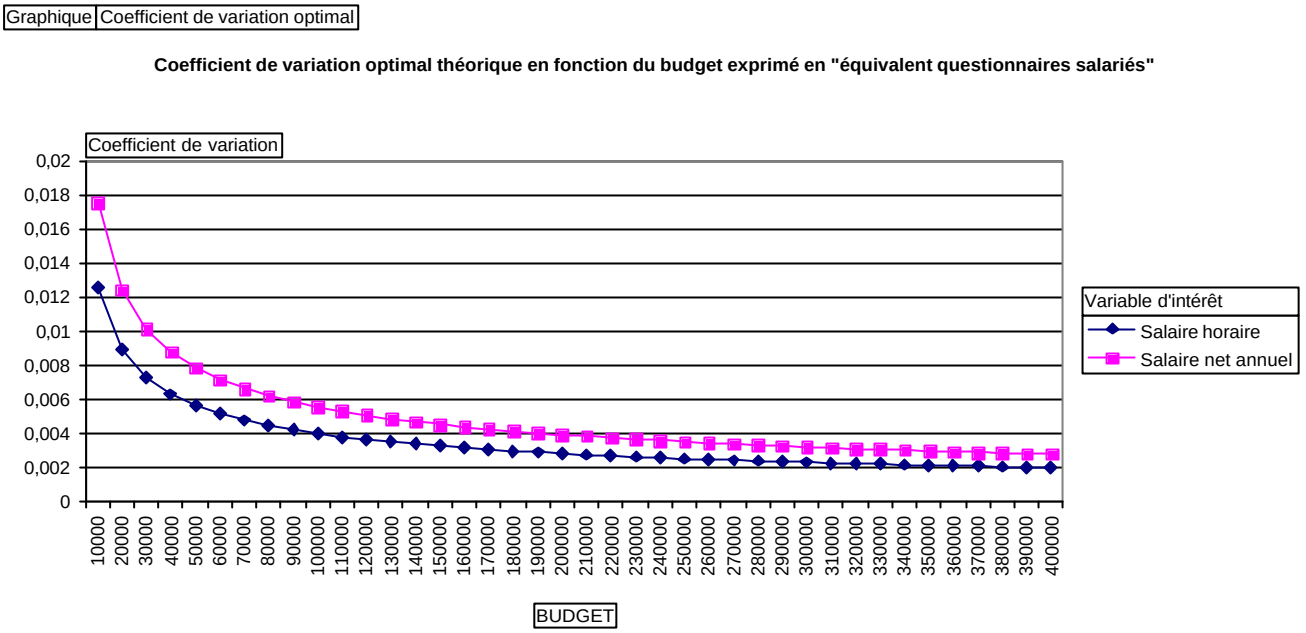
On a calculé, pour ce rapport de prix r=0,25, et en fonction du budget, la variance optimale qu'on peut obtenir ainsi que la taille d'échantillon correspondant à cet optimum pour le salaire annuel (appelé salaire net dans les graphiques) et le salaire horaire. Les graphiques suivants montrent que par rapport à la situation des années passées, **l'effort doit porter en premier lieu sur le nombre d'établissements sondés, ou, dit autrement, que l'économie doit avant tout se faire sur le nombre de salariés.**

Graphique 3

Taille de l'échantillon optimale en fonction du budget exprimé en "équivalent questionnaires salariés"

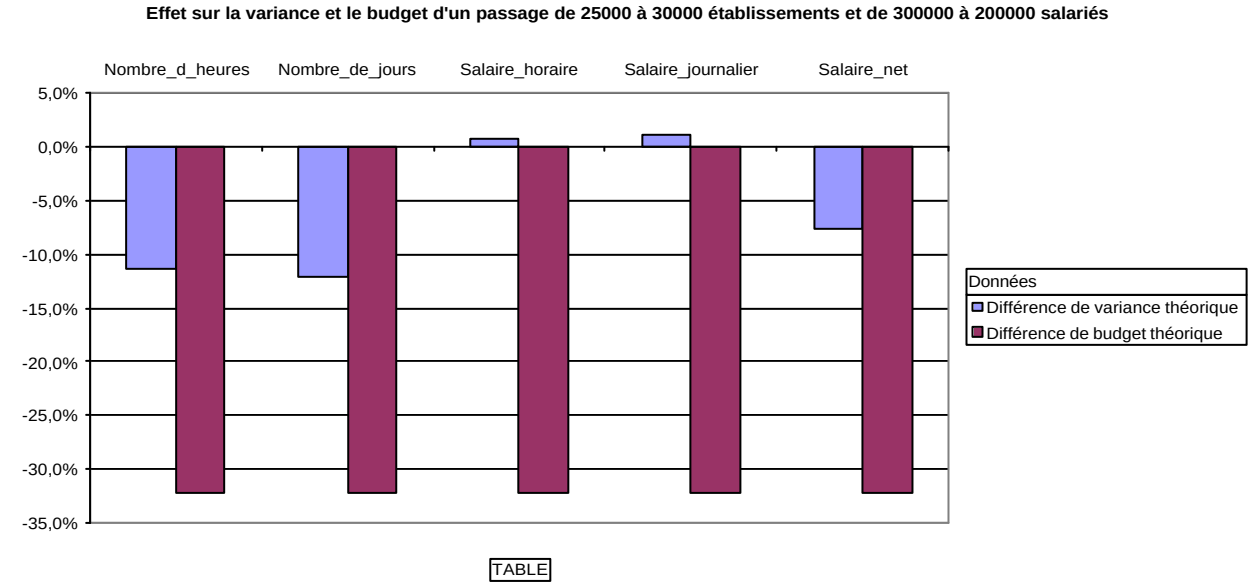


- Graphique 4 -



On peut incidemment se demander quel effet aurait une modification des tailles d'échantillon par rapport à la situation de référence de l'enquête précédente (ECMO 2000). On ne connaît de cette dernière que la taille globale de l'échantillon qui était de 300 000 salariés et 25 000 établissements. En prenant pour rapport de prix la valeur calculée r=0,25, et en choisissant de tirer 5 000 établissements en plus mais 100 000 salariés en moins, on constate que le budget total de l'enquête diminuerait de 30% sans détériorer la variance des principales variables - et même parfois en la diminuant ! Ce résultat prouve que les paramètres de l'enquête ECMO de 2000 ne correspondent pas à un optimum de Pareto : on peut gagner sur tous les plans, c'est-à-dire économiser sensiblement en tirant beaucoup moins de salariés (leur questionnaire est relativement cher par rapport aux établissements) et en compensant par une faible augmentation du nombre d'établissements - qui pèsent beaucoup sur la diminution de variance - tout en gagnant en précision (ou du moins en ne perdant pas).

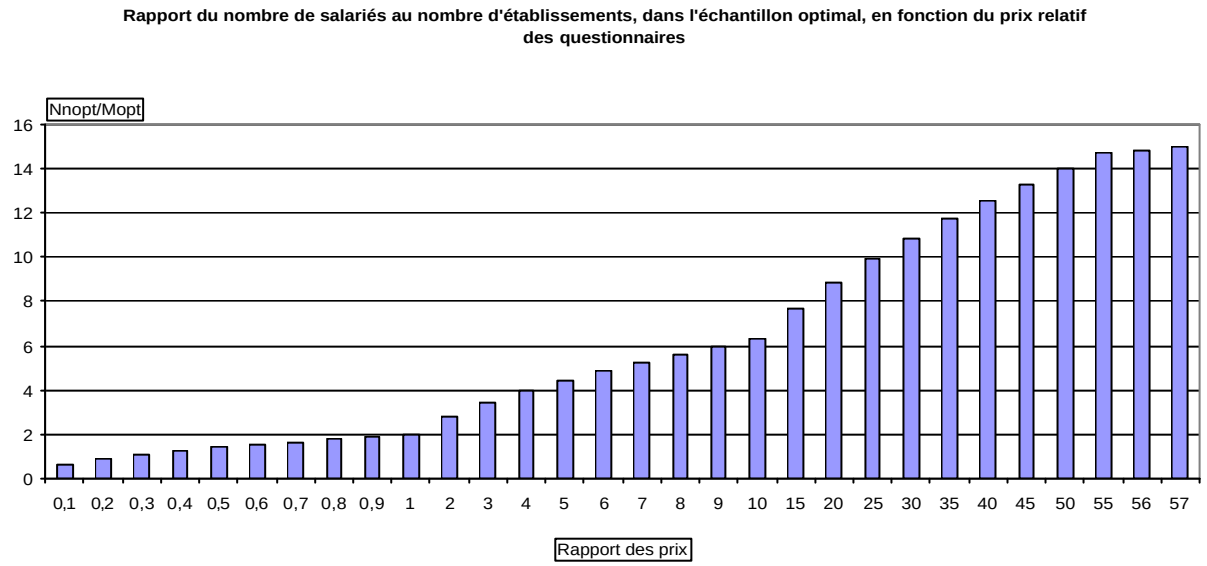
- Graphique 5 -



3.3.2 Sensibilité du résultat à la valeur de r

Comme on peut le constater en calculant la dérivée de $\frac{n^{opt}}{m^{opt}} = \frac{r + \sqrt{\frac{C_n}{C_m}}\sqrt{r}}{1 + \sqrt{\frac{C_m}{C_n}}\sqrt{r}}$ par rapport à r, ce rapport est sensible au rapport des prix des questionnaires, en tout cas avec les ordres de grandeur des paramètres C_m et C_n que l'on manipule. On a représenté ce rapport pour différentes valeurs de r.

- Graphique 6 -



On retiendra qu'au-delà de l'importance qu'il y aurait à affiner le rapport des prix des questionnaires, le nombre de salariés à tirer par établissement est faible, de l'ordre de quelques unités - et qu'avec le rapport de 0,25 qui semble le plus judicieux, on devrait tirer finalement presque autant d'établissements que de salariés.

4. Allocation optimale par strate

4.1 Prise en compte des contraintes supplémentaires

On suppose dans cette partie avoir déterminé m et n par le calcul d'optimisation précédent. Une fois déterminés m et n, l'allocation optimale théorique par strate est donnée par les formules (2). Mais comme nous l'avons déjà souligné, les formules (2) ne tiennent pas compte d'un certain nombre de contraintes, et l'allocation obtenue n'est que la solution d'un problème simplifié. La résolution complète, qui tient compte de toutes les contraintes, est en effet difficile à obtenir en toute généralité. Les contraintes omises dans le problème simplifié sont des contraintes de seuil : elles expriment, d'une part que le nombre d'établissements à tirer dans une strate ne doit pas être supérieur au nombre total d'établissements de la strate, et d'autre part que le nombre de salariés à tirer par établissement (en distinguant cadres et non cadres) ne doit pas dépasser le nombre total de salariés employés dans cet établissement. Or le calcul de l'allocation sans contrainte ne vérifie pas ce présupposé, ce qui signifie que certaines contraintes omises sont en fait saturées.

En la circonstance, le cas des contraintes sur le nombre de salariés est extrêmement confortable pour la simple raison qu'aucune contrainte de ce type n'est saturée : les conditions numériques sont telles, avec un faible nombre de salariés à tirer par établissement, que nous n'avons jamais rencontré ce cas.

Concernant la contrainte sur le nombre d'établissements, pour obtenir une solution qui tienne compte des contraintes de seuil, on part de la solution théorique du problème simplifié, et on réalloue les effectifs « excédentaires » des strates saturées (donc celles pour lesquelles la solution théorique donne un effectif à tirer supérieur à celui de la strate).

Classons les strates selon les rapports $\frac{\sqrt{a_h}}{M_h}$ (3) décroissants. La contrainte correspondant au nombre d'établissements de la strate h est saturée si et seulement si $m_h^{opt} \geq M_h$, c'est-à-dire d'après la formule (2) si $\frac{\sqrt{a_h}}{M_h \sum \sqrt{a_h}}.m \geq 1$ ou encore si $\frac{\sqrt{a_h}}{M_h} \geq \frac{\sum \sqrt{a_h}}{m}$, le deuxième membre ne dépendant pas de la strate. La contrainte saturant en premier correspond donc bien au plus grand rapport de type (3), et ainsi de suite.

On en déduit un algorithme permettant de re allouer les excédents des strates saturées sur les autres strates, jusqu'à ce que toutes les contraintes soient vérifiées. Cette première ré allocation peut provoquer de nouveaux dépassements, auquel cas il faut enchaîner par une deuxième ré allocation. Dans l'exemple ci-dessous (obtenu pour le salaire horaire, avec n=200 000 et m=20 000), il a fallu 2 étapes de ré allocation pour aboutir à une solution acceptable. Par rapport à la solution sans contrainte, la variance a augmenté d'environ 10,9% et la distribution des taux de sondage s'est trouvée légèrement modifiée.

- Tableau 2 -

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

Lecture. Colonne 3 : les strates représentant 1% des cadres qui sont le moins sondées sont sondées à moins de 0,15%.

4.2 Allocation optimale pour n=200 000 et m=20 000

La réalité diffère beaucoup de l'optimum : en effet, tirer un grand nombre d'établissements revient à faire porter une lourde charge sur les grandes entreprises, qui remplissent souvent les questionnaires pour leurs établissements (nombreux à être échantillonnés dans ces circonstances). De ce fait, les ordres de grandeur des tailles d'échantillon d'établissements préconisés par l'optimisation sont hors d'atteinte. Il faut donc se restreindre à utiliser l'enseignement de l'optimisation de manière qualitative : augmenter certes le nombre d'établissements à échantillonner mais de manière modérée.

Par ailleurs, les contraintes budgétaires conduisent à restreindre le nombre d'établissements interrogés, afin de dégager des moyens pour envoyer directement des questionnaires aux salariés : en effet, certaines informations individuelles ne sont pas connues, ou mal connues, par les établissements (tel le diplôme). Cette perspective conduit à envisager un échantillon d'établissements plus petit que pour l'enquête sur le coût de la main d'œuvre (ECMO) 2000[3]. Dans ces conditions, l'apport de l'optimisation consiste à limiter encore plus fortement le nombre de salariés. C'est pourquoi on partira concrètement sur la base d'un échantillon de 20 000 établissements et de 200 000 salariés « seulement ». En annexe, on décrit le plan de sondage obtenu avec ces effectifs. L'échantillon des salariés interrogés directement à leur domicile sera identique à celui des salariés interrogés par le biais de leur établissement, au problème des adresses absentes près.

On rappelle le contexte : la base de sondage de l'enquête sur la structure des salaires en 2002 est constituée par le fichier DADS stratifié suivant les critères de taille, activité et région. Le sondage est à deux degrés : établissements, puis salariés. La population des salariés est stratifiée en distinguant cadres et non-cadres. En considérant comme variable d'intérêt le salaire horaire, comme pour les enquêtes sur la structure des salaires (ESS) précédentes, en s'imposant les tailles $m = 20\,000$ et $n = 200\,000$, on détermine le nombre d'établissements et de salariés à tirer dans chaque strate par une allocation optimale, conformément aux calculs présentés dans la partie 2.1.

Afin de ne pas alourdir la charge pesant sur les établissements, on a écrêté, conventionnellement, le nombre maximum de salariés enquêtés par établissement à 28. Ce choix est un compromis entre l'ECMO 2000, pour laquelle on enquêtait jusqu'à 32 salariés, ce qui a paru lourd à certains établissements lors de la collecte et l'ESS 94 qui enquêtait 24 salariés au maximum. In fine, si on s'intéresse au salaire horaire, 7 strates (sur 373) dépassent le seuil des 26 salariés par établissement.

Conclusion

Cette optimisation s'appuie sur quelques hypothèses simplificatrices qui paraissent raisonnables. Ce n'est pas tant la méthodologie dont elle procède qui est remarquable que son application dans le cadre d'une enquête d'envergure : dans le monde «réel» il y a en effet extrêmement peu de plans de sondage qui sont optimisés de manière complète, et les démarches sérieuses de recherche d'un optimum sont suffisamment exceptionnelles pour qu'on mérite de les signaler lorsqu'elles existent.

Un des obstacles récurrents à la recherche d'optimum demeure l'établissement d'une contrainte de budget pertinente : comme souvent, on ne sait pas bien évaluer les coûts unitaires d'enquête, et hélas on constate que les optimum sont assez sensibles à ces coûts.

A contrario, on constate que certaines contraintes de nature stratégique, ici celles touchant à la charge des entreprises, conduisent à s'éloigner - très fortement dans le cas présent - de l'optimum. Cette constatation peut paraître décevante pour le technicien, mais on peut lui rétorquer que c'est parce que son optimisation est construite sur un principe partiel, qui s'appuie sur les seules contraintes de budget et qu'il faudrait en fait intégrer une contrainte supplémentaire de charge globale.

Bibliographie

[1] Ardilly P., « Les techniques de sondage », *Technip*, 1994

[3] On se souviendra (cf point 3.3.1) qu'avec 30 000 établissements et 200 000 salariés, on diminuait déjà le budget de 30% sans détériorer les variances !

5. ANNEXES

Annexe 1 : Formules d'allocation utilisées

Expression de la variance (dans le cas où on ne distinguerait pas selon la catégorie des salariés)

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

avec

$$\begin{cases} a_h = M_h^2 . S_{T,h}^2 - M_h (\sum_{i=1}^{M_h} N_i . S_i^2) \\ b_h = M_h (\sum_{i=1}^{M_h} N_i^2 . S_i^2) \end{cases}$$

M_h : nombre d'établissements dans la strate h.
N_i : effectif de l'établissement i.
m_h : nombre d'établissements tirés dans la strate h.
n_h : nombre de salariés tirés dans la strate h.
S_i^2 : dispersion des salaires dans l'établissement i.
$S_{T,h}^2$: dispersion des masses salariales entre les établissements de la strate h.

Remarque : les b_h sont toujours positifs alors que les a_h peuvent être négatifs.

Expression de la variance (dans le cas où on distinguerait selon la catégorie des salariés)

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

avec

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

$N_{i,cadres}$, $N_{i,non-cadres}$: nombre de cadres et de non-cadres dans l'établissement i.

$S^2_{i,cadres}$, $S^2_{i,non-cadres}$: dispersion des salaires des cadres et non-cadres dans l'établissement i.

$S_{h,cadres}$, $S_{h,non-cadres}$: moyennes des $S_{i,cadres}$ et $S_{i,non-cadres}$ dans la strate h.

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

Allocation optimale à m et n fixés

On cherche les m_h et n_h optimaux pour le programme :

$$\min \sum_h \left(\frac{a_h}{m_h} + \frac{d_h}{n_h}\right) \qquad (1)$$

sous contraintes $\sum m_h = m$

$\sum n_h = n$

On trouve

$$m_h^{opt} = m.\frac{\sqrt{a_h}}{\sum \sqrt{a_h}} \quad n_h^{opt} = n.\frac{\sqrt{d_h}}{\sum \sqrt{d_h}} \quad (2)$$

et ensuite la répartition selon cadres et non-cadres se fait selon la formule :

$$n^{opt}_{h,cadres} = n_h^{opt}.l_h \text{ et } n_{h,non-cadres} = n_h^{opt}.(1-l_h)$$

Expression de la variance optimale

La <u>valeur optimale de la variance</u> est obtenue en remplaçant dans (1) les valeurs m_h et n_h par leurs valeurs optimales données par (2). Elle est de la forme :

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

avec

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

Annexe 2 : Modalités en clair

Activité en nomenclature NACE (pour le sondage, on n'a retenu que les modalités C à K)

'A'="Agriculture, chasse, sylviculture"
'B'="Pêche, aquaculture"
'C'="Industries extractives"
'D'="Industrie manufacturière"
'E'="Production et distribution d'électricité, de gaz et d'eau"
'F'="Construction"
'G'="Commerce, réparations automobile et d'articles domestiques"
'H'="Hôtels et restaurants"
'I'="Transports et communications"
'J'="Activités financières"
'K'="Immobilier, locations et services aux entreprises"
'L'="Administration publique"
'M'="Education"
'N'="Santé et action sociale"
'O'="Services collectifs, sociaux et personnels"
'P'="Services domestiques"
'Q'="Activités extra-territoriales"

Tranches de taille d'établissement

- 0 : moins de 10
- 1 : de 10 à 19
- 2 : de 20 à 49
- 3 : de 50 à 99
- 4 : de 100 à 199
- 5 : de 200 à 500
- 6 : plus de 500

Zeat (on a exclu la Corse de la zeat Méditerranée)

- 1 : Région parisienne
- 2 : Bassin parisien
- 3 : Nord
- 4 : Est
- 5 : Ouest
- 7 : Sud-Ouest
- 8 : Centre-Est
- 9 : Méditerranée

Annexe 3 : description de l'allocation par strate obtenue avec n=200 000 et m=20 000 et pour le salaire horaire comme variable d'intérêt

1-Eléments de distribution de la différence entre les allocations obtenues pour le salaire annuel et le salaire horaire

2-Strates exclues du traitement

3-Strates les plus sondées et les moins sondées

4-Strates sondées exhaustivement (établissements)

5- Eléments de l'allocation optimale détaillée pour n=200 000 et m=20 000 et le salaire horaire comme variable d'intérêt (les strates sont ordonnées par nombre décroissant de salariés à tirer par établissement).

ANNEXE 3.1 : Eléments de distribution de la différence entre les allocations obtenues pour le salaire annuel et le salaire horaire

Distribution du rapport entre le taux de sondage par strate obtenu avec le salaire net (perçu) comme variable d'intérêt et le taux de sondage obtenu avec le salaire horaire montre qu'il y a peu de différence dans les allocations.

Données	Établissements	Salariés
C1	0,764562379	0,772874757
C5	0,846393206	0,849758290
C10	0,902349768	0,881130363
C25	0,928014504	0,929379027
C50	1,020798687	0,972426387
C75	1,158271405	1,011672098
C90	1,230959391	1,062845881
C95	1,391795243	1,088159403
C99	1,475494128	1,177965216

Lecture. Cn représente le nième centile. Colonne 1 : le rapport des taux de sondage (pour le salaire annuel et pour le salaire horaire) est inférieur à 0,76 pour 1% des établissements et supérieur à 1,47 pour 1% des établissements.

ANNEXE 3.2 : Strates exclues du traitement

Un certain nombre de strates ont été exclues du traitement : certaines en raison d'un effectif trop faible (1 établissement) et une strate (comportant 2 établissements) pour laquelle il n'a pas été possible de calculer l'allocation ($a_h < 0$). On choisira de sonder tous les établissements de ces strates.

NACE	TRANCHE	ZEAT	Nombre de salariés	Nombre d'établissements
C	4	1	194	1
		3	241	2
		8	132	1
	5	3	253	1
		4	418	1
	6	1	2053	1
		7	1575	1
		9	754	1
F	6	5	588	1
H	6	4	503	1
		8	603	1

ANNEXE 3.3 : Strates les plus sondées et les moins sondées

ANNEXE 3.3.1 : Strates les plus sondées

Lecture : dans la strate correspondant à la modalité (5,1,E) de la variable (TRANCHE, ZEAT, NACE), le taux de sondage des cadres est de 8,71%.

Taux de sondage CADRES			
NACE	TRANCHE	ZEAT	Strates les plus sondées : taux de sondage
E	5	1	8,71%
F	5	1	5,94%
G	6	1	5,64%
I	2	1	6,34%
	4	1	5,85%
J	2	1	5,90%
	3	1	8,57%
	4	1	7,24%
	5	1	9,39%
	6	1	8,99%

Taux de sondage ÉTABLISSEMENTS			
NACE	TRANCHE	ZEAT	Strates les plus sondées : taux de sondage
D	6	1	100,00%
		2	100,00%
		3	100,00%
		7	100,00%
		8	100,00%
		9	100,00%
E	5	1	100,00%
	6	1	100,00%
		2	100,00%
		3	100,00%

Taux de sondage NON-CADRES			
NACE	TRANCHE	ZEAT	Strates les plus sondées : taux de sondage
C	0	5	2,20%
E	5	1	2,80%
I	2	1	2,53%
J	0	1	2,20%
	1	1	2,53%
	2	1	2,01%
	3	1	2,61%
	4	1	2,41%
	5	1	2,80%
	6	1	2,06%

Taux de sondage SALARIÉS			
NACE	TRANCHE	ZEAT	Strates les plus sondées : taux de sondage
E	5	1	4,97%
I	2	1	2,96%
J	0	1	2,50%
	1	1	3,20%
	2	1	3,01%
	3	1	4,44%
	4	1	3,92%
	5	1	5,03%
	6	1	4,89%
K	0	1	2,45%

ANNEXE 3.3.2 : Strates les moins sondées

Taux de sondage CADRES			
TRANCHE	ZEAT	NACE	Strates les moins sondées : taux de sondage
0	2	H	0,10%
	3	F	0,12%
	4	H	0,11%
	5	H	0,10%
	7	F	0,10%
		H	0,10%
	8	H	0,10%
	9	H	0,08%
1	4	H	0,15%

Taux de sondage ÉTABLISSEMENTS			
TRANCHE	ZEAT	NACE	Strates les moins sondées : taux de sondage
0	1	H	0,27%
	2	E	0,24%
	3	H	0,29%
	5	E	0,29%
		H	0,28%
	7	H	0,28%
	8	E	0,29%
		H	0,26%
	9	F	0,30%
		H	0,25%

Taux de sondage NON-CADRES			
TRANCHE	ZEAT	NACE	Strates les moins sondées : taux de sondage
3	7	H	0,33%
4	5	H	0,31%
5	2	H	0,29%
	4	G	0,33%
		H	0,28%
	5	G	0,33%
		H	0,29%
6	5	H	0,21%
		K	0,28%

Taux de sondage SALARIÉS			
TRANCHE	ZEAT	NACE	Strates les moins sondées : taux de sondage
3	7	H	0,34%
4	4	H	0,35%
	5	H	0,34%
5	2	H	0,31%
	4	H	0,31%
	5	H	0,32%
6	4	K	0,36%
	5	H	0,23%
		K	0,29%

ANNEXE 3.4 : Strates sondées exhaustivement (établissements)

Dans certaines strates, tous les établissements sont sondés. Cela n'arrive pas pour les strates de salariés et il n'y a pas de strates dans lesquelles les salariés seraient sondés exhaustivement. Le tableau suivant présente toutes les strates dans lesquelles tous les établissements sont sondés. On donne aussi à titre indicatif le nombre de salariés total de ces établissements.

TRANCHE	NACE	ZEAT	Nombre d'établissements	Nombre total de salariés
5	E	1	37	10802
	J	1	163	51080
6	D	1	160	192936
		2	250	231638
		3	64	85605
		7	68	80638
		8	140	139003
		9	43	46536
	E	1	45	53676
		2	30	25235
		3	7	7852
		9	17	16655
	F	1	20	14466
	G	1	114	95200
	I	1	141	314245
		3	18	39316
		5	36	75106
		8	47	84551
		9	51	91904
	J	1	133	164276
		3	7	7657
	K	1	538	527912
		7	98	99301
		9	74	62871
Total			2301	2518461

ANNEXE 3.5 : Eléments de description de l'allocation

La variable considérée est le salaire horaire, la taille d'échantillon est fixée à 20 000 établissements et 200 000 salariés.

Aperçu des allocations ordonnées par nombre de salarié par établissement décroissant : le sondage théorique prévoit de sonder plus de 26 salariés par établissement pour 7 strates.

Lecture : les 2 premières colonnes donnent les nombres respectifs de salariés et d'établissements dans la strate.
NBETAB est le nombre d'établissements échantillonnés dans la strate, NBSAL, NCADRE, NNCADRE sont respectivement le nombre de salariés, de cadres et de non cadres à tirer dans chaque établissement (NBSAL=NCADRE+NNCADRE).

Le tableau présenté ici est incomplet : il s'agit d'un aperçu de la forme finale du résultat de l'échantillonnage.

TRANCHE	NACE	ZEAT	Nombre de salariés	Nombre d'établissements	NBETAB	NCADRE	NNCADRE	NBSAL
0	C	1	317	76	0,58	1,46	8,30	9,76
		2	1213	275	1,31	0,40	6,71	7,11
		3	151	40	0,17	0,00	0,00	0,00
		4	654	159	0,63	0,33	7,22	7,55
		5	1114	255	1,31	1,42	17,78	19,20
		7	1215	273	1,15	0,42	7,03	7,45
		8	1161	267	0,96	0,46	9,50	9,95
		9	921	217	0,92	0,23	7,75	7,98
	D	1	78887	20532	141,86	1,46	7,17	8,63
		2	74451	19674	82,15	0,33	8,43	8,76
		3	21036	5395	27,26	0,50	8,28	8,79
		4	37025	9579	41,67	0,34	7,86	8,20
		5	58080	15455	59,23	0,24	8,44	8,68
		7	49941	14094	52,00	0,29	8,00	8,29
		8	66714	17821	75,48	0,50	7,47	7,97
		9	52838	15386	58,33	0,51	9,19	9,70
.....	.....							
2	C	1	645	22	1,67	2,92	3,97	6,89
		2	2783	91	2,08	1,70	8,84	10,54
		3	207	7	0,24	0,46	5,51	5,96
		4	1381	48	1,13	1,26	7,37	8,63
		5	1776	60	0,98	1,58	10,05	11,64
		7	1596	53	1,53	1,31	6,55	7,86
		8	793	28	0,61	1,05	7,03	8,08
		9	1318	46	0,98	1,36	6,69	8,04
	D	1	103458	3396	150,77	2,81	5,50	8,32
		2	116635	3718	104,29	1,18	6,62	7,81
		3	42489	1342	32,80	1,31	7,41	8,72
		4	64074	2043	46,32	1,53	8,56	10,08
		5	87710	2788	56,56	1,33	8,22	9,55
		7	64273	2078	47,14	1,36	7,82	9,18
		8	114503	3687	92,87	1,92	8,69	10,61
		9	46119	1530	35,42	1,81	8,09	9,89
.....	.....							

(suite)

TRANCHE	NACE	ZEAT	Nombre de salariés	Nombre d'établissements	NBETAB	NCADRE	NNCADRE	NBSAL
4	C	2	219	2	0,40	0,48	2,95	3,43
		5	492	4	0,15	1,52	14,87	16,39
		7	398	3	1,12	2,44	3,78	6,22
		9	580	4	0,32	1,33	9,80	11,13
	D	1	80231	577	124,15	4,80	4,16	8,96
		2	143129	1019	128,41	1,66	5,14	6,80
		3	41178	291	30,65	1,88	6,34	8,21
		4	68514	493	48,54	2,26	6,78	9,04
		5	98774	698	82,90	1,82	5,91	7,73
		7	43263	312	35,76	2,15	5,95	8,10
		8	93962	669	84,61	2,40	5,99	8,40
		9	25213	185	26,34	2,51	5,69	8,20
	E	1	3034	20	3,38	3,09	5,40	8,49
		2	2386	19	1,34	3,20	9,79	12,98
		3	1675	13	0,93	2,12	8,77	10,89
		4	614	5	0,34	3,88	11,04	14,92
		5	1608	11	1,09	2,66	7,38	10,04
		7	968	6	1,26	0,71	4,68	5,39
		8	1197	8	0,59	1,59	8,72	10,31
		9	1918	13	1,19	3,39	11,80	15,19
.....	.....							
6	C	4	8464	2	0,55	8,42	86,07	94,49
	D	1	192936	160	160,00	17,79	8,52	26,31
		2	231638	250	250,00	1,95	5,22	7,17
		3	85605	64	64,00	2,63	8,97	11,60
		4	153770	124	110,11	3,81	9,57	13,38
		5	155893	160	150,04	1,96	5,43	7,39
		7	80638	68	68,00	8,40	9,93	18,33
		8	139003	140	140,00	6,10	8,94	15,04
		9	46536	43	43,00	6,44	7,39	13,83
	E	1	53676	45	45,00	7,48	8,09	15,57
		2	25235	30	30,00	1,84	5,12	6,96
		3	7852	7	7,00	1,89	6,57	8,46
		4	11058	14	9,90	1,92	6,08	8,00
		5	12361	14	6,95	2,22	9,08	11,30
		7	15842	17	11,53	2,52	7,21	9,74
		8	22067	22	21,29	2,31	6,00	8,31
		9	16655	17	17,00	2,01	5,61	7,62