Rapport de méthodes

Enquête suisse sur la structure des salaires 2000

0

Statistische Grundlagen und Übersichten
Bases statistiques et produits généraux
Basi statistiche e presentazioni generali

Plan d'échantillonnage, pondération et méthode d'estimation pour le secteur privé

Office fédéral de la statistique
Bundesamt für Statistik
Ufficio federale di statistica
Uffizi federal da statistica

OFS BFS UST

Neuchâtel, 2002

Statistik der Schweiz

Die vom Bundesamt für Statistik (BFS) herausgegebene Reihe «Statistik der Schweiz» gliedert sich in folgende Fachbereiche:

0 Statistische Grundlagen und Übersichten
1 Bevölkerung
2 Raum und Umwelt
3 Arbeit und Erwerb
4 Volkswirtschaft
5 Preise
6 Industrie und Dienstleistungen
7 Land- und Forstwirtschaft
8 Energie
9 Bau- und Wohnungswesen
10 Tourismus
11 Verkehr und Nachrichtenwesen
12 Geld, Banken, Versicherungen
13 Soziale Sicherheit
14 Gesundheit
15 Bildung und Wissenschaft
16 Kultur, Medien, Zeitverwendung
17 Politik
18 Öffentliche Verwaltung und Finanzen
19 Rechtspflege
20 Einkommen und Lebensqualität der Bevölkerung
21 Nachhaltige Entwicklung und regionale Disparitäten

Statistique de la Suisse

La série «Statistique de la Suisse» publiée par l'Office fédéral de la statistique (OFS) couvre les domaines suivants:

0 Bases statistiques et produits généraux
1 Population
2 Espace et environnement
3 Vie active et rémunération du travail
4 Economie nationale
5 Prix
6 Industrie et services
7 Agriculture et sylviculture
8 Energie
9 Construction et logement
10 Tourisme
11 Transports et communications
12 Monnaie, banques, assurances
13 Protection sociale
14 Santé
15 Education et science
16 Culture, médias, emploi du temps
17 Politique
18 Administration et finances publiques
19 Droit et justice
20 Revenus et qualité de vie de la population
21 Développement durable et disparités régionales

Statistik der Schweiz

Methodenbericht

Statistique de la Suisse

Rapport de méthodes

Enquête suisse sur la structure des salaires 2000

Plan d'échantillonnage, pondération et méthode d'estimation pour le secteur privé

Auteur **Monique Graf**
Office fédéral de la statistique

Editeur **Office fédéral de la statistique**

OFS BFS UST

Office fédéral de la statistique
Bundesamt für Statistik
Ufficio federale di statistica
Uffizi federal da statistica

Neuchâtel, 2002

Préambule

Le présent rapport décrit les aspects méthodologiques de l'enquête suisse sur la structure des salaires 2000, du plan d'échantillonnage à l'estimation de la précision. Le plan d'échantillonnage reprend pour l'essentiel celui de l'enquête de 1998. Nous avons tenu compte dans la pondération du changement apporté par l'introduction d'une enquête exhaustive dans deux cantons (Genève et Tessin). Pour la première fois, des calculs d'intervalles de confiance des médianes pondérées des salaires tenant compte de la variabilité due à l'échantillonnage et à la non-réponse sont publiés. Ce rapport décrit en détails la méthode utilisée.

Ce travail n'aurait pas vu le jour sans le soutien de Philippe Eichenberger et de Beat Hulliger (METH), l'esprit d'ouverture et le souci apporté à la qualité de Didier Froidevaux et Hans Mangold (LOHN) et la profonde connaissance des données sur les salaires de Jacques Méry (LOHN). Qu'ils soient remerciés pour les nombreuses discussions tout au long de ce travail. Mes remerciements vont aussi à Sabine Herbel (CSC) qui a préparé les fichiers de données.

Résumé

La présente enquête est une enquête auprès des entreprises dont le but principal est d'estimer des salaires médians pour un certain nombre de domaines (classe d'activité, niveau de qualification, sexe ...). Le plan d'échantillonnage prévoit un échantillon stratifié en grappes à deux niveaux. Il est conçu pour garantir un coefficient de variation de 5% sur le salaire moyen par strate. Les effectifs par strates sont calculés de deux manières: sur la base des salaires ou de leur logarithme. On constate qu'ils sont pratiquement identiques. La pondération reprend pour l'essentiel les poids d'extrapolation de Horvitz-Thompson. La non-réponse étant supposée ignorable dans les strates, les poids sont donc basés sur l'effectif par strate des entreprises répondantes. Les poids jugés trop grands sont alors tronqués et le déficit est distribué sur les autres classes de taille de même région et classe d'activité économique. On donne enfin la description complète de la méthode de calcul des intervalles de confiance des médianes pondérées. On définit un coefficient de variation synthétique de la médiane, calculé à partir de l'intervalle de confiance. Une comparaison de la précision de la moyenne et de la médiane est réalisée sur la base des coefficients de variation. Cette comparaison justifie l'établissement du plan d'échantillonnage sur la base des moyennes.

Mots-clé

rapport de méthodes; LSE; ESS; plan d'échantillonnage; médiane; non-paramétrique; intervalle de confiance; coefficient de variation; salaire.

Complément d'information:	Monique Graf, tél. 032 713 66 15 monique.graf@bfs.admin.ch
Réalisation:	Service de méthodes statistiques, OFS
Diffusion:	Office fédéral de la statistique CH-2010 Neuchâtel Tél. 032 713 60 60 / Fax 032 713 60 61 Order@bfs.admin.ch
Internet:	http://www.statistik.admin.ch
Numéro de commande:	338-0010
Prix:	gratuit
Série:	Statistique de la Suisse
Domaine:	0 Bases statistiques et produits généraux
Langue du texte original:	Français
Graphisme/Layout:	OFS
Copyright:	OFS, Neuchâtel 2002 La reproduction est autorisée, sauf à des fins commerciales, si la source est mentionnée.
ISBN:	3-303-00247-9

Table des matières

1 Plan d'échantillonnage **5**

1.1 Introduction 5
1.2 Changements relativement au plan d'échantillonnage 1998 5
1.3 Comparaison avec une évaluation basée sur la loi lognormale pour la répartition des salaires dans les entreprises 8

2 Pondération pour le secteur privé **11**
2.1 Comparaison du nombre de salariés déduit de la LSE00 avec d'autres statistiques . 11
2.2 Pondération finale 16

3 Estimation **18**
3.1 Introduction 18
3.2 Approche non-paramétrique 18
3.3 Intervalle de confiance non paramétrique 19

4 Mise en oeuvre dans le cas de la LSE00 **29**
4.1 Inversion de la fonction de répartition empirique 29
4.2 Non-réponse et imputation de la variance 30
4.3 Problème des entreprises supplémentaires de Genève et du Tessin 30
4.4 Détermination des paramètres 34
4.5 Résultats 35

Annexes **42**

Bibliographie **57**

1 Plan d'échantillonnage

1.1 Introduction

Le plan d'échantillonnage pour le secteur privé décrit ci-après n'a pas été complètement suivi, car en cours d'enquête, deux cantons, Genève et Tessin, ont obtenu de bénéficier d'une enquête exhaustive. Il était alors trop tard pour changer la planification. Nous avons donc fait les adaptations suivantes. La Suisse a été divisée en trois régions, Genève, Tessin et Reste. Les strates d'origine (catégories croisée de classes NOGA2 et de 5 classes de taille d'entreprises) ont encore été subdivisées en 3 selon ces régions. Le nombre d'entreprises par nouvelle strate a été obtenu du REE (registe des entreprises et des établissements). On a supposé que le plan d'échantillonnage prévu a été appliqué dans le Reste (qui représente le 90% environ de la population). Pour les deux cantons, l'enquête est exhaustive et le plan initial n'a donc pas été suivi. Le plan d'échantillonnage, comme celui des exercices précédents [5], prévoit un certain taux de non-réponse des entreprises, différent selon la classe de taille. On constatera que ce taux est approximativement le même dans les deux cantons à enquête exhaustive que dans le Reste. La non-réponse implique donc une incertitude sur les résultats qui peut être d'autant plus importante que le nombre d'entreprises de la nouvelle strate est faible. En effet dans ce cas une seule non-réponse peut diminuer le taux de réponse au-dessous du seuil acceptable pour une précision donnée.

1.2 Changements relativement au plan d'échantillonnage 1998

Le plan d'échantillonnage pour l'enquête suisse sur la structure des salaires de 1996 est décrit dans [5]. En 1998, R. Peters a utilisé la même méthode, avec quelques changements sur les paramètres d'entrée, tenant compte des résultats de 1996. Le tableau 1 résume ces modifications. Les taux de sondage du deuxième niveau correspondent à la proportion des salaires demandés à l'entreprise, et les taux de réponse, au taux de réponse des entreprises (pour la définition des classes de taille, voir [5]).

TAB. 1: Modification des paramètres entre les plans basés sur les échantillons de 1994 et 1996

	taux de sondage (%) 2ème niveau (f_{2h})		taux de réponse (α_h)	
classe de taille	1994	1996	1994	1996
1	100	100	60	60
2	100	100	75	70
3	72	79	75	70
4	34	34	80	80
5	62	62	85	85

Pour l'essentiel, la méthode pour le plan 2000 reste la même. Les taux de 1996 ci-dessus ont été conservés pour l'échantillon 1998. Voici les modifications introduites.

1.2.1 Méthode de fusion de l'échantillon 1998 avec le registre des entreprises et des établissements (REE) actualisé à 2000

Il y a deux variables communes à l'échantillon et au registre, le code entreprise (UNTKZ) et le numéro REE (BURNR). Comme nous ne distinguons pas les établissements à l'intérieur d'une même entreprise, le code entreprise est une variable plus sûre, car si un établissement se détache d'une entreprise, cette variable sera actualisée. Ainsi, nous ne voulons pas utiliser pour les calculs une entreprise qui aurait encore en 2000 le même numéro REE qu'en 1998, mais plus le même code entreprise. Par ailleurs, une entreprise simple n'a pas de code entreprise. Nous utilisons alors le numéro REE pour lier les caractéristiques des entreprises simples. Dans les plans d'échantillonnage précédents, la priorité était donnée au numéro REE. Les conséquences de ce changement sont données au tableau 2.

TAB. 2: Effectifs de l'échantillon 1998 du secteur privé, selon la méthode de fusion

	ancienne méthode	nouvelle méthode
nombre d'entreprises	6268	5965
nombre d'emplois	444689	397532

On constate qu'un peu moins d'entreprises sont prises en considération pour le calcul des taux d'échantillonnage.

1.2.2 Coefficient de variation

On observe une diminution du nombre total d'entreprises enregistrées dans le REE, ainsi que du nombre d'emplois (tableau 3).

TAB. 3: Diminution des effectifs dans le REE (secteur privé)

	REE 1998	REE 2000
nombre d'entreprises	168091	130731
nombre d'emplois	2'123904	1'965762

Le coefficient de variation nominal intra-strate (i.e. combinaison de NOGA 2 et de classe de taille) avait été fixé à 6%. Ce coefficient a pour conséquence que le nombre d'entreprises à interroger était tombé à 8901, alors qu'il y a la possibilité d'en interroger 10000. Aussi avons-nous posé ce coefficient de variation à 5%, ce qui donne 9724 à interroger.

Il y a peu de strates pour lesquelles le coefficient de variation nominal n'est pas garanti. Il s'agit des strates suivantes (table 4) :

TAB. 4: Strates pour lesquelles le CV nominal de 5% est dépassé de plus de 0.001

NOGA 2	taille	N	n LSE98	n net LSE00	n brut LSE00	M	m LSE98	m LSE00	CV
15	5	16	15	14	16	17017	8045	9232	0.069
17	5	5	4	4	5	2212	155	1097	0.104
18	5	4	2	3	4	1665	568	774	0.211
61	4	8	6	6	8	930	147	237	0.073
62	3	12	6	8	12	378	169	199	0.108
67	3	33	14	23	33	940	278	518	0.080
75	4	1	0	1	1	117	0	40	0.064
99	5	1	0	1	1	2	0	1	0.198

$n_{brut.LSE00}$ donne le nombre d'entreprises à interroger. On voit qu'il s'agit dans tous les cas de la totalité des entreprises de la strate. Le taux de réponse attendu, donné à la table 1, détermine l'effectif attendu de réponses ($n_{net.LSE00}$). Celui-ci est trop faible pour garantir le coefficient de variation. On constate que ces strates contiennent généralement très peu d'entreprises (N). Une solution pourrait être de les grouper avec celles d'une classe de taille adjacente. Une autre solution serait de trouver un moyen pour augmenter le taux de réponse.

Dans la classe de taille 5, où le nombre d'entreprises est généralement faible, pour les secteurs d'activité non mentionnés dans la table 4, le coefficient de variation est compris entre 0.2% et 4.2%. Il est "trop bon", mais ne peut être corrigé, à cause du petit nombre d'entreprises.

1.2.3 Modifications dues au changement de version du logiciel SAS

Entre le plan d'échantillonnage 1998 et celui de 2000, il y a eu le passage de la version 6 à la version 8.0 de SAS®. Il en découle une incompatibilité au niveau de la procédure UNIVARIATE

Dans la version 6, le paramètre ID a deux usages :

- Afficher les valeurs extrêmes des variables spécifiées sous VAR, avec la valeur correspondante des variables données sous ID, pour chaque combinaison de valeurs des variables données sous BY, définissant les groupes.
- Recopier dans le fichier de sortie spécifié sous OUT=, pour chaque groupe, la première occurrence des variables données sous ID.

Le deuxième usage a disparu dans la version 8, ce qui a entraîné la modification suivante du programme :

Les variables numériques ont été passées de ID à VAR, et remplacées par leur moyenne ; les variables caractères de ID à BY. Une vérification est faite que ces variables sont bien constantes par groupe.

1.2.4 Secteur public restant

Ce secteur voit de grands changements. En particulier, la classe NOGA ”00” (entreprises du secteur public restant n’appartenant pas aux classes Noga 85 et 90) a probablement changé de définition depuis 1998. Les effectifs sont tellement différents par rapport à l’exercice précédent, qu’ils sont rapportés intégralement ici :

TAB. 5: Plans d’échantillonnage pour les strates du secteur public

NOGA 2	N	n LSE98	n net LSE00	n brut LSE00	M	m LSE98	m LSE00	mean	std	CV	weight
Exercice 2000 (cv=5%)											
00	333	32	12	15	4831	487	174	6680	337	0.050	34.69
85	146	16	10	13	10516	1194	720	5726	275	0.048	18.25
90	145	8	19	24	982	51	129	6393	315	0.049	9.54
Exercice 1998 (cv=6%)											
NOGA 2	N	n LSE96	n net LSE98	n brut LSE98	M	m LSE96	m LSE98	mean	std	CV	weight
00	330	67	63	79	4423	1050	844	7317	438	0.060	6.55
85	114	42	16	20	8605	4967	1208	5806	350	0.060	8.91
90	189	39	10	13	848	731	45	6243	195	0.031	23.63

Ces différences tiennent à plusieurs facteurs.

- Le nombre de salariés (M) a augmenté dans la classe ”85”.
- Le nombre d’entreprises de la classe ”90” (N) a diminué, mais le nombre de salariés a augmenté.
- Le salaire moyen ($mean$) de la classe ”00” dans l’échantillon 96 était très haut (7317.-). Il est revenu à une valeur inférieure dans l’échantillon 98. Remarquons que les écarts-type (std) sont plus homogènes en 98 qu’en 96.

Comme l’effectif calculé de l’échantillon est une fonction croissante de l’écart-type, et décroissante du salaire moyen, ces effets s’amplifient mutuellement pour donner les différences constatées.

1.3 Comparaison avec une évaluation basée sur la loi lognormale pour la répartition des salaires dans les entreprises

La méthode usuelle des estimations de la moyenne et de la variance intra-entreprises est optimale pour une répartition normale des salaires. Or on sait qu’à l’intérieur d’une entreprise, les salaires ont souvent une répartition asymétrique avec étalement vers la droite. Une façon simple de rendre une telle répartition plus symétrique est d’utiliser la transformation logarithmique. Cela revient, du point de vue théorique, à préférer l’hypothèse d’une répartition lognormale pour les

salaires, au lieu d'une répartition normale. Ce qui suit est une analyse du bien-fondé de cette hypothèse, et une évaluation des conséquences sur le calcul du plan d'échantillonnage pour la LSE00.

La comparaison est basée sur le test de normalité de Shapiro-Wilk [7], pour une taille n d'échantillon $12 \leq n \leq 200$, et sur une modification de Royston [3] pour $4 \leq n \leq 11$ (voir Procédure SAS UNIVARIATE).

TAB. 6: Nombre d'entreprises par classe de taille et groupe de valeur des tests

classe de taille	deux lois acceptées	seult. loi lognormale	seult. loi normale	aucune loi acceptée
1	1437	53	7	215
2	985	266	21	344
3	211	136	11	302
4	230	164	9	318
5	154	185	5	587

Dans la table 6, les 5965 entreprises de l'échantillon 98 de la LSE sont réparties en 4 sous-groupes, selon que le test de normalité sur la répartition des salaires, d'une part, et sur les logarithmes des salaires, d'autre part, est en-dessus de la valeur critique de 0.05 ou non. Remarquons que, la classe de taille 1 correspondant aux entreprises de moins de 4 employés [5], les tests d'ajustement n'ont pas de sens dans ce cas. On voit que la proportion des très grandes entreprises (classe de taille 5, 500 employés ou plus), pour lesquelles l'une ou l'autre loi est acceptée au seuil de 5%, est nettement plus faible que celle des classes de taille inférieures.

TAB. 7: Nombre de strates non vides par classe de taille

classe de taille	deux lois acceptées	seult. loi lognormale	seult. loi normale	aucune loi acceptée
1	40	20	6	38
2	40	38	13	40
3	39	37	10	41
4	33	35	7	38
5	18	27	4	37

La table 7 montre la répartition des strates contenant au moins une entreprise dans la classe de taille et le sous-groupe de résultat des tests défini ci-dessus. La valeur moyenne de la probabilité significative par strate et sous-groupe est représentée à la fig 1. Les sous-groupes sont délimités par les lignes horizontales et verticales. Une strate de l'échantillon LSE98 peut donc être représentée par 4 points, si elle contient des entreprises dans chacun des 4 sous-groupes. Le graphique montre une majorité de points au-dessus de la diagonale principale, indiquant une adéquation meilleure (ou moins mauvaise) de la répartition des salaires à la loi lognormale relativement à la loi normale.

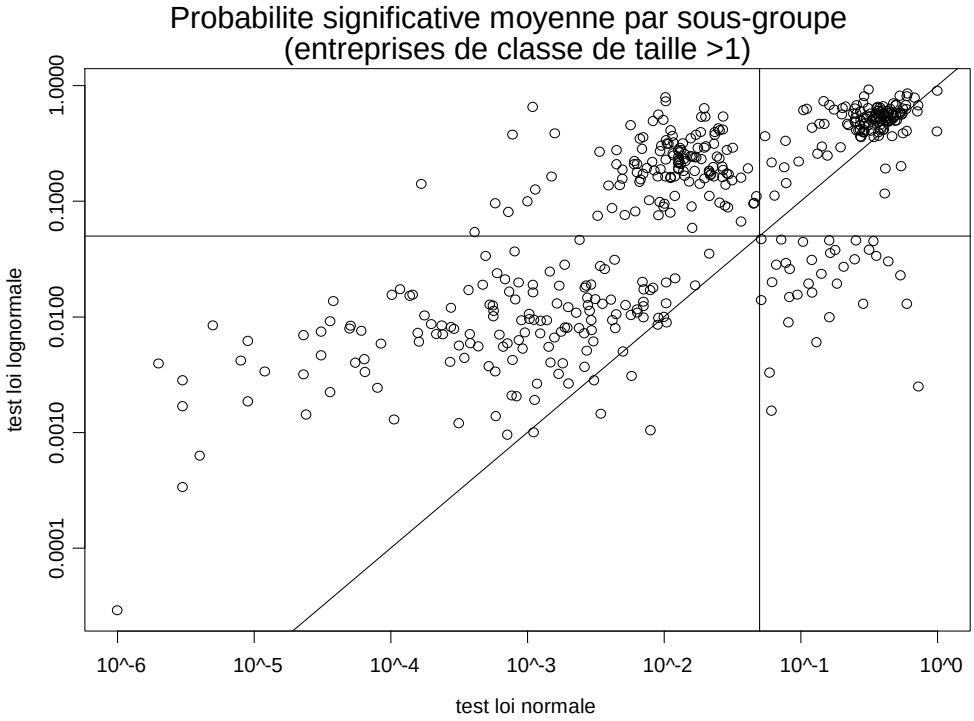


FIG. 1: Strates partitionnées en 4 sous-groupes d'entreprises, selon les résultats des tests (voir texte)

La méthode d'établissement du plan d'échantillonnage a été reprise, avec la modification suivante : la moyenne et la variance du salaire intra-entreprise ont été calculées sur la base des estimateurs du maximum de vraisemblance pour la loi lognormale. Si pour une entreprise, $\overline{y}_{\log}$ et $s^2_{\log}$ sont respectivement la moyenne arithmétique et la variance empirique du logarithme des salaires de l'entreprise, les estimations du salaire moyen $\overline{y}$, de la variance s^2 et du coefficient de variation c sont

$$\begin{aligned} \overline{y} &= \exp(\overline{y}_{\log} + \frac{1}{2} s^2_{\log}) \\ s^2 &= \overline{y}^2 \left(\exp(s^2_{\log}) - 1\right) \\ c &= \sqrt{\exp(s^2_{\log}) - 1} \end{aligned}$$

Le calcul de la variance inter-entreprise n'a pas été modifié. Le plan d'échantillonnage résultant est très peu influencé par ces modifications. La table 8 montre les strates pour lesquelles l'effectif varie de plus de 10 entreprises. La colonne n_{LSE98} indique le nombre total d'entreprises de la strate. La somme des effectifs des 4 groupes (deux lois acceptées,..., aucune) est égale au nombre d'entreprises de la strate ayant plus d'un employé, et peut donc être inférieure à n_{LSE98}. On voit

qu'à une exception-près, il s'agit de strates de classe de taille 1, pour lesquelles la modification de l'estimation par la loi lognormale a peu de raison d'être.

TAB. 8: Strates pour lesquelles le nombre d'entreprises à interroger varie de plus de 10

NOGA2	classe de taille	deux lois acceptées	seult. loi lognormale	seult. loi normale	aucune	n_{LSE98}	Δn
15	1	7	0	0	6	16	-10
18	1	3	0	1	2	14	-36
23-24	1	10	3	0	0	14	-10
30-32	1	9	0	1	1	15	-10
33	1	43	1	0	4	57	-10
36-37	1	8	1	0	0	12	+34
65	2	24	13	0	9	46	+13
67	1	26	1	0	0	33	-10
70-71	1	23	0	0	5	35	-49

Le détail de la comparaison des deux plans d'échantillonnage se trouve à la table 14. Dans cette table, le suffixe 1 correspond au plan d'échantillonnage basé sur les estimations ordinaires, décrites dans [5], et le suffixe 2 au plan basé sur les estimations intra-entreprises ci-dessus.

diffcv = différence des coefficients de variation (1-2) ;
diffmean = différence dans l'estimation des salaires moyens par strate ;
diffnb = différence dans l'effectif à tirer pour l'échantillon LSE00.

En conclusion, la présente analyse montre qu'il n'y a pas lieu de modifier la méthode de calcul du plan d'échantillonnage.

2 Pondération pour le secteur privé

2.1 Comparaison du nombre de salariés déduit de la LSE00 avec d'autres statistiques

Le calage sur marges est une méthode bien établie dont le but est d'atténuer la variance et le biais d'échantillonnage. Un autre avantage est l'amélioration de la comparabilité des statistiques. En principe, il est donc souhaitable d'appliquer un calage, ceci d'autant plus que la taille de l'échantillon est petite. Mentionnons qu'un calage sur les pleins-temps du RE avait été effectué en 94, 96 et 98. Dans le cas de la LSE, la taille correspond environ au quart de tous les salariés de Suisse. Elle est donc suffisamment importante pour qu'on puisse se passer d'un calage, si aucune variable vraiment satisfaisante n'est trouvée. Nous avons envisagé deux possibilités de calage, qui sont décrites ci-dessous, et y avons finalement renoncé.

2.1.1 Calage sur le RE

Les emplois à plein temps du RE sont les emplois correspondant à un taux d'activité $\geq 90\%$. Les entreprises fournissent pour le RE une répartition des salariés en 3 classes d'activité (limites 50% et 90%). Le nombre d'emplois à plein temps (variable ANZBE) est considéré comme la mesure la plus fiable de la taille de l'entreprise.

Dans la statistique des salaires, on dispose du taux d'activité individuel correspondant aux salaires relevés (variable IBGR) et du nombre total de salariés de l'entreprise interrogée (variable UNTGROE). En extrapolant la somme des emplois relevés correspondant à IBGR 90%, on obtient la taille estimée de l'entreprise en nombre de salariés à plein temps.

Cette mesure de taille ne correspond pas tout à fait à la définition de ANZBE, car les emplois du RE regroupent les salariés, les apprentis et les propriétaires. ANZBE a donc été corrigée par soustraction du nombre de propriétaires [1], qui sont exclus de la statistique des salaires. Cependant cette variable contient encore les apprentis, qui ne sont pas non plus considérés dans l'échantillon LSE. L'écart observé à la fig. 2 (en haut) provient dans une large mesure de cette différence de définition, comme on le voit plus bas.

Le nombre d'apprentis par classe Noga a été relevé en 1998, lors du recensement des entreprises. Ces données nous ont été transmises par la section UNT[1]. Un apprenti est toujours considéré comme un travailleur à plein temps.

Négligeant le fait qu'il ne s'agit pas de la même année, nous avons soustrait le nombre d'apprentis de ANZBE. On voit le résultat à la fig. 2 (en bas). En comparaison avec le panneau du haut, l'adéquation entre l'échantillon et le registre s'est nettement améliorée. La pente de la droite des moindres carrés est de 0.98. Un test de Student bilatéral que la pente vaut 1 fournit une probabilité significative de 0.96. Le léger écart entre la pente estimée et 1 pourrait provenir

1. d'une différence entre le nombre d'apprentis en 1998 et en 2000,

2. d'un changement de la taille des entreprises depuis la dernière mise à jour de ANZBE,

3. du fait qu'on a soustrait le nombre total d'apprentis de la classe NOG_2 considérée, et pas seulement ceux appartenant à des entreprises du secteur privé (cela concerne principalement les classes Noga 60, 64, 75, 80, 85),

4. de la variabilité d'échantillonnage. Comme notre échantillon contient environ le quart de tous les salariés de Suisse, on peut considérer les valeurs extrapolées comme assez sûres.

Nous en concluons qu'il n'y a pas d'utilité évidente à caler les données sur le nombre d'emplois à plein temps du registre.

[1] fichier "Unternehmen Noga 2.xls".

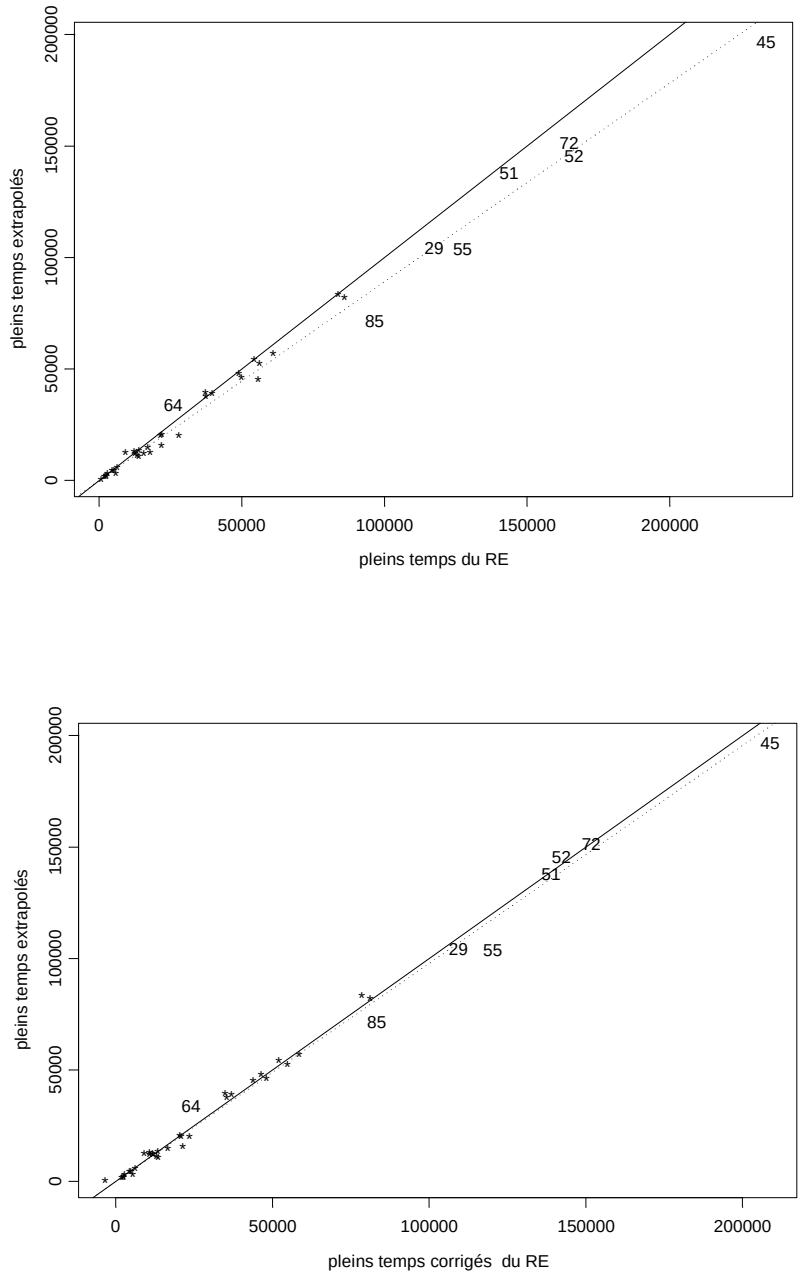
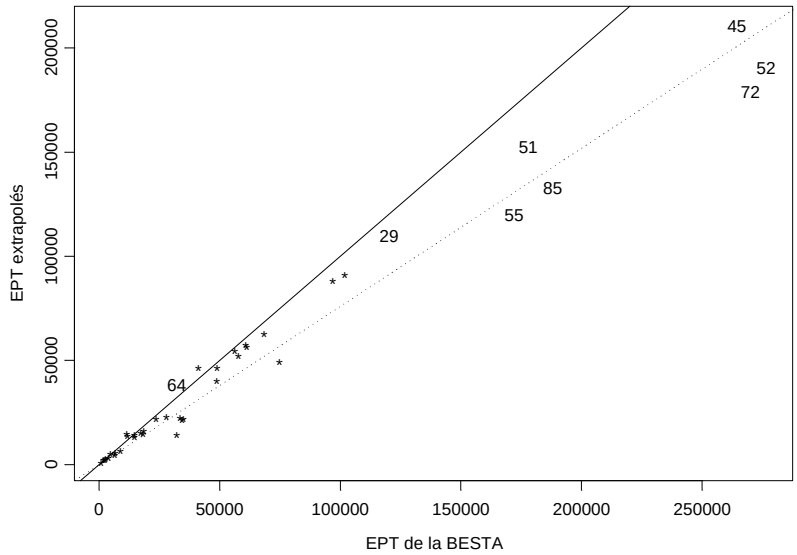


FIG. 2: Postes à plein-temps extrapolés de l'échantillon sur la base de IBGR ≥ 0.9 en fonction du nombre de plein temps du RE par classe d'activité NOG_2 (en haut). Postes à plein-temps extrapolés de l'échantillon sur la base de IBGR ≥ 0.9 en fonction du nombre de plein temps du RE par classe d'activité (NOG_2) corrigé des apprentis (en bas). Les valeurs sont les no de classes Noga, selon le regroupement LSE. La bissectrice est en trait plein et la droite des moindres carrés passant par l'origine en traitillé.



Postes en équivalent plein-temps de l'échantillon sur la base de IBGR en fonction du nombre d'équivalents plein temps de la BESTA par classe d'activité (NOG_2)

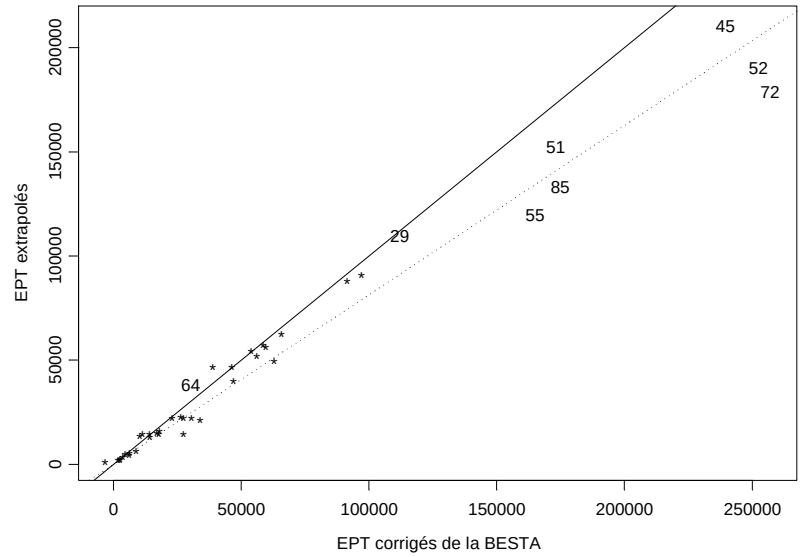


FIG. 3: Postes en équivalent plein-temps de l'échantillon sur la base de IBGR en fonction du nombre d'équivalents plein temps de la BESTA par classe d'activité (NOG_2) corrigé des apprentis.

2.1.2 Calage sur la BESTA

On peut comparer de manière analogue les équivalents plein temps (EPT) de la statistique de l'emploi (BESTA) et les données LSE. Ces EPT sont calculés par une combinaison linéaire des effectifs des 3 classes de taux d'activité, les coefficients de cette combinaison linéaire dépendant du secteur économique. Cette formule fournit donc une approximation des équivalents plein temps que l'on calculerait, si on pouvait - comme dans la LSE - sommer les taux d'activité individuels. La BESTA est trimestrielle, ce qui a l'avantage de permettre de tenir compte de la saisonnalité de l'emploi, présente dans certaines branches. Un obstacle pour la comparaison est qu'il s'agit d'une statistique des établissements, alors que la LSE s'adresse aux entreprises. Appliquant la méthode du passage du concept établissement au concept entreprise [2], et ne conservant que les emplois correspondant au secteur privé, nous obtenons un résultat que l'on peut comparer aux données LSE. Les données, fournies par la section EWL, sont reproduites au tableau 15.

Pour obtenir la taille de l'entreprise en EPT dans la LSE, nous extrapolons le total cumulé des valeurs de ibgr. Cela a l'avantage d'utiliser plus complètement l'information fournie par l'échantillon que précédemment. En effet, dans la première méthode, en relation avec les données du RE, seule la variable indicatrice des pleins temps est utilisée.

La fig.3 montre les résultats avec (en haut) et sans (en bas) les apprentis. On constate une nette différence entre la BESTA et la LSE. Cette différence peut provenir de plusieurs facteurs :

1. On n'a pas soustrait les propriétaires des effectifs donnés par la BESTA.
2. Les EPT sont calculés de deux façons différentes : le taux d'occupation est réparti en 3 classes dans la BESTA et le total est reconstitué par la formule de conversion, alors que dans la LSE, il est observé directement.
3. Les deux statistiques peuvent comporter une certaine variabilité due à l'échantillonnage. La BESTA inclut environ 4 fois plus d'entreprises que la LSE, mais il n'est pas certain qu'il y ait un réel avantage à caler la LSE sur la BESTA.

Il est difficile, sans faire une étude ad hoc, de déterminer l'importance relative des causes possibles de différence.

Dans les deux approches, le fait de ne pas connaître le nombre d'apprentis du secteur privé, et de soustraire le nombre total d'apprentis devrait théoriquement amener à une correction excessive. A la table 15 (en Annexe), on voit la différence entre les EPT de la BESTA au 4e trimestre 1998 et 2000, lorsqu'on ôte ou non les formes juridiques >20, ainsi que le nombre d'apprentis en 1998 correspondant au total des formes juridiques. La classe la plus problématique est la classe 85 : les 13'000 apprentis se rapportent à un total de 288'000, et non de 180'000. On voit pourtant à la fig. 3 que malgré cette correction théoriquement excessive, les EPT de la BESTA sont encore supérieurs à ceux de la LSE. Ce qui nous ramène aux 3 causes possibles énumérées ci-dessus. Nous renonçons donc également à un calage sur les valeurs de la BESTA, car trop d'inconnues subsistent quant à la comparabilité avec la LSE.

2.2 Pondération finale

Considérons les 3 variables de stratification NOG_2, GRCLT, REGION, ainsi que GRAN.

NOG_2 = classes Noga 2 selon le regroupement LSE.

GRAN = classes de taille fixes (2-4, 5-19, 20-49, 50-499, 500+).

GRCLT = classes de taille déterminées par p99, le 99e percentile des tailles des entreprises d'une classe NOG_2 (2-4, 5-19, 20-min(49, p99), 50-min(499, p99), min(499, p99)).

REGION = RS ("reste de la Suisse"), GE et TI. Ces deux cantons ont demandé une enquête exhaustive.

Soit INITGEW1 et INITGEW2 les poids d'extrapolation inter et intra-entreprise respectivement.

INITGEW1 = rapport du nombre d'entreprises dans la strate STRAK (combinaison de nog_2, grclt, region) donné par le RE, sur le nombre d'entreprises répondantes.

INITGEW2 = rapport du nombre de salariés déclarés (UNTGROE), sur le nombre de salaires fournis (ANZLOHN) par l'entreprise.

La pondération finale est obtenue par les étapes suivantes (seul le point 3 est modifié par rapport à 1998, voir plus haut) :

1. FINGEW2 = min(INITGEW2 , 20) : le poids initial intra-entreprise est tronqué à 20.
   La valeur de 20 est arbitraire ; nous avons repris ce qui avait été fait en 1998. On peut justifier cette mesure de la façon suivante : La proportion des salaires à fournir dépend de la taille de l'entreprise. Elle est au minimum de 1/6 pour les plus grandes. Tronquer le poids à 20 revient à accepter un taux de réponse de 30%. En dessous, l'influence de l'entreprise est limitée.
2. GEWICHT = min(INITGEW1*FINGEW2 , 60) : le poids total est tronqué à 60. Cela avait aussi été fait en 1998. La justification est qu'un poids de 60 correspond environ au 99e percentile des poids.
3. GEWICHT est calé sur la somme des poids d'extrapolation sur marges de NOG_2, GRAN, REGION, c'est-à-dire les variables utilisées dans la publication.
   Le calage est fait avec la macro SAS CALMAR, avec l'option dite "méthode linéaire". Cette option revient à ajuster un modèle linéaire (spécifié par les variables dans les marges de calage) avec contraintes. Son avantage sur le raking ratio est qu'elle converge à la première itération. Elle peut aboutir à des poids ajustés négatifs, mais ce n'est pas le cas ici. Le raking ratio converge dans notre cas en 3 itérations, mais un message d'erreur indique que "le calage ne s'est pas fait parfaitement".
4. FINGEW1 = INITGEW1 : le poids initial inter-entreprise est conservé.
5. FINGEW2 = GEWICHT/ FINGEW1.

La transformation des poids est illustrée à la fig. 4. En haut on voit les poids initiaux (INITGEW) en fonction des poids finaux (GEWICHT). Ces poids, identiques pour tous les salaires d'une même entreprise, sont obtenus par winsorisation des poids d'extrapolation intra-entreprises (INITGEW2) et calage sur les marges des poids initiaux (INITGEW) par (NOG_2, GRAN, REGION). Les poids intra-entreprises sont représentés à la fig. 4 (en bas).

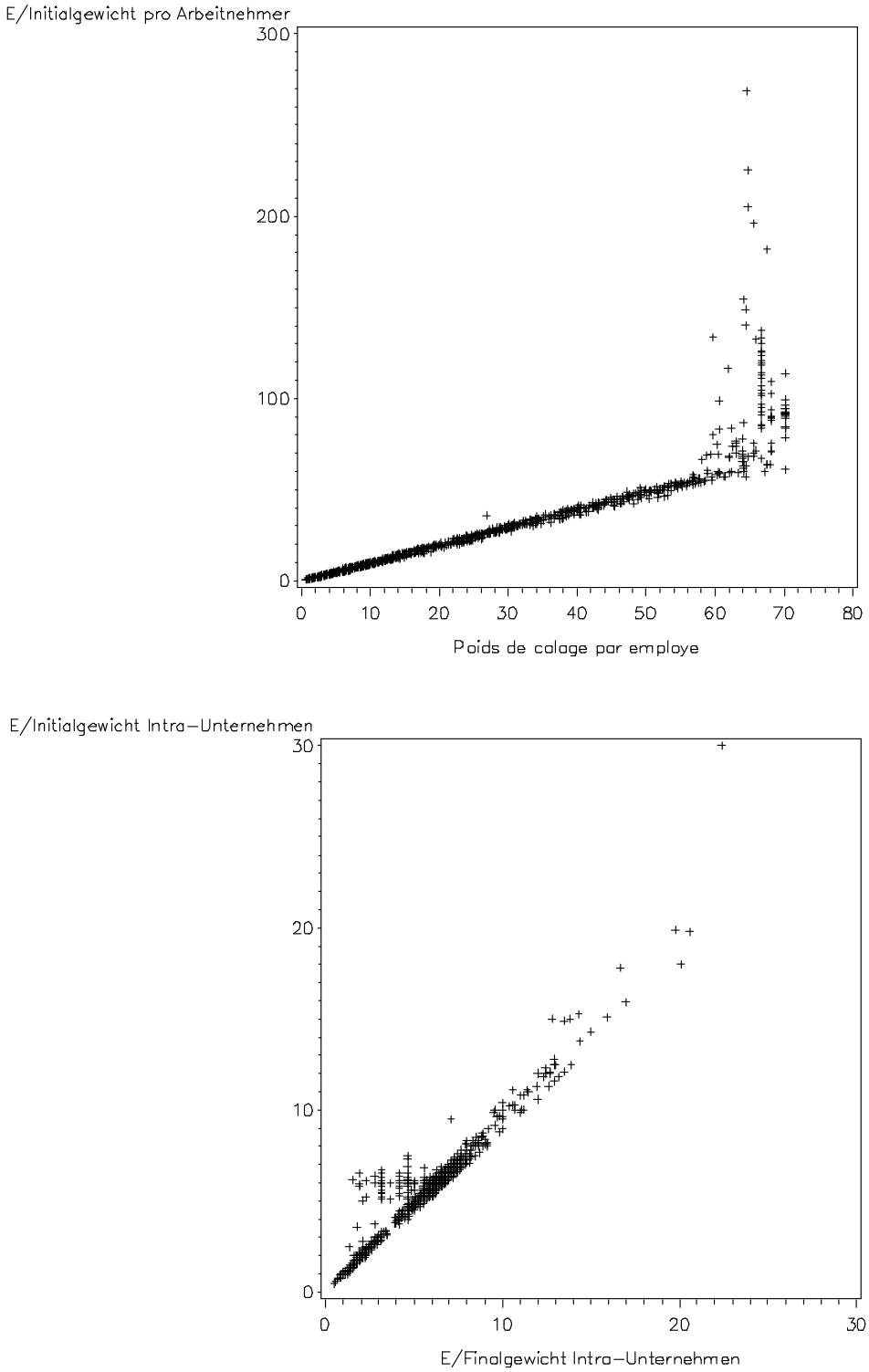


FIG. 4: Poids initiaux INITGEW en fonction des poids finaux GEWICHT (en haut). Poids initiaux intra-entreprises INITGEW2 en fonction des poids finaux FINGEW2 (en bas).

3 Estimation

3.1 Introduction

La médiane est une mesure de position bien adaptée aux répartitions asymétriques, comme des salaires. Elle est utilisée depuis 1994 [5] dans l'enquête sur la structure des salaires (LSE). De 1994 à 1998 [4], les mesures de précision pour le salaire médian ont été basées sur une approximation normale, c'est-à-dire que l'écart-type de la médiane est évalué par celui de la moyenne, multiplié par $\sqrt{\pi/2}$. L'approche non-paramérique décrite ci-après montre que, dans le cas de la LSE, cette approximation a tendance à surestimer la variabilité de la médiane. La nouvelle méthode apporte donc un réel avantage. Mentionnons qu'une approche de ce type avait été envisagée dans [4], mais non implémentée.

3.2 Approche non-paramétrique

L'estimation de la médiane peut être vue comme l'estimation d'un point particulier de la fonction de répartition de la population. Une procédure (Woodruff, 1952), bien connue dans le cas d'observations indépendantes et de même loi, consiste à obtenir un intervalle de confiance sur l'échelle des pourcentages, et à en prendre l'image inverse par la fonction de répartition (f.r.) empirique. Särndal et al. [6, 5.11] décrivent le principe de la méthode appliqué aux plans d'échantillonnages simples sans remise et stratifiés. Cette méthode repose sur un certain nombre d'hypothèses que l'on rappelle maintenant :

- La fonction de répartition de la population devrait être continue. Dans le cas d'une loi discrète, l'intervalle de confiance obtenu a un niveau de confiance supérieur au niveau de confiance nominal. Bien sûr, dans le cas de populations finies, la fonction de répartition théorique n'est jamais continue, mais si la population est suffisamment grande et pour une variable comme le salaire, l'effet devrait être négligeable.
- L'estimation de la variance est effectuée sur l'échelle des images par la f.r., c'est-à-dire des pourcentages. Dans le cas de l'inférence pour une population finie, la méthode consiste à calculer l'estimateur linéarisé de la variance d'un quotient. La distribution de la proportion des observations inférieures à la vraie médiane étant à support borné, l'approximation normale devrait donner un intervalle conservatif. En effet le coefficient d'aplatissement est négatif dans ce cas. Pour des échantillons de taille modérée, l'approximation normale devrait être suffisante. Pour d'autres quantiles, l'asymétrie peut-être plus prononcée rend la convergence plus lente.

Les deux conditions mènent à une surestimation de la longueur de l'intervalle de confiance dans le cas d'une enquête en population finie. Il se peut que la procédure soit inefficace pour de petits échantillons.

Enumérons les principaux avantages

- Cette méthode est en adéquation parfaite avec le calcul de la variance d'échantillonnage. En effet, on ne fait pas d'hypothèses de modèle sur la population. C'est donc la contrepartie pour la médiane de l'estimation de la variance basée sur le plan d'échantillonnage.

– L'utilisation de formules fermées pour l'estimation de la variance facilite la généralisation à des enquêtes complexes et de grands échantillons.

– La méthode est moins exigeante en temps de calcul que les techniques de rééchantillonnage.

Rappelons (voir Chap. 1) que les strates sont les catégories croisées de classes NOGA2, de 5 classes de taille d'entreprises et de 3 régions. Dans chaque strate h, on a un échantillonnage par grappes à deux niveaux :

Au premier niveau, on tire un échantillon aléatoire simple d'entreprises.
Au second niveau, on tire un échantillon aléatoire simple d'employés.

Notations :

N	nombre d'entreprises
N_h	nombre d'entreprises dans la strate h
U_h	univers des entreprises de la strate h
s_h	échantillon d'entreprises de la strate h
n_h	taille de s_h
$f_h = n_h/N_h$	taux de sondage dans la strate h
U_{hi}	univers des emplois de l'entreprise i de la strate h
M_{hi}	nombre d'emplois dans U_{hi}
s_{hi}	échantillon d'emplois dans l'entreprise hi
m_{hi}	taille de s_{hi}
$f_{hi} = m_{hi}/M_{hi}$	taux de sondage dans l'entreprise hi
b_j	taux d'occupation standardisé de l'emploi j de s_{hi}
w_{hi}	poids d'extrapolation des emplois de s_{hi}
$\overline{b}_{hi} = \sum_{s_{hi}} b_j/m_{hi}$	taux d'occupation standardisé moyen dans s_{hi}
$v_{hi} = w_{hi}\overline{b}_{hi}$	poids d'extrapolation pondéré par le taux d'occupation moyen
$\overline{v}_h = \sum_{s_h} v_{hi}/n_h$	poids d'extrapolation pondéré moyen dans s_h
$\overline{b}_{U,hi} = \sum_{U_{hi}} b_j/M_{hi}$	taux d'occupation standardisé moyen dans U_{hi}

3.3 Intervalle de confiance non paramétrique

Ce paragraphe expose la méthode d'estimation de la variance par linéarisation pour un échantillon stratifié par grappes à deux niveaux appliquée au cas des fonctions de répartition (f.r.). Les propriétés de bases des f.r. (non décroissantes et comprises entre 0 et 1) dictent en partie la forme de la pondération. La f.r. à un niveau supérieur d'agrégation est ainsi toujours calculée par le quotient combiné. On commence par présenter le calcul au niveau de la grappe, puis de la strate, puis on en déduit l'intervalle de confiance de la médiane d'une strate. On passe ensuite au niveau d'un agrégat de strates (par exemple la population tout entière), et enfin au calcul de l'intervalle de confiance pour un domaine quelconque.

3.3.1 Fonction de répartition des salaires dans une grappe

Soit y_{hij} le j-ème salaire communiqué par l'entreprise $hi, j = 1, ..., m_{hi}$ et y une valeur de référence quelconque. Posons

$$z_{hij} = \begin{cases} 1 & \text{si } y_{hij} \leq y \\ 0 & \text{sinon} \end{cases}$$

Soit $F_{hi}(.)$ la fonction de répartition des salaires dans U_{hi} et $\widehat{F}_{hi}(.)$ la fonction de répartition empirique dans s_{hi}, pondérée par b_j. Notons par $\overline{b}_{U,hi}$ le taux d'occupation standardisé moyen des emplois de l'entreprise i de la strate h. Alors, pour un salaire Y tiré au hasard dans U_{hi}

$$F_{hi}(y) = \mathbf{P}(Y \leq y) = \frac{1}{M_{hi}} \sum_{j \in U_{hi}} \frac{b_j}{\overline{b}_{U,hi}} z_{hij} \tag{1}$$

$$\widehat{F}_{hi}(y) = \frac{1}{m_{hi}} \sum_{j \in s_{hi}} \frac{b_j}{\overline{b}_{hi}} z_{hij} = \overline{z}_{hi} \tag{2}$$

$\widehat{F}_{hi}(y)$ est en fait le quotient de deux quantités aléatoires ; on peut récrire (2) sous la forme :

$$\widehat{F}_{hi}(y) = \frac{\sum_{j \in s_{hi}} b_j \; z_{hij}}{\sum_{j \in s_{hi}} b_j}$$

Ainsi $\widehat{F}_{hi}(y)$ estime $F_{hi}(y)$ approximativement sans biais[2] :

$$\mathbf{E}(\widehat{F}_{hi}(y)) \simeq F_{hi}(y) \tag{3}$$

Une approximation de sa variance est donnée par la méthode de linéarisation [6, formule (5.6.6)] appliquée à l'échantillonnage simple sans remise avec $\widehat{R} = \widehat{F}_{hi}(y)$:

$$\begin{aligned} \mathbf{AV}(\widehat{F}_{hi}(y)) &= \frac{1}{\overline{b}_{U,hi}^2} \frac{1 - f_{hi}}{m_{hi}} \frac{1}{M_{hi} - 1} \sum_{j \in U_{hi}} \left(b_j \; z_{hij} - F_{hi}(y) b_j\right)^2 \\ &= \frac{1 - f_{hi}}{m_{hi}} \frac{1}{M_{hi} - 1} \sum_{j \in U_{hi}} \frac{b_j^2}{\overline{b}_{U,hi}^2} \left(z_{hij} - F_{hi}(y)\right)^2 \end{aligned} \tag{4}$$

Posons

$$S_{hi}^2 = \frac{1}{m_{hi} - 1} \sum_{j \in s_{hi}} \frac{b_j^2}{\overline{b}_{hi}^2} \left(z_{hij} - \widehat{F}_{hi}(y)\right)^2 \tag{5}$$

[2] Le biais peut être évalué par un développement de Taylor.

Remarque Si les poids b_j étaient constants, on aurait simplement

$$S_{hi}^2 = \frac{m_{hi}}{(m_{hi}-1)} \widehat{F}_{hi}(y) \left(1 - \widehat{F}_{hi}(y)\right)$$

On estime la variance approchée (4) par

$$\widehat{\mathbf{V}}(\widehat{F}_{hi}(y)) = \frac{1-f_{hi}}{m_{hi}} S_{hi}^2 \tag{6}$$

3.3.2 Fonction de répartition des salaires dans une strate

Au niveau de la strate, la fonction de répartition des salaires est définie par

$$F_h(y) = \frac{\sum_{i \in U_h} M_{hi} \overline{b}_{U,hi} F_{hi}(y)}{\sum_{i \in U_h} M_{hi} \overline{b}_{U,hi}}$$

Posons $v_{hi} = w_{hi} m_{hi} \overline{b}_{hi}$ et $\overline{v}_h$ la moyenne sur s_h des v_{hi}. Rappelons que w_{hi} est le poids d'extrapolation M_{hi}/m_{hi} "lissé" au niveau de la strate[3]. Ainsi le poids v_{hi} de l'entreprise i de la strate h est le total des taux d'occupation extrapolé à tous les employés. C'est une mesure de la taille de l'entreprise.

La f.r. estimée de la strate h est donnée par

$$\widehat{F}_h(y) = \frac{\sum_{i \in s_h} v_{hi} \widehat{F}_{hi}(y)}{\sum_{i \in s_h} v_{hi}} = \frac{1}{n_h} \sum_{i \in s_h} \frac{v_{hi}}{\overline{v}_h} \widehat{F}_{hi}(y) \tag{7}$$

Comme moyenne pondérée des $\widehat{F}_{hi}(y)$, elle vérifie les conditions d'une fonction de répartition. Le membre du centre de (7) montre que $\widehat{F}_h(y)$ est un quotient de 2 quantités aléatoires. La variance approchée par la linéarisation du quotient est donnée par

$$\mathbf{AV}\left(\widehat{F}_h(y)\right) = \frac{1}{\left(\sum_{i \in U_h} M_{hi} \overline{b}_{U,hi}\right)^2} \mathbf{V}\left(\sum_{i \in s_h} v_{hi} \left(\widehat{F}_{hi}(y) - F_h(y)\right)\right) \tag{8}$$

La formule [6, p. 180] donne une estimation de cette variance pour un plan d'échantillonnage en grappes à deux niveaux. Ici, cette expression fournira une approximation de la variance de la f.r.

[3] Techniquement, $w_{hi} =$ FINGEW2

dans une strate. Posons donc dans cette formule $\breve{e}_j$ pour la valeur extrapolée et pondérée des écarts $e_j = z_{hij} - \widehat{F}_h(y)$:

$$\begin{aligned}
\breve{e}_j &= w_{hi} b_j \left(z_{hij} - \widehat{F}_h(y) \right) \\
\overline{\breve{e}}_{hi} &= \frac{1}{m_{hi}} \sum_{j \in s_{hi}} \breve{e}_j = w_{hi} \overline{b}_{hi} \frac{1}{m_{hi}} \sum_{j \in s_{hi}} \frac{b_j}{\overline{b}_{hi}} \left(z_{hij} - \widehat{F}_h(y) \right) = w_{hi} \overline{b}_{hi} \left(\widehat{F}_{hi}(y) - \widehat{F}_h(y) \right) \\
\sum_{j \in s_{hi}} \breve{e}_j &= m_{hi} \overline{\breve{e}}_{hi} = v_{hi} \left(\widehat{F}_{hi}(y) - \widehat{F}_h(y) \right) \\
\sum_{i \in s_h} m_{hi} \overline{\breve{e}}_{hi} &= \sum_{i \in s_h} v_{hi} \left(\widehat{F}_{hi}(y) - \widehat{F}_h(y) \right) = 0
\end{aligned}$$

Soit A_h la variance empirique des totaux $m_{hi} \overline{\breve{e}}_{hi}$ dans s_h, et B_{hi} la variance empirique des écarts extrapolés $\breve{e}_j$ dans s_{hi}.

$$\begin{aligned}
A_h &= \frac{1}{n_h - 1} \sum_{i \in s_h} m_{hi}^2 \overline{\breve{e}}_{hi}^2 = \frac{1}{n_h - 1} \sum_{i \in s_h} v_{hi}^2 \left(\widehat{F}_{hi}(y) - \widehat{F}_h(y) \right)^2 \qquad (9) \\
B_{hi} &= \frac{1}{m_{hi} - 1} \sum_{j \in s_{hi}} \left(\breve{e}_j - \overline{\breve{e}}_{hi} \right)^2 = \frac{1}{m_{hi} - 1} \left(\sum_{j \in s_{hi}} \breve{e}_j^2 - m_{hi} \overline{\breve{e}}_{hi}^2 \right) \\
&= \frac{1}{m_{hi} - 1} w_{hi}^2 \overline{b}_{hi}^2 \left\{ \sum_{j \in s_{hi}} \frac{b_j^2}{\overline{b}_{hi}^2} \left(z_{hij} - \widehat{F}_h(y) \right)^2 - m_{hi} \left(\widehat{F}_{hi}(y) - \widehat{F}_h(y) \right)^2 \right\} \\
&= \frac{1}{m_{hi} - 1} \frac{v_{hi}^2}{m_{hi}^2} \left\{ \sum_{j \in s_{hi}} \frac{b_j^2}{\overline{b}_{hi}^2} \left(z_{hij} - \widehat{F}_h(y) \right)^2 - m_{hi} \left(\widehat{F}_{hi}(y) - \widehat{F}_h(y) \right)^2 \right\} \\
B_h &= \sum_{i \in s_h} m_{hi} \left(1 - f_{hi} \right) B_{hi} \qquad (10)
\end{aligned}$$

La variance (8) est estimée par

$$\begin{aligned}
\widehat{\mathbf{V}}_{2\text{st},h} \left(\widehat{F}_h(y) \right) &= \left(\frac{1}{n_h \overline{v}_h} \right)^2 \left\{ n_h \left(1 - f_h \right) A_h + f_h B_h \right\} \\
&= \left(\frac{1}{\overline{v}_h} \right)^2 \frac{1}{n_h} \left\{ \left(1 - f_h \right) A_h + \frac{f_h}{n_h} B_h \right\} \qquad (11)
\end{aligned}$$

Remarque Si les poids b_j étaient constants sur s_{hi}, l'expression B_{hi} ne dépendrait pas de $\widehat{F}_h(y)$et serait égale à $w_{hi}^2 \overline{b}_{hi}^2 S_{hi}^2$, où S_{hi}^2 est donnée par la remarque suivant (5).

Après substitution et simplification, on obtient

$$\begin{aligned} & \widehat{\mathbf{V}}_{\text{2st},h}\left(\widehat{F}_h(y)\right) \qquad (12) \\ = \; & \frac{1}{n_h}\frac{1-f_h}{n_h-1}\sum_{i\in s_h}\frac{v_{hi}^2}{\overline{v}_h^2}\left(\widehat{F}_{hi}(y)-\widehat{F}_h(y)\right)^2 \\ & +\frac{f_h}{n_h^2}\sum_{i\in s_h}\frac{v_{hi}^2}{\overline{v}_h^2}\frac{1}{m_{hi}}\frac{1-f_{hi}}{(m_{hi}-1)}\left\{\sum_{j\in s_{hi}}\frac{b_j^2}{\overline{b}_{hi}^2}\left(z_{hij}-\widehat{F}_h(y)\right)^2 - m_{hi}\left(\widehat{F}_{hi}(y)-\widehat{F}_h(y)\right)^2\right\} \end{aligned}$$

Cette expression donne la variance estimée de l'estimateur $\widehat{F}_h(y)$ dans la strate h. Si la strate h est exhaustive au niveau des entreprises ($f_h = 1$) comme c'est le cas pour les grandes tailles, la variance est donnée par la deuxième ligne. Si les entreprises de la strate fournissent tous les salaires ($f_{hi} = 1$) comme c'est le cas pour les petites tailles, la variance est donnée par la première ligne.

Remarque Pour la mise en oeuvre informatique on s'en tient à la forme (11), avec A_h sous la première forme indiquée dans (9) et B_h donné par (10).

3.3.3 Intervalle de confiance du salaire médian d'une strate

Soit $\widetilde{y}_{U,h}$ la médiane des salaires de la population U_h, et $\widetilde{y}_h$ le salaire médian observé dans la strate h.

$$\widetilde{y}_h = \widehat{F}_h^{-1}(0.5) \qquad (13)$$

Pour la définition opérationnelle de l'inverse de la fonction de répartition empirique, voir § 4.1.

La valeur de référence $\widetilde{y}_{U,h}$ est inconnue. La variance $\widehat{\mathbf{V}}_{\text{2st},h}(\widehat{F}_h(\widetilde{y}_{U,h}))$ du quantile correspondant est alors évaluée en $\widetilde{y}_h$.

$$\widehat{\mathbf{V}}_{\text{2st},h}(\widehat{F}_h(\widetilde{y}_{U,h})) \cong \widehat{\mathbf{V}}_{\text{2st},h}(\widehat{F}_h(\widetilde{y}_h)) \qquad (14)$$

L'intervalle de confiance asymptotique à 95% sur le quantile 0.5 de $\widehat{F}_h$ est donné par l'intervalle $[c_{1,0.95}, c_{2,0.95}]$ où

$$\begin{aligned} c_{1,0.95} &= 0.5 - 1.96\left\{\widehat{\mathbf{V}}_{\text{2st},h}(\widehat{F}_h(\widetilde{y}_h))\right\}^{1/2} \\ c_{2,0.95} &= 0.5 + 1.96\left\{\widehat{\mathbf{V}}_{\text{2st},h}(\widehat{F}_h(\widetilde{y}_h))\right\}^{1/2} \end{aligned}$$

et $\widehat{\mathbf{V}}_{\text{2st}}(\widehat{F}_h(\widetilde{y}_h))$ est donné par (12) avec $y = \widetilde{y}_h$. L'intervalle de confiance du salaire médian $\widetilde{y}_{U,h}$ est donc

$$\left[\widehat{F}_h^{-1}(c_{1,0.95})\ ,\ \widehat{F}_h^{-1}(c_{2,0.95})\right] \qquad (15)$$

3.3.4 Estimation et intervalle de confiance de la médiane générale

Notons sans indices la f.r. générale des salaires au niveau suisse.

$$F(y) = \frac{\sum_h N_h \overline{v}_{U,h} F_h(y)}{\sum_h N_h \overline{v}_{U,h}}$$

où $\overline{v}_{U,h} = N_h^{-1} \sum_{i \in U_h} M_{hi} \overline{b}_{U,hi}$ représente la taille moyenne des entreprises de U_h en équivalents plein-temps.

La f.r. empirique est donnée par la moyenne pondérée des fonctions de répartition empiriques $\widehat{F}_h(y)$ (7) :

$$\widehat{F}(y) = \frac{\sum_h N_h \overline{v}_h \widehat{F}_h(y)}{\sum_h N_h \overline{v}_h} = \sum_h p_h \widehat{F}_h(y) \tag{16}$$

où $\overline{v}_h = n_h^{-1} \sum_{i \in s_h} w_{hi} m_{hi} \overline{b}_{hi}$ est la taille moyenne extrapolée à toutes les entreprises de s_h[4].

La variance approchée est donnée par la méthode de linéarisation

$$\begin{aligned} \mathbf{AV}\left(\widehat{F}(y)\right) &= \frac{1}{\left(\sum_h N_h \overline{v}_{U,h}\right)^2} \mathbf{V}\left(\sum_h N_h \overline{v}_h \left(\widehat{F}_h(y) - F(y)\right)\right) \\ &= \frac{1}{\left(\sum_h N_h \overline{v}_{U,h}\right)^2} \sum_h \left(\frac{N_h}{n_h}\right)^2 \mathbf{V}\left(\sum_{i \in s_h} v_{hi} \left(\widehat{F}_h(y) - F(y)\right)\right) \end{aligned} \tag{17}$$

par l'indépendance des strates.

Considérons la contribution de la strate h à la variance :

$$\begin{aligned} \mathbf{V}\left(\sum_{i \in s_h} v_{hi} \left(\widehat{F}_h(y) - F(y)\right)\right) &= \mathbf{V}\left(\sum_{i \in s_h} v_{hi} \left(\widehat{F}_{hi}(y) - F(y)\right)\right) \\ &= \mathbf{V}\left(\sum_{i \in s_h} \sum_{j \in s_{hi}} w_{hi} b_j \left(z_{hij} - F(y)\right)\right) \end{aligned} \tag{18}$$

Comparant (8) et (18), on voit que les variances estimées auront la même forme, avec $\widehat{F}(y)$ remplaçant $\widehat{F}_h(y)$.

[4]On multiplie le nombre total d'emplois estimés de s_h par le poids d'extrapolation N_h/n_h, noté FINGEW1 dans les programmes.

Posons donc

$$\begin{aligned}
\breve{e}_j &= w_{hi} b_j \left(z_{hij} - \widehat{F}(y) \right) \qquad (19)\\
\overline{\breve{e}}_{hi} &= w_{hi} \overline{b}_{hi} \frac{1}{m_{hi}} \sum_{j \in s_{hi}} \frac{b_j}{\overline{b}_{hi}} \left(z_{hij} - \widehat{F}(y) \right) = w_{hi} \overline{b}_{hi} \left(\widehat{F}_{hi}(y) - \widehat{F}(y) \right)\\
\sum_{j \in s_{hi}} \breve{e}_j &= m_{hi} \overline{\breve{e}}_{hi} = v_{hi} \left(\widehat{F}_{hi}(y) - \widehat{F}(y) \right) \qquad (20)\\
\overline{\breve{e}}_h &= \frac{1}{n_h} \sum_{i \in s_h} m_{hi} \overline{\breve{e}}_{hi} = \frac{1}{n_h} \sum_{i \in s_h} v_{hi} \left(\widehat{F}_{hi}(y) - \widehat{F}(y) \right) = \overline{v}_h \left(\widehat{F}_h(y) - \widehat{F}(y) \right)
\end{aligned}$$

La seule différence avec les expressions similaires de la variance estimée de $\widehat{F}_h(y)$ dans une strate est que maintenant $\overline{\breve{e}}_h \neq 0$. Remarquons qu'on sépare les contributions inter et intra-strates en écrivant

$$\begin{aligned}
\breve{e}_j &= w_{hi} b_j \left(z_{hij} - \widehat{F}_h(y) \right) + w_{hi} b_j \left(\widehat{F}_h(y) - \widehat{F}(y) \right)\\
\overline{\breve{e}}_{hi} &= w_{hi} \overline{b}_{hi} \left(\widehat{F}_{hi}(y) - \widehat{F}_h(y) \right) + w_{hi} \overline{b}_{hi} \left(\widehat{F}_h(y) - \widehat{F}(y) \right)
\end{aligned}$$

La variance estimée de $\widehat{F}(y)$ dans la strate h est ainsi donnée par

$$\widehat{\mathbf{V}}_{\text{2st},h} \left(\sum_{i \in s_h} v_{hi} \left(\widehat{F}_h(y) - \widehat{F}(y) \right) \right) = n_h (1 - f_h) A_h + f_h \sum_{i \in s_h} m_{hi} \left(1 - f_{hi} \right) B_{hi} \qquad (21)$$

où A_h et B_{hi} sont les variances empiriques dans s_h et s_{hi} respectivement :

$$\begin{aligned}
A_h &= \frac{1}{n_h - 1} \left(\sum_{i \in s_h} m_{hi}^2 \overline{\breve{e}}_{hi}^2 - n_h \overline{\breve{e}}_h^2 \right) \qquad (22)\\
B_{hi} &= \frac{1}{m_{hi} - 1} \sum_{j \in s_{hi}} \left(\breve{e}_j - \overline{\breve{e}}_{hi} \right)^2 \qquad (23)\\
B_h &= \sum_{i \in s_h} m_{hi} \left(1 - f_{hi} \right) B_{hi}
\end{aligned}$$

On en déduit la variance estimée globale :

$$
\begin{aligned}
& \widehat{\mathbf{V}}_{\text{str,2st}}(\widehat{F}(y)) \\
= & \frac{1}{\left(\sum_h N_h \overline{v}_h\right)^2} \sum_h \left(\frac{N_h}{n_h}\right)^2 \widehat{\mathbf{V}}_{\text{2st},h} \left(\sum_{i \in s_h} v_{hi} \left(\widehat{F}_h(y) - \widehat{F}(y)\right)\right) \qquad (24) \\
= & \frac{1}{\left(\sum_h N_h \overline{v}_h\right)^2} \sum_h \left(\frac{N_h \sum_{s_h} v_{hi}}{n_h}\right)^2 \frac{1}{\left(\sum_{s_h} v_{hi}\right)^2} \widehat{\mathbf{V}}_{\text{2st},h} \left(\sum_{i \in s_h} v_{hi} \left(\widehat{F}_h(y) - \widehat{F}(y)\right)\right) \\
= & \sum_h (p_h)^2 \frac{1}{\left(n_h \overline{v}_h\right)^2} \widehat{\mathbf{V}}_{\text{2st},h} \left(\sum_{i \in s_h} v_{hi} \left(\widehat{F}_h(y) - \widehat{F}(y)\right)\right) \\
= & \sum_h (p_h)^2 \frac{1}{\left(n_h \overline{v}_h\right)^2} \left(n_h(1-f_h)A_h + f_h B_h\right) \qquad (25)
\end{aligned}
$$

Cette expression, avec A_h et B_{hi} donnés respectivement par (22) et (23), représente la variance estimée de l'estimateur de la fonction de répartition évaluée en y. La linéarisation de l'estimateur du quotient en fait une expression additive dans les strates.

La méthode pour trouver l'intervalle de confiance de la médiane $\widetilde{y}_U$ est la même que précédemment.

Soit $\widetilde{y}$ la médiane selon la distribution $\widehat{F}(.)$. Alors bien sûr $\widehat{F}(\widetilde{y}) = 0.5$ et les bornes de l'intervalle sont :

$$
\begin{aligned}
c_{1,0.95} &= 0.5 - 1.96 \left\{\widehat{\mathbf{V}}_{\text{str,2st}}(\widehat{F}(\widetilde{y}))\right\}^{1/2} \\
c_{2,0.95} &= 0.5 + 1.96 \left\{\widehat{\mathbf{V}}_{\text{str,2st}}(\widehat{F}(\widetilde{y}))\right\}^{1/2}
\end{aligned}
$$

où $\widehat{\mathbf{V}}_{\text{str,2st}}(\widehat{F}(\widetilde{y}))$ est donné par (25), avec $y = \widetilde{y}$. L'intervalle de confiance sur le salaire médian $\widetilde{y}_U$ est l'image inverse des deux bornes ci-dessus par $\widehat{F}$:

$$
\left[\widehat{F}^{-1}(c_{1,0.95}) \ , \ \widehat{F}^{-1}(c_{2,0.95})\right] \qquad (26)
$$

3.3.5 Coefficient de variation synthétique

La méthode non-paramétrique de calcul d'un intervalle de confiance n'étant pas basée sur l'évaluation de l'écart-type, on ne peut donc pas en déduire un coefficient de variation au sens strict. De plus comme cet intervalle de confiance n'est pas nécessairement symétrique autour de la médiane, il faut spécifier les deux bornes de l'intervalle pour le définir complètement. Cela peut être un peu lourd. Il peut parfois être utile de disposer d'une mesure synthétique pour comparer l'ordre de grandeur de la précision de cet estimateur avec d'autres. Nous proposons donc un coefficient de

variation (CV) synthétique, basé sur l'intervalle de confiance à 95% :

$$CV_{syn95} = \frac{\max\left\{\widehat{F}^{-1}(c_{2,0.95}) - \widetilde{y}\ ,\ \widetilde{y} - \widehat{F}^{-1}(c_{1,0.95})\right\}}{1.96\ \widetilde{y}} \tag{27}$$

On pourrait aussi définir un CV à partir de l'intervalle $[c'_1, c'_2] = 0.5 \pm \sqrt{\widehat{\mathbf{V}}_{\text{str,2st}}(\widehat{F}(\widetilde{y}))}$:

$$CV_{syn} = \frac{\max\left\{\widehat{F}^{-1}(c'_2) - \widetilde{y}\ ,\ \widetilde{y} - \widehat{F}^{-1}(c'_1)\right\}}{\widetilde{y}}$$

Ces deux CV sont comparés pour la Suisse au niveau NOGA2 au chapitre 4.

3.3.6 Estimation pour un agrégat de strates

Soit $\mathcal{H}$ un agrégat de strates. Par exemple, une combinaison d'une classe NOGA2 et d'une région (RS, GE, TI) est formée de 5 strates correspondant aux 5 classes de taille (pour autant qu'elles ne soient pas vides).

La fonction de répartition $F_{\mathcal{H}}(y)$ est définie par une moyenne pondérée des f.r. $F_h(y)$ des strates concernées.

$$\widehat{F}_{\mathcal{H}}(y) = \frac{\sum_{k\in\mathcal{H}} N_h \overline{v}_h \widehat{F}_h(y)}{\sum_{k\in\mathcal{H}} N_h \overline{v}_h} = \sum_{k\in\mathcal{H}} p_{\mathcal{H},h} \widehat{F}_h(y) \tag{28}$$

Puisque $\overline{v}_h$ dépend des unités tirées, $\widehat{F}_{\mathcal{H}}(y)$ est à nouveau un quotient de variables aléatoires.

L'estimateur de la variance linéarisée fournit

$$\widehat{\mathbf{V}}_{\text{str,2st}}\left(\widehat{F}_{\mathcal{H}}(y)\right) = \sum_{k\in\mathcal{H}} p^2_{\mathcal{H},h} \widehat{\mathbf{V}}_{\text{2st},h}\left(\widehat{F}_{\mathcal{H}}(y)\right) \tag{29}$$

où $\widehat{\mathbf{V}}_{\text{2st},h}\left(\widehat{F}_{\mathcal{H}}(y)\right)$ est obtenue en substituant $\widehat{F}_{\mathcal{H}}(y)$ à $\widehat{F}(y)$ dans (25). L'intervalle de confiance ou le coefficient de variation synthétique se calculent de manière analogue.

3.3.7 Estimation dans un domaine

La méthode adoptée est la plus simple. Elle est décrite dans [6, paragraphe 10.3]. Nous l'adaptons ici au contexte des fonctions de répartition.

Supposons que l'on veuille estimer le salaire médian d'une partie U_d de la population, par exemple de la sous-population des emplois correspondant à un certain niveau de qualification.

Soit $s_{d,hi} \subset s_{hi}$ le sous-échantillon ainsi défini. Posons

$$\begin{aligned}
I_{d,hij} &= \begin{cases} 1 & \text{si } j \in s_{d,hi} \\ 0 & \text{si } j \in s_{hi} - s_{d,hi} \end{cases} \\
m_{d,hi} &= \sum\nolimits_{s_{hi}} I_{d,hij} \\
b_{d,j} &= I_{d,hij}\ b_j \\
\overline{b}_{d,hi} &= \frac{1}{m_{hi}} \sum\nolimits_{s_{hi}} b_{d,j} \\
v_{d,hi} &= w_{hi} m_{hi} \overline{b}_{d,hi} \\
\overline{v}_{d,h} &= \frac{1}{n_h} \sum_{i \in s_h} v_{d,hi}
\end{aligned}$$

Pour un salaire Y tiré au hasard dans $U_{d,hi}$, on a les probabilités conditionnelles

$$\begin{aligned}
F_{d,hi}(y) &= \mathbf{P}(Y \leq y \mid U_{d,hi}) \\
\widehat{F}_{d,hi}(y) &= \frac{\sum_{j \in s_{hi}} b_{d,j}\ z_{hij}}{\sum_{j \in s_{hi}} b_{d,j}} \\
\widehat{F}_{d,h}(y) &= \frac{\sum_{i \in s_h} w_{hi} m_{hi} \overline{b}_{d,hi} \widehat{F}_{d,hi}(y)}{\sum_{i \in s_h} w_{hi} m_{hi} \overline{b}_{d,hi}} = \frac{\sum_{i \in s_h} v_{d,hi} \widehat{F}_{d,hi}(y)}{\sum_{i \in s_h} v_{d,hi}} \\
\widehat{F}_{d}(y) &= \frac{\sum_h N_h \overline{v}_{d,h} \widehat{F}_{d,h}(y)}{\sum_h N_h \overline{v}_{d,h}} = \sum_h p_{d,h} \widehat{F}_{d,h}(y)
\end{aligned}$$

La f.r. $\widehat{F}_{d,hi}(y)$ pouvant s'exprimer comme un quotient ordinaire sur s_{hi}, $\widehat{\mathbf{V}}_{\text{2st},h}\left(\widehat{F}_{d,h}(y)\right)$ a donc une forme analogue à (12) et l'estimation sur un agrégat de strates a une forme analogue à (29). Dans le calcul concret, on ne garde que les éléments de $s_{d,hi}$, les autres étant nuls par la multiplication par $I_{d,hij}$. Si bien que, avec $\breve{e}_j = w_{hi} b_{d,j} \left(z_{hij} - \widehat{F}_d(y)\right)$ et $\overline{\breve{e}}_{d,hi}$ la moyenne des $\breve{e}_j$ sur $s_{d,hi}$, on a $m_{d,hi}\overline{\breve{e}}_{d,hi} = m_{hi}\overline{\breve{e}}_{hi}$. De même, $\sum_{s_d} m_{d,hi}\overline{\breve{e}}_{d,hi} = n_{d,h}\overline{\breve{e}}_{d,h}$, où $n_{d,h}$ est la taille de $s_h \cap U_d = s_{d,h}$.(23) et (22) deviennent donc

$$B_{hi} = \frac{1}{m_{hi} - 1} \left\{ \sum_{j \in s_{d,hi}} \left(\breve{e}_j - \overline{\breve{e}}_{d,hi}\right)^2 + \left(1 - \frac{m_{d,hi}}{m_{hi}}\right) m_{d,hi} \overline{\breve{e}}^2_{d,hi} \right\} \tag{30}$$

$$A_h = \frac{1}{n_h - 1} \left\{ \sum_{i \in s_{d,h}} \left(m_{d,hi}\overline{\breve{e}}_{d,hi} - \overline{\breve{e}}_{d,h}\right)^2 + \left(1 - \frac{n_{d,h}}{n_h}\right) n_{d,h} \overline{\breve{e}}^2_{d,h} \right\} \tag{31}$$

Ces formules coïncident avec (23) et (22) si tous les cas sont considérés, c'est-à-dire, s'il n'y a pas de domaine. Ce sont les formules introduites dans le programme.

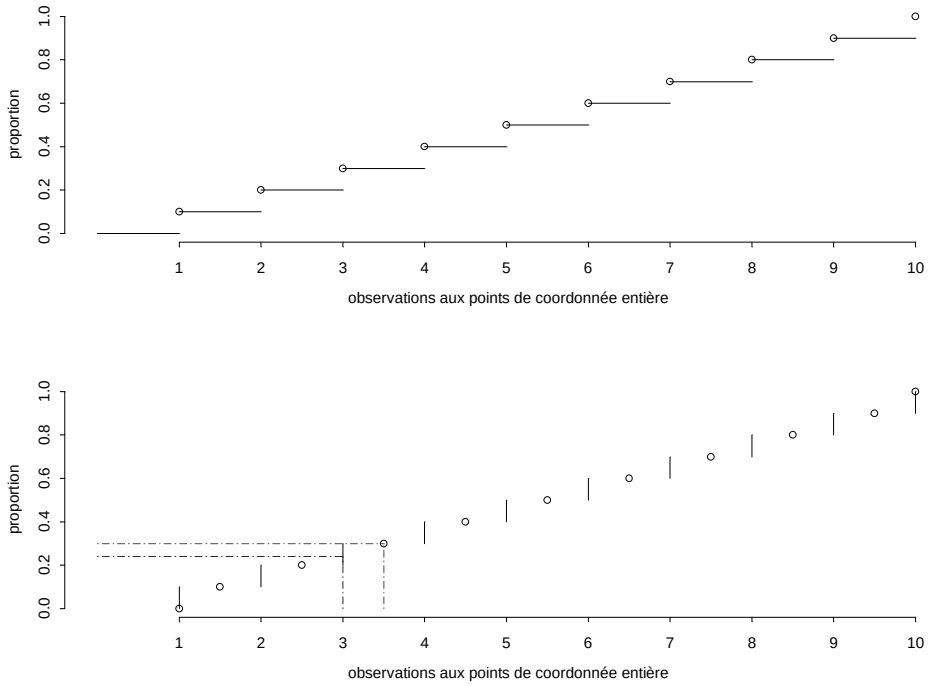


FIG. 5: Fonction de répartition : graphe de $\left(y, \widehat{F}(y)\right)$ (en haut) ; fonction de répartiton inverse : graphe de $\left(\widehat{F}^{-1}(p), p\right)$ (en bas).

4 Mise en oeuvre dans le cas de la LSE00

4.1 Inversion de la fonction de répartition empirique

La fonction de répartition empirique étant une fonction en escaliers, sa réciproque n'est pas définie de façon unique. Le choix effectué correspond à l'unique option de la procédure SAS® UNIVARIATE applicable aux quantiles pondérés. La fig. 5 en explique le principe. Supposons que l'on dispose de 10 observations 1,2,...,10. Le panneau du haut montre la fonction de répartition empirique, et celui du bas la fonction inverse. Ainsi, le percentile 24% est donné 3, alors que le percentile 30% vaut 3.5.

Plus précisément, si $y_{[1]} \leq y_{[2]} \leq ... \leq y_{[n]}$ est l'échantillon ordonné, et $w_1, w_2, ..., w_n$ sont les poids correspondants, on a

$$\widehat{F}(y_{[i]}) = \begin{cases} \frac{\sum_{j=1}^{i} w_j}{\sum_{j=1}^{n} w_j} & \text{si } y_{[i]} < y_{[i+1]} \\ \widehat{F}(y_{[i+1]}) & \text{si } y_{[i]} = y_{[i+1]} \end{cases}$$

et

$$\widehat{F}^{-1}(p) = \begin{cases} \frac{1}{2}(y_{[i]} + y_{[i+1]}) & \text{si } \widehat{F}(y_{[i]}) = p \\ y_{[i+1]} & \text{si } \widehat{F}(y_{[i]}) < p < \widehat{F}(y_{[i+1]}) \end{cases} \tag{32}$$

4.2 Non-réponse et imputation de la variance

Nous supposons qu'une non-réponse totale est ignorable, étant donné la strate.

- Au niveau de l'unité primaire, on pourrait peut-être améliorer le traitement par une analyse de la non-réponse. Le taux de sondage utilisé est le taux de sondage effectif des entreprises (f_h), c'est-à-dire le rapport du nombre d'entreprises ayant fourni au moins 1 salaire (NREPK) sur le nombre d'entreprises de la strate (BURNHK).
- Rappelons qu'une strate est une combinaison d'une classe NOGA (41 modalités), d'une classe de taille (4 ou 5 modalités), et d'une région (3 modalités). Ces strates sont au total 509 car certaines sont vides, et 31 ne contiennent qu'une seule entreprise. Parmi les strates à 2 ou 3 entreprises, 16 n'ont qu'une entreprise répondante. Dans ces 16 cas, la variance inter-entreprises n'est donc pas estimable. On l'a donc imputée par la moyenne des variances relatives dans les strates de la même région est de la même classe NOGA, plus précisément, la valeur de $A_h/\left(N_h\overline{v}_h\right)$ (22) a été imputée par la moyenne des $A_k/\left(N_k\overline{v}_k\right)$ où k parcourt l'ensemble des strates de même classe NOGA et de même région.
- Au niveau de l'unité secondaire, nous supposons une non-réponse ignorable, à l'intérieur de l'entreprise. Cette hypothèse est vérifiée, si on admet que les entreprises ont suivi les consignes de tirage au hasard. Pour le calcul de la variance, nous prenons donc comme taux de réponse intra-entreprises (f_{hi}) le rapport du nombre de salaires fournis (ANZLOHN) sur le nombre de salaires annoncés (UNTGROE).
- En 2000, 1317 entreprises sur 17'665 n'ont fourni qu'un seul salaire. Parmi ces entreprises, 19 annoncent UNTGROE$>$1 (4 fois 3 et 15 fois 2) et ont donc une variance intra entreprise (6) non estimable. Elle pourrait être imputée par $(1 - f_{hi})\,\frac{1}{4}$, qui représenterait une borne supérieure si on n'avait pas de poids différent pour les salaires d'une même entreprise. Comme il n'est pas certain que cette borne vaille encore dans ce cas, on a ignoré ces 19 entreprises dans calcul de la variance. Pour les autres, UNTGROE=1 et la variance est nulle.
- La pratique d'estimation pour la production est de supprimer de l'analyse les non-réponses partielles sur les variables de sexe (GESCHLE) et niveau d'exigeance (ANFORNI). Cela revient, en 2000, à supprimer environ 27'000 salaires sur un total de environ 500'000. Pour le calcul de la variance, nous traiterons les cas considérés pour la production comme un domaine.

4.3 Problème des entreprises supplémentaires de Genève et du Tessin

Comme nous l'avons dit, le plan d'échantillonnage n'a pas été suivi dans les cantons de GE et TI, où une enquête exhaustive a été entreprise. Pour avoir des estimations sans biais au niveau CH,

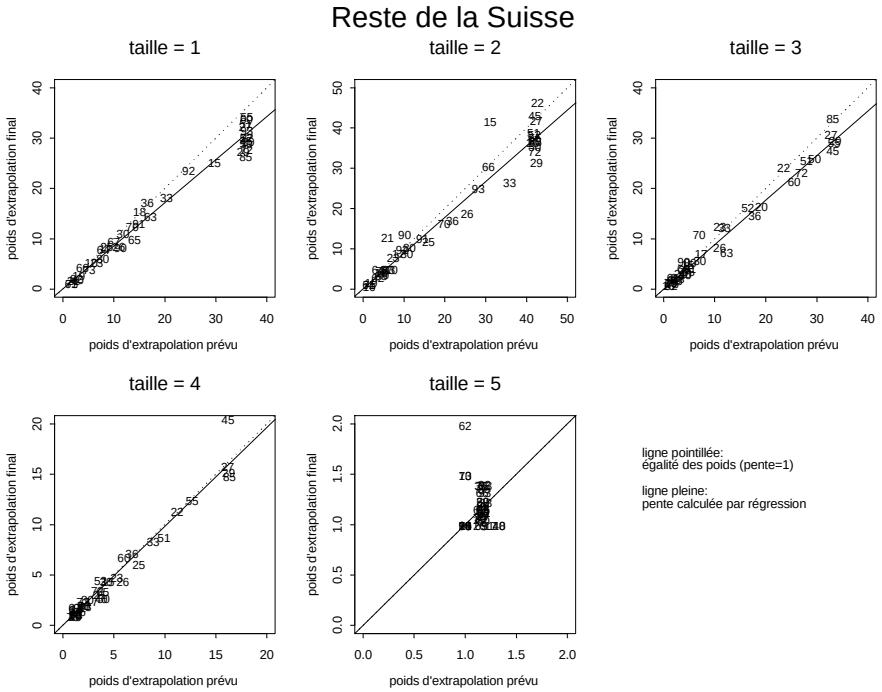


FIG. 6: Partie du pays où le plan de sondage a été suivi : Poids d'extrapolation final par strates en fonction du poids d'extrapolation initial tenant compte de la non-réponse prévue.

on a introduit une stratification a posteriori subdivisant la Suisse en 3 parties disjointes GE,TI et RS (pour "reste de la Suisse"). Cela revient à considérer que les entreprises de l'échantillon initial provenant de RS forment un échantillon aléatoire de RS. Cette hypothèse n'est pas très gênante, car RS contient environ 90% du total ; elle a l'avantage de conserver l'additivité des résultats des 3 parties au niveau CH. Une alternative serait de pondérer l'ensemble des entreprises de GE et TI, de telle sorte qu'elles aient globalement le même poids que les entreprises GE et TI de l'échantillon suisse. Cette dernière méthode permettrait de conserver les poids de l'enquête initiale pour les entreprises de RS, mais comme le plan d'échantillonnage ne contrôle pas la provenance cantonale, on introduirait un élément aléatoire dans la pondération de GE et TI, cantons pour lesquels on dispose pourtant d'une enquête exhaustive.

Nous adoptons donc la première méthode, et choisissons des poids d'extrapolation différents pour les entreprises de GE, TI et RS, reflétant les taux de sondage différents (voir chapitre 2). La fig 6 analyse les poids d'extrapolation tels qu'ils avaient été prévus initialement (effectif de la strate / effectif net de l'échantillon) pour la Suisse, et tels qu'ils ont été calculés en fonction du taux de sondage et de la non-réponse réelle de l'enquête 2000 dans la partie RS (effectif de la strate dans RS / nb d'entreprises répondantes). Il y a une très bonne adéquation générale. Les entreprises de taille 2 et 3 ont un peu moins bien répondu que prévu (écart entre la ligne pointillée correspondant à l'égalité parfaite entre les poids d'extrapolation prévus et finaux, et la ligne pleine calculée par

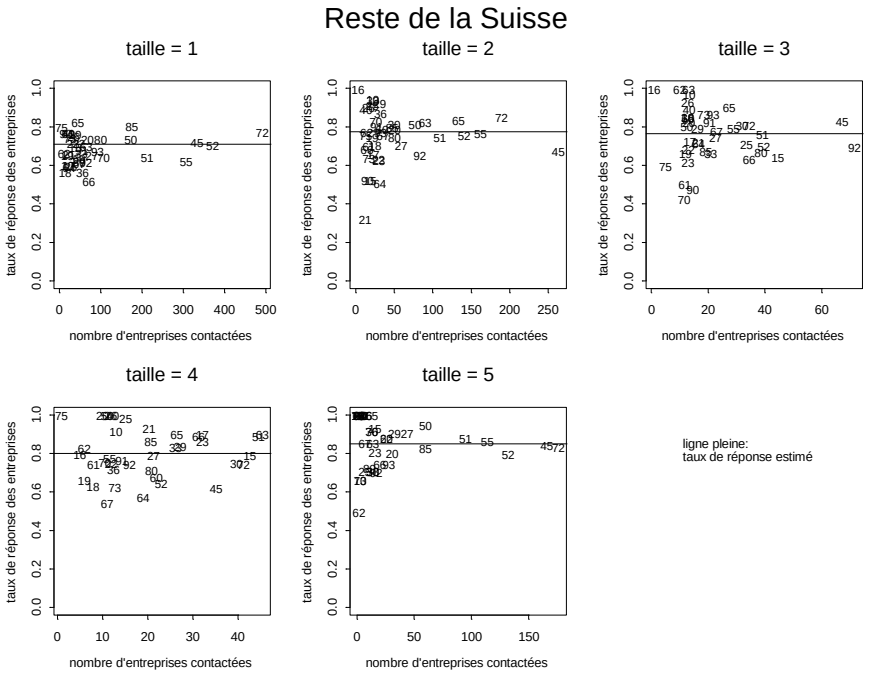


FIG. 7: Partie du pays où le plan de sondage a été suivi : Analyse du taux de réponse.

régression L1). Au niveau des plus grandes entreprises (classe de taille 5) qui forment des strates exhaustives, l'écart provient essentiellement du petit nombre d'entreprises concernées.Les fig. 7 et 8 montrent le taux de réponse des entreprises par région, en fonction du nombre d'entreprises dans la strate pour GE et TI et du nombre d'entreprises contactées estimé pour RS[5]. La ligne montre le taux de réponse estimé, calculé en minimisant la somme des valeurs absolues de l'erreur (”norme L1”). Pour les classes de taille 5 à la fig. 8, la valeur moyenne semble visuellement très haute. Cela montre que la majorité des strates a un taux de réponse égal à 1. On voit que le taux de réponse prévu (tableau 1) est prudemment en-dessous du taux moyen estimé, mais que certaines strates ont un taux inférieur, sans qu'on puisse déceler un lien avec l'une ou l'autre des variables de stratification.

[5]Ce nombre est estimé en multipliant l'effectif des entreprises donné par le registre pour la strate et pour RS par l'inverse du taux de sondage prévu dans le plan d'échantillonnage.

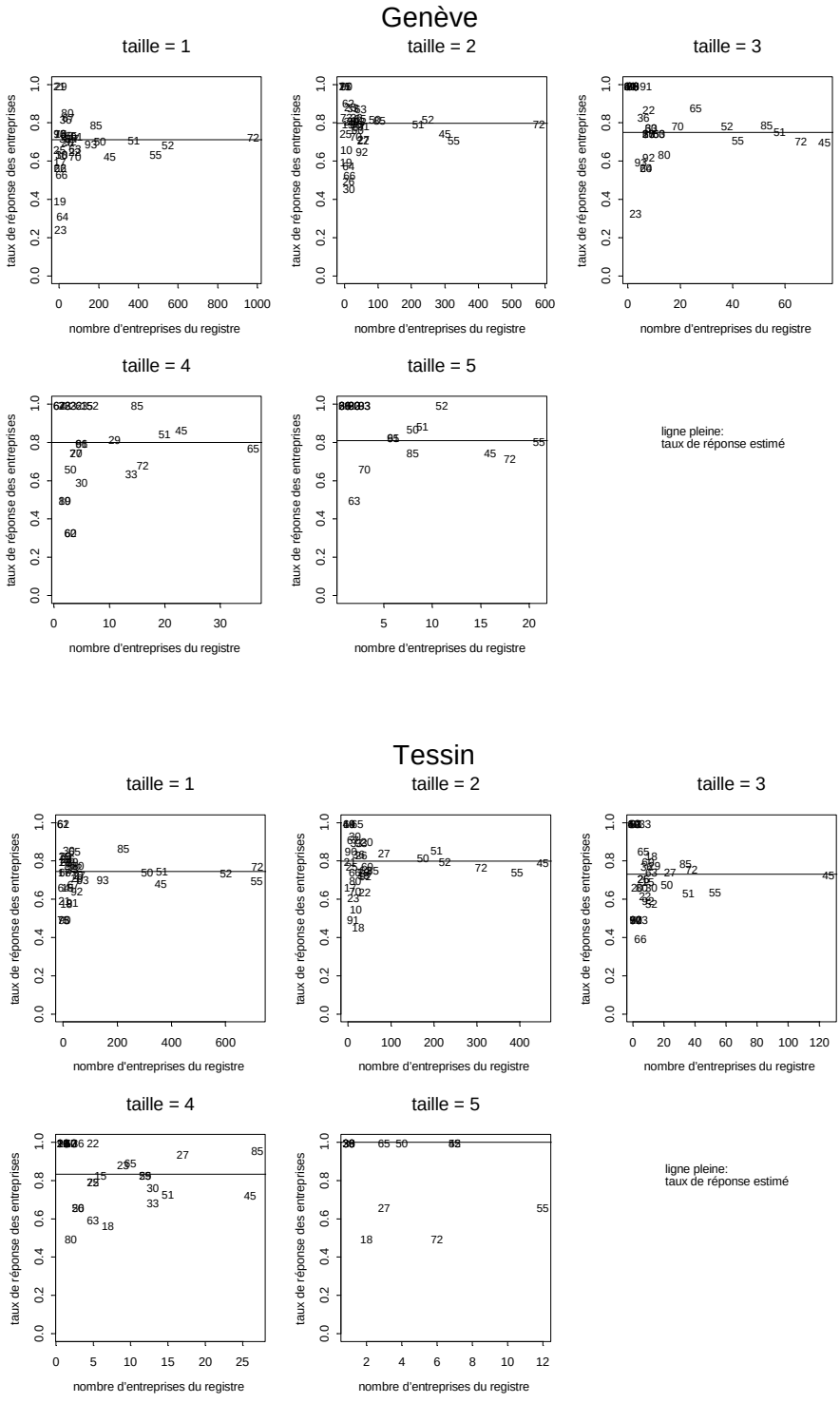


FIG. 8: Analyse du taux de réponse par strates à Genève et au Tessin.

4.4 Détermination des paramètres

- L'effectif N_h de la strate h (STRAK) est donné par BURNHK, la valeur du REE au moment du tirage.
- La taille n_h de s_h est prise comme la taille effective de l'échantillon des entreprises répondantes de la strate h.
- Le taux $f_h = n_h/N_h$ est donc le produit du taux de sondage et du taux de réponse des unités primaires.
- L'effectif M_{hi} est la variable UNTGROE extraite du questionnaire. La taille ANZBE de l'entreprise donnée dans le REE n'est pas pertinente, puisque elle représente le nombre de pleins-temps de l'entreprise, voir chap.2.
- Le poids entreprise w_{hi} est donné par le poids de calage GEWICHT[6]. Celui-ci est décomposé dans le produit du poids d'extrapolation inter-entreprises $1/f_h$ (FINGEW1) et du poids intra-entreprise lissé FINGEW2. Cette décomposition n'a d'influence ni sur l'estimation de la médiane, ni sur son intervalle de confiance.
- La taille m_{hi} de s_{hi} est donnée par le nombre de salaires ANZLOHN effectivement transmis par l'entreprise.
- Le taux $f_{hi} = m_{hi}/M_{hi}$ est donc le taux de sondage effectif dans l'entreprise.

L'un des salaires publiés est le salaire mensuel brut standardisé MBLS, c'est-à-dire le salaire mensuel brut ramené à un temps de travail de 40h/semaine et de $4\frac{1}{2}$ semaines par mois. A ce salaire est associé le taux d'occupation standardisé IBGRS, c'est-à-dire le taux d'occupation ramené à un plein-temps standard de 40h/semaine. En général, le taux d'occupation standardisé est supérieur au taux d'occupation IBGR donné relativement au plein-temps effectif de la branche d'activité. Dans la pratique, le salaire brut standardisé médian publié est la médiane de MBLS pondérée par IBGRS.

- Le poids b_j du salarié dans l'entreprise est donc IBGRS.

Nous présenterons dans ce rapport, pour illustration de la méthode, quelques estimations liées au salaire mensuel brut standardisé.

4.4.1 Estimation pour les cantons de Genève et du Tessin

La méthode est la même, ces cantons représentant des agrégats particuliers de strates.

La comparaison entre les résultats de l'enquête exhaustive et de l'échantillon original pourrait être faite grâce à l'analyse par domaines. Considérant les entreprises tessinoises (par exemple) de cet échantillon comme un domaine des entreprises suisses, on peut calculer l'intervalle de confiance du salaire médian basé sur les entreprises tirées pour le plan d'échantillonnage. La comparaison avec l'intervalle de confiance déduit de l'enquête exhaustive permettrait d'en évaluer le bénéfice.

[6] GEWICHT représente $N_h/n_h w_{hi} = f_h^{-1} w_{hi}$

4.4.2 Tableau de correspondance entre le texte et le programme

Voir les programmes en annexe.

$\frac{N_h}{n_h} w_{hi} b_j$	GEWIBGRS
m_{hi}	ANZLOHN
$\frac{N_h}{n_h} w_{hi} m_{hi} \overline{b}_{hi}$	SVHI
$\frac{N_h}{n_h} \check{e}_j$	EJ
$\frac{N_h}{n_h} \overline{\check{e}}_{hi}$	E_HI
$\left(\frac{N_h}{n_h}\right)^2 B_{hi}$	BHI
$\left(\frac{N_h}{n_h}\right)^2 \overline{\check{e}}_{hi}^2$	E_HI2
$\frac{N_h}{n_h} n_h \overline{v}_h = N_h \overline{v}_h$	SVH
$\left(\frac{N_h}{n_h}\right)^2 (1 - f_{hi})\, m_{hi} B_{hi}$	BHI -> (1-THI)*ANZLOHN*BHI
$\left(\frac{N_h}{n_h}\right)^2 \sum_{s_h} (1 - f_{hi})\, m_{hi} B_{hi}$	BH
$\left(\frac{N_h}{n_h}\right)^2 A_h$	AH
$\left(\frac{N_h}{n_h}\right)^2 \left[n_h (1 - f_h) A_h + f_h \sum_{s_h} (1 - f_{hi})\, m_{hi} B_{hi}\right]$	V2STH
$\sum_h N_h \overline{v}_h$	DENOM
$\sum_{s_{hi}} I_{d,hij}$	NDHI
$\sum_{s_h} \sum_{s_{hi}} I_{d,hij}$	TOTH
$\sum_{s_h} \max\left\{\left(\sum_{s_{hi}} I_{d,hij}\right) - 1, 0\right\}$	DLH
$\sum_h \left(\frac{N_h}{n_h}\right)^2 \left[n_h (1 - f_h) A_h + f_h \sum_{s_h} (1 - f_{hi})\, m_{hi} B_{hi}\right]$	SV2ST

4.5 Résultats

Voici quelques exemples des résultats obtenus pour MBLS, le salaire mensuel brut standardisé.

4.5.1 Exemples pour des agrégats de strates

Estimations par région La fig. 9 montre les médianes et les intervalles de confiance à 95% pour la Suisse et les deux cantons de Genève et du Tessin, toutes strates confondues (tot) et par classe d'activité. Le panneau du haut représente la production et celui du bas les services. Les valeurs manquantes correspondent à des effectifs jugés trop faibles. La limite a été fixée à 7 entreprises au moins dans l'agrégat ou le domaine, et le nb de salaires moins le nb d'entreprises doit être >54. Il y a deux raisons à cela :

1. par souci de protection des données, la pratique a été jusqu'à maintenant de masquer les résultats correspondant à moins de 6 entreprises ou moins de 60 salaires.
2. pour pouvoir utiliser les approximations normales, il ne faut pas que le nombre de "degrés de liberté" donné par le nombre d'unités moins le nombre de grappes devienne trop petit.

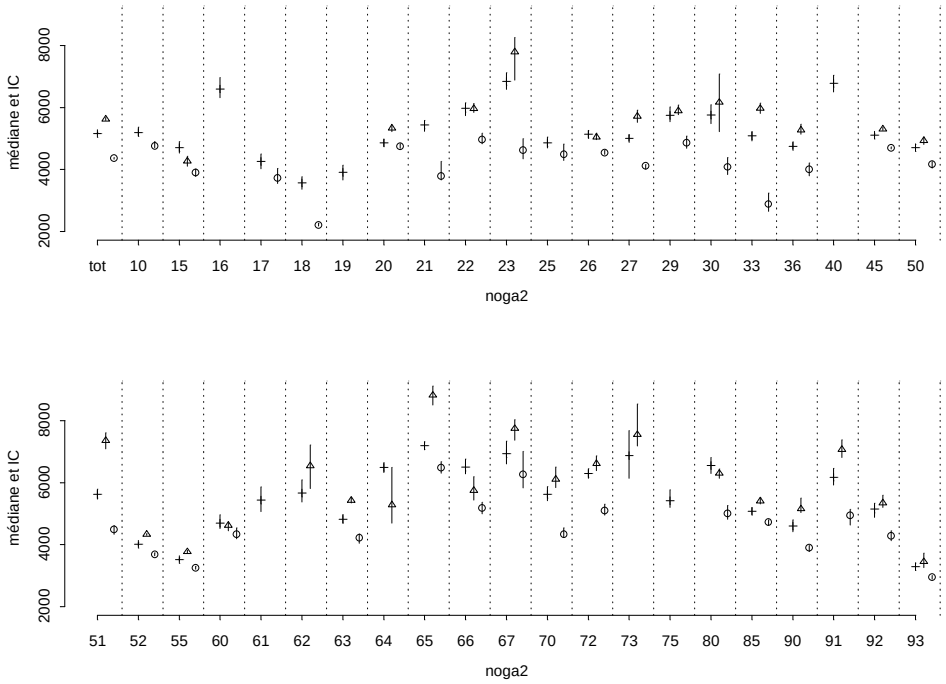


FIG. 9: Médiane et intervalles de confiance à 95% par classe NOGA2 selon les regroupements du tableau 16. + = CH Δ = GE o = TI

> Ainsi, au lieu de fixer le nombre limite de salaires, on fixe une limite à l'éparpillement des salaires.

Cela dit, cette procédure est très empirique et pourrait certainement être améliorée.

On constate à l'examen de la fig. 9 que le caractère exhaustif de l'enquête dans les deux cantons a quelque-chose d'illusoire. Remarquons tout d'abord qu'on n'a visé que l'exhaustivité des grappes, et non des unités dans les grappes. C'est-à-dire que le taux de sondage dans les entreprises est le même dans ces cantons que pour le reste de la Suisse. Certaines strates ont très peu d'entreprises au niveau du canton, si bien qu'une seule non-réponse peut avoir une très grande influence. L'incertitude sur la médiane provient donc d'une part du taux de sondage intra-entreprises qui est inférieur à 1 pour les plus grandes, et d'autre part de la non-réponse totale de certaines entreprises. Certains domaines ont un résultat masqué. C'est le cas par exemple à Genève pour les classes 18 et 19. Les résultats sont masqués à Genève (fig. 9) à cause de la 2e condition ci-dessus pour la classe 18 et à cause de la 1ère pour la classe 19, voir table 11.

Décomposition de la variance de la classe Noga 73 On constate à la fig. 9 que l'intervalle de confiance du salaire médian pour la Suisse dans la classe 73 (recherche et développement) est relativement grand. Analysons donc les composantes de la variance pour cet agrégat de strates. Les

TAB. 9: Détail du calcul de la variance pour une classe Noga

STRATE	NREPK	TH	DLH	SVH	V2STH	INTER	INTRA
7311	20	0.2381	105	395	719.69	720	0
7321	11	0.2	111	544.21	758.29	758	0
7331	16	0.69565	413	786.54	1788.27	1712	76
7341	8	0.4	355	2417.55	73020.43	71567	1453
7351	2	0.66667	378	1037.83	8226.13	7649	577
7312	6	0.75	22	33.49	12.55	12	1
7322	5	0.83333	59	62.71	9.05	9	0
7332	2	1	42	74.1	13.6	0	14
7342	2	1	28	173.28	187.06	0	187
7313	1	0.5	1	3.83	.	.	.
total				5528.54	84735.07	82427	2308

résultats sont présentés à la table 9. Les deux première positions du code de la strate correspondent à la classe d'activité, la 3e à la classe de taille et la dernière à la région (1=RS, 2=GE, 3=TI). . Les variables qui suivent sont expliquées dans le tableau de correspondance ci-dessus. INTER et INTRA correspondent à la décomposition de V2STH en deux termes, facteurs de TH et (1-TH) respectivement. SVH donne les équivalents plein-temps standardisés extrapolés pour la strate. Le total correspondant (5528.54) est la variable DENOM. On voit que la composante INTER de la strate 7341 forme le 84% de SV2ST (=84735). C'est donc cette strate qui est responsable du grand intervalle de confiance. Le taux de réponse effectif (TH) de cette strate ne vaut que 0.4. Le plan d'échantillonnage tablait sur un taux effectif de 0.5 : des 14 entreprises contactées sur les 22 existant en Suisse dans cette classe de taille, on s'attendait à 11 réponses. Il y en a eu 10, mais seulement 8 sur un total de 20 dans la strate 7341. La façon la plus efficace de diminuer la variance dans cet agrégat serait donc d'améliorer le taux de réponse effectif de la strate 7341.

Remarquons encore que la composante INTRA ne forme au total dans cet agrégat que le 3% de la variance totale.

Exemple d'analyse des variation d'une enquête à l'autres : Classes Noga 18 et 19 En comparant les estimations pour la Suisse de 1998 et 2000 (table 10), on observe une augmentation importante du salaire médian dans les classes NOGA2 18 et 19. Les estimations de 1998 [1] sont en effet en-dessous de la borne inférieure de l'intervalle de confiance calculé pour 2000 (table 16, en Annexe).

Les données du REE montrent que le nombre d'entreprises a fortement diminué entre 1998 et 2000 (classe de taille 1 : de 255 à 151 pour la classe 18, et de 100 à 48 pour la classe 19). Cela peut contribuer à l'augmentation du salaire. Le taux de réponse a été médiocre dans les deux cantons dans toutes les strates concernées (table 11).

TAB. 10: Médianes et intervalle de confiance pour NOGA2 18 et 19

NOGA	1998	2000		
	médiane	médiane	IC à 95%	
18	3294	3569	3366	3759
19	3534	3912	3662	4140

TAB. 11: Effectifs et nombre de réponses pour NOGA2 18 et 19

		nb d'entreprises				nb de réponses			
		1998	2000			1998	2000		
NOGA2	taille		GE	RS	TI		GE	RS	TI
18	1	255	14	125	12	8	9	8	7
18	2	139	0	79	24	20	0	18	11
18	3	45	0	25	12	26	0	11	10
18	4	28	0	9	7	10	0	5	4
18	5	5	0	2	2	3	0	2	1
19	1	100	5	38	5	8	2	13	4
19	2	45	5	30	3	11	3	16	3
19	3	17	0	12	1	9	0	8	1
19	4	10	2	6	1	9	1	4	1
19	5	2	0	1	0	2	0	1	0

4.5.2 Analyse des résultats sur le coefficient de variation synthétique pour les données agrégées par classes NOGA2

La table 16 (Annexe) présente les résultats complets pour les données sur toute la Suisse présentées la fig. 9. Le nombre total d'entreprises répondantes et de salaires observés par classe NOGA2 sont notés respect. n_e et n_s. Les deux bornes de confiance, b_{i95} et b_{s95} donnent une information complète sur la précision de la médiane estimée, mais ne permettent pas une comparaison rapide de la précision de plusieurs médianes. Pour cela, un coefficient de variation est utile, puisqu'il synthétise l'information en une seule valeur. Nous avons défini plus haut (chap. 3) deux coefficients de variation synthétiques, CV_{syn95} et CV_{syn}. Il est intéressant de comparer ces coefficients avec CV_{moy}, celui de la moyenne des MBLS et CV_{perc}, celui du pourcentage (la moyenne des z_{hij}) sur lequel l'intervalle de confiance de la médiane est basé. On constate (table 16) que CV_{perc} est en général supérieur aux autres coefficients de variation. Les coefficients de variation synthétiques de la médiane $\widetilde{y}$ sont du même ordre de grandeur que CV_{moy}. Cela justifie a posteriori la planification de l'enquête dont le but est de limiter le CV des moyennes par strate à 5%.

La table 12 présente les coefficients de corrélation simple entre ces statistiques. On voit que les deux CV synthétiques sont bien corrélés, ce qui montre qu'ils sont des mesures cohérentes de la précision. Comme CV_{syn95} permet de reconstruire un intervalle qui contient toujours l'in-

TAB. 12: Coefficients de corrélation entre les CV de la table 16

	CV_{moy}	CV_{syn95}	CV_{syn}	CV_{perc}
CV_{moy}	1.00	*	*	*
CV_{syn95}	0.79	1.00	*	*
CV_{syn}	0.73	0.93	1.00	*
CV_{perc}	0.76	0.83	0.84	1.00

tervalle de confiance, on le préférera à CV_{syn}. L'intervalle reconstruit est symétrique autour de la médiane, alors que l'intervalle de confiance original peut avoir l'une des bornes moins éloignée de la médiane que l'autre.

Il y a (table 16) 13 classes NOGA2 avec moins de 100 entreprises répondantes ; 7 d'entre elles ont un $CV_{perc} > 6\%$. Cette valeur n'est dépassée dans les autres cas que pour la classe 60. Remarquons que ces grands CV sur l'échelle du pourcentage n'aboutissent pas à des CV_{syn95} nécessairement très grands. Ces derniers ne dépassent 3% que 5 fois et 4 % une seule fois : pour la classe 73 (6%) qui a aussi le plus grand CV_{perc} (11%). Pour la classe 60, le $CV_{syn95} = 3\%$, malgré son grand $CV_{perc} = 7\%$.

Exemple d'analyse d'un domaine Considérons ANFORNI=3, c'est-à-dire les travaux demandant des connaissances professionnelles spécialisées. Comparons les intervalles de confiance pour ce domaine et les deux sous-domaines correpondant aux deux sexes. La fig. 10 présente les résultats pour les mêmes subdivisions en classes d'activité NOGA2 que précédemment. On constate une moins grande variabilité d'une classe NOGA à l'autre que lorsque tous les niveaux d'exigeance sont pris ensemble et une différence d'amplitude variable, mais systématique entre les deux sexes.

Répartition des coefficients de variation synthétiques pour les domaines La fig. 11 donne la répartition des coefficients de variation synthétiques du salaire par classe Noga et pour les catégories d'emploi : tous (TOT), hommes (M), femmes (F), ANFORNI = 1+2 (A12), 3 (A3), 4+5 (A45), et les combinaisons. Plusieurs constatations intéressantes s'imposent :

- Les boîtes à moustaches sont systématiquement plus étalées pour les femmes que pour les hommes.
- Certaines classes Noga2 sont omises (table 13) pour des raisons de protection des données (effectifs insuffisants). La catégorie la plus touchée est A12F, celle des femmes les plus qualifiées.
- Lorsqu'elle est présente, la classe Noga2 61 (transports par eau) a toujours le plus grand CV_{syn95}.
- Tous les domaines considérés ont un CV médian inférieur à 4% et un 3e quartile inférieur à 6%.

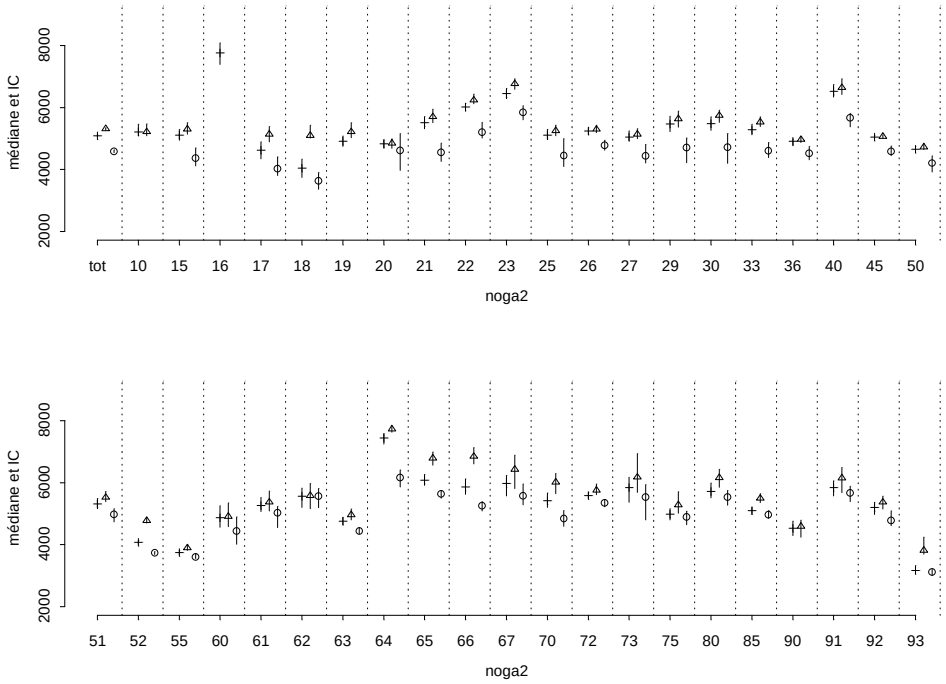


FIG. 10: Médiane et intervalles de confiance à 95% par classe NOGA2, ANFORNI=3. + = tous les cas Δ = hommes o = femmes

– Les classes Noga2 dont le CV est supérieur à 10% sont indiquées par leur numéro à la fig. 11. Il serait sage de considérer comme peu sûres les estimations correspondant à un CV$>$10%.

4.5.3 Conclusion

L'estimation non-paramétrique des intervalles de confiance de la médiane est une méthode flexible, utile pour les répartitions asymétriques et bien adaptée au contexte de l'estimation basée sur le plan d'échantillonnage. L'expérience pratique découlant de l'analyse de la présente enquête suggère que.

– Les plus grands intervalles de confiance surviennent lorsque la répartition empirique est bimodale. Dans ce cas, bien sûr la médiane n'a pas grand sens.

– De très grands poids d'extrapolation rendent l'estimation de la médiane instable (comme celle de tout autre paramètre). Il est donc important de limiter les poids les plus grands.

– Une forte concentration des observations autour de la médiane induit de petits intervalles de confiance, même si l'intervalle sur l'échelle des pourcentages est grand. Cette concentration se voit généralement pour les salaires, ce qui explique le fait que le CV sur l'échelle des

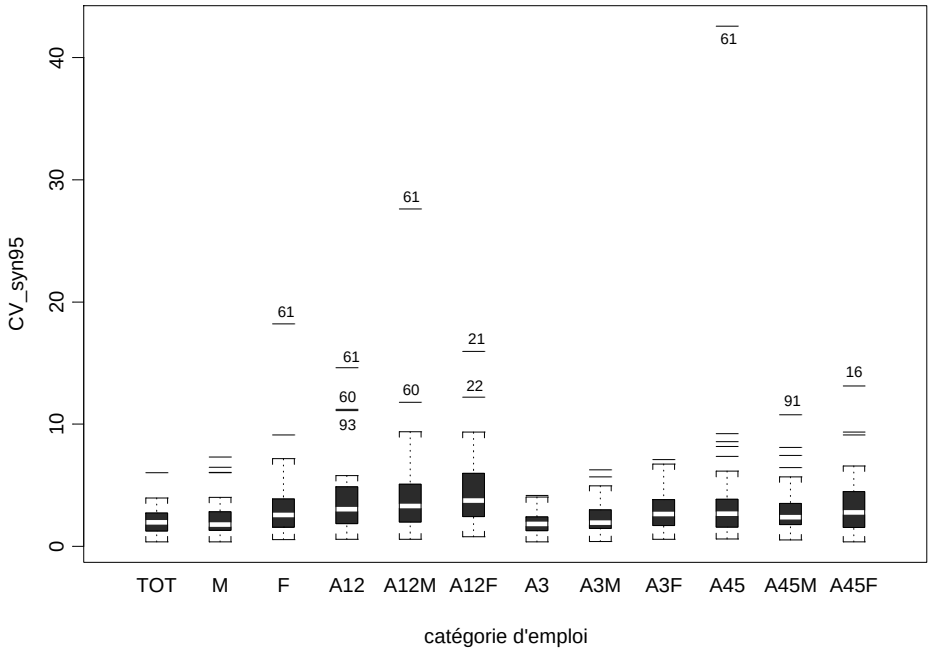


FIG. 11: Boîtes à moustaches du coefficient de variation synthétique par catégorie de salaire. Les numéros correpondent aux classes Noga2.

pourcentages soit généralement plus grand que sur l'échelle des salaires. C'est cette caractéristique qui rend la méthode particulièrement efficace dans cette enquête.

– Si le domaine est de petite taille, les CV par classe Noga2 ont tendance à être plus variables.

TAB. 13: Classes Noga2 non représentées par domaine

TOT															
M															
F	10														
A12															
A12M				19											
A12F	10	16	17	19	20	25	26	36	40	60	61			75	90
A3															
A3M		16													
A3F	10	16		19											90
A45														75	
A45M											61	62	67	75	
A45F	10										61	62		75	90

ANNEXE 1 : Tables

Les tables 14 à 16 sont commentées dans le texte.

TAB. 14: Comparaison des plans d'échantillonnage basés sur deux estimations des paramètres

Obs	NOGA2	taille	cv1	nb1	cv2	nb2	diffcv	diffmean	diffnb
1	.	.	0.004	9722	0.004	9839	0.000	-5	-117
2	10-14	.	0.018	80	0.018	79	0.000	2	1
3	10-14	1	0.051	22	0.050	22	0.001	-32	0
4	10-14	2	0.050	27	0.050	26	0.000	6	1
5	10-14	3	0.033	14	0.033	14	0.000	5	0
6	10-14	4	0.027	14	0.027	14	0.000	-4	0
7	10-14	5	0.007	3	0.006	3	0.001	2	0
8	15	.	0.035	173	0.035	179	0.000	17	-6
9	15	1	0.050	40	0.050	50	0.000	60	-10
10	15	2	0.050	21	0.050	21	0.000	-9	0
11	15	3	0.050	50	0.050	47	0.000	35	3
12	15	4	0.050	46	0.050	45	0.000	16	1
13	15	5	0.069	16	0.069	16	0.000	15	0
14	16	.	0.025	11	0.023	11	0.002	30	0
15	16	2	0.052	3	0.053	3	-.001	22	0
16	16	3	0.024	1	0.023	1	0.001	4	0
17	16	4	0.044	6	0.043	6	0.001	37	0
18	16	5	0.016	1	0.011	1	0.005	22	0
19	17	.	0.034	101	0.034	97	0.000	44	4
20	17	1	0.050	25	0.050	25	0.000	-12	0
21	17	2	0.049	24	0.050	24	-.001	-12	0
22	17	3	0.038	14	0.037	14	0.001	17	0
23	17	4	0.050	33	0.050	29	0.000	41	4
24	17	5	0.104	5	0.099	5	0.005	129	0
25	18	.	0.079	87	0.080	123	-.001	42	-36
26	18	1	0.051	17	0.050	53	0.001	-12	-36
27	18	2	0.050	33	0.050	34	0.000	-1	-1
28	18	3	0.050	19	0.050	19	0.000	42	0
29	18	4	0.048	14	0.052	13	-.004	13	1
30	18	5	0.211	4	0.215	4	-.004	108	0
31	19	.	0.014	76	0.013	76	0.001	12	0
32	19	1	0.050	25	0.050	25	0.000	-36	0
33	19	2	0.049	27	0.049	27	0.000	0	0
34	19	3	0.049	13	0.048	13	0.001	4	0
35	19	4	0.038	9	0.031	9	0.007	20	0
36	19	5	0.006	2	0.006	2	0.000	22	0
37	20	.	0.013	185	0.013	185	0.000	-1	0
38	20	1	0.025	73	0.031	73	-.006	2	0
39	20	2	0.016	53	0.015	53	0.001	-9	0
40	20	3	0.046	14	0.046	14	0.000	-6	0
41	20	4	0.022	11	0.019	11	0.003	12	0
42	20	5	0.011	34	0.011	34	0.000	29	0
43	21	.	0.038	83	0.038	85	0.000	9	-2
44	21	1	0.051	28	0.049	30	0.002	-63	-2
45	21	2	0.042	14	0.042	14	0.000	-1	0

Obs	NOGA2	taille	cv1	nb1	cv2	nb2	diffcv	diffmean	diffnb
46	21	3	0.049	17	0.050	17	-.001	10	0
47	21	4	0.050	21	0.050	21	0.000	9	0
48	21	5	0.005	3	0.004	3	0.001	9	0
49	22	.	0.017	139	0.017	139	0.000	2	0
50	22	1	0.046	53	0.047	53	-.001	-71	0
51	22	2	0.041	33	0.041	33	0.000	-14	0
52	22	3	0.047	14	0.046	14	0.001	18	0
53	22	4	0.044	13	0.044	13	0.000	-2	0
54	22	5	0.017	26	0.016	26	0.001	29	0
55	23-24	.	0.018	141	0.018	151	0.000	23	-10
56	23-24	1	0.050	38	0.050	48	0.000	-109	-10
57	23-24	2	0.050	36	0.050	37	0.000	-14	-1
58	23-24	3	0.051	14	0.050	14	0.001	33	0
59	23-24	4	0.050	35	0.050	34	0.000	19	1
60	23-24	5	0.014	18	0.014	18	0.000	30	0
61	25	.	0.028	112	0.028	114	0.000	21	-2
62	25	1	0.050	30	0.050	33	0.000	-99	-3
63	25	2	0.051	23	0.051	23	0.000	-5	0
64	25	3	0.050	36	0.049	36	0.001	27	0
65	25	4	0.049	16	0.049	15	0.000	32	1
66	25	5	0.016	7	0.015	7	0.001	18	0
67	26	.	0.017	91	0.017	92	0.000	-1	-1
68	26	1	0.050	37	0.051	37	-0.001	-63	0
69	26	2	0.051	19	0.049	20	0.002	-11	-1
70	26	3	0.040	14	0.040	14	0.000	4	0
71	26	4	0.032	13	0.032	13	0.000	-1	0
72	26	5	0.015	8	0.015	8	0.000	13	0
73	27-28	.	0.016	243	0.016	243	0.000	2	0
74	27-28	1	0.039	85	0.038	85	0.001	23	0
75	27-28	2	0.029	64	0.029	64	0.000	-12	0
76	27-28	3	0.031	24	0.031	24	0.000	-1	0
77	27-28	4	0.042	23	0.041	23	0.001	6	0
78	27-28	5	0.007	47	0.007	47	0.000	10	0
79	29,34	.	0.023	154	0.023	154	0.000	7	0
80	29,34	1	0.042	40	0.043	40	-0.001	-25	0
81	29,34	2	0.039	33	0.039	33	0.000	-18	0
82	29,34	3	0.045	17	0.045	17	0.000	-7	0
83	29,34	4	0.048	29	0.047	29	0.001	11	0
84	29,34	5	0.009	35	0.009	35	0.000	16	0
85	30-32	.	0.022	203	0.022	214	0.000	8	-11
86	30-32	1	0.050	53	0.050	63	0.000	-142	-10
87	30-32	2	0.050	54	0.050	56	0.000	-20	-2
88	30-32	3	0.050	36	0.050	36	0.000	17	0
89	30-32	4	0.050	45	0.050	44	0.000	23	1
90	30-32	5	0.011	15	0.011	15	0.000	2	0
91	33	.	0.024	153	0.024	158	0.000	30	-5
92	33	1	0.050	58	0.050	68	0.000	-158	-10
93	33	2	0.050	24	0.050	23	0.000	1	1
94	33	3	0.050	23	0.049	23	0.001	9	0

Obs	NOGA2	taille	cv1	nb1	cv2	nb2	diffcv	diffmean	diffnb
95	33	4	0.050	30	0.051	26	-0.001	61	4
96	33	5	0.017	18	0.016	18	0.001	18	0
97	36-37	.	0.018	140	0.017	109	0.001	-6	31
98	36-37	1	0.050	62	0.035	28	0.015	-37	34
99	36-37	2	0.050	36	0.050	39	0.000	-34	-3
100	36-37	3	0.032	14	0.031	14	0.001	10	0
101	36-37	4	0.038	13	0.037	13	0.001	3	0
102	36-37	5	0.009	15	0.009	15	0.000	8	0
103	40-41	.	0.020	67	0.020	68	0.000	2	-1
104	40-41	1	0.049	22	0.049	23	0.000	-140	-1
105	40-41	2	0.025	14	0.024	14	0.001	-5	0
106	40-41	3	0.029	14	0.028	14	0.001	4	0
107	40-41	4	0.036	13	0.035	13	0.001	-1	0
108	40-41	5	0.024	4	0.025	4	-0.001	11	0
109	45	.	0.009	955	0.009	955	0.000	-10	0
110	45	1	0.013	362	0.014	362	-0.001	-56	0
111	45	2	0.011	289	0.011	289	0.000	-18	0
112	45	3	0.030	76	0.031	76	-0.001	1	0
113	45	4	0.021	39	0.021	39	0.000	0	0
114	45	5	0.005	189	0.005	189	0.000	7	0
115	50	.	0.013	382	0.013	382	0.000	-42	0
116	50	1	0.023	197	0.026	197	-0.003	-166	0
117	50	2	0.026	86	0.027	86	-0.001	-31	0
118	50	3	0.046	14	0.045	14	0.001	-2	0
119	50	4	0.051	13	0.051	13	0.000	-1	0
120	50	5	0.009	72	0.009	72	0.000	17	0
121	51	.	0.017	568	0.017	567	0.000	-9	1
122	51	1	0.041	247	0.044	247	-0.003	-284	0
123	51	2	0.035	124	0.035	124	0.000	-4	0
124	51	3	0.050	44	0.050	46	0.000	27	-2
125	51	4	0.050	49	0.050	46	0.000	31	3
126	51	5	0.018	104	0.018	104	0.000	35	0
127	52	.	0.008	802	0.008	802	0.000	-14	0
128	52	1	0.019	425	0.022	425	-.003	-151	0
129	52	2	0.022	157	0.022	157	0.000	-7	0
130	52	3	0.050	44	0.050	44	0.000	7	0
131	52	4	0.050	26	0.050	26	0.000	16	0
132	52	5	0.005	150	0.005	150	0.000	19	0
133	55	.	0.014	743	0.014	743	0.000	-6	0
134	55	1	0.020	363	0.021	363	-.001	-41	0
135	55	2	0.025	187	0.025	187	0.000	-3	0
136	55	3	0.050	33	0.050	33	0.000	-6	0
137	55	4	0.048	13	0.047	13	0.001	-2	0
138	55	5	0.012	147	0.012	147	0.000	14	0
139	60	.	0.010	154	0.010	154	0.000	-6	0
140	60	1	0.030	53	0.032	53	-.002	-37	0
141	60	2	0.043	37	0.043	37	0.000	-39	0
142	60	3	0.046	14	0.046	14	0.000	0	0
143	60	4	0.050	23	0.051	23	-.001	-3	0
144	60	5	0.002	27	0.002	27	0.000	3	0

Obs	NOGA2	taille	cv1	nb1	cv2	nb2	diffcv	diffmean	diffnb
145	61	.	0.033	65	0.033	65	0.000	0	0
146	61	1	0.044	24	0.044	24	0.000	-38	0
147	61	2	0.050	18	0.052	18	-.002	32	0
148	61	3	0.035	14	0.036	14	-.001	-27	0
149	61	4	0.073	8	0.072	8	0.001	9	0
150	61	5	0.031	1	0.030	1	0.001	22	0
151	62	.	0.009	59	0.008	59	0.001	20	0
152	62	1	0.044	17	0.045	17	-.001	-326	0
153	62	2	0.051	19	0.050	19	0.001	10	0
154	62	3	0.108	12	0.109	12	-.001	-18	0
155	62	4	0.041	9	0.038	9	0.003	55	0
156	62	5	0.007	2	0.006	2	0.001	22	0
157	63	.	0.022	274	0.022	284	0.000	9	-10
158	63	1	0.050	80	0.050	83	0.000	-46	-3
159	63	2	0.050	111	0.050	119	0.000	-53	-8
160	63	3	0.049	16	0.049	16	0.000	31	0
161	63	4	0.050	51	0.050	50	0.000	21	1
162	63	5	0.031	16	0.031	16	0.000	18	0
163	64	.	0.005	110	0.005	119	0.000	5	-9
164	64	1	0.050	28	0.050	35	0.000	-181	-7
165	64	2	0.050	37	0.050	39	0.000	-12	-2
166	64	3	0.050	21	0.049	21	0.001	-1	0
167	64	4	0.050	21	0.050	21	0.000	5	0
168	64	5	0.002	3	0.002	3	0.000	10	0
169	65	.	0.016	313	0.016	297	0.000	71	16
170	65	1	0.050	57	0.050	65	0.000	-192	-8
171	65	2	0.050	156	0.050	143	0.000	38	13
172	65	3	0.050	37	0.051	31	-.001	197	6
173	65	4	0.050	41	0.051	36	-.001	141	5
174	65	5	0.012	22	0.012	22	0.000	47	0
175	66	.	0.011	189	0.011	191	0.000	28	-2
176	66	1	0.050	80	0.050	83	0.000	-592	-3
177	66	2	0.049	16	0.051	16	-.002	-33	0
178	66	3	0.050	37	0.050	37	0.000	4	0
179	66	4	0.049	35	0.051	34	-.002	14	1
180	66	5	0.010	21	0.011	21	-.001	50	0
181	67	.	0.023	164	0.025	176	-.002	-146	-12
182	67	1	0.050	63	0.050	73	0.000	-263	-10
183	67	2	0.050	49	0.050	51	0.000	-180	-2
184	67	3	0.080	33	0.091	33	-0.011	-402	0
185	67	4	0.049	12	0.046	12	0.003	33	0
186	67	5	0.027	7	0.027	7	0.000	0	0
187	70-71	.	0.021	198	0.021	245	0.000	-14	-47
188	70-71	1	0.050	123	0.050	172	0.000	-142	-49
189	70-71	2	0.050	30	0.050	30	0.000	-5	0
190	70-71	3	0.050	16	0.050	14	0.000	76	2
191	70-71	4	0.050	13	0.051	13	-0.001	5	0
192	70-71	5	0.018	16	0.018	16	0.000	15	0
193	72-74	.	0.012	1079	0.012	1079	0.000	-70	0
194	72-74	1	0.025	570	0.029	570	-0.004	-372	0
195	72-74	2	0.023	220	0.023	220	0.000	-2	0

Obs	NOGA2	taille	cv1	nb1	cv2	nb2	diffcv	diffmean	diffnb
196	72-74	3	0.050	40	0.050	41	0.000	3	-1
197	72-74	4	0.050	49	0.050	48	0.000	7	1
198	72-74	5	0.011	200	0.011	200	0.000	19	0
199	73	.	0.028	86	0.027	84	0.001	7	2
200	73	1	0.050	30	0.051	28	-0.001	-64	2
201	73	2	0.049	19	0.049	19	0.000	-30	0
202	73	3	0.052	20	0.052	20	0.000	-2	0
203	73	4	0.050	14	0.048	14	0.002	20	0
204	73	5	0.010	3	0.010	3	0.000	22	0
205	75	.	0.018	31	0.019	31	-0.001	-7	0
206	75	1	0.055	8	0.068	8	-0.013	-148	0
207	75	2	0.033	14	0.034	14	-0.001	-22	0
208	75	3	0.045	7	0.045	7	0.000	41	0
209	75	4	0.040	1	0.046	1	-0.006	-29	0
210	75	5	0.026	1	0.025	1	0.001	-14	0
211	80	.	0.022	252	0.022	258	0.000	-9	-6
212	80	1	0.050	115	0.050	117	0.000	-54	-2
213	80	2	0.050	57	0.050	59	0.000	-13	-2
214	80	3	0.050	44	0.050	46	0.000	-23	-2
215	80	4	0.051	23	0.051	23	0.000	5	0
216	80	5	0.024	13	0.024	13	0.000	18	0
217	85	.	0.015	363	0.014	363	0.001	-23	0
218	85	1	0.026	195	0.027	195	-0.001	-158	0
219	85	2	0.027	53	0.026	53	0.001	-64	0
220	85	3	0.050	23	0.050	23	0.000	-1	0
221	85	4	0.034	24	0.033	24	0.001	4	0
222	85	5	0.013	68	0.013	68	0.000	10	0
223	90	.	0.030	53	0.030	53	0.000	3	0
224	90	1	0.030	17	0.033	17	-0.003	20	0
225	90	2	0.051	17	0.051	17	0.000	-3	0
226	90	3	0.049	16	0.049	16	0.000	5	0
227	90	5	0.023	3	0.020	3	0.003	7	0
228	91	.	0.021	154	0.021	165	0.000	-24	-11
229	91	1	0.050	68	0.050	77	0.000	-177	-9
230	91	2	0.050	34	0.050	36	0.000	-18	-2
231	91	3	0.051	23	0.050	23	0.001	-1	0
232	91	4	0.049	18	0.049	18	0.000	2	0
233	91	5	0.012	11	0.011	11	0.001	8	0
234	92	.	0.019	288	0.019	291	0.000	-24	-3
235	92	1	0.050	72	0.050	72	0.000	-73	0
236	92	2	0.050	97	0.050	97	0.000	-59	0
237	92	3	0.050	84	0.050	87	0.000	-13	-3
238	92	4	0.052	16	0.052	16	0.000	2	0
239	92	5	0.007	19	0.006	19	0.001	8	0
240	93	.	0.022	200	0.023	198	-0.001	-29	2
241	93	1	0.032	107	0.039	107	-0.007	-72	0
242	93	2	0.050	36	0.051	36	-0.001	-13	0
243	93	3	0.050	26	0.051	24	-0.001	25	2
244	93	5	0.021	31	0.021	31	0.000	8	0
245	99	.	0.206	1	0.171	1	0.035	8	0
246	99	5	0.206	1	0.171	1	0.035	8	0

TAB. 15: Equivalents plein-temps selon la BESTA au 4e trimestre et nb. d'apprentis selon le recensement des entreprises 98

Noga 2	EPT secteur privé		EPT total		différences		apprentis
LSE	1998	2000	1998	2000	1998	2000	1998
10-14	4694	4638	4741	4685	47	47	106
15	56067	57689	56093	57715	26	26	1731
16	2655	2657	2655	2657	0	0	50
17	15710	14603	15712	14605	2	2	378
18	7599	6567	7741	6690	142	123	394
19	2818	2485	2818	2485	0	0	97
20	36237	34864	36257	34884	20	20	4430
21	13930	14542	13930	14542	0	0	473
22	48438	48967	48441	48970	3	3	2760
23-24	58658	61135	58666	61143	8	8	1466
25	23836	23650	23836	23650	0	0	696
26	18433	18462	18433	18462	0	0	524
27-28	94701	96860	94701	96860	0	0	5301
29,34,35	120003	120241	124317	124634	4314	4393	8080
30	58880	60839	59256	61231	376	392	2300
33	66343	68349	66343	68349	0	0	2585
36-37	28140	27905	28140	27905	0	0	1576
40-41	12300	11707	22437	21396	10137	9689	1432
45	252985	264460	260342	272074	7357	7614	24921
50	73236	74792	73252	74808	16	16	11921
51	170112	177885	170117	177890	5	5	4777
52	269906	276488	269916	276498	10	10	24152
55	179138	172104	179372	172328	234	224	7165
60	48120	48890	85825	87135	37705	38245	1912
61	1780	1707	1780	1707	0	0	27
62	10484	11394	10484	11394	0	0	81
63	38094	41108	39282	42391	1188	1283	2204
64	31581	32183	72194	73553	40613	41370	1931
65	94263	101812	106558	115092	12295	13280	4765
66	55659	56317	57120	57795	1461	1478	2502
67	8009	8905	8033	8932	24	27	133
70-71	18243	18264	18380	18401	137	137	606
72,74	249372	270041	252164	272976	2792	2935	13025
73	5614	6601	8976	10518	3362	3917	282
75	764	776	112001	113781	111237	113005	4043
80	32706	33517	161280	165301	128574	131784	6152
85	180105	188192	287887	300548	107782	112356	13272
90	3632	3768	11134	11515	7502	7747	119
91	16893	17274	26448	27046	9555	9772	439
92	32443	34419	35911	38090	3468	3671	535
93	33713	32228	34382	32876	669	648	4906

TAB. 16: Résultats pour la Suisse

NOGA2	n_e	n_s	$\overline{y}$	CV_{moy}	$\widetilde{y}$	b_{i95}	b_{s95}	CV_{syn95}	CV_{syn}	CV_{perc}
0	17610	489373	5902	0.45	5163	5136	5200	0.37	0.41	0.79
10-45	4695	150424	5754	0.49	5253	5200	5299	0.51	0.49	1.31
50-93	12915	338949	6003	0.67	5082	5029	5129	0.53	0.53	0.98
10-14	81	961	5601	1.70	5190	5089	5365	1.72	1.98	6.04
15	189	10495	5158	1.89	4703	4530	4901	2.15	2.62	4.63
16	10	1753	7506	3.02	6602	6315	6962	2.78	3.44	5.75
17	88	1551	4683	2.22	4262	4024	4496	2.85	3.12	5.74
18	86	1581	4128	2.92	3569	3366	3759	2.90	2.91	5.41
19	57	571	4434	2.58	3912	3662	4140	3.26	3.22	6.38
20	277	2952	5134	1.52	4862	4746	4973	1.22	1.15	5.72
21	66	3441	5676	1.86	5433	5232	5585	1.89	1.40	5.54
22	266	8269	6379	2.08	5975	5742	6150	1.99	2.09	5.00
23-24	142	23471	7390	1.78	6842	6583	7122	2.09	2.09	4.72
25	106	3178	5337	2.42	4855	4693	5042	1.97	1.98	4.95
26	128	2703	5482	1.30	5138	5005	5251	1.32	1.62	5.30
27-28	421	14273	5396	1.13	5006	4924	5124	1.20	1.34	3.46
29,34,35	232	16991	6222	1.81	5750	5542	6014	2.34	2.54	5.72
30	214	11927	6322	2.11	5761	5483	6085	2.87	2.71	5.16
33	229	12584	5610	1.62	5092	4922	5222	1.70	1.63	3.48
36-37	204	4244	5241	1.77	4754	4664	4887	1.43	1.22	4.22
40-41	60	3588	7332	2.09	6780	6505	7034	2.07	2.30	6.55
45	1839	25891	5382	0.53	5108	5062	5150	0.46	0.45	2.07
50	907	10192	5141	1.30	4706	4622	4811	1.14	1.10	3.14
51	1369	29195	6720	1.57	5620	5463	5786	1.51	1.65	3.04
52	1792	95791	4545	0.86	4009	3953	4075	0.84	0.75	2.03
55	1891	19043	3859	1.66	3519	3467	3578	0.86	1.05	2.31
60	265	6049	5055	1.55	4691	4519	4964	2.97	2.71	6.85
61	45	676	5855	7.16	5435	5067	5856	3.95	3.99	9.67
62	53	1659	7263	2.14	5662	5372	6093	3.88	3.02	5.38
63	393	18368	5358	1.43	4824	4690	4974	1.59	1.47	3.98
64	78	22554	7125	1.30	6487	6328	6646	1.25	1.13	2.26
65	424	29742	8923	1.65	7190	7059	7314	0.93	0.95	1.78
66	151	22089	7377	1.75	6504	6281	6752	1.95	1.89	3.62
67	205	1880	8859	2.65	6937	6608	7332	2.91	2.88	4.63
70-71	261	3924	6646	3.43	5622	5412	5873	2.28	1.87	5.06
72,74	2833	35834	7570	2.35	6290	6139	6448	1.28	1.26	2.36
73	73	1587	7664	5.85	6870	6134	7679	6.01	4.80	10.53
75	24	490	6176	2.06	5414	5200	5770	3.35	4.80	7.02
80	275	6063	7116	2.73	6550	6292	6812	2.04	2.29	4.67
85	845	20822	5418	1.05	5076	4971	5188	1.13	1.22	3.12
90	51	707	5007	2.62	4599	4419	4804	2.27	2.78	9.36
91	245	4470	6927	2.46	6175	5919	6458	2.34	2.38	5.46
92	320	4983	5708	1.93	5145	4878	5335	2.65	2.82	4.82
93	415	2831	3670	1.76	3287	3172	3400	1.79	2.37	4.95

ANNEXE 2 : Principales procédures SAS

	Commentaires	*Si non spécifié, fichiers de la libname ess*	
		input	output
	1998		
compar.sas compar.out	Comparaison des pondérations initiale et finale.	test98.provgtot	
ponder.sas ponder.out	Recherche de la méthode utilisée en 1998.	test98.provgtot	
ibgrs.sas ibgrs.out	Comparaison des médianes de mbls pondérées avec et sans multiplication par ibgrs.	test98.provgtot	
entrepr.sas	Statistiques au niveau des entreprises	test98.provgtot	
noga2.sas noga2.out	Statistiques au niveau des strates		
macro calcpoids.sas	Normalisation par strate d'une variable de poids.		
macro extrapol.sas	Extrapolation par strate avec 2 options.		
	2000, échantillon CH		
choix.sas choix.out	Choix d'une pondération avec ibgr: Relation entre les poids d'extrapolation avec les PT du RE et les EPT de la BESTA.	test00.provgtot	ptstra strates ptn ptnog eptn noga2
taille.sas taille.out	Etude de la corrélation entre les plein-temps du RE et estimés.	ptstra	
choixs.sas choixs.out	Choix d'une pondération avec ibgrs: Relation entre les poids d'extrapolation avec les PT du RE et les EPT de la BESTA.	test00.provgtot	ptstras strates ptn ptnogs eptn noga2s
taux00.sas taux00.out	Comparaison de ibgr et ibgrs en 2000: étude de taux = ibgr/ibgrs.	test00.provgtot	
compa00.sas compa00.out	Test de l'effet de la troncature du poids intra-entreprise	test00.provgtot	
	2000, échantillon augmenté		
compa001.sas compa001.out compa001b,c,d,e.out	Comparaison de l'échantillon CH et de l'échantillon augmenté sur la base de la pondération initiale.	test00.provgtot test00.provgtot_kalt	priv00 ech
anforni.sas anforni.out	Etude des anforni manquant: anzlohn = nb de salaires fournis nmiss = nb de salaires pour lesquels anforni est donné.	priv00 ech	
compa003.sas	Etude des variables liées à la classe Noga. Comparaison des médianes obtenues avec ou sans troncature du poids intra-entreprise sur les cas retenus pour le communiqué de presse. priv00 = échantillon augmenté ech00 = échantillon CH.	test00.provgtot_k (fichier définitif)	priv00 ech00
p00std.sas	Comparaison des médianes provenant de p00std et p00stdch.	datdyn.p00std datdyn.p00stdch	

Nom	*Commentaires*	*input*	*output*
	Comparaison avec le registre et la BESTA		
lectept.sas lectept.out	Création des EPT du privé à partir de la BESTA du 4Q98 et 4Q00 dans le concept entreprise par classe Noga2 avec le regroupement LSE.	ept9800.prn	eptmh.txt
lectappr.sas	Regroupement LSE du fichier des apprensis	unternehmen noga 2.xls	appnog.txt
apprentis.ssc	Dessins faits dans S-Plus	appnog.txt	appnog98 (S-Plus)
	Programmes établissant la pondération finale		
cal.sas cal4.out	Calcul de la pondération finale. Troncature et extrapolation des poids initiaux par calage sur marges. priv01 est test00.provgtot_k modifié comme suit: gewicht (modifié, pondération finale) fingew1 (modifié, pondération finale) fingew2 (modifié, pondération finale) rukt (nouveau) = RS, GE, TI ("régions" de ukt) nukt (nouveau) = 1, 2, 3 (rukt recodé)	test00.provgtot_k	priv01 (voir ci-contre)
compa002.sas compa002.out	Comparaison des médianes avec la pondération initiale et finale de 2000. priv01 est modifié comme suit: gewibgr1 = initgew*ibgrs gewibgrs = gewicht*ibgrs.	priv01	priv01 (voir ci-contre)
fingew00.sas fin.out fingew00.out	Stockage dans un fichier entreprises. Analyse des poids finaux > 50. Description de fingew00.	priv01	fingew00
	Calcul et représentation de la précision		
icmed.sas icmoy.sas icmed02.sas icmoy02.sas	macro de calcul d'un IC pour la médiane macro de calcul d'un IC pour la moyenne idem, mais avec paramètres d'appel permettant de généraliser à d'autres jeux de données.	-	-
appelic.sas appelicmoy.sas	création des tableaux - par région, noga2, anforni et sexe, - par région, noga2 et classe de taille appellent icmed et icmoy respect.	priv01	 conf00 conf00b
totRS.ssc	programme S-plus dessinant la fig.6		
foncplot.ssc	programme S-plus dessinant les fig. 7 et 8		
dessIC.ssc	programme S-plus dessinant les fig. 9 et 10		
boxsyn.ssc	programme S-plus dessinant la fig. 13		

ANNEXE 3 : Macro SAS

```
rsubmit;
************************************************************************** ;
* icmed02.sas                                                             ;
* macro de calcul de la médiane de mbls                                   ;
*              de son intervalle de confiance non-paramétrique à 95%      ;
*              du coefficient de variation du percentile 50%              ;
*              du coefficient de variation synthétique de la médiane      ;
************************************************************************** ;
/*
                                        17 mai 2002
                                        M. Graf


Cette macro est appelée par les macros appelas et appelg définies et utilisées dans le programme
appelic.sas.

Les no de formules se réfèrent au rapport de méthode:
Enquête suisse sur la structure des salaires 2000.
Plan d échantillonnage, pondération et méthode d'estimation pour le secteur privé.
-------------------------------------------------------------------------------------------
Input : fichier priv contenant les variables
          id2 burnr          identificateur de l entreprise
             strate          strak numérique calculée dans appelic.sas
          varet mbls         salaire mensuel standardisé
          poids gewibgrs     = gewicht*ibgrs, poids du salarié pondéré par le taux d activité
           anz anzlohn       nombre de salaires fournis
             nrepk           nombre d entreprises fournissant au moins 1 salaire
             thi             taux de réponse intra-entreprise calculé dans appelvar
             th              taux de réponse des entreprises  calculé dans appelvar.
             nukt            région pour l'imputation (combinaison de strates)
             nog2            classe Noga pour l'imputation (combinaison de strates)
             pub1            borne inf. du nb d entreprises pour la publication
             pub2            borne inf. du nb de salaires - nb d entreprises pour la public.

Output: fichier conf2 contenant les variables
             mediane         la médiane des salaires de priv, pondérée par gewibgrs
             b_inf et b_sup   les bornes de l IC non-paramétrique à 95%
             CV_syn95        le CV synthétique basé sur l IC à 95%
             CV_syn          le CV synthétique basé sur l IC à 68% (i.e. + ou - 1 écart-type)
             CV_perc         le CV du percentile 50%.
             n_e             nb d entreprises pour la base du calcul
             n_s             nb de salaires pour la base du calcul

Ce fichier est complété dans appelas et appelg par les variables de stratification et de domaine.

Les valeurs calculées dans conf2 ne sont publiées que si n_e > 6  et n_s-n_e > 54 pour le domaine concerné.


Exemple d appel:
--------------------
%icmed02(priv, conf2, mbls, gewibgrs, burnr, anzlohn, thi, th, strate, nrepk, nukt, nog_2, 6, 54);

*/
```

```
%macro icmed02(priv, conf2, varet, poids, id2, anz, thi, th, strate, nrepk, nukt, nog2, pub1, pub2);
*-------------------------------------------------------------------------------------------------------------------------------;
proc sort data=&priv out=priv;
 by &id2;
run;

* médiane générale ;
* ------------------- ;
proc univariate data=priv noprint;
 var &varet;
 weight &poids;
 output out=medg median=y;
run;
data _NULL_; set medg;
call symput("y",y);                                  * médiane en variable globale;
run;

* Calcul du paramètre theta, de ej et e_hi ;
* ------------------------------------------------ ;
%let theta=0.5 ;                                     * quantile théorique de la médiane ;
proc univariate data=priv noprint;
 var &poids;
 by &id2;
 output out=entr1 sum=svhi ;                         * somme des taux d activité standardisés;
run;
data priv; merge entr1 priv; by &id2;
 zhij=1;
 if &varet > &y then zhij=0  ;                       * calcul des indicatrices ;
 ej=&poids*(zhij-&theta) ;                           * par rapport à (19), ej est ici multipliée ;
                                                     * par fingew1=Nh/nh;
run;

* Remarque;
* ------------;
* e_hi est donc aussi multipliée par fingew1 ;
* Bhi et Ah par (fingew1)**2 ;

proc univariate data=priv noprint;
 var ej ;
 by &id2 ;
 output out= entr1 sum=e_hi var= Bhi N=NDhi;          * formules (20) et (23) ;
run;


* Niveau entreprises ;
* ---------------------- ;
data entr; set priv;
 if &id2 ne lag(&id2) ;
 drop &varet &poids zhij ej;
run;

data entr; merge entr1  entr ; by &id2 ;
 NDhi1=NDhi-1;
Bhi=(NDhi1*Bhi + NDhi*(1-NDhi/&anz)*(e_hi/NDhi)**2);
                                                     * modification de la variance intra-entreprises ;
 Bhi=Bhi/(&anz-1);                                   * pour un domaine de taille inconnue (30);
                                                     * si tous les cas sont considérés, NDhi=&anz;
```

```
 Bhi=Bhi*(1-&thi)*&anz ;                          *  composantes de (21) ;
 if &thi=1 then Bhi=0;
run;


* Niveau strates ;
* ---------------- ;
proc sort data=entr; by &strate;
run;
proc univariate data=entr noprint;
 by &strate;
 var Bhi svhi NDhi1 NDhi;
 output out=stra1 sum= Bh svh dlh toth;          * facteur de fh dans le second membre de (21);
                                                 * et dénominateur, 1ère ligne de (24) ;
                                                 * dlh = nb de salaires-nb d entreprises;
                                                 * toth= nb de salaires du domaine dans la strate;
run;
proc univariate data=entr noprint;
 by &strate;
 var e_hi;
 output out=stra2 var=Ah sum=e_h;                        * formule (22) ;

data stra; set entr;
 if &strate ne lag(&strate);
 drop e_hi Bhi svhi &anz  &thi &id2  NDhi NDhi1;
run;

* traitement du cas &nrepk=1;
* on impute la valeur de Ah/(svh*svh) ;
* par la moyenne des Ah/(svh*svh) sur la région et la classe &nog2;

data stra; merge stra stra1 stra2;
 by &strate;
 n_e=toth-dlh;                                           * nombre d entreprises dans le domaine;
 if n_e > 1 then do;
Ah=(n_e-1)*Ah + n_e*(1-n_e/&nrepk)*(e_h/n_e)**2;         * modification de la variance inter-entreprise;
  Ah=Ah/(&nrepk-1);                                      * pour un domaine de taille inconnue (31);
 end;                                                    * si tous les cas sont considérés, n_e=&nrepk;
 Ahrel=Ah/(svh*svh);
run;

proc sort data=stra; by &nukt &nog2;
run;
proc univariate data=stra noprint;
 var Ahrel; by &nukt &nog2;
 output out= stra3 mean=m_Ahrel;

data stra; merge stra stra3; by &nukt &nog2;
 if n_e = 1 then do;
   Ah = svh*svh*m_Ahrel;                                 * imputation;
   if &th =1 then Ah=0;                                  * tirage exhaustif ;
 end;
 V2sth = &nrepk*(1-&th)*Ah + &th *Bh;                    * formule (21) mult. par (svh*svh);
run;

* Niveau agrégat ;
* ------------------ ;
```

```
proc univariate data=stra noprint;
 var V2sth svh dlh toth;
 output out= tot sum= SV2st denom dl totd;          * numérateur et racine du dénomin. de (24) ;
                                                    * dl = nb de salaires - nb d entreprises;
                                                    * totd = nb de salaires du domaine;
run;

* Politique de publication;
* ----------------------------;
data _NULL_; set tot;
 call symput("dl",dl);
 call symput("totd",totd);
run;
%if ((&dl>&pub2)*(&totd-&dl>&pub1)) %then %do;
*-------------------------------------------------------------;
data tot; set tot;
 SV2st=SV2st/(denom*denom);                         * variance formule (24) ;
 demi95=sqrt(SV2st)*1.96;                           * demi-longueur de l IC à 95% ;
 demi=sqrt(SV2st);                                   * écart-type ou demi-long. de I lC à 68.26% ;
 c1= (&theta - demi95)*100;                         * limites de confiance à 95% ;
 c2= (&theta + demi95)*100;
 c3= (&theta - demi)*100;                           * limites de confiance à 68.26% ;
 c4= (&theta + demi)*100;
 c3=max(c3,0);                                      * domaine d existence des percentiles ;
 c1=max(c1,0);
 c4=min(c4,100);
 c2=min(c2,100);
 CVperc = 100*sqrt(SV2st)/&theta;                   * CV du percentile 50% ;

* Calcul des quantiles ;                            * formule (25) ;
* ------------------------ ;

 call symput("c1",c1);
 call symput("c2",c2);
 call symput("c3",c3);
 call symput("c4",c4);
 call symput("CVperc",CVperc);
run;

proc univariate data=priv noprint;
 var &varet;
 weight &poids;
 output out=confg  pctlpre=b_
                pctlname=i95 i68 s68 s95
                pctlpts= &c1 &c3 &c4 &c2 ;
run;

data &conf2; set confg;
 mediane=&y;
 CV_syn95=max(mediane-b_i95, b_s95-mediane);
 CV_syn95=100*CV_syn95/(1.96*mediane);              * CV  synthétique à 95% (en %) ;
 CV_syn=max(mediane-b_i68, b_s68-mediane);
 CV_syn=100*CV_syn/mediane;                         * CV synth. à 68.26% (en %) ;
 CV_perc = &CVperc;                                 * CV du percentile 50% ;
 n_e = &totd-&dl;
 n_s = &totd;
 drop b_i68 b_s68;
```

```
run;
*-----------------------------------------------;
%end;

%if ^((&dl>&pub2)*(&totd-&dl>&pub1)) %then %do;                    * alors on ne publie pas;
*-----------------------------------------------------------------;
data &conf2;
 mediane=&y;
 CV_syn95=.;
 CV_syn=.;
 CV_perc =.;
 n_e = &totd-&dl;
 n_s = &totd;
run;
*-----------------------------------------------------------------;
%end;

%mend icmed02;
*-------------------------------------------------------------------------------------------------------------------;
endrsubmit;
```

Références

[1] (2000). L'enquête suisse sur la structure des salaires 1998. Résultats commentés et tableaux. *Statistique de la Suisse*, OFS.

[2] Graf, M. (2002). Passage du concept établissement au concept entreprise. *Rapport de méthode*, OFS.

[3] Royston, J.P. (1992). Approximating the Shapiro-Wilk's W-Test for Non-normality. *Statistics and Computing*, **2,** 117-119.

[4] Peters, R. & Hulliger, B. (1996). Schätzverfahren für die Lohnstruktur erhebung 1994 / Procédure d'estimation pour l'enquête 1994 sur la structure des salaires. *Rapport de méthode*, OFS.

[5] Peters, R. (1997). Enquête 1996 sur la structure des salaires en Suisse ; établissement du plan d'échantillonnage. *Rapport de méthodes,* OFS,

[6] Särndal, C.-E., Swensson, B. & Wretman, J. (1992). *Model Assisted Survey Sampling.* Springer Series in Statistics.

[7] Shapiro, S.S. et Wilk, M.B. (1965). An Analysis of Variance Test for Normality (complete samples). *Biometrika,* **52**, 591-611.

Methodenberichte des Dienstes Statistische Methoden des BFS
Rapports de méthodes du Service de méthodes statistiques de l'OFS
Methodological reports of the Statistical Methods Unit of SFSO

Hulliger, B., Peters, R. (1994). Enquête sur la structure des salaires en Suisse: stratégie d'échantillonnage pour le secteur privé.

Peters, R., Hulliger, B. (1994). La technique de pondération des données: application à l'enquête suisse sur la santé.

Schwendener, P. (1995). Verbrauchserhebung 1990 - Vertrauensintervalle.

Hulliger, B. (1995). Konjunkturelle Mietpreiserhebung: Stichprobenplan und Schätzverfahren.

Peters, R. (1995). Enquête de structure sur les loyers: établissement d'un plan d'échantillonnage stratifié.

Salamin, P.A. (1995). Estimation des flux pour le module II des comptes globaux du marché de travail.

Hulliger, B. (1996). Gütertransportstatistik 1993: Schätzverfahren mit Kompensation der Antwortausfälle.

Hulliger, B., Peters, R. (1996). Enquête sur le comportement de la population suisse en matière de transport en 1994: plan d'échantillonnage et pondération.

Peters, R. (1996). Schéma de pondération des indices PAUL.

Peters, R., Hulliger, B. (1996). Schätzverfahren für die Lohnstruktur-Erhebung 1994 / Procédure d'estimation pour l'enquête de 1994 sur la structure des salaires.

Hulliger, B. (1996). Haushalterhebung Familie 1994: Stichprobenplan, Stichprobenziehung und Reservestichproben.

Comment, T., Hulliger, B., Ries, A. (1996). Gewichtungsverfahren für die Schweizerische Arbeitskräfteerhebung (1991-1995).

Peters, R. (1996). Pondération des données de l'enquête sur la famille en Suisse.

Peters, R. (1997). Enquête 1996 sur la structure des salaires en Suisse: établissement du plan d'échantillonnage.

Peters, R. (1997). Etablissement du plan d'échantillonnage pour l'enquête 1996 sur la recherche et le développement dans l'économie privée en Suisse.

Salamin, P.A. (1997). Evaluation de la Statistique de l'emploi.

Bender, A., Hulliger, B. (1997). Enquête suisse sur la population active: rapport de pondération pour 1996.

Peters, R. (1998). Extrapolation des données de l'enquête de structure sur les loyers.

Salamin, P.A. (1998). Etablissement d'une clef de passage pondérée entre l'ancienne (NGAE 85) et la nouvelle nomenclature (NOGA 95) générale des activités économiques.

Hulliger, B., Kassab, M. (1998). Evaluation of Estimation Methods for the Survey on Environment Protection Expenditures of Swiss Communes.

de Rossi, F.-X. (1998). Méthodes statistiques pour le compte routier suisse.

Hulliger, B., Eichenberger, P. (2000). Stichprobenregister für Haushalterhebungen: Umstellung auf Telefonnummern ohne Namen und Adressen, Abläufe für Erstellung und Stichprobenziehung. Bestellnummer: 338-0000

Renaud, A. (2001). Statistique suisse des bénéficiaires de l'aide sociale. Plan d'échantillonnage des communes. Numéro de commande: 338-0001

Hüsler, J., Müller, S. (2001). Schlussbericht Betriebszählung 1995 (BZ 95), Mehrfach imputierte Umsatzzahlen. Bestellnummer: 338-0002

Graf, M. (2001). Désaisonnalisation. Aspects méthodologiques et application à la statistique de l'emploi. Numéro de commande: 338-0003

Potterat, J., Hulliger, B. (2001). Schätzung der Sägereiproduktion mit der Sägerei-Erhebung PAUL. Bestellnummer: 338-0004

Peters, R., Renfer, J.-P. et Hulliger, B. (2001). Statistique de la valeur ajoutée 1997-1998. Procédure d'extrapolation des données. Numéro de commande: 338-0005

Salamin, P.A. (2001). La technique de la double enquête pour la statistique du transport routier de marchandise. Numéro de commande: 338-0006

Graf, M. (2002). Passage du concept établissement au concept entreprise. Numéro de commande: 338-0007

Kilchmann, D., Hulliger, B. (2002). Stichprobenplan für die Obstbaumzählung 2001, Bestellnummer: 338-0008

Renaud, A., Eichenberger P. (2002). Estimation de la couverture du recensement de la population de l'an 2000. Procédure d'enquête et plan d'échantillonnage de l'enquête de couverture. Numéro de commande: 338-0009

Graf, M. (2002). Enquête suisse sur la structure des salaires 2000. Plan d'échantillonnage, pondération et méthode d'estimation pour le secteur privé. Numéro de commande: 338-0010

Publikationsprogramm BFS

Das Bundesamt für Statistik (BFS) hat – als zentrale Statistikstelle des Bundes – die Aufgabe, statistische Informationen breiten Benutzerkreisen zur Verfügung zu stellen.

Die Verbreitung der statistischen Information geschieht gegliedert nach Fachbereichen (vgl. Umschlagseite 2) und mit verschiedenen Mitteln:

Programme des publications de l'OFS

En sa qualité de service central de statistique de la Confédération, l'Office fédéral de la statistique (OFS) a pour tâche de rendre les informations statistiques accessibles à un large public.

L'information statistique est diffusée par domaine (cf. verso de la première page de couverture); elle emprunte diverses voies:

Diffusionsmittel	*Kontakt* *N° à composer*	*Moyen de diffusion*
Individuelle Auskünfte	032 713 60 11	Service de renseignements individuels
Das BFS im Internet	http://www.statistik.admin.ch	L'OFS sur Internet
Medienmitteilungen zur raschen Information der Öffentlichkeit über die neusten Ergebnisse	032 713 60 11	Communiqués de presse: information rapide concernant les résultats les plus récents
Publikationen zur vertieften Information (zum Teil auch als Diskette)	032 713 60 60	Publications: information approfondie (certaines sont disponibles sur disquette)
Online-Datenbank	032 713 60 86	Banque de données (accessible en ligne)

Nähere Angaben zu den verschiedenen Diffusionsmitteln liefert das alle 2 Jahre nachgeführte **Publikationsverzeichnis.** Es ist auf dem Internet abzurufen oder kann gratis über Tel. 032 713 60 60 bezogen werden.

La **Liste des publications**, qui est mise à jour tous les deux ans, donne davantage de détails sur les divers moyens de diffusion. Elle est proposée sur Internet et peut être obtenue gratuitement en téléphonant au n° 032 713 60 60.

Methodenberichte des Dienstes Statistische Methoden
Rapports de méthodes du Service de méthodes statistiques
Methodological reports of the Statistical Methods Unit

Die Methodenberichte beschreiben die mathematischen und statistischen Methoden, die den Resultaten und Analysen der öffentlichen Statistik zu Grunde liegen. Sie enthalten ausserdem die Evaluation und Entwicklung von neuen Methoden im Hinblick auf eine zukünftige Anwendung. Diese Publikationen sollen einerseits die verwendeten Methoden dokumentieren, um Transparenz und Wissenschaftlichkeit sicher zu stellen, und sie sollen andererseits die Zusammenarbeit mit den Hochschulen und der Wissenschaft fördern.

Zur Illustration der beschriebenen mathematischen Konzepte werden im Bericht numerische Resultate aufgeführt. Diese sind allerdings nicht als offizielle Resultate der betreffenden Erhebungen zu verstehen. Ebenfalls können die tatsächlich angewendeten Methoden leicht von den hier beschriebenen abweichen.

Les rapports de méthodes décrivent les méthodes mathématiques et statistiques à la base des résultats et des analyses de la statistique publique. Ils présentent également l'évaluation et le développement de nouvelles méthodes en vue d'une application future. Ces publications visent d'une part à documenter les méthodes utilisées ou envisagées dans un souci de transparence et de rigueur scientifique, et d'autre part à favoriser la collaboration avec le monde scientifique et universitaire.

Les résultats numériques présentés dans les rapports de méthodes illustrent les concepts mathématiques décrits, mais ne sont pas des résultats officiels des enquêtes concernées. De même, les méthodes réellement appliquées peuvent différer légèrement de celles décrites dans ces rapports.

La présente enquête est une enquête auprès des entreprises dont le but principal est d'estimer des salaires médians pour un certain nombre de domaines (classe d'activité, niveau de qualification, sexe ...). Le plan d'échantillonnage prévoit un échantillon stratifié en grappes à deux niveaux. Il est conçu pour garantir un coefficient de variation de 5% sur le salaire moyen par strate. Les effectifs par strates sont calculés de deux manières: sur la base des salaires ou de leur logarithme. On constate qu'ils sont pratiquement identiques. La pondération reprend pour l'essentiel les poids d'extrapolation de Horvitz-Thompson. La non-réponse étant supposée ignorable dans les strates, les poids sont donc basés sur l'effectif par strate des entreprises répondantes. Les poids jugés trop grands sont alors tronqués et le déficit est distribué sur les autres classes de taille de même région et classe d'activité économique. On donne enfin la description complète de la méthode de calcul des intervalles de confiance des médianes pondérées. On définit un coefficient de variation synthétique de la médiane, calculé à partir de l'intervalle de confiance. Une comparaison de la précision de la moyenne et de la médiane est réalisée sur la base des coefficients de variation. Cette comparaison justifie l'établissement du plan d'échantillonnage sur la base des moyennes.

Nº de commande:
338-0010

Commandes:
032 713 60 60 Fax: 032 713 60 61

Prix:
gratuit

ISBN 3-303-00247-9