Rapport de méthodes

Enquête suisse sur la santé 2002

0

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

Plan d'échantillonnage, pondération et estimation de la précision

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

OFS BFS UST

Neuchâtel, 2005

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, Informationsgesellschaft, Sport
17 Politik
18 Öffentliche Verwaltung und Finanzen
19 Kriminalität und Strafrecht
20 Wirtschaftliche und soziale Situation der Bevölkerung
21 Nachhaltige Entwicklung und Disparitäten auf regionaler und internationaler Ebene

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, société de l'information, sport
17 Politique
18 Administration et finances publiques
19 Criminalité et droit pénal
20 Situation économique et sociale de la population
21 Développement durable et disparités régionales et internationales

Statistik der Schweiz | **Statistique de la Suisse**

Methodenbericht | **Rapport de méthodes**

Enquête suisse sur la santé 2002

Plan d'échantillonnage, pondération et estimation de la précision

Auteurs **Eric Graf et Jean-Pierre Renfer**
Office fédéral de la statistique

Editeur **Office fédéral de la statistique**

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

OFS BFS UST

Neuchâtel, 2005

Préambule

L'enquête suisse sur la santé 2002 (ESS02) est une enquête quinquennale réalisée pour la troisième fois consécutive par l'Office fédéral de la statistique (OFS) sur mandat du conseil fédéral. Le service des méthodes statistiques (METH) de l'OFS a été mandaté pour réaliser les travaux méthodologiques, notamment le tirage et le suivi des échantillons durant l'année d'enquête, puis le développement de modèles de pondération et le calcul de la précision, une fois les données plausibilisées disponibles. Ces travaux ont été menés pour la première partie par M. Jean-Pierre Renfer qui a ensuite accompagné M. Eric Graf en charge du développement des travaux relatifs à l'extrapolation. Les travaux préparatoires relatifs au plan d'échantillonnage ont été menés par M. Christian Vonlanthen de la section de la santé (GES) qui a largement contribué aux méthodes utilisées dans les enquêtes précédentes et à leur documentation.

Que soient ici remerciées les personnes ayant pris une part active dans les discussions et la collaboration avec METH: M. Roland Calmonte responsable de l'ESS au sein de la section GES ainsi que Mmes Marilina Galati et Karin Sätteli.

Toute notre gratitude va également à Paul André Salamin (METH) pour ses remarques avisées après une lecture approfondie de ce rapport.

Résumé

L'ESS02 a pour objectif principal de fournir, à partir d'un échantillon de près de 20'000 personnes résidant en Suisse et âgées de 15 ans et plus, des résultats sur l'état de santé de la population, les comportements en matière de santé, l'assurance-maladie et le recours aux prestations du système de santé.

Ce rapport décrit les méthodes utilisées pour établir le plan d'échantillonnage en abordant les questions relatives à la délimitation de la population, au cadre de sondage et aux unités d'échantillonnage. Le tirage, la gestion et l'érosion des échantillons font également partie du chapitre consacré à l'échantillonnage.

L'ESS se compose en fait de deux enquêtes: la première est une enquête téléphonique alors que la seconde, sous forme écrite, ne s'adresse qu'aux personnes ayant répondu à la première. La pondération relative à chacune de ces enquêtes y est décrite dans les deux chapitres suivants.

Enfin les méthodes pour calculer la précision des résultats obtenus sont décrites et des exemples de calculs d'intervalles de confiance sont fournis.

Mots-clé

rapport de méthodes; cadre de sondage; intervalle de confiance; Jackknife; non-réponse; plan d'échantillonnage; pondération; stratification; winsorisation.

Complément d'information: Eric Graf, tél. 032 713 64 01
Eric.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: 338-0030

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 2005
La reproduction est autorisée, sauf à des fins commerciales,
si la source est mentionnée.

ISBN: 3-303-00314-9

Table des matières

1 Introduction **5**
1.1 Délimitation de la population 5
1.2 Cadre de sondage et unités d'échantillonnage 5
1.3 Déroulement de l'enquête et types d'interview 6

2 Plan d'échantillonage et suivi des échantillons **7**
2.1 Stratification 7
2.2 Tailles initiales des échantillons par vague 8
2.3 Tirage des échantillons et adaptation de leurs tailles 10
2.4 Réalisation des interviews 12
2.5 Erosion des échantillons et raisons de non-participation 12

3 Pondération de l'enquête téléphonique **14**
3.1 Pondération initiale 14
3.2 Forme des estimateurs 18
3.3 Le calage sur marge 19
3.4 Choix des variables de calage 20
3.5 Pondération finale 21

4 Pondération de l'enquête écrite **24**
4.1 Mécanisme de non réponse 24
4.2 Modèle retenu pour la pondération initiale 25
4.3 Détermination des variables de calage 26
4.4 Winsorisation des poids 28
4.5 Effets de la pondération initiale et finale 31

5 Estimation de la précision **32**
5.1 Intervalles de confiance 32
5.2 L'estimation jackknife de la variance 33
5.3 Présentation des macros `SAS` utilisant le jackknife 34
5.4 Exemples de calculs d'intervalles de confiance 37
5.5 Fiabilité des résultats et limites d'exploitation 38

6 Conclusions **42**

A Tailles des échantillons bruts et nets par grandes régions, par strates et par vagues **43**

B Les effets du calage sur marge **48**

C Tests relatifs au paramètre réplicat des macros jackknife **57**

1 Introduction

1.1 Délimitation de la population

Après les enquêtes de 1992 et 1997, l'enquête suisse sur la santé menée durant l'année 2002 (ESS02) a permis pour la troisième fois consécutive de fournir des résultats aussi bien sur l'état de santé de la population que sur les comportements en matière de santé ou sur le recours aux prestations du système de santé suisse. La taille de cette enquête par échantillonnage en termes d'interviews réalisées a évolué au cours du temps, notamment du fait que certains cantons aient demandé et financé un échantillon supplémentaire à des fins d'analyse au niveau cantonal. Pour l'ESS02, parmi les 26 cantons et demi-cantons, 16 ont fait cette démarche (contre 7 en 1992 et 5 en 1997). Le nombre d'interviews réalisées a ainsi passé de 15'000 en 1992, à 13'000 en 1997 et 20'000 en 2002. Comme pour les deux enquêtes précédentes, la population (cible) que l'on désire étudier est constituée par l'ensemble des **personnes âgées de 15 ans et plus résidant de manière permanente en Suisse**. A noter que les personnes vivant dans des homes pour personnes âgées, dans les prisons ou d'autres institutions collectives ne font pas partie de la population étudiée.

1.2 Cadre de sondage et unités d'échantillonnage

Pour mener les enquêtes par échantillonnage auprès des ménages et des personnes en Suisse, l'OFS dispose d'un registre téléphonique (SRH) qui est réactualisé par une livraison trimestrielle de la part de Swisscom Directories. Il contient les numéros de téléphone des ménages inscrits dans ce registre et fournit ainsi le cadre de sondage pour le tirage d'échantillons de ménages privés. Une fois le ménage contacté, ses caractéristiques peuvent être déterminées (comme la taille du ménage, l'âge, le sexe ou la nationalité de ses membres) et un échantillon aléatoire de personnes de la population cible (âgée de 15 ans ou plus) peut être interrogé pour répondre au questionnaire. Bien que la plupart des personnes âgées de 15 ans ou plus résidant de manière permanente en Suisse vivent dans des ménages privés disposant d'un raccordement téléphonique inscrit dans ce registre, le recouvrement entre la population cible et le cadre de sondage n'est pas parfait. Il convient de traiter les effets de ces défauts de couverture (comme ceux liés à la non-réponse au questionnaire) par des méthodes appropriées permettant de réduire les risques de biais. Dans cette enquête, le recours au calage sur marges à partir de la statistique annuelle sur la population (ESPOP) s'est avéré judicieux.
Le fait de disposer d'un cadre de sondage conduisant à des ménages privés ne permet pas de procéder directement à un échantillonnage de personnes. Il faut donc considérer un tirage aléatoire à deux niveaux. Au **premier niveau**, l'unité d'échantillonnage est le ménage privé (l'unité de tirage correspond au raccordement téléphonique du ménage). On parle dans ce cas d'**unités primaires (ménages)**. Pour chaque ménage de l'échantillon tiré au premier niveau, on procède à un tirage aléatoire d'une (et une seule !) personne âgée de 15 ans ou plus. Il s'agit du **deuxième niveau** dont les éléments sont appelées **unités secondaires (personnes)**.

1.3 Déroulement de l'enquête et types d'interview

De manière à pouvoir bénéficier de la réactualisation du registre téléphonique chaque trimestre, l'échantillon a été tiré et l'enquête réalisée en quatre vagues successives correspondant aux quatre saisons. En effet, l'échantillon de la première vague (janvier, février, mars 2002) a été tiré sur la base du registre réactualisé en décembre 2001, alors que les échantillons des trois vagues suivantes l'ont été sur la base du registre réactualisé en mars, juin et septembre 2002 respectivement. La réalisation sur le terrain de l'ESS02 a été confiée à l'institut IHA-GfK (après l'institut LINK en 1997 et IPSO en 1992) dont les centres à Lucerne et à Lausanne ont permis de gérer les questionnaires en trois langues (allemand, français et italien).
Les techniques d'enquête utilisées sont les mêmes que pour les enquêtes précédentes. En effet, l'enquête se déroule en **deux phases** : dans la **première phase**, l'échantillon de ménages a été contacté par téléphone après avoir reçu quelques jours auparavant une lettre les invitant à participer. Une personne cible (âgée de 15 ans ou plus) a ensuite été tirée aléatoirement pour répondre à l'**enquête téléphonique** [1]. La **seconde phase** consiste en un **questionnaire écrit** envoyé uniquement aux personnes ayant répondu à l'enquête téléphonique (les personnes ayant réalisé une interview Proxy n'ont cependant pas la possibilité d'y participer). Contrairement à l'enquête de 1992 où le questionnaire écrit n'avait été envoyé qu'aux personnes âgées de moins de 75 ans, en 1997 et en 2002 aucune limite d'âge n'a été fixée.
Dans la première phase, les trois types d'interview suivants ont été utilisés pour récolter les informations :

1. **Interview par téléphone** CATI ("Computer Assisted Telephone Interview").
2. **Interview face-à-face** (pour les personnes de plus de 75 ans incapables de faire une interview CATI, méthode CAPI ("Computer Assisted Personal Interview").
3. **Interview CATI-Proxy** [2] (une tierce personne répond pour la personne cible).

Le tableau 1 fournit en moyenne les tentatives de contact et la durée par vague des interviews CATI (y compris PROXY) et CAPI. On peut notamment en déduire la durée totale des interviews qui représente plus de 1500 jours de travail (8h./jour) !

	CATI Tentatives de contact	CATI personnes interrogées	CATI durée moyenne	CAPI personnes interrogées	CAPI durée moyenne
1ère vague	8.2	4651	40 minutes	60	45 minutes
2ème vague	9.4	5046	37 minutes	55	46 minutes
3ème vague	9.2	5004	39 minutes	48	47 minutes
4ème vague	8.7	4800	36 minutes	42	43 minutes
Total	8.7	19'501	38 minutes	205	45 minutes

TAB. 1 – **Tentatives de contact et durées des interviews CATI et CAPI**

[1] Une minorité de l'échantillon a pu choisir de répondre au questionnaire en face-à-face.

[2] Une interview PROXY a lieu si la personne cible est absente et inatteignable pendant plus de 4 semaines ou qu'elle ne peut s'exprimer dans une des trois langues d'interview ou que ses capacités ne lui permettent pas de comprendre et répondre correctement.

2 Plan d'échantillonage et suivi des échantillons

2.1 Stratification

Les échantillons de ménages privés dans l'ESS02 ont été tirés pour chacune des vagues selon un plan d'échantillonnage stratifié par canton. Cette stratification cantonale se justifie notamment par le fait que la plupart des cantons ou demi-cantons ont demandé d'augmenter la taille de l'échantillon à des fins d'analyse au niveau cantonal. A noter que les deux demi-cantons d'Appenzel AI et AR (qui n'ont pas demandé une augmentation de la taille de leur échantillon) ont été regroupés, portant ainsi le nombre de strates à 25. Les cantons avaient ainsi la possibilité d'augmenter la taille de l'échantillon dans leur strate en finançant le nombre d'interviews réalisés (au minimum 400) de manière à disposer globalement pour leur canton d'une taille nette estimée à 1000 interviews pour le questionnaire téléphonique (tableau 2).

CANTON	Taille nette prévue	Interviews réalisées	Taux de réponse ESS97	Taille brute estimée	Taille brute effective
AG	1000	1016	68%	1479	1557
AI/AR	284	280	71%	402	452
BE	1588	1606	66%	2407	2523
BL	1000	1019	67%	1483	1619
BS	1000	1017	61%	1635	1670
FR	1000	1005	70%	1436	1680
GE	1000	1017	62%	1617	1886
GL	256	257	64%	401	451
GR	329	328	47%	701	668
JU	658	650	64%	1021	1162
LU	1000	1021	73%	1369	1653
NE	691	696	73%	951	1107
NW	191	197	64%	301	331
OW	192	190	64%	301	301
SG	1000	1013	63%	1584	1626
SH	254	255	64%	401	384
SO	1000	1019	64%	1556	1604
SZ	269	273	67%	401	489
TG	404	405	67%	601	598
TI	1510	1529	62%	2450	2838
UR	272	262	68%	401	429
VD	1021	1042	69%	1481	1847
VS	1000	970	56%	1786	1833
ZG	1000	1009	66%	1526	1648
ZH	1590	1630	66%	2406	2512
Total	19'508	19'706	65%	30'098	32'868

Tab. 2 – **Comparaison des tailles estimées et réelles de l'échantillon**

Le tableau 2 fournit globalement sur l'ensemble des quatre vagues la comparaison pour chaque canton (strate) entre les tailles nettes prévues initialement et les interviews effectivement réalisées. Pour cela, les adresses nécessaires (tailles brutes) ont été initialement estimées à partir des taux de réponse de la dernière enquête (ESS97) en divisant les tailles nettes prévues par les taux de réponse correspondant de l'ESS97. Les tailles brutes effectivement utilisées figurent également dans le tableau 2 et leur calcul est décrit dans les deux paragraphes suivants.

2.2 Tailles initiales des échantillons par vague

Avant le début de l'enquête, il s'agit d'estimer les tailles brutes des différentes vagues et notamment de la vague 1 à partir des éléments à disposition. En l'occurence, pour chaque canton, la taille brute initiale par vague a simplement été obtenue en divisant par 4 celle pour l'ensemble de l'année du tableau 2 (les arrondis a été traités de manière conservatrice en rajoutant une unité au besoin). Considérons les temps t_0 et t_1 correspondant respectivement aux périodes avant l'enquête et après la première vague et les notations suivantes :

$\widehat{m}_{total}^{n,t_0}$: taille nette estimée avant l'enquête pour l'ensemble de l'année ;

$\widehat{m}_{v_l}^{n,t_0}$: taille nette estimée avant l'enquête pour la vague $l = 1, 2, 3, 4$;

$m_{v_1}^{n,t_1}$: taille nette réalisée après la 1ère vague ;

Les notations suivantes sont utilisées pour les tailles brutes : $\widehat{m}_{total}^{b,t_0}$, $\widehat{m}_{v_l}^{b,t_0}$ et $m_{v_1}^{b,t_1}$.

Initialement, on a donc :

$$\widehat{m}_{v_l}^{n,t_0} = \frac{1}{4}\widehat{m}_{total}^{n,t_0} \texttt{ et } \widehat{m}_{v_l}^{b,t_0} = \frac{1}{4}\widehat{m}_{total}^{b,t_0}, l = 1, 2, 3, 4$$

Après la 1ère vague, on peut ainsi calculer la différence (bilan) entre les interviews effectivement réalisées et prévues : $m_{v_1}^{n,t_1}$-$\widehat{m}_{v_1}^{n,t_0}$. Dans la pratique, on tire en plus de la taille brute initiale un échantillon de réserve dont une partie peut être activée (envoi d'une lettre à l'avance, suivie quelques jours plus tard d'une interview téléphonique). En effet, lors de chaque vague, les lettres à l'avance sont envoyées en deux demi-vagues, dont la taille de la seconde peut être légèrement adaptée en fonction des interviews déjà réalisées dans la première demi-vague. On obtient ainsi après la première vague les résultats par strate du tableau 3 en termes de tailles brutes initiales, effectives et interviews réalisées de même que la différence entre la taille réalisée et prévue. A noter qu'après la première vague, les taux de réponse de l'enquête en cours peuvent être recalculés (par strate) de la manière suivante :

$$t_{v_1}^{rep,t_1} = \frac{m_{v_1}^{n,t_1}}{m_{v_1}^{b,t_1}}$$

Les taux de réponse de la 1ère vague peuvent désormais être utilisés pour l'estimation des tailles brutes de la 2ème vague. Après la deuxième vague, on peut intégrer l'information disponible de l'ensemble des deux premières vagues pour estimer les tailles brutes de la 3ème vague et se baser sur l'information à disposition des trois premières vagues pour finalement obtenir une estimation relativement précise de la taille de la 4ème et dernière vague comme décrit dans le prochain paragraphe.

CANTON	Taille brute initiale	Taille brute effective	Interviews réalisées	Bilan : Interviews réalisées-prévues
AG	371	398	249	-1
AI/AR	102	101	58	-13
BE	603	624	387	-10
BL	372	464	274	+24
BS	410	445	264	+14
FR	360	433	251	+1
GE	406	414	227	-23
GL	101	101	55	-9
GR	176	175	84	+2
JU	256	283	136	-28
LU	343	404	239	-11
NE	239	260	167	-6
NW	76	74	39	-9
OW	76	75	42	-6
SG	398	426	243	-7
SH	101	103	71	+8
SO	391	394	243	-7
SZ	101	101	51	-16
TG	151	151	91	-10
TI	613	630	332	-45
UR	101	101	53	-15
VD	371	370	240	-15
VS	447	438	248	-2
ZG	383	435	263	+13
ZH	602	649	404	+7
Total	7'550	8'049	4'711	-164

TAB. 3 – **Vague 1 : tailles brutes initiales, effectives et interviews réalisées**

2.3 Tirage des échantillons et adaptation de leurs tailles

La gestion des échantillons n'est pas triviale et doit être conduite avec le plus grand soin de manière à satisfaire au mieux le but recherché. Dans notre cas, il s'agit de garantir un nombre minimum d'interviews réalisées réparties aussi uniformément que possible sur les quatre vagues de l'enquête, sans pour autant sacrifier les taux de réponse en activant un nombre trop élevé d'adresses pour cela. Un avantage indéniable de procéder en quatre vagues en tirant un nouvel échantillon chaque trimestre durant une année est de pouvoir gérer les tailles brutes de ces échantillons en les adaptant en fonction de l'information obtenue sur les vagues précédentes. En fait, il s'agit de considérer les temps t_0, t_1, t_2, t_3 et t_4 correspondant respectivement aux périodes avant l'enquête, après la 1ère, 2ème, 3ème vague et 4ème vague. Au temps t_0, c'est-à-dire avant le début de l'enquête, seule l'information de la dernière enquête (ESS97) est disponible pour estimer la taille brute de l'échantillon à partir de la taille nette (nombre prévu d'interviews réalisées). Cependant, dès la fin de la première vague, les estimations peuvent être successivement affinées en répartissant le solde des interviews réalisées sur le reste des vagues. Les estimations des tailles des vagues restantes peuvent désormais être calculées, et ceci pour chaque canton, en considérant les notations suivantes :

$\widehat{m}_{v_l}^{n,t_q}$: taille nette estimée en $t_q = t_0, t_1, t_2, t_3$ pour la vague $l = 1, 2, 3, 4$ $(l \geq q + 1)$.

$m_{v_l}^{n,t_q}$: taille nette réalisée en $t_q = t_1, t_2, t_3, t_4$ pour la vague $l = 1, 2, 3, 4$ $(l = q)$.

$\Delta_{v_l}^{t_l} = \widehat{m}_{v_l}^{n,t_{l-1}} - m_{v_l}^{n,t_l}$: solde (interviews estimées - réalisées) en t_l de la vague $l \geq 1$.

Après la 1ère vague : $\widehat{m}_{v_l}^{n,t_1} = \widehat{m}_{v_2}^{n,t_0} + \dfrac{1}{3}\Delta_{v_1}^{t_1}, l = 2, 3, 4$;

Après la 2ème vague : $\widehat{m}_{v_l}^{n,t_2} = \widehat{m}_{v_3}^{n,t_1} + \dfrac{1}{2}\Delta_{v_2}^{t_2}, l = 3, 4$;

Après la 3ème vague : $\widehat{m}_{v_l}^{n,t_3} = \widehat{m}_{v_4}^{n,t_2} + \dfrac{1}{1}\Delta_{v_3}^{t_3}, l = 4.$

L'estimation des tailles brutes s'obtenant en divisant les tailles nettes estimées par les taux de réponse recalculés sur la base des derniers éléments connus.
Pour illustrer les formules décrites dans ce paragraphe, nous présentons les calculs effectués pour l'adaptation des tailles nettes du canton d'Argovie.

Avant l'enquête : Comme l'indique la première ligne du tableau 2, la taille nette prévue pour l'ensemble de l'année est de 1000 observations, on a donc initialement :

$\widehat{m}_{v_l}^{n,t_0} = \dfrac{1}{4}\widehat{m}_{total}^{n,t_0} = 250, l = 1, 2, 3, 4.$

Après la 1ère vague :

$m_{v_1}^{n,t_1} = 249$;
$\Delta_{v_1}^{t_1} = \widehat{m}_{v_1}^{n,t_0} - m_{v_1}^{n,t_1} = 250 - 249 = 1$;

D'où : $\widehat{m}_{v_l}^{n,t_1} = \widehat{m}_{v_2}^{n,t_0} + \frac{1}{3}\Delta_{v_1}^{t_1} = 250 + \frac{1}{3}(1) \cong 250, l = 2,3,4\,;$

Après la 2ème vague :

$m_{v_2}^{n,t_2} = 270\,;$
$\Delta_{v_2}^{t_2} = \widehat{m}_{v_2}^{n,t_1} - m_{v_2}^{n,t_2} = 250 - 270 = -20\,;$
D'où : $\widehat{m}_{v_l}^{n,t_2} = \widehat{m}_{v_3}^{n,t_1} + \frac{1}{2}\Delta_{v_2}^{t_2} = 250 + \frac{1}{2}(-20) = 240, l = 3,4\,;$

Après la 3ème vague :

$m_{v_3}^{n,t_3} = 250\,;$
$\Delta_{v_3}^{t_3} = \widehat{m}_{v_3}^{n,t_2} - m_{v_3}^{n,t_3} = 240 - 250 = -10\,;$
D'où : $\widehat{m}_{v_l}^{n,t_3} = \widehat{m}_{v_4}^{n,t_2} + \frac{1}{1}\Delta_{v_3}^{t_3} = 240 + 1(-10) = 230, l = 4\,;$

Après la 4ème vague :

$m_{v_4}^{n,t_4} = 247\,;$

En adaptant ainsi successivement les tailles de l'échantillon, on obtient une taille finale proche de celle prévue initialement, en l'occurence pour le canton d'Argovie au lieu des 1000 observations prévues on a : 249+270+250+247=1016.
Le tableau 4 fournit les résultats des estimations successives pour les quatre vagues de l'enquête.

VAGUE	Taille initiale	Adaptation selon vague 1	selon vague 2	selon vague 3	Taille effective	Interviews réalisées
1ère vague	7'550	8'049	8'049	8'049	8'049	4'711
2ème vague	7'550	8'144	8'266	8'266	8'266	5'101
3ème vague	7'550	8'144	8'258	8'308	8'308	5'052
4ème vague	7'550	8'144	8'258	8'308	8'245	4'842
Total	30'200	32'481	32'831	32'773	32'868	19'706

TAB. 4 – **tailles brutes initiales, adaptées et effectives par vague**

2.4 Réalisation des interviews

La réalisation des interviews a été confiée aux enquêtrices et enquêteurs qui ont reçu une formation de base avant le début de l'enquête, complétée par des séances d'instructions durant toute la durée de l'enquête. Dans la première phase (enquête téléphonique auprès d'une personne cible tirée aléatoirement dans le ménage de l'échantillon), l'interview était réalisée (une fois la lettre d'invitation à participer à l'enquête envoyée) selon le type d'interview le plus approprié. Les personnes ayant participé à cette première phase, pour autant qu'il ne s'agissait pas d'une interview Proxy (une tierce personne répond pour la personne cible), ont toutes été invitées à participer à la deuxième phase, celle du questionnaire écrit. Pour cela, les personnes qui ne renvoyaient pas le questionnaire écrit dans le mois qui suivait leur interview CATI ou CAPI, recevaient un premier rappel par courrier accompagné d'un nouveau questionnaire. Sans nouvelle deux semaines après ce premier rappel, un second rappel, téléphonique cette fois, a été mené pour réduire autant que possible la non-réponse au questionnaire écrit.

2.5 Erosion des échantillons et raisons de non-participation

Pour les quatre vagues de l'enquête, l'OFS a livré au total 32'868 adresses à l'institut de sondage. L'étude de l'érosion des échantillons est particulièrement importante et se situe à différents niveaux dans l'enquête : au niveau des adresses elles-mêmes qui ne sont pas parfaites (701 adresses se sont avérées être invalides et 1'338 ne pas correspondre à des ménages privés (maisons de vacances, ménages collectifs, entreprises)), puis au niveau des ménages privés (qui ne peuvent ou ne veulent pas répondre, en l'occurence 7'668), puis au niveau des personnes cible de l'enquête téléphonique (dont 3'455 n'ont pas participé) et enfin au niveau de l'enquête par écrit (avec 2'618 pertes supplémentaires). Cette érosion de l'échantillon est résumée dans le tableau 5 avec les pourcentages correspondants à la base y relative.
A partir des informations contenues dans le tableau 5, il est possible de définir différents taux dont l'interprétation permet de quantifier la qualité des résultats obtenus. On peut ainsi définir les *taux de validité*, *taux d'éligibilité*, *taux de participation des ménages*, *taux de participation des personnes (tél.)* et *taux de participation des personnes (écrit)* suivants :

$$t_v = \text{taux de validité} = \frac{\texttt{adresses valides}}{\texttt{total des adresses}} = \frac{32167}{32868} \cong 98\% \,;$$

$$t_e = \text{taux d'éligibilité} = \frac{\texttt{ménages privés}}{\texttt{adresses valides}} = \frac{30829}{32167} \cong 96\% \,;$$

$$t_p^M = \text{taux de participation (ménages)} = \frac{\texttt{ménages participants}}{\texttt{ménages privés}} = \frac{23161}{30829} \cong 75\% \,;$$

$$t_{p,tel}^P = \text{taux de participation (tél.)} = \frac{\texttt{personnes cible participantes}}{\texttt{ménages participants}} = \frac{19706}{23161} \cong 85\% \,;$$

$$t_{p,ecrit}^P = \text{taux de participation (écrit)} = \frac{\texttt{personnes cible participantes}}{\texttt{personnes participantes CATI+CAPI}} = \frac{16141}{18759} \cong 86\%.$$

Echantillon brut (Lettres envoyées)	32'868	100%					
Adresses contactées (Classification en ménages privés)	32'167	98%	100%				
Ménages privés	30'829	94%	96%	100%			
Ménages privés participants	23'161	70%	72%	75%	100%		
Personnes cible participantes (tél.)	19'706	60%	61%	64%	85%	100%	
Personnes cible interviewées en CATI et CAPI	18'759	57%	58%	61%	81%	95%	100%
Personnes cible participantes (écrit)	16'141	49%	50%	52%	70%	82%	86%

TAB. 5 – **Erosion de l'échantillon au niveau des ménages et des personnes**

Les raisons de non-participation sont détaillées dans le tableau 6, aussi bien au niveau des ménages que des personnes pour l'enquête téléphonique.

	Unités primaires **ménages**		Unités secondaires **personnes cible**	
TOTAL	**7'668**	100%	**3'455**	100%
Pas d'intérêt	1'993	26.0%	796	23.0%
Longue absence	1'586	20.7%	827	23.9%
Contre les enquêtes	1'229	16.0%	573	16.6%
Autres raisons	950	12.4%	646	18.7%
Problèmes de santé	798	10.4%	129	3.7%
Problèmes de langue	548	7.1%	111	3.2%
Pas de temps	365	4.8%	296	8.6%
Motifs familiaux	110	1.4%	39	1.1%
Pas de réponse	89	1.2%	38	1.1%

TAB. 6 – Raisons de non-participation au niveau des unités primaires et secondaires

3 Pondération de l'enquête téléphonique

De manière à tenir compte à la fois du plan d'échantillonnage et de la non-réponse, une pondération initiale a été effectuée. Ces poids initiaux comportent l'information sur les probabilités d'inclusion et une première correction pour la non-réponse. A partir de cette pondération initiale, le recours à des données auxiliaires de référence a permis de procéder à un calage fournissant la pondération finale utilisée dans le calcul des différents estimateurs. Ce chapitre est consacré au développement du modèle de pondération de l'enquête téléphonique depuis la pondération initiale jusqu'à la pondération finale en passant par la forme des estimateurs, le calage sur marge et le choix des variables de calage.

3.1 Pondération initiale

Chaque individu ayant participé à l'enquête se voit attribuer un poids initial qui n'est rien d'autre que le produit de l'inverse de sa probabilité d'inclusion et du taux de réponse par strate. Nous distinguerons ainsi les *poids d'échantillonnage* des *poids dûs à la non-réponse* dont le produit fournit les *poids initiaux*. Pour cela, fixons les notations suivantes.
Soit $U = \{1, ..., k, ..., N\}$ la population des personnes de 15 ans et plus résidant dans des ménages privés en Suisse. Cette population est partitionnée en M unités primaires (ménages) notées $U_1, ..., U_i, ..., U_M$. L'ensemble des unités primaires (la population des ménages privés) est notée par $U^M = \{1, ..., i, ..., M\}$. La taille d'un ménage U_i (le nombre de personnes de 15 ans ou plus qu'il contient) est désignée par N_i. Le tableau 7 résume ces notations.

k	indice utilisé pour désigner des personnes
i	indice utilisé pour désigner des ménages
U_i	ensemble des personnes d'un même ménage
N_i	taille du ménage U_i

Tab. 7 – Notations.

Comme dans cette enquête il s'agit d'un tirage aléatoire stratifié à deux niveaux pour chacune des 4 vagues, on peut établir, sans restreindre la généralité, la pondération initiale pour une strate et une vague fixée. En effet, on se restreint ainsi au cas d'un tirage aléatoire simple aux deux niveaux suivants :

Premier niveau : Un échantillon s^M de ménages est tiré dans U^M ($s^M \subset U^M$) selon le plan $p_M(\cdot)$.

Deuxième niveau : Dans chaque ménage $i \in s^M$, on tire un échantillon s_i^P de taille 1 (une personne), $s_i^P \subset U_i$, selon le plan $p_i(\cdot|s^M)$.

La probabilité d'inclusion π_k d'un individu k dans l'échantillon s'écrit :

$$\begin{aligned} \pi_k &= Pr(k \in s^P) \\ &= Pr(i \in s^M) \cdot Pr(k \in U_i | i \in s^M) \\ &= \pi_i^M \cdot \pi_{k|i} \\ &= \pi_i^M \cdot \frac{1}{N_i} \qquad \text{avec } k \in U_i. \end{aligned} \tag{1}$$

Le poids d'échantillonnage attribué à la personne k est inversément proportionnel à la probabilité qu'a celle-ci d'appartenir à l'échantillon :

$$w_k^E = \frac{1}{\pi_k} = \frac{N_i}{\pi_i^M} \qquad \text{avec } k \in U_i \tag{2}$$

Cependant, toutes les personnes (respectivement ménages) de l'échantillon ne participent pas à l'enquête pour différentes raisons évoquées au chapitre 2.
Si τ_k est le taux de réponse ou plus exactement la probabilité de participer à l'enquête (conditionnellement au fait que la personne k a été tirée dans l'échantillon), $\tau_k = Pr(k \text{ participe } | k \in s^P)$, alors le poids dû à la non réponse w_k^{NR} est donné par

$$w_k^{NR} = \frac{1}{\tau_k}$$

Finalement, on obtient le poids initial noté w_k^{IN} :

$$w_k^{IN} = w_k^E \cdot w_k^{NR} = \frac{1}{\pi_k \cdot \tau_k} \tag{3}$$

Pour obtenir le jeu de poids initiaux, il convient d'introduire les notations suivantes (tableau 8). Le développement ci-dessus est valable pour chaque strate. Nous introduisons ici l'indice h qui désigne la strate.

$s_h^{b,M}$	Échantillon <u>brut</u> de ménages dans la strate h, de taille m_h^b
$s_h^{n,M}$	Échantillon <u>net</u> de ménages dans la strate h, de taille m_h^n (interviews réalisées)
$s_h^{b,P}$	Échantillon <u>brut</u> de personnes dans la strate h
M_h	Nombre de ménages dans la population de la strate h.

TAB. 8 – Notations pour la pondération initiale de l'ESS02.

Dans chaque strate (canton), un sondage aléatoire simple à deux niveaux a été effectué. La probabilité d'inclusion π_i^M qu'un ménage soit dans l'échantillon est la même pour tous les ménages au sein d'une même strate :

$$\pi_i^M = Pr(i \in s_h^{b,M})$$

La probabilité d'inclusion π_i^M des ménages de la strate h est donnée par :

$$\pi_h^M i = \frac{m_h^b}{M_h} \qquad \text{avec } i \in \text{strate } h$$

On peut ainsi en déduire le poids d'échantillonnage de l'individu k du ménage U_i dans la strate h qui est proportionnel au nombre de personnes éligibles du ménage U_i :

$$w_{k,h}^E = \frac{M_h}{m_h^b} \cdot N_i \qquad \text{avec } k \in U_i \text{ et } i \in \text{strate } h \tag{4}$$

En faisant une hypothèse d'homogénéité de la non réponse par strate aussi bien au niveau des ménages que des personnes, on a :

$$\tau_k = \tau_h, \forall k \in s_h^{b,P}$$

Comme une seule personne par ménage a participé à l'enquête, on peut calculer τ_h par

$$\tau_h = \frac{m_h^n}{m_h^b}$$

On obtient ainsi les poids dus à la non réponse par strate.
Remarque : Une seule personne par ménage peut participer à l'enquête, de ce fait, nous nous permettons de comparer le m_h^b qui représente en fait un nombre d'adresses de *ménages* par strate au m_h^n qui représente, par strate toujours, un nombre de *personnes* qui ont répondu au questionnaire téléphonique :

$$w_h^{NR} = \frac{1}{\tau_h} = \frac{m_h^b}{m_h^n} \tag{5}$$

Ce poids ne dépend pas de l'individu k ni de son ménage U_i, il dépend du nombre d'unités primaires (ménages) existantes dans la strate h et du nombre de personnes (cf. remarques ci-dessus) qui ont participé parmi celles qui ont été sélectionnées pour l'enquête.

Finalement, le poids initial par strate est donné par :

$$w_{k,h}^{IN} = w_{k,h}^{E} \cdot w_{k,h}^{IN} = \frac{M_h}{m_h^b} N_i \cdot \frac{m_h^b}{m_h^n} = \frac{M_h}{m_h^n} N_i \qquad \text{avec } k \in U_i \text{ et } i \in \text{strate } h \qquad (6)$$

Afin de ne pas alourdir inutilement le texte nous dénoterons désormais par N_k la taille du ménage de la personne k ($N_k = N_i$ si $k \in U_i$). Le poids initial se notera alors :

$$w_{k,h}^{IN} = \frac{M_h}{m_h^n} N_k$$

Le poids initial $w_{k,h}^{IN}$ d'un individu k dépend :
- de la strate h (canton) dans laquelle il/elle se trouve,
- du nombre d'unités primaires (ménages) existantes dans cette strate (M_h),
- du nombre d'unités primaires sélectionnées et répondantes dans la strate h (m_h^n)
- du nombre d'unités secondaires par unité primaire N_k, c'est-à-dire du nombre de personnes éligibles dans le ménage de k.

On peut interpréter la somme des poids initiaux en considérant le quotient $\frac{M_h}{m_h^n}$ qui par expansion fournit une estimation du nombre de ménages par canton. La multiplication par N_k fournissant une estimation du nombre de personnes de 15ans+ vivant en Suisse.
Rappelons que 4 sous-échantillons, un par saison, définissent l'échantillon total. Le calcul des poids initiaux a été mené pour chaque vague séparément. Ce qui a pour effet que la somme $\sum_i w_{k,h}^{IN}$ des poids sur les quatre vagues sera 4 fois trop grande. Les poids initiaux désirés ont été obtenus en considérant la moyenne des poids initiaux sur les quatres vagues.

Dans les données de l'ESS02, N_k s'appelle `tkont03`. Le tableau 9 fournit pour l'échantillon des répondants à l'enquête téléphonique le nombre de ménages avec $1, 2, ..., q$ personnes de 15+ ans.
Rappelons encore que le registre dans lequel on a "tiré"[3] les adresses est mis à jour trimestriellement. Il est donc inévitable d'avoir quelques cas de déménagements entre les strates dans notre échantillon stratifié. Ces mouvements de population ont

[3]En fait on tire des numéros de téléphone dans le SRH (StichprobenRegister für Haushalterhebungen).

tkont03 $= N_k$	Fréquence	%	Fréquence cumulée	%
1	5635	28.60	5635	28.60
2	10392	52.74	16027	81.33
3	2158	10.95	18185	92.28
4	1195	6.06	19380	98.35
5	269	1.37	19649	99.71
6	46	0.23	19695	99.94
7	8	0.04	19703	99.98
8	2	0.01	19705	99.99
9	1	0.01	19706	100.00

TAB. 9 – N_k : nbr. de personnes de 15ans+ vivant dans les ménages ayant participé à l'enquête téléphonique.

été pris en compte de la manière suivante : le calcul des poids initiaux a été fait en gardant la strate initiale. L'exploitation des données a cependant été faite dans les cantons où habitaient les personnes au moment de l'interview. Ainsi les données de l'ESS02 contenaient deux variables pour le canton :

- La variable `Kanton` correspond au canton dans lequel la personne interviewée habite au moment de l'interview,
- La variable `Kanton_1` correspond au canton dans lequel la personne habite (habitait) selon le plan d'échantillonnage.

Les tableaux 22 à 28 de l'annexe A donnent ces tailles par strate et par vague.

3.2 Forme des estimateurs

Soit $p(\cdot)$ un plan d'échantillonnage fixé. L'inclusion d'un élément k dans l'échantillon est un événement aléatoire décrit par la variable aléatoire I_k définie comme suit :

$$I_k = \begin{cases} 1 & \text{si } k \in s \\ 0 & \text{sinon} \end{cases}$$

Notons que $I_k = I_k(s)$ est une fonction de l'échantillon s, nous appellerons I_k l'indicatrice d'appartenance à l'échantillon de l'élément k.
La probabilité d'inclusion (π_k) d'ordre 1 associée à l'individu k, i.e. la probabilité que l'individu k se trouve dans un échantillon sélectionné est donnée par

$$\pi_k = Pr(k \in s) = Pr(I_k = 1) = \sum_{k \in s} p(s)$$

où $k \in s$ signifie qu'il s'agit de la somme sur les échantillons s qui contiennent l'individu k. De manière analogue, la probabilité pour que deux individus k et l soient inclus dans un échantillon est dénotée par π_{kl} et se dérive de $p(\cdot)$ comme suit :

$$\pi_{kl} = Pr(k\&l \in s) = Pr(I_k \cdot I_l = 1) = \sum_{k\&l \in s} p(s).$$

Si $k = l$ on obtient que $\pi_{kk} = \pi_k$.
L'*estimateur Horvitz-Thompson*, appelé également π-estimateur, pour un total de la variable d'intérêt Y est défini et sa variance obtenue dans le tableau ci-dessous :

	valeur prise sur la population U	estimateur Horvitz-Thompson
Total	$t = \sum_{k \in U} y_k$	$\hat{t}_\pi = \sum_s \frac{y_k}{\pi_k} = \sum_U \frac{y_k}{\pi_k} I_k$
Variance	$V(\hat{t}) = \sum \sum_U (\frac{\pi_{kl}}{\pi_k \pi_l} - 1) y_k y_l$	$\hat{V}(\hat{t}) = \sum \sum_s \frac{1}{\pi_{kl}} (\frac{\pi_{kl}}{\pi_k \pi_l} - 1) y_k y_l$

A noter que le π-estimateur est sans biais puisque

$$E[\hat{t}_\pi] = E[\sum_U I_k \frac{y_k}{\pi_k}] = \sum_U \frac{y_k}{\pi_k} E[I_k] = \sum_U \frac{y_k}{\pi_k} \pi_k = \sum_U y_k = t.$$

3.3 Le calage sur marge

Après avoir effectué une pondération initiale (sous l'hypothèse d'homogénéité de la non-réponse par strate est uniforme et qu'elle ne dépend d'aucun paramètre en particulier), on souhaite apporter une correction à ces poids de telle manière que leur somme satisfasse certaines valeurs connues de la population.

Le but d'une correction à posteriori comme le *calage sur marges* est d'améliorer l'estimation en diminuant les biais dus à la non-réponse. Toutes les informations disponibles simultanément sur l'échantillon et la population peuvent alors être utilisées.
Rappelons ici brièvement la technique du calage sur marges en introduisant les notations suivantes :
Soient U la population de N individus et Y une variable d'intérêt pour laquelle on désire estimer le total sur la population (avec le π-estimateur par exemple). Soient encore $X_1, X_2, ..., X_J$ J variables connues sur l'échantillon s et dont on connaît les totaux sur la population :

$$t_{X_j} = \sum_{k \in U} x_{jk}.$$

Ces variables sont appelées des *variables auxiliaires*. Les totaux sur la population sont une information dont on aimerait tenir compte. Pour cela, on cherche à estimer le total $\hat{t}_Y$ de Y à l'aide d'un estimateur de la forme :

$$\hat{t}_{Y_w} = \sum_{k \in s} w_k y_k,$$

où les poids w_k sont proches (dans un sens à préciser) des poids initiaux w_k^{IN}, et vérifient les *contraintes de calage* :

$$\sum_{k \in s} w_k x_{jk} = t_{X_j} \ , \ j = 1, ..., J.$$

On cherche donc un estimateur "peu différent" de l'estimateur de Horvitz-Thompson qui *cale* l'échantillon sur les totaux des variables auxiliaires.

Avant de réaliser le calage, il faut choisir une *fonction de distance* G, d'argument w_k/w_k^{IN}, pour mesurer les distances entre w_k et w_k^{IN}. La fonction G doit vérifier les conditions suivantes : elle doit être positive et convexe avec $G(1) = G'(1) = 0$.
Une fois la fonction G choisie, le problème consiste à déterminer les poids $w_k (k \in s)$ solutions du problème d'optimisation sous contraintes suivant (en notant les vecteurs $x'_k = (x_{1k}, ..., x_{Jk})$ et $t'_X = (t_{X_1}, ..., t_{X_J})$) :

$$\min_{w_k} \sum_{k \in s} w_k^{IN} G(\frac{w_k}{w_k^{IN}}) \text{ sous la contrainte } \sum_{k \in s} w_k x_k = t_X. \tag{7}$$

On cherche ainsi à minimiser une somme pondérée (par les w_k^{IN}) des "distances" entre les poids initiaux w_k^{IN} et les pondérations recherchées w_k, sous les contraintes de calage.
La fonction de distance utilisée ici est celle correspondant à la méthode *raking ratio* et les résultats sont obtenus par l'algorithme *iterative proportional fitting* (*Sautory, 2004*).

3.4 Choix des variables de calage

Le choix des variables auxiliaires X_j retenues pour effectuer le calage sur marge est basé sur les considérations suivantes : ces variables doivent être étroitement corrélées aux variables d'intérêt. On espère ainsi que le lien entre variables de calage et variable d'analyse fera que les corrections apportées sur les variables de calage se répercuteront sur la variable d'analyse. De plus, les X_j doivent expliquer au mieux la probabilité de réponse.

Remarquons qu'il est très important de posséder des informations (auxiliaires) récentes sur la population. En effet, si une grande évolution avait lieu, le fait de caler l'échantillon selon une information périmée introduirait un biais et il est dans ce cas plus judicieux de ne pas faire de correction avec un calage sur marges.

Pour définir les marges de calage nous avons besoin d'une source d'information sur la population. Le fichier ESPOP (Evaluation de l'évolution de la population selon les recensements, état fin 2001) a été utilisé comme population. Ce fichier ESPOP nous donne le nombre de personnes vivant en Suisse pour tous les croisements possibles entre les variables sexe (2 valeurs), nationalité (2 valeurs), canton (26 valeurs) et âge (de 0 à 99+, 100 valeurs). Il s'agit de l'information la plus fiable sur la population suisse disponible au moment du développement des modèles de pondérations.
Les critères de calage pour la pondération des résultats du questionnaire téléphonique en 2002 sont les suivants :
- le sexe : homme, femme ;
- la nationalité : suisse, étranger ;
- le lieu : voir tableau 10 ;
- les classes d'âge : voir tableau 10.

Le calage sur marge garantit que le nombre d'hommes et de femmes des données pondérées de notre enquête soit égal au nombre d'hommes et de femme de +15 ans vivant effectivement en Suisse (d'après les données de l'ESPOP pour fin 2001).
Il en va de même pour le nombre d'étrangers et de suisses, ainsi que pour certaines classes d'âge et régions.

3.5 Pondération finale

Idéalement on aimerait caler sur les 4 variables en même temps :

```
SEX x NATIO x AGE x LIEU
```

Ainsi les données pondérées redonneraient vraiment, si on les sommait, par exemple, le nombre d'hommes étrangers entre 35 et 45 ans vivant dans le canton de St Gall. Malheureusement la taille de notre échantillon est trop petite pour avoir une taille suffisante dans chaque cellule. Il est *nécessaire* que le nombre d'observations par cellule soit suffisamment grand pour assurer que le rôle de l'unité individuelle soit négligeable vis-à-vis de l'ensemble. Dans la pratique, il est raisonnable de considérer des cellules avec au moins une trentaine d'observations .
En calant sur les quatre critères en même temps on aurait $2 \cdot 2 \cdot 6 \cdot 18 = 432$ cellules. Mais nous avons par exemple aucune femme étrangère de plus de 70 ans dans le canton d'Argovie dans les données récoltées en 2002.

On doit donc caler en deux étapes :

On aurait presque assez de valeur par cellules (par rapport au critère d'un mi-

le lieu (18 modalités)	
lieu=1	KANTON_1=1 (Zurich)
lieu=2	KANTON_1=2 (Berne)
lieu=3	KANTON_1=3 (Lucerne)
lieu=9	KANTON_1=9 (Zoug)
lieu=10	KANTON_1=10 (Fribourg)
lieu=11	KANTON_1=11 (Soleure)
lieu=12	KANTON_1=12 (Bâle-ville)
lieu=13	KANTON_11=13 (Bâle-campagne)
lieu=17	KANTON_1=17 (St Gall)
lieu=19	KANTON_1=19 (Argovie)
lieu=21	KANTON_1=21 (Tessin)
lieu=22	KANTON_1=22 (Vaud)
lieu=23	KANTON_1=23 (Valais)
lieu=24	KANTON_1=24 (Neuchâtel)
lieu=25	KANTON_1=25 (Genève)
lieu=26	KANTON_1=26 (Jura)
lieu=100	KANTON_1=8(Glarus), 14(Schaffhouse), 15(Appenzell Rh. Ext.), 16(Appenzell Rh. Int.), 18(Grisons)), 20(Turgovie) la région "Suisse orientale" sans le canton de Zoug
lieu=200	KANTON_1=4(Uri), 5(Schwyz), 6(Obwald), 7(Nidwald) la région "Suisse centrale" sans le canton de Lucerne.
les classes d'âge (6 modalités)	
âge=1	15-24 ans
âge=2	25-34 ans
âge=3	35-44 ans
âge=4	45-59 ans
âge=5	60-69 ans
âge=6	70 ans et plus.

TAB. 10 – Définition des variables de calage "âge" et "lieu" dans la pondération finale des résultats de l'enquête téléphonique.

nimum de 30) pour caler d'abord sur NATIO puis sur SEX x AGE x LIEU [4]. Cela signifie pour la deuxième étape $2 \cdot 6 \cdot 18 = 216$ valeurs de calage. Malheureusement on a quand-même encore 4 cellules délicates contenant respectivement que 30, 30, 29 et 23 observations. Le cas est discutable, mais nous avons préféré opter pour la

[4]Notons que nous avons constaté durant les tests de modèles pour le calage sur marges en utilisant la macro *%CALMAR* dans SAS que cette dernière ne supporte pas plus de 99 variables de calage. On ne peut donc pas calculer directement un tel modèle sur la base de 216 variables indicatrices. Il est nécessaire d'utiliser la macro *%INTCAL* qui nous crée, par exemple, une seule variable de calage avec 216 marges.

solution exposée ci-après.

Pour cela, on a donc d'abord ajusté les totaux pour

`NATIO x LIEU` (étape 1),

puis ajusté pour

`SEXE x AGE` (étape 2).

Ce qui ne garantit pas l'exactitude des totaux de `NATIO` et `LIEU` ; par contre les totaux sont exacts pour `SEX` et `AGE`. On pourrait choisir de caler d'abord sur `SEX x AGE` puis sur `NATIO x LIEU`. On aurait alors des totaux exacts pour `NATIO` et `LIEU` mais plus pour `SEX` et `AGE`... Les valeurs pour le sexe et les classes d'âge étant plus importantes dans une enquête sur la santé que celles pour la nationalité et la région, on a choisit la première variante pour calculer les poids finaux et définitifs du questionnaire téléphonique.
Les graphiques 4 et 5 de l'annexe A montrent que les poids finaux se comportent d'une manière semblable par rapport aux poids initiaux que l'on cale d'abord sur `NATIO x LIEU` et ensuite sur `SEX x AGE` (→ poids finaux NK SA) ou l'inverse (→ Poids finaux SA NK).
Le tableau 11 montre que le calage sur marges menant aux poids finaux choisit (NK SA), que nous désignerons désormais par $w_{k,h}^{FIN}$, n'a pas apporté de corrections trop importantes. En effet si on définit le facteur de correction C_k par

$$C_k = \frac{w_{k,h}^{FIN}}{w_{k,h}^{IN}},$$

on voit que sa moyenne, sa médiane et son mode sont proches de 1 avec une variance empirique faible.

Moyenne	1.0	Minimum	0.6
Médiane	0.9	Maximum	2.7
Mode	0.8	Intervalle interquartile	0.3
Variance	0.1	qt. 99%	2.0
		qt. 95%	1.6
		qt. 5%	0.8
		qt. 1%	0.7

Tab. 11 – Caractéristiques des rapports C_k.

Le tableau 12 donne les statistiques des poids initiaux et finaux du questionnaire téléphonique. En comparant les caractéristiques des poids initiaux à celles des poids finaux, on voit que le calage sur marges a fait augmenter la variance empirique des poids. L'effet est plus prononcé sur la droite de la queue de distribution des poids, en particulier au-delà du 99^e percentile. Le rapport entre le plus grand et le plus

petit des poids finaux est presque d'un facteur 100. Malgré cela nous n'avons pas winsorisé les poids w^{FIN} comme cela a été fait pour le poids finaux du questionnaire écrit ; en effet, une telle opération apportait de trop grandes distorsion dans les biais (comparaison de totaux avec les valeurs connues de l'ESPOP).

Poids initiaux w^{IN}			
Somme	5957661	Intervalle interquartile	247.2
Moyenne	302.32	qt. 99%	1060.2
Médiane	246.6	qt. 95%	701.8
Mode	493.2	qt. 5%	74.9
Ecart-type	222.0	qt. 1%	38.1
Minimum	37.1	Max/Min	53.9
Maximum	2000.1	qt. 99%/qt. 1%	27.8
		qt. 95%/qt. 5%	9.4
Poids finaux w^{FIN}			
Somme	6017638	Intervalle interquartile	265.1
Moyenne	305.37	qt. 99%	1217.7
Médiane	230.3	qt. 95%	780.1
Mode	98.8	qt. 5%	69.1
Ecart-type	247.0	qt. 1%	35.8
Minimum	29.7	Max/Min	101.5
Maximum	3011.2	qt. 99%/qt. 1%	34.0
		qt. 95%/qt. 5%	11.3

Tab. 12 – Statistiques des poids initiaux w^{IN} et finaux w^{FIN} de l'enquête téléphonique.

4 Pondération de l'enquête écrite

4.1 Mécanisme de non réponse

Dans l'ESS on a d'une part un échantillon relatif à l'enquête téléphonique d'autre part celui du relevé écrit. Ces deux échantillons sont liés, puisque tous les individus ayant participé au relevé écrit ont dans un premier temps été interviewés par téléphone. Ainsi, comme le montre la tableau 13, nous avons de la non réponse possible à plusieurs niveaux.

Au niveau de la pondération, ceci induit deux poids :

1. w^{FIN} le poids final issu d'un calage sur marges pour l'échantillon du questionnaire téléphonique,
2. w_e^{FIN} le poids final obtenu après un calage sur marges pour l'échantillon du relevé écrit.

cadre de sondage (numéros de téléphone)
↓
échantillon brut de ménages
↓
échantillon brut de personnes
↓
échantillon net de personnes relatif au questionnaire téléphonique
↓
échantillon net de personnes relatif au questionnaire écrit

TAB. 13 – Le mécanisme de la non réponse du cadre de sondage à l'échantillon net du questionnaire écrit.

Pour les enquêtes réalisées en 1997 et 2002, l'échantillon du relevé écrit est considéré comme échantillon provenant d'un sondage aléatoire simple de l'échantillon du questionnaire par téléphone. On peut ensuite ajuster l'échantillon écrit de manière assez fine grâce à un calage sur l'enquête téléphonique.

4.2 Modèle retenu pour la pondération initiale

Pour la pondération initiale de l'échantillon provenant de l'enquête écrite 2002 nous avons testé 3 modèles :
- un modèle dit de non-réponse ignorable qui est celui qui a été retenu,
- le modèle simple,
- un modèle LOGIT modélisant la non-réponse.

Modèle 1 : non-réponse ignorable

Il s'agit d'ajouter une étape intermédiaire supplémentaire par rapport à ce qu'il s'est fait en 1997 : afin de compenser l'érosion entre les deux échantillons, on réajuste les sommes des poids par strates (cantons) pour obtenir les totaux de la population des personnes par strate. Cela revient à considérer que le probabilité de réponse est uniforme par strate. Pour cela, on détermine un facteur f_h pour chaque strate tel que :

$$f_h \cdot \sum_{k \in s_{e_h}^{n,P}} w_{k,h}^{FIN} = N_h \ , \forall h,$$

avec $s_{e_h}^{n,P}$ l'échantillon net de personnes de la strate h ayant participé à l'enquête écrite.
Notons que $f_h \geq 1$, pour tout h puisque $s_{e_h}^{n,P} \subseteq s_h^{n,P}$. Les poids initiaux du questionnaire écrit sont donc donnés par :

$$we_{k,h}^{IN} = f_h \cdot w_{k,h}^{FIN} \ , \forall h, \forall k \in s_{e_h}^{n,P}.$$

En procédant ainsi, la variance empirique des poids finaux a légèrement diminué par rapport au modèle 2.

Modèle 2 : modèle simple

Comme en 1997, on prend comme poids initiaux simplement les poids finaux $w_{k,h}^{FIN}$ de l'enquête téléphonique.
Les analyses de biais (voir Annexe B) des poids finaux écrits qu'on obtient par la suite montrent que ce modèle plus simple que le premier donne de moins bons résultats. Nous y avons donc renoncé.

Modèle 3 : modèle LOGIT

Une tentative de modélisation de la non-réponse entre le questionnaire écrit et le questionnaire téléphonique a été faite de manière à tenir compte des caractéristiques (vue, santé, âge, nationalité, lieu, ...) des non-répondants au questionnaire écrit pour calculer les poids initiaux écrits (basés sur les poids finaux téléphoniques). La procédure `logistic` de SAS a été utilisée pour cela.
Les analyses de biais (voir Annexe B) des poids finaux qu'on obtient par la suite montrent que ce modèle donne de moins bons résultats que les deux premiers. Nous y avons donc renoncé.

4.3 Détermination des variables de calage

Les critères de calage de l'échantillon écrit sont les suivants :
1. nationalité × niveau de formation achevée
2. sexe × classes d'âge
3. santé autoévaluée
4. handicap de la vue

On motive ce choix par le fait que les individus ayant des problèmes de vue seront moins enclins à remplir un questionnaire écrit. D'autre part, il est ressorti des enquêtes précédentes que l'attitude vis-à-vis de la santé ainsi que l'état de santé sont des facteurs explicatifs importants de la non-réponse. En particulier, les personnes jugeant leur état de santé comme étant mauvais ont un taux de réponse plus faible que les individus se sentant subjectivement bien.

En calculant le nombre de personnes de l'échantillon provenant du questionnaire écrit satisfaisant aux critères définis par les croisements des variables `SEXE` (2 modalités), `AGE` (6 modalités), `TSUBG01` (5 modalités), `TSBHD03` (4 modalités), `NATIO` (2 modalités) et `BILDUN4` (4 modalités), on arrive à la conclusion que, sous la condition d'avoir au moins une trentaine de personnes par cellule, l'ordre cité plus haut est la meilleure manière de faire les calages sur marges. Le tableau 14 donne des détails

sur les variables auxiliaires.

`Tsubg01`	Comment allez-vous en ce moment ?	1= "Très bien" 2= "bien" 3= "Comme ci, comme ça (moyen)" 4= "mal" 5= "très mal"
`Bildun4`	Formation achevée	0= sans formation, indéfini 1= scolarité obligatoire 2= degré secondaire 3= degré tertiaire
`Age`	Classes d'âge	1= 15 à 24 ans 2= 25 à 34 ans 3= 35 à 44 ans 4= 45 à 59 ans 5= 60 à 69 ans 6= 70 ans et +
`Tsbhd03`	Voyez-vous suffisamment bien pour lire un livre ou un journal ?	1= "oui, sans difficulté" 2= "oui, sans trop de difficulté" 3= "oui, mais avec beaucoup de difficulté" 4= "non"

TAB. 14 – Variables de calage pour l'échantillon provenant du questionnaire écrit.

Imputations

Les variables `TSUBG01` et `TSBHD03` sont sujettes à quelques non réponses partielles :
- `TSUBG01` : 4 non-réponses dans le questionnaire téléphonique (il s'agit de personnes n'ayant pas répondu à cette question précise mais qui ont pris part à l'enquête écrite) ;
- `TSBHD03` : 77 personnes n'y ont pas répondu ($77/19706 = 0.39\%$) pour la partie téléphonique dont 59 ont rempli un questionnaire écrit.

Or un calage sur marge n'est possible que s'il n'y a aucune non-réponse parmi les variables considérées. Pour parer à ce problème, une solution possible est d'imputer aléatoirement des valeurs à ces données manquantes en respectant la distribution qu'on a des données connues. C'est ce que l'on a fait. On a ensuite procédé au calage. Remarquons encore que la variable construite `BILDUN4` fournie dans le fichier indicateurs 2002 a 25 valeurs manquantes. En 1997 un '0' ($\equiv$ 'sans formation, indéfini') avait été attribué aux non réponses de ce style. On a procédé de la même manière pour 2002, ainsi `BILDUN4` ne contenait plus de valeurs manquantes pour le calage.
Les poids écrits après calages sont dénotés par w_e^{CAL}.

4.4 Winsorisation des poids

Après avoir effectué les calages sur marges, il s'est avéré qu'il existe une grande disproportionnalité entre les poids obtenus.
En effet, $max_{k,h}(w_{e_{k,h}}^{CAL}) \cong 170 \cdot min_{k,h}(w_{e_{k,h}}^{CAL})$. Cette situation peut conduire à des problèmes importants lors des exploitations du fait que la variabilité des estimateurs aura tendance à augmenter. Dans le tableau 15 on constate que la variance des poids w_e^{CAL} a augmenté de 22% par rapport à celle des poids w_e^{IN}. On va donc corriger une dernière fois le jeu de poids. On effectue une coupure au 2ème et au 98ème percentiles sur l'ensemble des poids w_e^{CAL} :

$$w_e^{COUP} = \begin{cases} p_2 & \text{si } w_e^{CAL} \leq p_2 \\ w_e^{CAL} & \text{si } p_2 < w_e^{CAL} < p_{98} \\ p_{98} & \text{si } w_e^{CAL} \geq p_{98} \end{cases}$$

Avec

$$A = \{w | w_e^{CAL} \leq p_2\},\ B = \{w | p_2 < w_e^{CAL} < p_{98}\} \text{ et } C = \{w | w_e^{CAL} \geq p_{98}\},$$

on a

$$\begin{aligned} \sum_{A \cup B \cup C} w_{e_{k,h}}^{COUP} &= \sum_A w_{e_{k,h}}^{COUP} + \sum_B w_{e_{k,h}}^{CAL} + \sum_C w_{e_{k,h}}^{COUP} \\ &= \sum_A p_2 + \sum_B w_{e_{k,h}}^{CAL} + \sum_C p_{98} \\ &= |A| \cdot p_2 + \sum_B w_{e_{k,h}}^{CAL} + |C| \cdot p_{98} \\ &\neq N = \sum_{A \cup B \cup C} w_{e_{k,h}}^{CAL}. \end{aligned}$$

Afin de corriger cela, on multiplie les poids w_e^{COUP} par le facteur $f^{WIN} = \frac{N}{\sum w_e^{COUP}}$. Les nouveaux poids sont donc donnés par

$$w_e^{FIN} = f^{WIN} \cdot w_e^{COUP}.$$

Ce sont les poids finaux pour l'échantillon provenant de l'enquête écrite.
Cette winsorisation des poids a provoqué une légère perturbation sur les valeurs des marges que nous ne corrigeons pas.
La figure 1 donne une vue de la répartition des poids après calages w_e^{CAL} (poids encore non winsorisés). Les percentiles valent $p_2 = 59.9$ et $p_{98} = 1320.6$. La figure 2 illustre les poids finaux w_e^{FIN}.

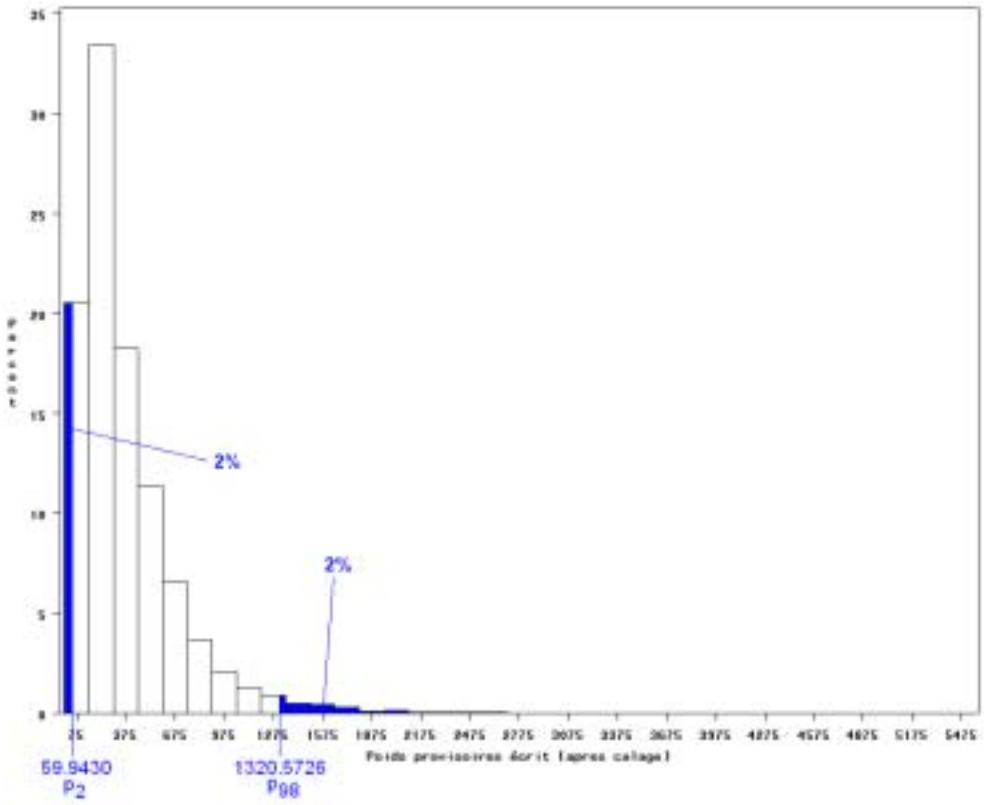


FIG. 1 – Histogramme des poids du questionnaire écrit w_e^{CAL} après calages, mais avant la winsorisation. On note la longue queue de la distribution.

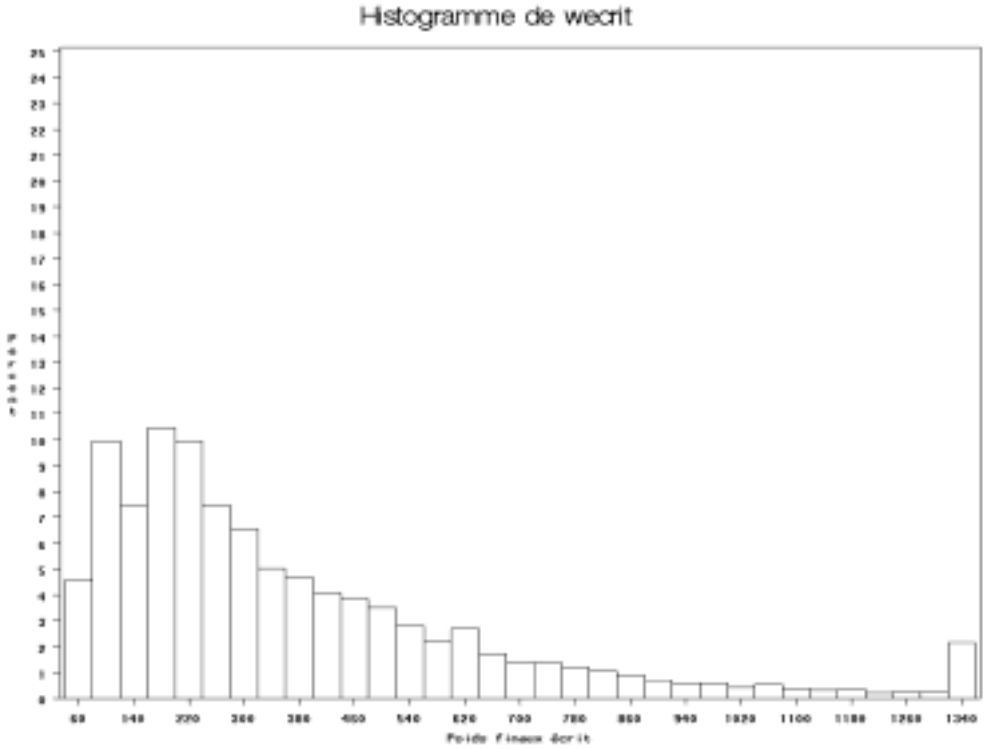


FIG. 2 – Histogramme des poids du questionnaire écrit w_e^{FIN} après winsorisation.

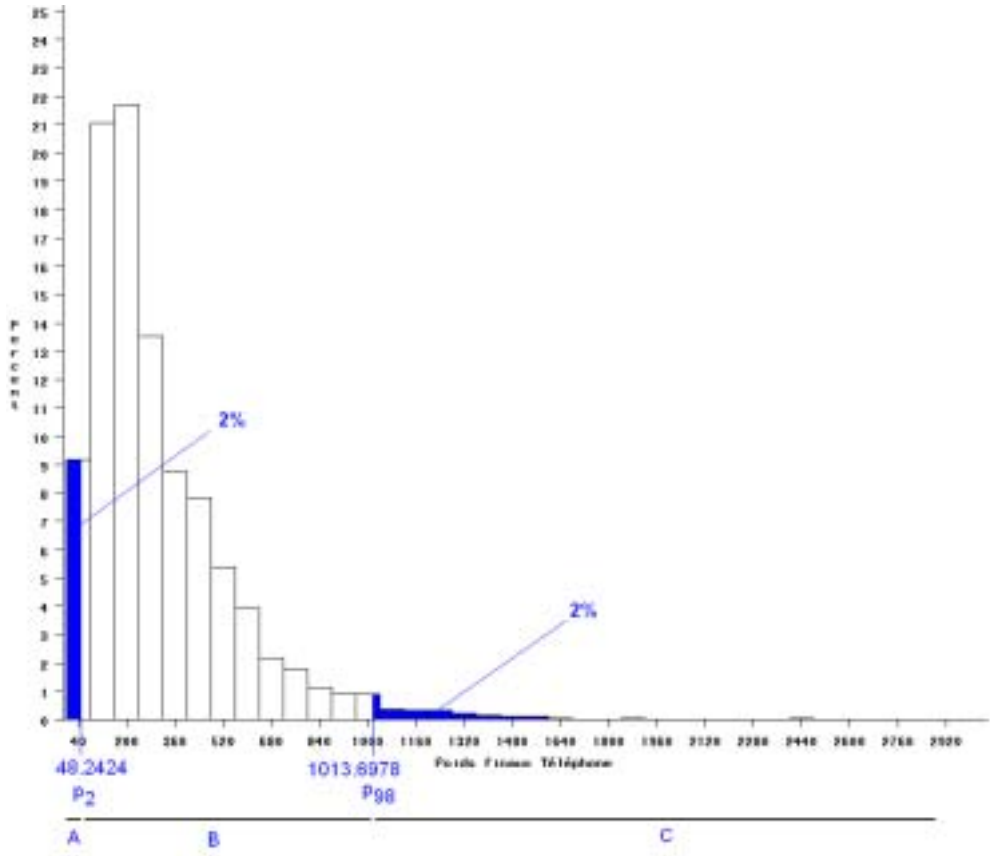


FIG. 3 – Histogramme des poids finaux de l'enquête téléphoniques w^{FIN}.

A noter que dans le cas du questionnaire téléphonique, une tentative de winsorisation a été menée, sans que cette dernière se soit avérée indiquée. En effet, La figure 3 illustre la répartition des poids finaux du questionnaire téléphonique. Dans ce cas les percentiles valent $p_2 = 48.2$ et $p_{98} = 1013.7$. Bien que les distorsions ne soient que peu affectées par le sectionnement des queues de la distribution, certains biais pour les totaux sur les variables de calage deviennent très grands (jusqu'à 69% de la valeur estimée). Nous ne pouvons pas nous permettre de si grands biais même si cela nous réduirait l'écart-type des poids qui passerait de 247 à 226.
Notons qu'en 1997 on n'a non plus pas mené de winsorisation sur les poids du questionnaire téléphonique.

4.5 Effets de la pondération initiale et finale

Le tableau 15 donne un apercu des caractéristiques principales des trois étapes de la pondération de l'échantillon provenant du l'enquête écrite.

Poids écrits initiaux w_e^{IN}			
Somme	6017638	Intervalle interquartile	317.9
Moyenne	372.82	qt. 99%	1448.4
Médiane	282.1	qt. 95%	945.9
Mode	513.6	qt. 5%	86.6
Ecart-type	296.0	qt. 1%	44.8
Minimum	36.8	Max/Min	103.0
Maximum	3796.0	qt. 99%/qt. 1%	32.3
		qt. 95%/qt. 5%	10.9
Poids écrit après calages w_e^{CAL}			
Somme	6017638	Intervalle interquartile	308.1
Moyenne	372.82	qt. 99%	1613.1
Médiane	274.4	qt. 95%	973.9
Mode	211.1	qt. 5%	80.2
Ecart-type	327.5	qt. 1%	42.2
Minimum	31.7	Max/Min	171.7
Maximum	5470.3	qt. 99%/qt. 1%	38.2
		qt. 95%/qt. 5%	12.2
Poids écrit finaux w_e^{FIN}			
Somme	6017638	Intervalle interquartile	315.8
Moyenne	372.82	qt. 99%	1353.5
Médiane	281.3	qt. 95%	998.1
Mode	61.4	qt. 5%	82.1
Ecart-type	287.2	qt. 1%	61.4
Minimum	61.4	Max/Min	22.0
Maximum	1353.5	qt. 99%/qt. 1%	22.0
		qt. 95%/qt. 5%	12.2

TAB. 15 – Statistiques des poids initiaux, poids après calages et poids finaux (winsorisés) du questionnaire écrit. On note l'augmentation de la variabilité provoquée par le calage, augmentation que l'on contrôle grâce à la winsorisation.

5 Estimation de la précision

5.1 Intervalles de confiance

Soit un paramètre θ (par exemple l'âge moyen ou l'effectif d'une sous-population) de la population qu'il s'agit d'estimer à l'aide de notre échantillon .

On appelle intervalle de confiance tout intervalle construit autour d'un estimateur ayant une certaine probabilité de contenir la valeur de θ.

Pour construire un intervalle de confiance qui contienne la valeur du paramètre θ avec une probabilité donnée, il s'agit de résoudre une équation de la forme :

$$P(B_{inf} \leq \theta \leq B_{sup}) = 1 - \alpha$$

où

θ est le paramètre à estimer,
B_{inf} la borne inférieure de l'intervalle,
B_{sup} la borne supérieure de l'intervalle ;
$1 - \alpha$: la probabilité donnée (niveau de confiance).

La probabilité α mesure le risque d'erreur de l'intervalle, soit la probabilité que l'intervalle ne contienne pas la vraie valeur du paramètre θ.
Dans la pratique, on cherche à calculer $\widehat{\theta}$, l'estimateur de θ. Si l'estimateur $\widehat{\theta}$ suit (au moins approximativement) une loi normale, on peut définir un intervalle de confiance I_α à $(1 - \alpha)\%$ ainsi :

$$I_\alpha = [\widehat{\theta} - z_{1-\alpha/2} \cdot \widehat{\sigma}(\widehat{\theta}); \widehat{\theta} + z_{1-\alpha/2} \cdot \widehat{\sigma}(\widehat{\theta})]$$

où $\widehat{\sigma}(\widehat{\theta})$ représente l'estimateur de l'écart-type associé à l'estimateur $\widehat{\theta}$ de θ et $z_{1-\alpha/2}$ le $\alpha/2$-quantile de la loi normale. Si le risque de première espèce est de 5% ($\alpha = 5\%$), on retrouve l'intervalle de confiance à 95% $[\widehat{\theta} - 1.96 \cdot \widehat{\sigma}(\widehat{\theta}); \widehat{\theta} + 1.96 \cdot \widehat{\sigma}(\widehat{\theta})]$.

Dans l'enquête suisse sur la santé, l'estimation de la variance $\hat{\sigma}^2$ d'un estimateur $\widehat{\theta}$ s'est faite à l'aide d'une méthode basée sur un certain type de *jackknife* présentée dans le prochain paragraphe.

5.2 L'estimation jackknife de la variance

Dans la littérature sur les enquêtes par échantillonnage (par exemple, *Cochran 1977)*, on trouve des formules d'estimation de la variance applicables pour plusieurs types de plans d'échantillonnage (tirages aléatoires simples, stratifiés, systématiques, à plusieurs degrés, en grappes). Ces formules d'estimation de la variance sont généralement valides lorsqu'il n'y a pas de non-réponse. Malheureusement, elles ne permettent pas de prendre en compte la variabilité des poids induite notamment par le calage.

De ce fait les estimateurs de la variance, ainsi que les intervalles de confiance obtenus par les procédures standards de logiciels statistiques tels que `SAS`, `SPSS` ou `EXCEL` sont incorrects dans notre situation si ces procédures ne tiennent pas compte des poids.
Pour cela, il faut se tourner vers d'autres méthodes d'estimations de la variance (*Särndal, 1992, Peters, 1994*) comme les méthodes de réplications dont un cas particulier est l'estimation par le jackknife de la variance. Fixons quelques notations afin de présenter le type de jackknife utilisé ici. Soit s un échantillon que l'on divise aléatoirement en g sous-échantillons $s_j, j = 1, ..., g$, de taille identique aux arrondis près, appelés *répliques de population.*

Si s est un échantillon issu d'un tirage aléatoire stratifié, chaque strate h est partitionnée en g groupes de taille identique, disons $s_{j,h}$. Ces groupes de taille $s_{j,h}$ sont appelés *répliques de strates.*

Pour chacune des g répliques de strates, on construit les g pseudo-valeurs

$$\widehat{Y}_j = g \cdot \widehat{Y} - (g-1) \cdot \widehat{Y}_{(j)}, j = 1, 2, ..., g$$

où :
- $\widehat{Y}_j$ est appelée la j-ième pseudo-valeur ;
- $\widehat{Y}$ est l'estimateur d'une variable d'étude Y ;
- $\widehat{Y}_{(j)}$ est l'estimateur de Y après avoir enlevé la j-ième réplique de population.

Pour l'estimateur du total $\widehat{Y} = \sum_{k \in S} w_k y_k$, où w_k est le poids associé à l'individu k, les $\widehat{Y}_{(j)}$ peuvent être estimés par l'approximation suivante

$$\widehat{Y}_{(j)} = \frac{\sum_{k \in s} w_k}{\sum_{k \in s \setminus s_j} w_k} \sum_{k \in s \setminus s_j} w_k y_k.$$

Pour être précis, on devrait écrire $\widehat{Y}_{(j)} = \sum_{k \in s \setminus s_j} w_k^n y_k$ où w_k^n = poids associé à l'individu k, recalculé par la même méthode de pondération que w_k, mais, cette fois, sur la base de l'échantillon $s \setminus s_j$.

On définit alors l'estimateur *jackknife* $\widehat{\bar{Y}}_{JK}$ de la variable d'étude Y par la moyenne des $\widehat{Y}_j$, c'est-à-dire

$$\widehat{\bar{Y}}_{JK} = \frac{1}{g}\sum_{j=1}^{g}\widehat{Y}_j$$

Si nous supposons que l'estimateur jackknife $\widehat{\bar{Y}}_{JK}$ est presque égal à l'estimateur $\widehat{Y}$ (ce qui est vérifié dans le cas des tirages aléatoires, avec probabilité égales ou non, avec ou sans remises), l'estimateur $var(\widehat{\bar{Y}}_{JK})$ de sa variance pourra être utilisé comme estimateur $var(\widehat{Y})$ de la variance de $\widehat{Y}$.
La méthode jackknife fait comme si les variables $\widehat{Y}_j$ $(j = 1, ..., g)$ étaient indépendantes :

$$\begin{aligned} var(\widehat{Y}) &\cong var(\widehat{\bar{Y}}_{JK}) \\ &= \frac{1}{g^2}\cdot var(\sum_{j=1}^{g}\widehat{Y}_j) = \frac{1}{g^2}\cdot\sum_{j=1}^{g} var(\widehat{Y}_j) \\ &\cong \frac{var(\widehat{Y}_j)}{g} = \frac{1}{g(g-1)}\sum_{j=1,2,...,g}(\widehat{Y}_j - \widehat{\bar{Y}}_{JK})^2 \end{aligned}$$

Un programme, écrit dans `SAS`, a été conçu afin de calculer cet estimateur jackknife de la variance (cf. annexe C).

5.3 Présentation des macros `SAS` utilisant le jackknife

Dans le cas de l'ESS, nous avons constaté après divers essais que les procédures standard (dans SAS par exemple `MEANS`, `UNIVARIATE`) ont tendance à sous-estimer l'écart-type accompagnant les estimations parce qu'elles ne tiennent pas compte de la variabilité due à la non-réponse et des traitements que nous avons appliqué aux poids (calage sur marge, winsorisation). De ce fait ces procédures ne devraient pas être utilisées car elles donnent une vue trop optimiste de la précision des estimations.
On dispose de quatre macros utilisant la méthode de réplication jackknife pour estimer la variance. Nous allons les exposer ci-après. Ces macros ont été implémentées sur SAS par Markus Eichenberger (BFI) en 1997-98 (*Eichenberger, 1997*).
Signalons encore que le prof. Davison de l'EPFL avait été mandaté par le service de méthodes statistiques de l'OFS pour comparer les différentes méthodes d'estimation de la variance (*Davison, 1998*). Les conclusions de cette expertise sont parues en juin 1998 : le bootstrap et la méthode jackknife linéarisé sont moins onéreuses en temps de calcul que d'autres méthodes de réplications. Le bootstrap et le jackknife linéarisé donnent des erreurs standards plus correctes : les méthodes standards comme l'estimation directe de la variance pouvant donner des estimations sérieusement sous-estimées ; il est aussi essentiel de prendre en compte la variabilité des poids.

Les intervalles de confiance calculés par ces macros sont par défaut des intervalles à 95% (risque α de 5%). On peut modifier le risque α en allant changer la valeur attribuée à `alpha` dans la première ligne du programme des macros (`%let alpha=0.05`).
Les paramètre renseignant ces macros sont les suivant :

- `data` : fichier de données à exploiter (obligatoire). Il contient les variables d'étude ainsi que les poids et la variable définissant les réplicats si ces paramètres sont fournis.

- `var` : la variable sur laquelle on veut les intervalles de confiance (obligatoire).

- `gewvar` : les poids (optionnel). Si ce paramètre n'est pas renseigné, chaque observation reçoit un poids de 1.
  Rappelons que le jeux de poids, ou plus précisément le modèle de pondération, a une influence sur ces estimations de la variance. En principe, pour chaque $Y_{(j)}$, on devrait procéder à une repondération. Par soucis de simplification, cette procédure n'a pas été effectuée ici et les poids sont considérés comme fixe.
  Les tests que nous avons menés sur les données de l'ESS02 montrent qu'il est essentiel de renseigner ce paramètre avec la variables des poids téléphoniques ou écrits selon les données étudiées (voir Annexe C).

- `class` : variables pour la construction de sous-groupes (par exemple, sexe, classes d'âge, etc.), on sépare ces variables de regroupement par un espace si on en utilise plusieurs (optionnel). Si ce paramètre n'est pas renseigné, l'estimation portera sur tout l'échantillon.

- `repl` : nom d'une variable contenue dans `data` pour l'identification de réplicats (optionnel).
  Par défaut, si ce paramètre n'est pas donné, la macro utilise `repl=_n_`, i.e. `repl` $= 1, 2, 3, ..., n$, n étant le nombre d'observations. Dans le cadre de l'enquête suisse sur la santé, laisser '`repl=`' revient au même que d'entrer `repl=idno` (`idno` étant le numéro d'identification personnel, il est unique pour chaque personne ayant participé à l'enquête). Dans les trois cas identiques mentionnés ci-dessus on impose le nombre maximum de réplicats qui correspond ici au nombre d'observations. Les auteurs de [*Davison, 1998*] mentionnent que la méthode d'estimation de la variance par le jackknife n'est pas la plus stable des méthodes, c'est pour cette raison qu'ils conseillent que le nombre de réplicats soit le plus élevé possible afin de réduire cette instabilité. Il est toutefois possible d'utiliser les macros avec `repl=2`, mais en pratique, de plus grands nombres semblent préférables.
  Nous avons testé les macros jackknife avec les données de l'ESS02. Ces tests montrent clairement que les macros donnent de meilleurs résultats lorsque le paramètre `repl` n'est pas renseigné. Nous recommandons donc de laisser `repl=` (vide). L'annexe C motive notre recommandation et donne des détails quant à ces tests.

Les variables `var`, `gewvar`, `class` et `repl` sont toutes supposées être numériques.

Les quatre macros proposées ci-après permettent de calculer les variances jackknife sur les estimations de proportions ou d'effectifs totaux, ainsi que sur les estimations de moyennes ou de totaux.

1. La macro jackknife "moyenne" :
`%jackmit(data=,var=,gewvar=,class=,repl=)`
jackknife pour le calcul d'intervalles de confiance sur des moyennes avec élimination de réplicats.
Le tableau 16 donne à titre d'exemple, l'output qu'on obtient lorsqu'on entre
`%jackmit(data=ESS02, var=poids_en_kg, gewvar=wght`[5] `, class= sex, repl=)`.

output `%jackmit`						
SEX	nb obs	poids moyen	erreur standard	borne inf	borne sup	coeff de variation (%)
hommes	8859	77.3	0.16	77.0	77.7	0.21
femmes	10673	63.2	0.15	62.9	63.5	0.23

TAB. 16 – Poids moyen en kilos par genre avec les bornes de l'intervalles de confiance calaculé par la macro `%jackmit`.

2. La macro jackknife "nb. d'observation" :
`%jackanz(data=,var=,gewvar=,class=,repl=)`
Méthode du jackknife pour le calcul d'intervalles de confiance sur des nombres d'observations avec élimination de réplicats.
Le tableau 17 montre un output de cette macro à titre d'exemple.

3. La Macro jackknife "proportions" :
`%jackprop(data=,var=,gewvar=,class=,repl=)`.
Méthode du jackknife pour le calcul d'intervalles de confiance sur des proportions avec élimination d'un réplicat.
Le tableau 18 montre un output de cette macro à titre d'exemple.

4. La Macro jackknife "somme" :
`%jacksum(data=,var=,gewvar=,class=,repl=)`.
Méthode du jackknife pour le calcul d'intervalles de confiance sur des totaux avec élimination d'un réplicat.
La macro `%jacksum` est surtout utilisée avec des variables indicatrices [6]. Sinon en entrant par exemple `%jacksum(data=ess,var=sex,gewvar=wght,class=,repl=)`, comme sex prend les valeurs 1 et 2, la macro calculera $1*$(somme des poids avec $sex =$

[5] Dans les données de l'ESS02, les poids finaux w^{FIN} de l'enquête téléphonique sont dans la variable dénommée $wght$.

[6] variable indicatrice : variable prenant les valeurs 0 et 1, 1 lorsqu'une propriété est satisfaite, 0 sinon.

output %jackanz						
sans les poids						
Nbr. de pièces	nb obs	moyenne	erreur standard	borne inf	borne sup	coeff de variation (%)
1	352	352	18.6	315.6	388.4	5.3
2	1702	1702	39.4	1624.7	1779.3	2.3
3	4757	4757	60.1	4639.3	4874.7	1.3
4	5799	5799	63.9	5673.7	5924.3	1.1
5	3600	3600	54.2	3493.7	3706.3	1.5
6	1855	1855	41.0	1774.7	1935.3	2.2
7	920	920	29.6	862.0	978.0	3.2
8	667	667	25.4	617.3	716.8	3.8
avec les poids						
1	352	72747.7	4742.0	63453.5	82041.9	6.5
2	1702	352044.1	10575.6	331316.6	372771.9	3.0
3	4757	1326802.8	22447.0	1282807.6	1370798.1	1.7
4	5799	1854047.1	25755.2	1803567.9	1904526.3	1.4
5	3600	1179248.0	21923.8	1136278.1	1222217.9	1.9
6	1855	625857.2	17189.6	592166.1	659548.6	2.8
7	920	326919.9	13001.9	301436.6	352403.2	4.0
8	667	258160.8	12340.6	233973.6	282348.0	4.8

TAB. 17 – **Nombre d'observations** du "Nombre de pièces dans l'appartement/maison dans le/la quel(le) vous vivez sans compter la cuisine ?" avec les bornes de l'intervalles de confiance calaculé par la macro `%jackanz` (variable `TWOVE01`).

1) + 2 ∗ (somme des poids avec $sex = 2$) ce qui n'a pas beaucoup de sens.
Autrement, on peut utiliser la variable 'un' (1 partout) et mettre une condition pour `class`, `%jacksum(data=ess,var=un,gewvar=wght,class=sex,repl=)`, cela revient à utiliser la macro `%Jackanz` avec `var=sex` et rien pour `class`.

5.4 Exemples de calculs d'intervalles de confiance

Considérons la question relative au poids (en Kg) de la personne interviewée.
Après avoir estimé la variance à l'aide de la méthode du jackknife, on obtient les intervalles de confiance à 95% présentés dans le tableau 16.
L'intervalle à 95% pour le poids moyen des hommes de plus de 15 ans en Suisse est donné par [77.0,77.7] et celui du poids moyen des femmes par : [62.9, 63.5].

Passons maintenant à la question relative à l'état subjectif de santé de la personne interviewée.
Toujours après avoir estimé la variance à l'aide de la méthode du jackknife, on obtient les intervalles de confiance à 95% indiqués dans le tableau 19.

output %jackprop						
sans les poids						
Nbr. de pièces	nb obs	moyenne	erreur standard	borne inf	borne sup	coeff de variation
1	352	1.8	0.1	1.6	2.0	5.3
2	1702	8.7	0.2	8.3	9.1	2.3
3	4757	24.2	0.3	23.6	24.8	1.3
4	5799	29.5	0.3	28.9	30.2	1.1
5	3600	18.3	0.3	17.8	18.9	1.5
6	1855	9.4	0.2	9.0	9.9	2.2
7	920	4.7	0.2	4.4	5.0	3.2
8	667	3.4	0.1	3.1	3.7	3.8
avec les poids						
1	352	1.2	0.1	1.1	1.4	6.5
2	1702	5.9	0.2	5.5	6.2	3.0
3	4757	22.1	0.4	21.4	22.9	1.7
4	5799	30.9	0.4	30.1	31.8	1.4
5	3600	19.7	0.4	19.0	20.4	1.9
6	1855	10.4	0.3	9.9	11.0	2.8
7	920	5.4	0.2	5.0	5.9	4.0
8	667	4.3	0.2	3.9	4.7	4.8

TAB. 18 – **Proportions** pour "Nombre de pièces dans l'appartement/maison dans le/la quel(le) vous visez sans compter la cuisine ?" avec les bornes de l'intervalles de confiance calaculé par la macro `%jackprop` (variable `TWOVE01`).

L'intervalle à 95% pour la proportion des hommes se sentant très bien est est donné par [23.1%, 26.2%] et celui des femmes par [61.5%, 64.8%].
On peut aussi soutenir, avec un risque d'erreur de 5%, que la différence, entre hommes et femmes, de pourcentage ayant répondu "très bien" à la question relative à l'état subjectif de santé **n'est pas** statistiquement significative. Par contre on peut dire, avec un risque d'erreur de 5%, que le nombre de femmes se sentant moyennement bien est significativement plus élevé que le nombre d'hommes se sentant de même.
Remarquons encore qu'en 1997, la proportion d'hommes se sentant très bien était estimée à 24.2% contre 22.4% en 2002, mais cette différence n'est pas statistiquement significative.

5.5 Fiabilité des résultats et limites d'exploitation

Nous allons terminer ce rapport par quelques remarques dont il faudrait toujours tenir compte lors d'exploitations et d'analyses des données.
Les pondérations qui ont été développées sont surtout faites pour estimer des proportions de la population possédant telle ou telle caractéristique d'après les questions

	tsubg01	%	erreur standard	borne inf	borne sup	coeff de variation (%)
hommes	Très bien	24.7	0.6	23.5	25.8	2.4
	bien	63.2	0.7	61.9	64.4	1.0
	moyen	9.1	0.4	8.4	9.8	4.2
	mal	2.6	0.2	2.2	3.0	8.1
	très mal	0.5	0.1	0.3	0.6	17.1
femmes	Très bien	22.5	0.5	21.4	23.5	2.3
	bien	61.5	0.6	60.4	62.7	1.0
	moyen	12.3	0.4	11.5	13.1	3.2
	mal	3.4	0.2	2.9	3.8	6.8
	très mal	0.4	0.1	0.3	0.5	16.5

TAB. 19 – Intervalle de confiance à 95% pour l'état subjectif de santé.

posées dans l'ESS. Elle sont moins faites pour estimer des totaux, d'autant moins lorsqu'on s'intéresse à des totaux de variables faiblement corrélés aux variables de calage.
Les pondérations sont faites pour des exploitations utilisant la totalité des observations (téléphoniques ou écrites), i.e. tous les répondants qu'ils aient ou non répondu à la ou les questions des variables étudiée. Formellement lorsqu'on étudie des variables filtrées (des questions posées qu'à une partie des répondants), on doit toujours tenir compte également des personnes qui ont été filtrées de manière à ce qu'on ait constamment en tout 19706 personnes pour le questionnaire téléphonique, 16141 pour le questionnaire écrit. Ces pondérations ne sont pas idéales pour estimer des totaux sur des sous-populations.

Toute étude de proportions, totaux, moyennes, etc. devrait toujours être accompagnée d'études de variances de manière à être conscient de l'imprécision qui accompagne les nombres que l'on publie. Terminons par un exemple qui permet bien de juger de la prudence dont il faut faire preuve :
La première partie du tableau 19 donne une estimation du nombre total de personnes en Suisse d'après les données CATI pondérées de l'ESS02 pour chacune des classes d'âge utilisées dans les tableaux standards. La remarque cruciale est que pas toutes ces classes d'âge correspondent aux classes d'âge selon lesquelles on a procédé aux calages dans les modèles de pondérations. Les classes 1,2,3 sont les mêmes, mais les définitions des classes 4,5,6,7 diffèrent. La deuxième partie du tableau donne les "vrais" nombres tirés de l'ESPOP01 pour la population. Pour les classes 1,2,3 on a les mêmes nombres puisque c'est sur ceux-là qu'on a calé. Par contre les estimations de totaux des classes 4 à 7 diffèrent des nombres de l'ESPOP, malgré les calages. Les nombres de l'ESPOP pour les classes 4 et 5 font encore partie de l'intervalle de confiance qu'on estime, mais ceux des classes 6 et 7 n'ont même plus cette qualité ! Une même comparaison (tableau 20 où l'on sépare encore par le sexe montre qu'on est hors de l'intervalle de confiance que pour le nombre de femmes (`sex=2`) des classes d'âge 6 et 7.

Nous considérons ici un indicateur (classes d'âge) construit sur la variable âge. Nous avons calé l'échantillon selon certaines classes d'âge, mais on constate ici que déjà pour un autre indicateur (autres classes d'âge), fortement corrélé au premier, reposant sur la même variable âge, l'imprécision peut être grande.
On imagine alors bien que pour des variables peu corrélées aux variables de calage, le biais et les variances peuvent devenir non négligeables et sont des informations essentielles à communiquer avec des résultats. Il faut également en tenir compte lors de l'interprétation.

ESS02, CATI

Classe d'âge	Nbr. Obs.	Total pondéré	Std_err	Borne inf.	Borne Sup.
1 : 15-24 ans	1701	844807	22708	800300	889314
2 : 25-34 ans	3022	1027184	21301	985436	1068932
3 : 35-44 ans	4202	1200514	21269	1158828	1242200
4 : 45-54 ans	3175	992818	20649	952347	1033288
5 : 55-64 ans	3244	834623	17509	800305	868940
6 : 65-74 ans	2538	642804	15316	612786	672823
7 : 75+ ans	1824	474888	13396	448633	501144
Total	19706	6017638			

ESPOP01

Classe d'âge	Percent	Population 15+
1 : 15-24 ans	14.04	844807
2 : 25-34 ans	17.07	1027184
3 : 35-44 ans	19.95	1200514
4 : 45-54 ans	16.65	1001928
5 : 55-64 ans	13.61	819066
6 : 65-74 ans	9.87	594240
7 : 75+ ans	8.81	529899
Total	100	6017638

Tab. 20 – Précision d'estimations de totaux.

ESS02, CATI

Hommes

Classe d'âge	Nbr. Obs.	Total pondéré	Std_err	Borne inf.	Borne Sup.
1 : 15-24 ans	857	430144	17824	395210	465078
2 : 25-34 ans	1361	507207	17017	473854	540560
3 : 35-44 ans	2059	605685	16565	573218	638152
4 : 45-54 ans	1449	499569	16795	466652	532486
5 : 55-64 ans	1409	412634	13939	385314	439955
6 : 65-74 ans	1065	274556	10147	254668	294444
7 : 75+ ans	709	179390	8168	163382	195398
Total	8909	2909185			

Femmes

1 : 15-24 ans	844	414663	17590	380188	449138
2 : 25-34 ans	1661	519977	15952	488712	551243
3 : 35-44 ans	2143	594829	16225	563029	626629
4 : 45-54 ans	1726	493249	14891	464063	522435
5 : 55-64 ans	1835	421989	11965	398538	445440
6 : 65-74 ans	1473	368248	12110	344514	391982
7 : 75+ ans	1115	295499	11027	273886	317111
Total	10797	3108453			

ESPOP01

	Hommes		**femmes**	
Classe d'âge	Percent	Population 15+	Population 15+	Percent
1 : 15-24 ans	14.79	430144	414663	13.34
2 : 25-34 ans	17.43	507207	519977	16.73
3 : 35-44 ans	20.82	605685	594829	19.14
4 : 45-54 ans	17.30	503144	498784	16.05
5 : 55-64 ans	13.89	404040	415026	13.35
6 : 65-74 ans	9.22	268217	326023	10.49
7 : 75+ ans	6.56	190748	339151	10.91
Total	100	2909185	3108453	100

Tab. 21 – Précision d'estimations de totaux.

6 Conclusions

L'objectif principal de l'ESS02 est de fournir, à partir des données récoltées sur un échantillon aléatoire stratifié par canton de personnes résidant en Suisse et âgées de 15 ans et plus, des résultats extrapolés aussi bien sur l'état de santé de la population que sur les comportements en matière de santé ou le recours aux prestations du système de santé.
Le but de ce rapport est de décrire les travaux méthodologiques conduisant aux résultats définitifs de l'ESS02. Que ce soit dans les estimations ponctuelles, les estimations de variance ou les analyses, les poids et le plan d'échantillonnage doivent être pris en compte. Le chapitre 1 est consacré au rappel de la délimitation de la population, du cadre de sondage et de l'unité d'échantillonnage. Le chapitre 2 est consacré au plan d'échantillonnage et au suivi des échantillons.
Pour obtenir ces résultats extrapolés à la population, une pondération différente a été élaborée selon le type de questionnaire. En effet, pour le questionnaire téléphonique qui fait l'objet du chapitre 3, la pondération finale a été obtenue après calage des poids initiaux sur les variables sexe, nationalité, canton (ou groupe de canton) et classes d'âge. Pour cela, le recours aux données de l'ESPOP (état fin 2001) comprenant le nombre de personnes de 15 ans et plus vivant en Suisse pour le croisement des variables considérées, a été nécessaire. Concrètement, les totaux ont d'abord été ajustés pour le croisement des variables nationalité et lieu puis sur le croisement des variables sexe et âge.

Pour la pondération de l'enquête écrite (menée sur un sous-échantillon de personnes ayant participé à l'enquête téléphonique) élaborée au chapitre 4, une pondération initiale a également été développée tout comme un calage sur marges mais cette fois sur l'enquête téléphonique. Les critères retenus pour le calage sur marges sont respectivement le croisement des variables nationalité et niveau de formation achevée, puis le croisement sexe et classes d'âge et enfin l'état subjectif de santé et le handicap de la vue. Le jeu de poids ainsi obtenu a encore subi une winsorisation des poids en ramenant au 2ème percentile les poids inférieurs à celui-ci et au 98ème percentile les poids supérieurs à cette dernière limite.

Enfin, le chapitre 5 est consacré à l'estimation de la précision des estimateurs développés dans cette enquête et a été calculée en tenant compte des poids finaux pour chacune des enquêtes. Il apparaît qu'une méthode de rééchantillonnage basée sur le jackknife fournit des résultats appropriés du point de vue de l'estimation de variance pour le calcul des intervalles de confiance et que l'utilisation de la procédure SURVEYMEANS (avec les jeux de poids correspondants) fournit des résultats très proches de ceux obtenus par la macro jackknife. L'estimation de la précision des résultats obtenus dans cette enquête est particulièrement importante pour établir la fiabilité des résultats et poser ainsi les limites d'exploitation de cette enquête.

A Tailles des échantillons bruts et nets par grandes régions, par strates et par vagues

Région 1 : Région Lémanique					
strate (h)	M_h	vague	m_h^n (`kanton_1`)	taille selon `kanton`	m_h^b
Vaud	266780	1	241	240	370
	264008	2	248	253	445
	265121	3	250	253	490
	264838	4	295	296	542
Valais	134302	1	257	248	438
	134034	2	230	221	437
	133766	3	264	257	473
	133384	4	248	244	485
Genève	160925	1	219	227	414
	160113	2	246	249	424
	159246	3	254	258	501
	158431	4	281	283	547

Tab. 22 – Tailles des échantillons bruts et nets par strates et par vagues pour la grande région 1.

Région 2 : Espace Mittelland					
strate (h)	M_h	vague	m_h^n (`kanton_1`)	taille selon `kanton`	m_h^b
Berne	409229	1	386	387	624
	407364	2	413	416	618
	406076	3	394	397	620
	404983	4	400	406	661
Fribourg	94776	1	253	251	433
	94619	2	265	261	425
	95190	3	252	250	415
	94807	4	243	243	407
Soleure	97734	1	240	243	394
	97201	2	274	274	395
	97113	3	237	237	375
	96893	4	266	265	440
Neuchâtel	69160	1	166	167	260
	68915	2	177	177	276
	68616	3	179	180	283
	67787	4	174	172	288
Jura	27384	1	138	136	283
	27311	2	184	183	300
	27248	3	182	179	307
	27191	4	154	152	272

TAB. 23 – Tailles des échantillons bruts et nets par strates et par vagues pour la grande région 2.

Région 3 : Suisse du Nord-Ouest					
strate (h)	M_h	vague	m_h^n (`kanton_1`)	taille selon `kanton`	m_h^b
Bâle Ville	83895	1	264	264	445
	82969	2	277	275	429
	82164	3	225	228	378
	81473	4	251	250	418
Bâle Campagne	106618	1	272	274	464
	106213	2	272	273	449
	105910	3	272	272	394
	105731	4	200	200	312
Aargau	212847	1	247	249	398
	211372	2	268	270	396
	211195	3	249	250	374
	211001	4	243	247	389

TAB. 24 – Tailles des échantillons bruts et nets par strates et par vagues pour la grande région 3.

Région 4 : Zürich					
strate (h)	M_h	vague	m_h^n (`kanton_1`)	taille selon `kanton`	m_h^b
Zurich	516621	1	397	404	649
	512194	2	417	422	638
	510581	3	417	423	627
	508019	4	381	381	598

TAB. 25 – Tailles des échantillons bruts et nets par strates et par vagues pour la grande région 4.

Région 5 : Suisse Orientale					
strate (h)	M_h	vague	m_h^n (`kanton_1`)	taille selon `kanton`	m_h^b
Glarus	15920	1	55	55	101
	15828	2	56	56	100
	15800	3	76	74	129
	15733	4	72	72	121
Schaffhouse	30803	1	69	71	103
	30671	2	72	72	100
	30555	3	60	60	85
	30481	4	52	52	96
St Gall	176528	1	245	243	426
	175371	2	275	276	428
	174689	3	273	272	403
	174135	4	219	222	369
Grisons	103731	1	85	84	175
	103118	2	83	84	176
	102585	3	87	86	170
	101838	4	77	74	147
Thurgovie	86804	1	91	91	151
	86289	2	106	106	155
	86089	3	121	121	160
	85892	4	87	87	132
Appenzell	26990	1	59	58	101
	26808	2	75	75	115
	26650	3	76	74	120
	26542	4	74	73	116

TAB. 26 – Tailles des échantillons bruts et nets par strates et par vagues pour la grande région 5.

Région 6 : Suisse Centrale					
strate (h)	M_h	vague	m_h^n (`kanton_1`)	taille selon `kanton`	m_h^b
Lucerne	133624	1	238	239	404
	132770	2	272	274	439
	132289	3	257	258	409
	132054	4	248	250	401
Uri	13662	1	53	53	101
	13602	2	84	82	123
	13576	3	71	70	110
	13564	4	57	57	95
Schwyz	50073	1	51	51	101
	49930	2	68	67	124
	49914	3	82	82	144
	49907	4	74	73	120
Obwald	13686	1	43	42	75
	13610	2	52	53	81
	13581	3	53	53	80
	13556	4	43	42	65
Niedwald	15162	1	41	39	74
	15074	2	44	44	79
	15039	3	59	59	97
	15024	4	55	55	81
Zouge	39802	1	261	263	435
	39468	2	266	263	423
	39368	3	246	246	392
	39438	4	237	237	398

TAB. 27 – Tailles des échantillons bruts et nets par strates et par vagues pour la grande région 6.

Région 7 : Tessin					
strate (h)	M_h	vague	m_h^n (`kanton_1`)	taille selon `kanton`	m_h^b
Tessin	142643	1	340	332	630
	141654	2	377	375	691
	141301	3	416	413	772
	140589	4	411	409	745

TAB. 28 – Tailles des échantillons bruts et nets par strates et par vagues pour la grande région 7.

B Les effets du calage sur marge

Cette section commence par présenter quelques graphes donnant un résumé visuel des poids initiaux et finaux en permettant de les comparer entre eux.

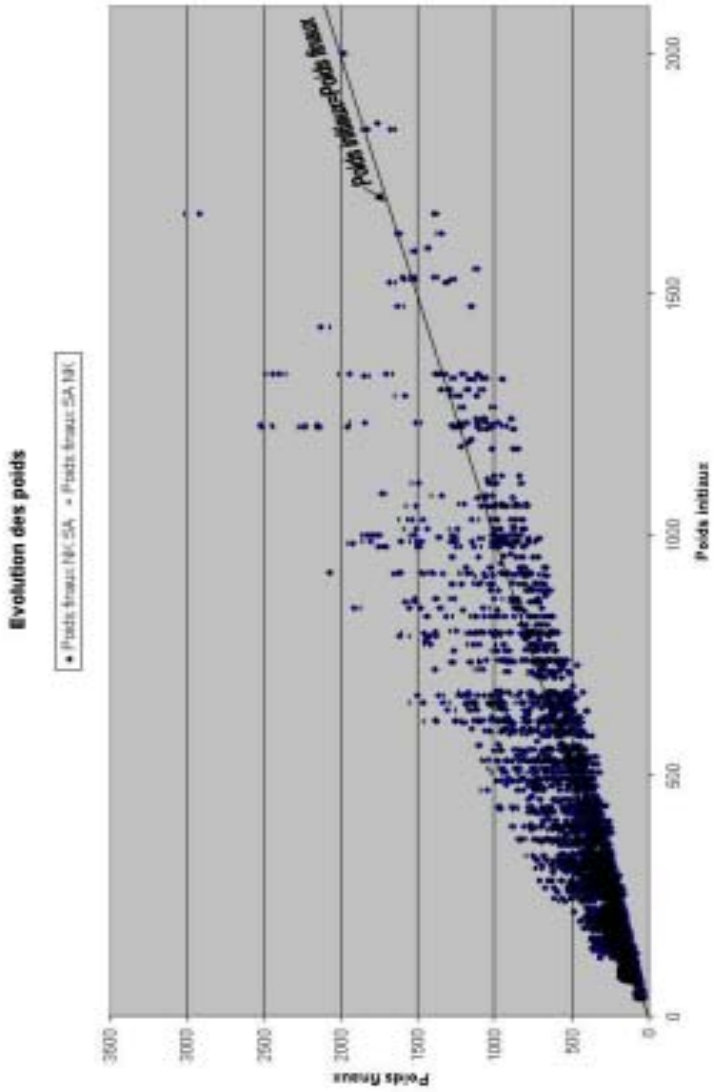


FIG. 4 – Poids finaux NK SA et SA NK (voir paragraphe 3.5) par rapport aux poids initiaux. On observe que les deux versions de poids finaux sont très proches l'une de l'autre et se comportent de la même manière par rapport aux poids initiaux. Les points sur la droite en traitillés indiquent les cas où poids final = poids initial.

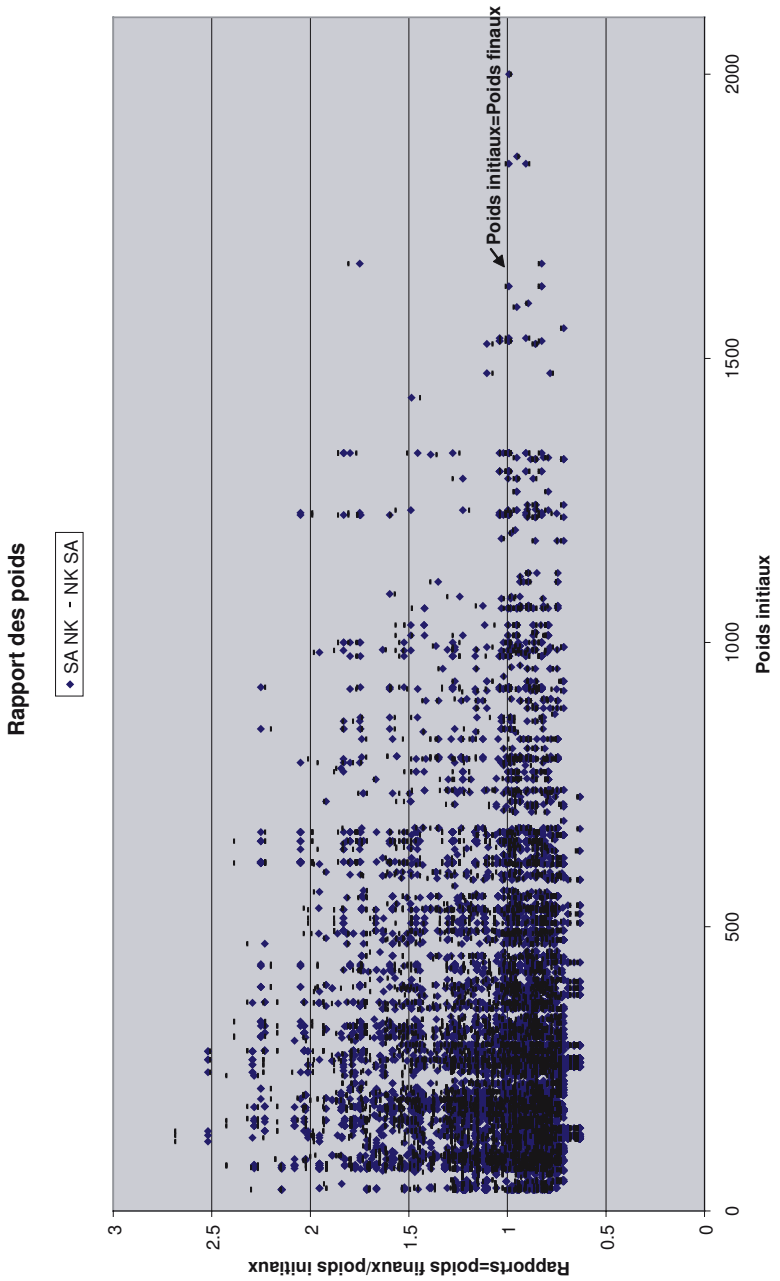


FIG. 5 – Rapports $\frac{\texttt{Poids finaux}_i}{\texttt{Poids initiaux}_i}$ pour tous les individus i ayant participés à l'enquête. On observe que les deux versions de poids finaux sont très proches l'une de l'autre et se comportent de la même manière par rapport aux poids initiaux. Les variations de poids sont plus prononcées à la hausse (le poids final est plus grands que le poids initial) qu'à la baisse. Les points sur l'horizontale à 1 indiquent les cas où poids final = poids initial. On devine comme une grille ou un quadrillage selon lesquels les poids sont distribués. Cela est dû au fait que le nombre de poids différents est limité (nombre de strates, de personnes par ménages, de cellules de calage limités).

Suivent maintenant toute une série d'histogrammes permettant d'apprécier les effets, en général bénéfiques, du calage sur marge. Dans chaque graphe on représente l'écart (le biais) entre les nombres de l'ESPOP et les estimations qu'on obtient avec les données de l'ESS en usant de la pondération initiale respectivement finale.

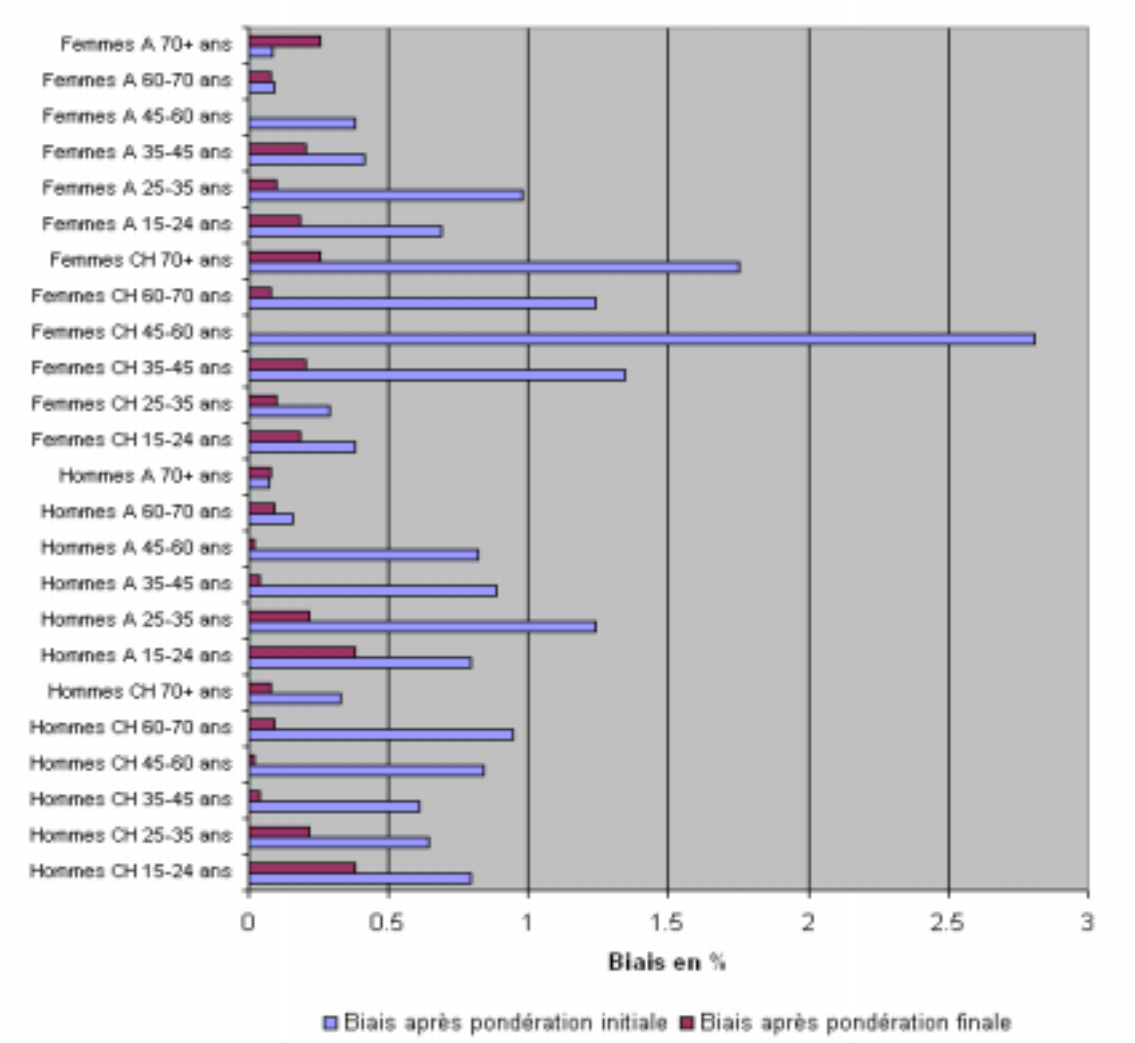


FIG. 6 – Biais entre la population (ESPOP01) et les échantillons pondérés (questionnaire téléphonique) (poids initiaux et poids finaux) pour le croisement `sex` × `natio` × `age`.

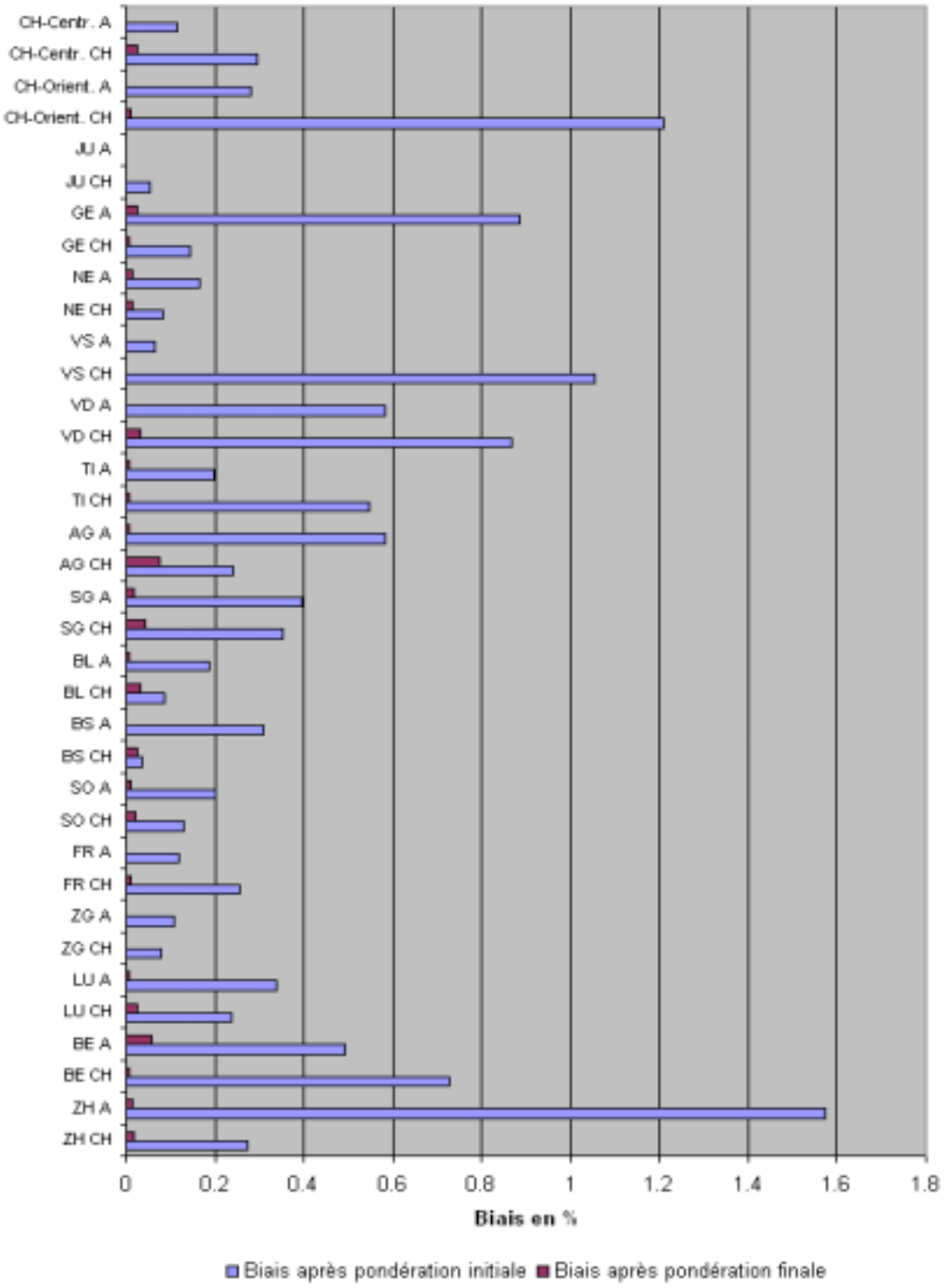


FIG. 7 – Biais entre la population (ESPOP01) et les échantillons pondérés (questionnaire téléphonique) (poids initiaux et poids finaux) pour le croisement `lieu` × `natio`, avec la variable `lieu` définie par rapport à `kanton1` (même construction que pour les calages).

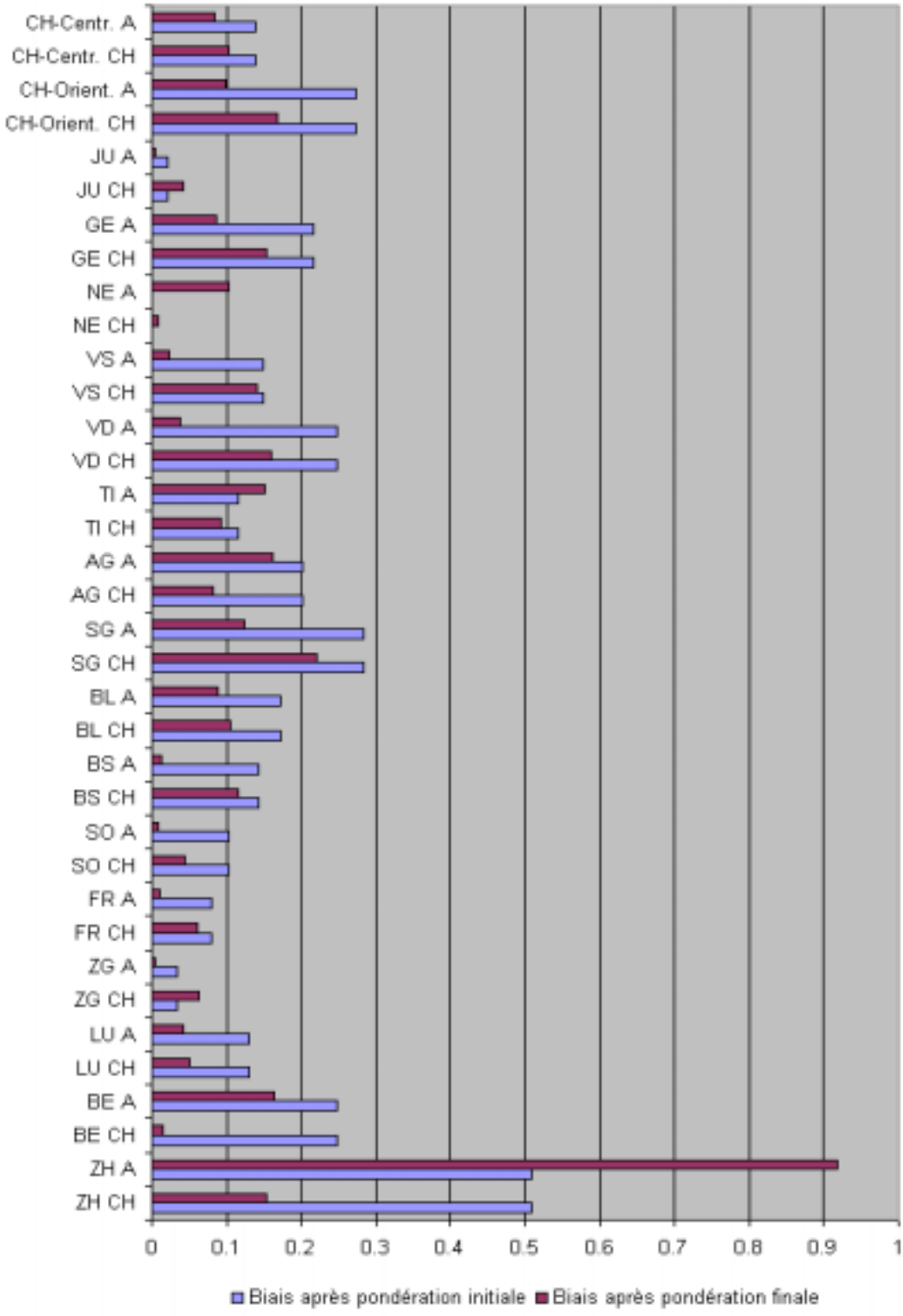


FIG. 8 – Biais entre la population (ESPOP01) et les échantillons pondérés (questionnaire écrit) (poids initiaux et poids finaux) pour le croisement `lieu` × `natio`, avec la variable `lieu` définie par rapport à `kanton1` (même construction que pour les calages).

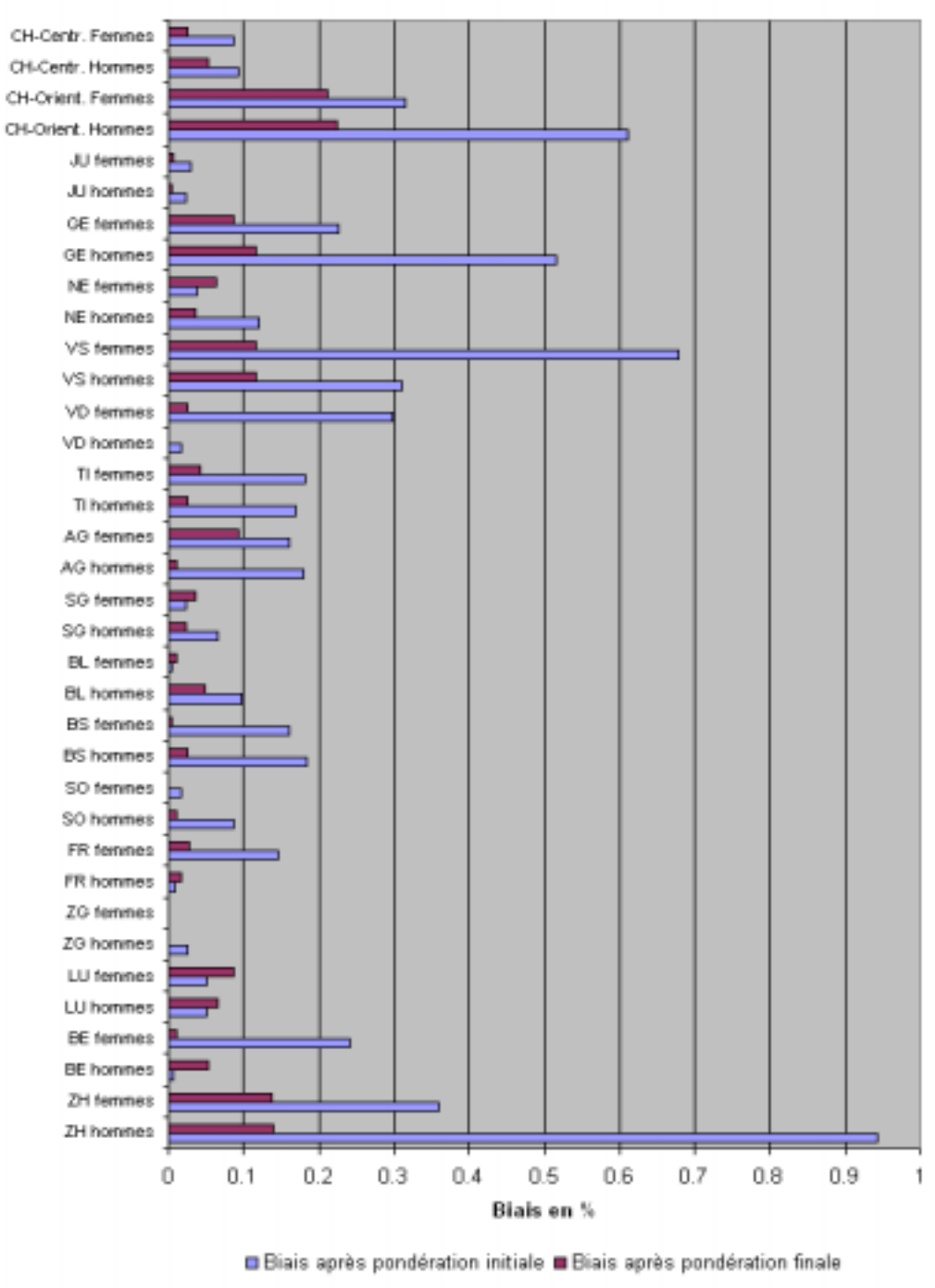


FIG. 9 – Biais entre la population (ESPOP01) et les échantillons pondérés (questionnaire téléphonique) (poids initiaux et poids finaux) pour le croisement `lieu` × `sex`, avec la variable `lieu` définie par rapport à `kanton1` (même construction que pour les calages).

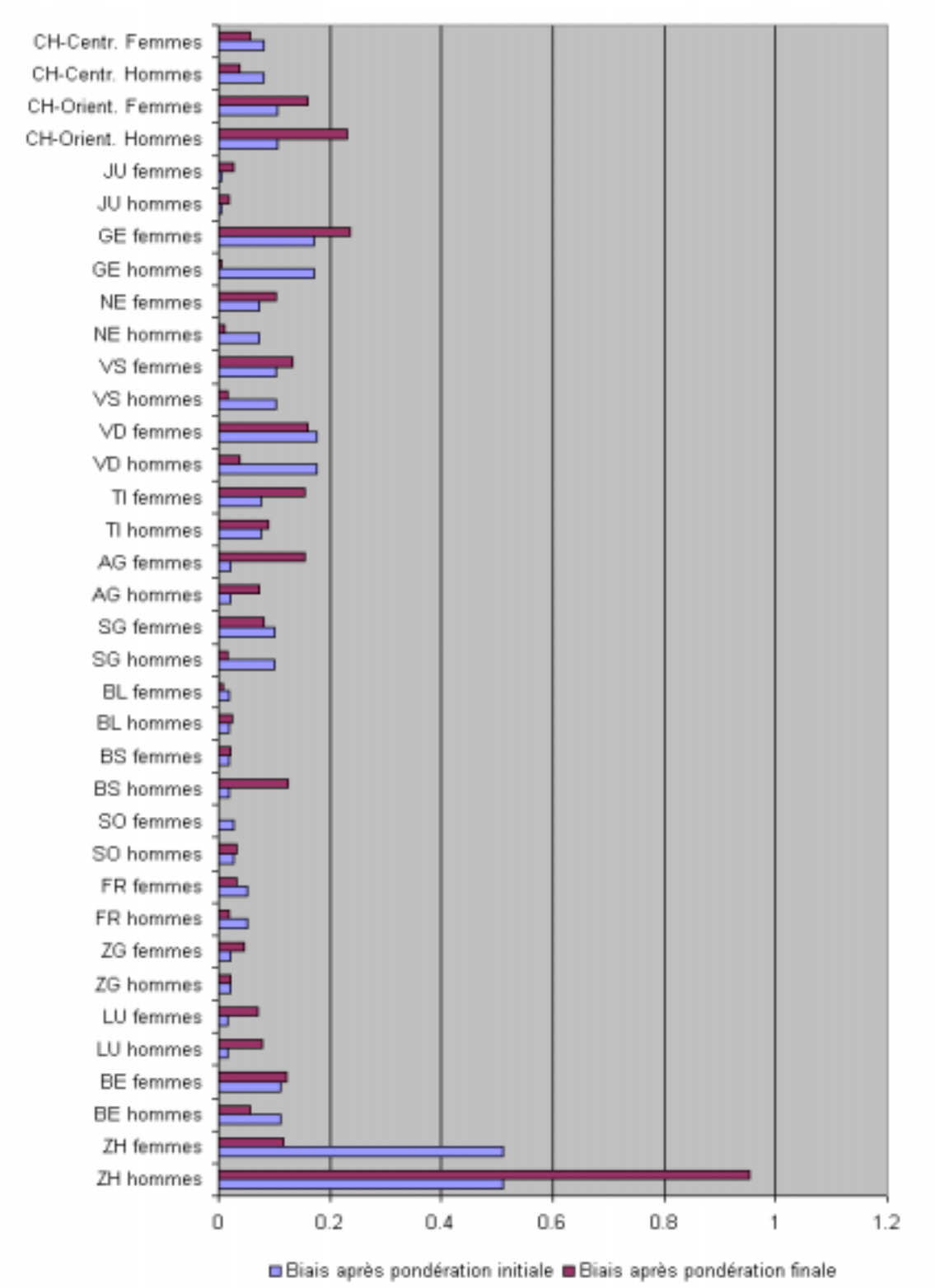


FIG. 10 – Biais entre la population (ESPOP01) et les échantillons pondérés (questionnaire écrit) (poids initiaux et poids finaux) pour le croisement `lieu` × `sex`, avec la variable `lieu` définie par rapport à `kanton1` (même construction que pour les calages).

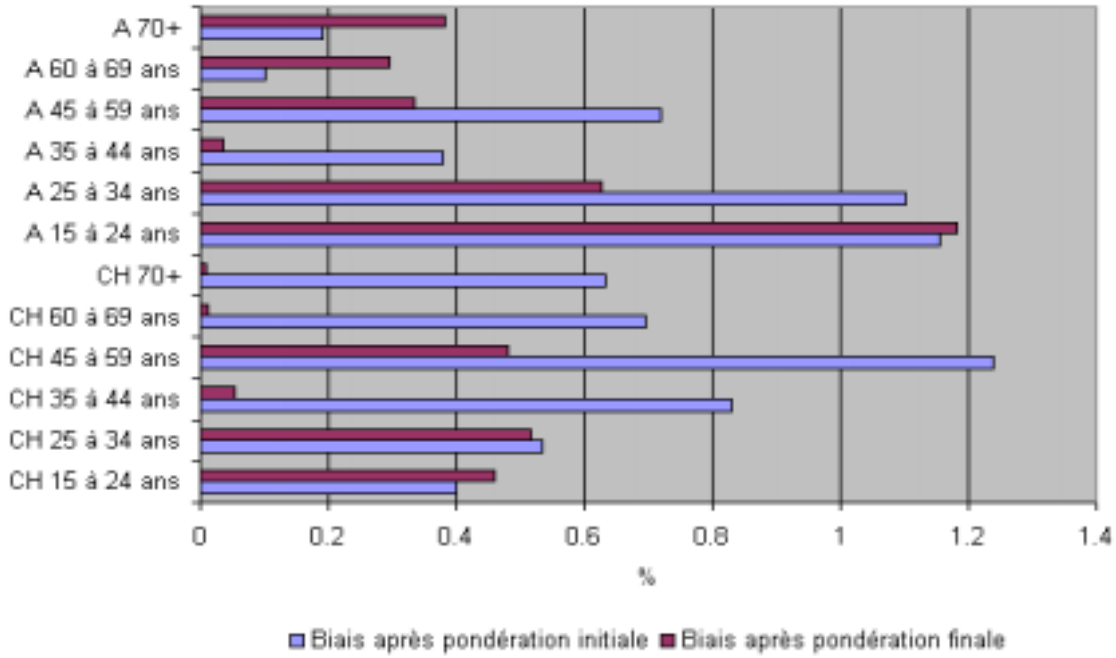


FIG. 11 – Biais entre la population (ESPOP01) et les échantillons pondérés (questionnaire écrit) (poids initiaux et poids finaux) pour le croisement `natio` × `age`.

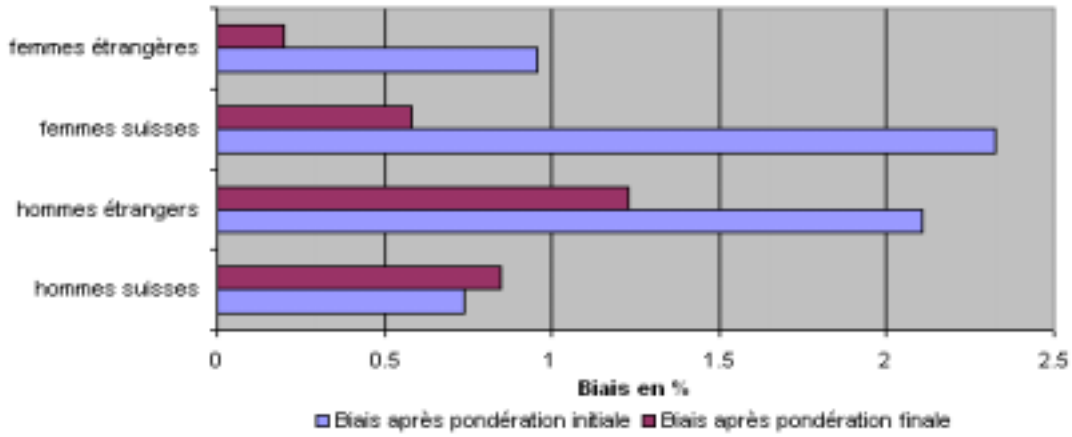


FIG. 12 – Biais entre la population (ESPOP01) et les échantillons pondérés (questionnaire téléphonique) (poids initiaux et poids finaux) pour le croisement `sex` × `natio`.

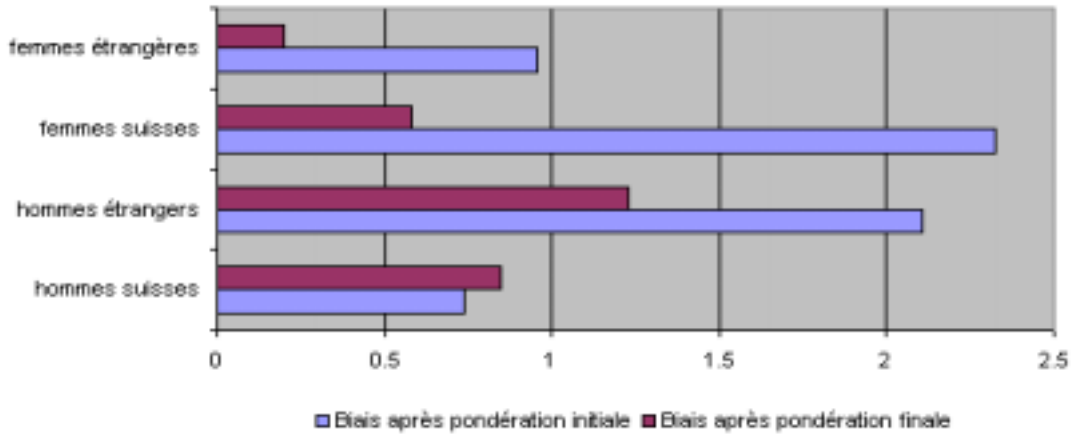


FIG. 13 – Biais entre la population (ESPOP01) et les échantillons pondérés (questionnaire écrit) (poids initiaux et poids finaux) pour le croisement `sex` × `natio`.

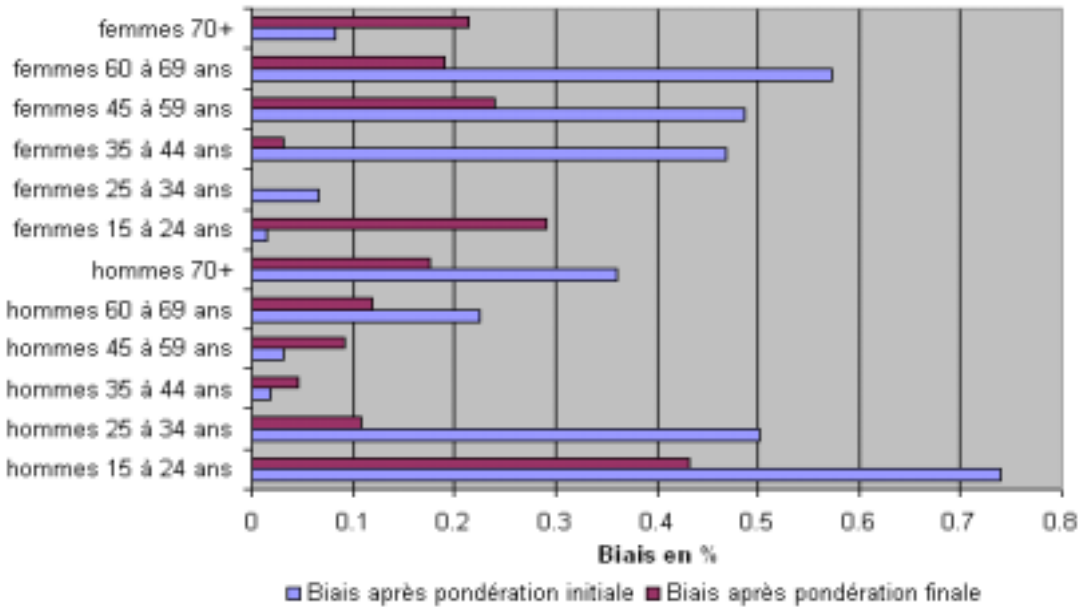


FIG. 14 – Biais entre la population (ESPOP01) et les échantillons pondérés (questionnaire écrit) (poids initiaux et poids finaux) pour le croisement `sex` × `age`.

C Tests relatifs au paramètre réplicat des macros jackknife

Comme expliqué au paragraphe 5.3, le paramètre `repl=` indique à la macro comment former les g sous-échantillons définis dans le paragraphe 5.2. Le rapport *(Davison, 1998)* indique que plus g est grand, plus l'estimation de la variance fournie par la macro sera stable. Les tests que nous avons conduits ici permettent de le vérifier dans le cadre de l'ESS02.

Parmi les questions que nous nous posions, citons les suivantes :

- Les conclusions (voir paragraphe 5.3) du rapport Davison s'appliquent-elles aussi dans le cadre de l'ESS02 ?
- Devrait-on créer une variable réplicats qui respecte les strates, cela fournirait-il des estimations plus stables ?
- Le fait d'utiliser ou non les poids a-t-il un influence sur la stabilité des estimations de la variance ?

Les graphiques G1 à G12 ont en abscisse le nombre de réplicats utilisés et en ordonnée l'estimation de l'écart-type (losanges) et l'estimation du coefficient de variation (carrés). Les graphiques G1, G3, G5, G7, G9 et G11 présentent les mêmes données que respectivement les graphiques G2, G4, G6, G8, G10 et G12 ; le premier groupe utilise une échelle conventionnelle, tandis que le deuxième utilise une échelle racine cubique.

Nos tests consistaient à étudier les estimations fournies par la macro `%jackprop` en faisant varier certains de ses paramètres :

`data=` : on a utilisé les données CATI de l'ESS02 ;

`var=` : Les tests ont été menés sur les variables `sex`,`tsubg01` et sur une variable représentant les classes d'âge utilisées dans les tableaux standards de l'ESS. Les graphiques présentés dans ce rapport illustrent les résultats pour la variable `sex`. Comme cette variable ne prend que les valeurs '1' ou '2', il suffit de présenter les résultats pour `sex=1`.

`gewvar=` : Des tests ont été effectués sans les poids (graphiques G1-2-5-6-9-10) et avec les poids finaux[7] w^{FIN} (graphiques G3-4-7-8-11-12).

`class=` : Nous n'avons pas renseigné ce paramètre dans nos tests.

`repl=` : C'est sur ce paramètre et la construction de la variable qui le renseigne que nous allons maintenant porter notre attention.

Deux familles de variables réplicats ont été générées : une respectant les strates et l'autre sans cette contrainte. La variable réplicats définit les g groupes d'observation(s) définis au paragraphe 5.2.

Prenons les deux exemples les plus extrêmes pour expliquer le fonctionnement de la macro par rapport au paramètre réplicat (`repl=`) :

(i) Si le paramètre réplicat n'est pas renseigné, qu'on l'identifie à un compteur de 1 à n (`repl=_n_`) où n est le nombre d'observations, ou à une variable contenant un identifiant unique pour chaque observation (pour l'ESS par exemple, `repl=idno`), alors on aura le nombre maximum de groupes : n.

[7] Dans le fichier de données de l'ESS, les poids finaux téléphonique w^{FIN} sont contenus dans la variable `wght`, les poids finaux écrits ont contenus dans la variable `wecrit`

(ii) A l'autre extrême si on identifie le paramètre réplicat à une variable réplicats qui vaudra 1 pour une moitié des observations et 2 pour l'autre moitié ne définira que 2 groupes.

Dans le cas *(i)* la macro écarte tour à tour, une à une, chaque observation et estime une variance sur les $(n-1)$ observations restantes. En fin de boucle, la macro a calculé n estimations. Le résultat fournit par la macro est la moyenne arithmétique de ces n variances estimées.

Dans le cas *(ii)* la macro calcule une première estimation de variance sur le groupe des observations pour lesquelles `repl=1`, puis une deuxième estimation pour celles où `repl=2`. Le résultat fournit par la macro est la moyenne arithmétique de ces 2 variances estimées. Intuitivement on conçoit facilement que le résultat fournit par la variante *(ii)* a une variabilité bien plus grande que celui de la variante *(i)*.

Le cas général consiste a former g groupes d'observations de tailles identiques ou presque, à les écarter les uns après les autres et chaque fois faire une estimation de variance sur $(g-1)$ groupes, puis faire la moyenne des g estimations. La question est maintenant : comment construire la variable réplicats qui renseignera le paramètre `repl` de la macro ? Est-il important de construire une variable réplicats qui respecte les strates définies par le plan d'échantillonnage ?

Construction de variables réplicats sans respecter les strates

Par "ne pas respecter les strates" nous entendons *définir les réplicats sans prendre en compte la structure stratifiée de l'échantillon.* Nous procédons comme si il n'y avait qu'une seule strate. Les graphiques G1-G4 illustrent les résultats de variances estimées pour 200 valeurs de g entre 2 et 19706. On observe qu'indépendamment du fait de tenir compte des poids, les estimations se stabilisent à partir de $g = 1000$ groupes. La variabilité grandit avec des g diminuant inférieurs à 1000 et reste à peu près constante pour des $g > 1000$. On constate également sur ces quatre premiers graphiques que le fait d'inclure les poids est très important : la variance et le coefficient de variation sont systématiquement sous-estimés dans les calculs sans les poids.

Construction de variables réplicats respectant les strates

La construction de la variable réplicats sous la contrainte des strates limite le nombre maximum de réplicats au nombre d'observations contenues dans la plus petite des strates, soit 190 pour l'ESS02. Dans ce cas on attribue aléatoirement un nombre entre 1 et 190 à chaque observation à l'intérieur des strates tout en veillant à avoir, dans la mesure du possible, le même nombre de 1, de 2, ..., de 190 dans chacune des strates. A l'autre extrême on attribuera dans chaque strate des 1 et des 2 en nombres égaux. Dans le cas intermédiaire, on attribuera dans chaque strate des 1, des 2, ..., des g $(1 < g < 190)$ en nombres égaux.

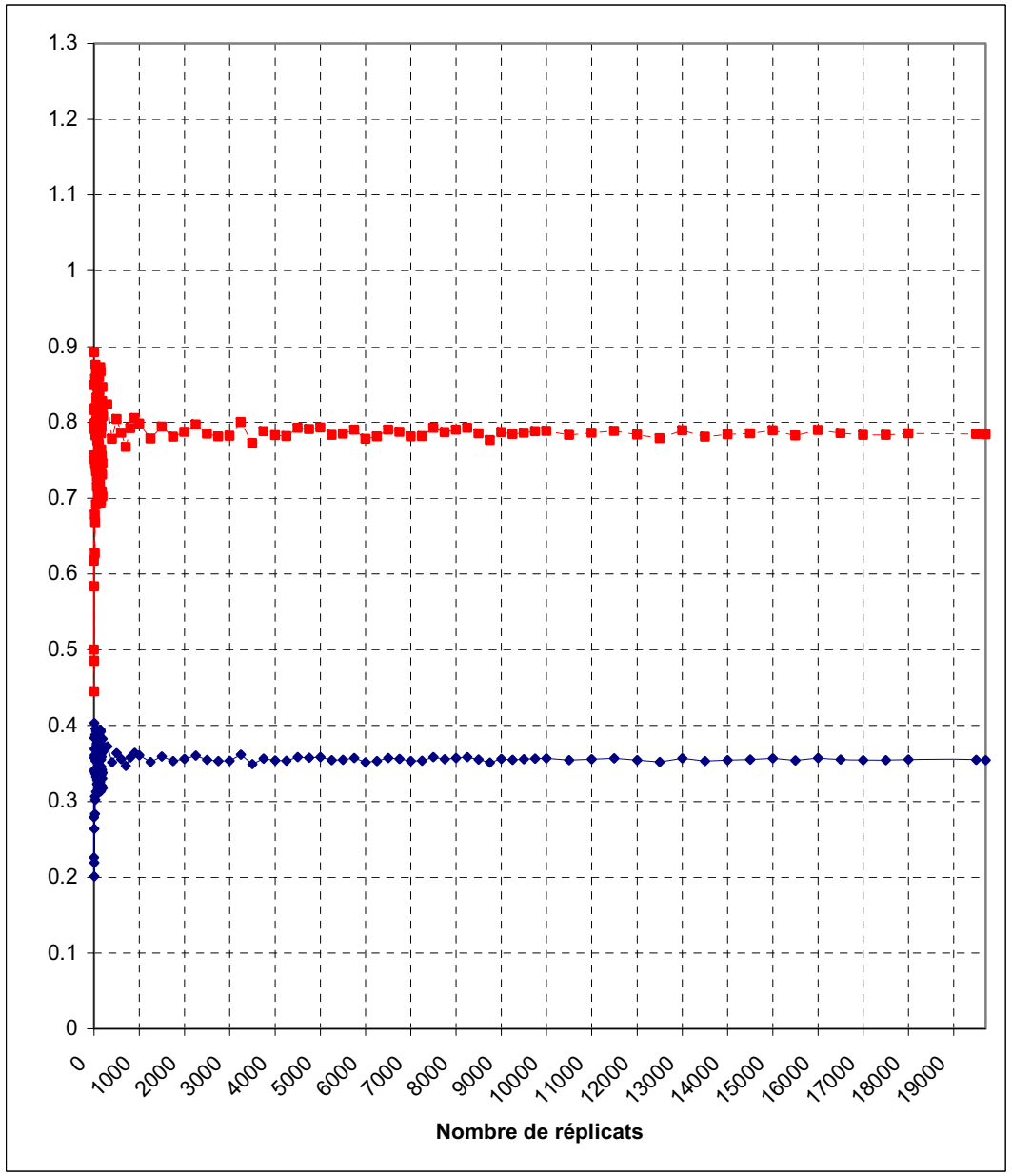
Les graphiques G5-G8 montrent une diminution de la variabilité des estimations avec un nombre croissant de réplicats, mais les résultats restent nettement moins stables que lorsqu'on ne respecte pas les strates.

La méthode présentée ici écarte à chaque itération $(\frac{1}{g})^{\text{ième}}$ des observations. En principe on devrait recalculer une pondération pour les données restantes à chaque fois

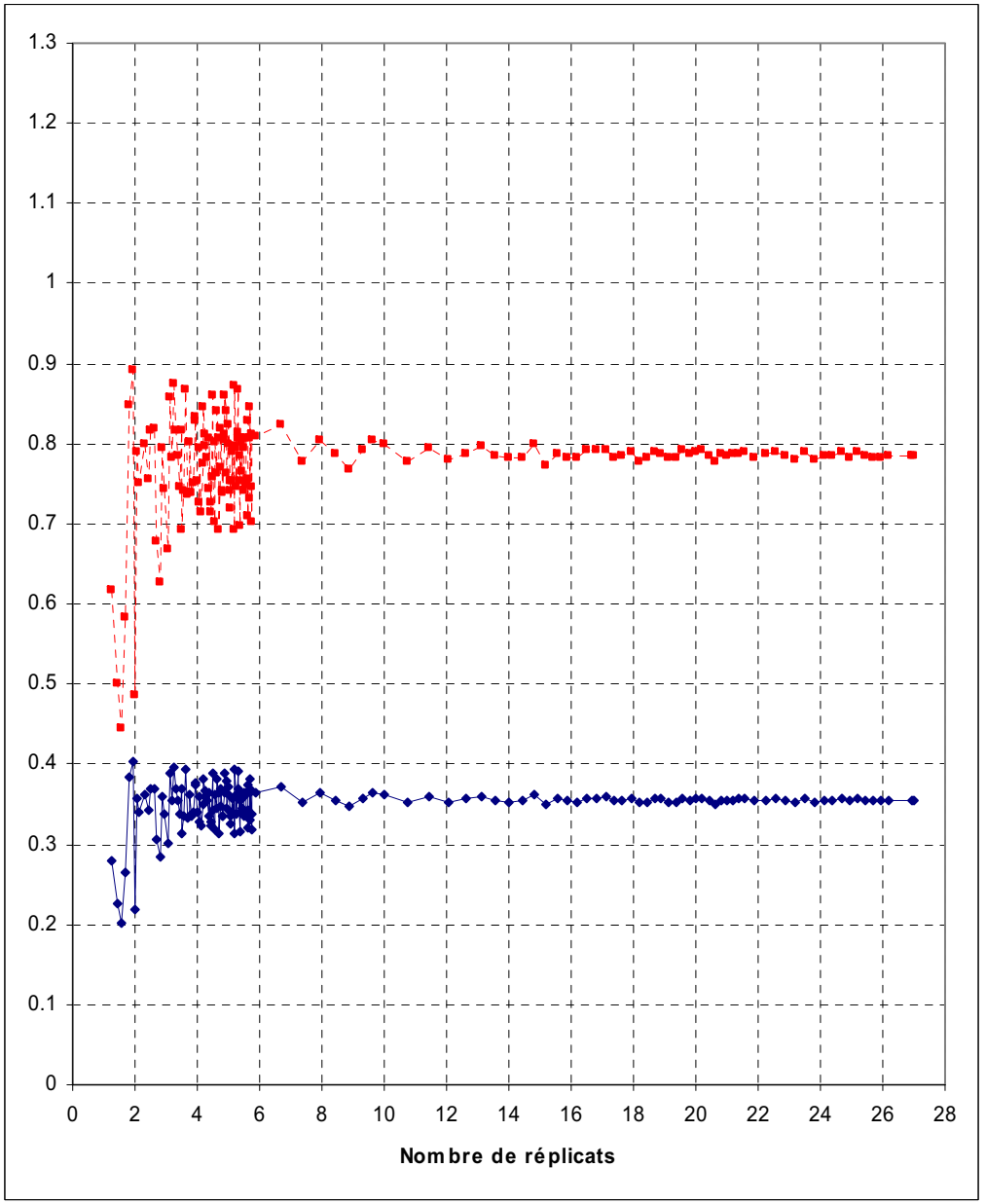
qu'on écarte un groupe avant d'estimer la variance. On pourrait penser que le fait de ne pas recalculer ces jeux de poids est à l'origine de la grande variabilité que l'on observe dans ces tests. Mais les graphiques G5-G8 montrent que la variabilité des estimations n'est que faiblement influencée par les poids. On voit également que, comme pour les graphiques G1-G4, il est nécessaire de prendre les poids en compte si l'on ne veut pas sous-estimer la variance. De plus on vérifie en observant les graphiques G9-G12 que ce ne sont pas tant les poids qui sont à l'origine de la grande variabilité, mais bien le nombre d'observations sur lesquelles on peut baser nos estimations. En effet, dans ces graphes, pour calculer le premier point on a à chaque itération de la macro retiré une observation par strate ; pour calculer le 2^e point du graphe on enlevait à chaque fois 2 observations par strate ; ainsi de suite jusqu'au dernier point du graphe représentant une estimation de variance où à chaque itération de la macro on retirait 189 observations dans chaque strate. On constate clairement que plus on retire d'observations, moins l'estimation de variance fournie par la macro est stable.

Conclusions

Nous avons vérifié que les conclusions de l'étude *(Davison, 1998)* s'appliquent dans le cadre de l'ESS02. Il apparaît que la méthode de ré-échantillonnage basée sur le jackknife fournit des résultats appropriés du point de vue de l'estimation de variance pour le calcul des intervalles de confiance. Il est donc vivement recommandé d'utiliser les macros jackknife pour le calcul de variances. Il faut renseigner le paramètre poids mais pas celui des réplicats (la macro fait alors le nombre maximum de réplicats, sans respecter les startes). Le recours à d'autre méthodes d'estimations de la variance doit se faire avec le plus grand soin.
Nous avons brièvement testé la procédure SURVEYMEANS de SAS sur quelques exemples de l'ESS02. Pour les cas considérés, SURVEYMEANS fournit des résultats très proches de ceux obtenus par les macros jackknife. Néanmoins, comme le temps ne nous a pas permis de tout tester, selon l'exploitation considérée, nous recommandons de toujours comparer les résultats de SURVEYMEANS avec ceux obtenus à l'aide des macros jackknife.
L'estimation de la précision des résultats obtenus dans cette enquête est particulièrement importante pour établir la fiabilité des résultats et poser ainsi les limites d'exploitation de cette enquête.

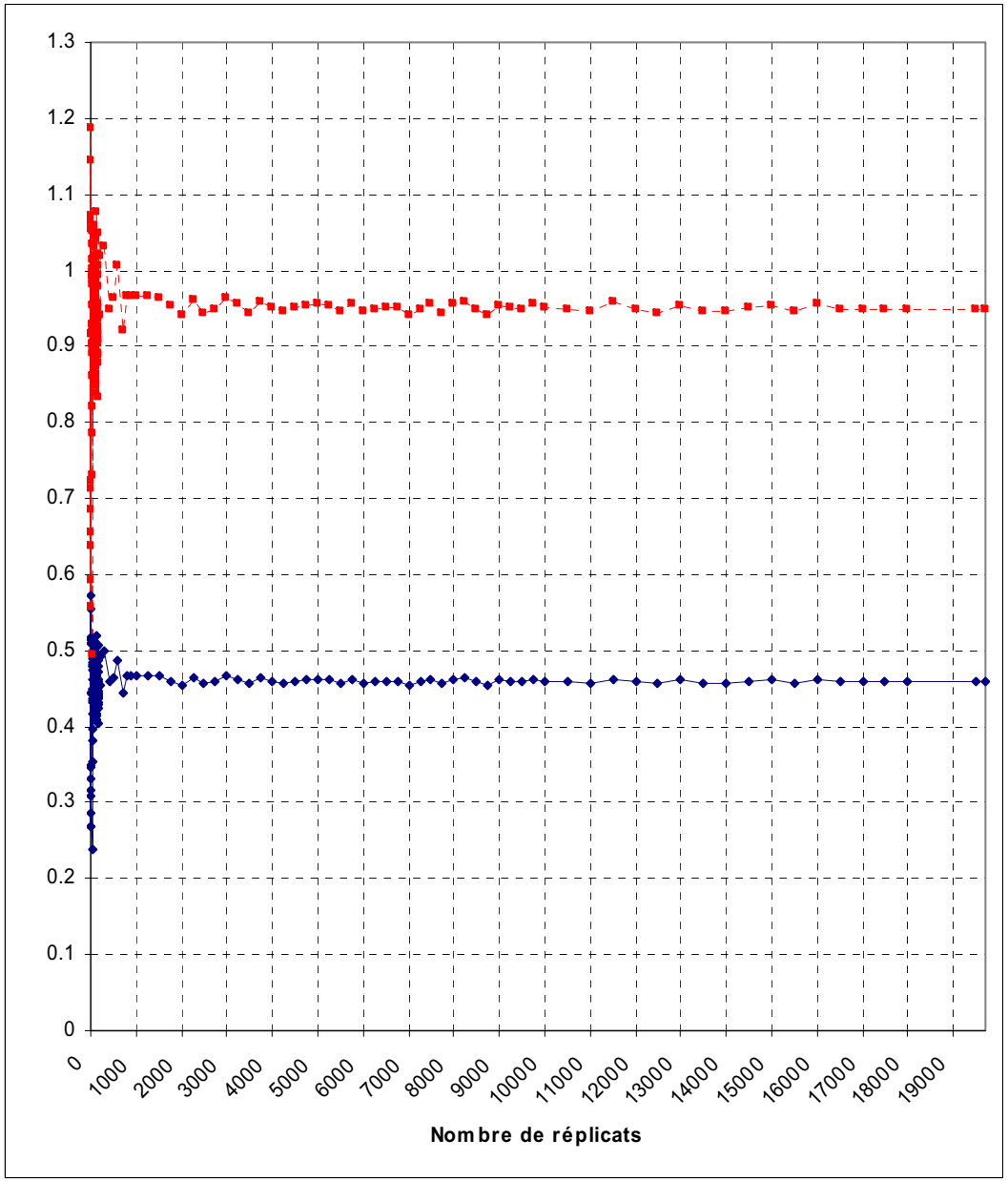


G1 : Sans poids, sans respecter les strates, échelle conventionnelle, sex=1

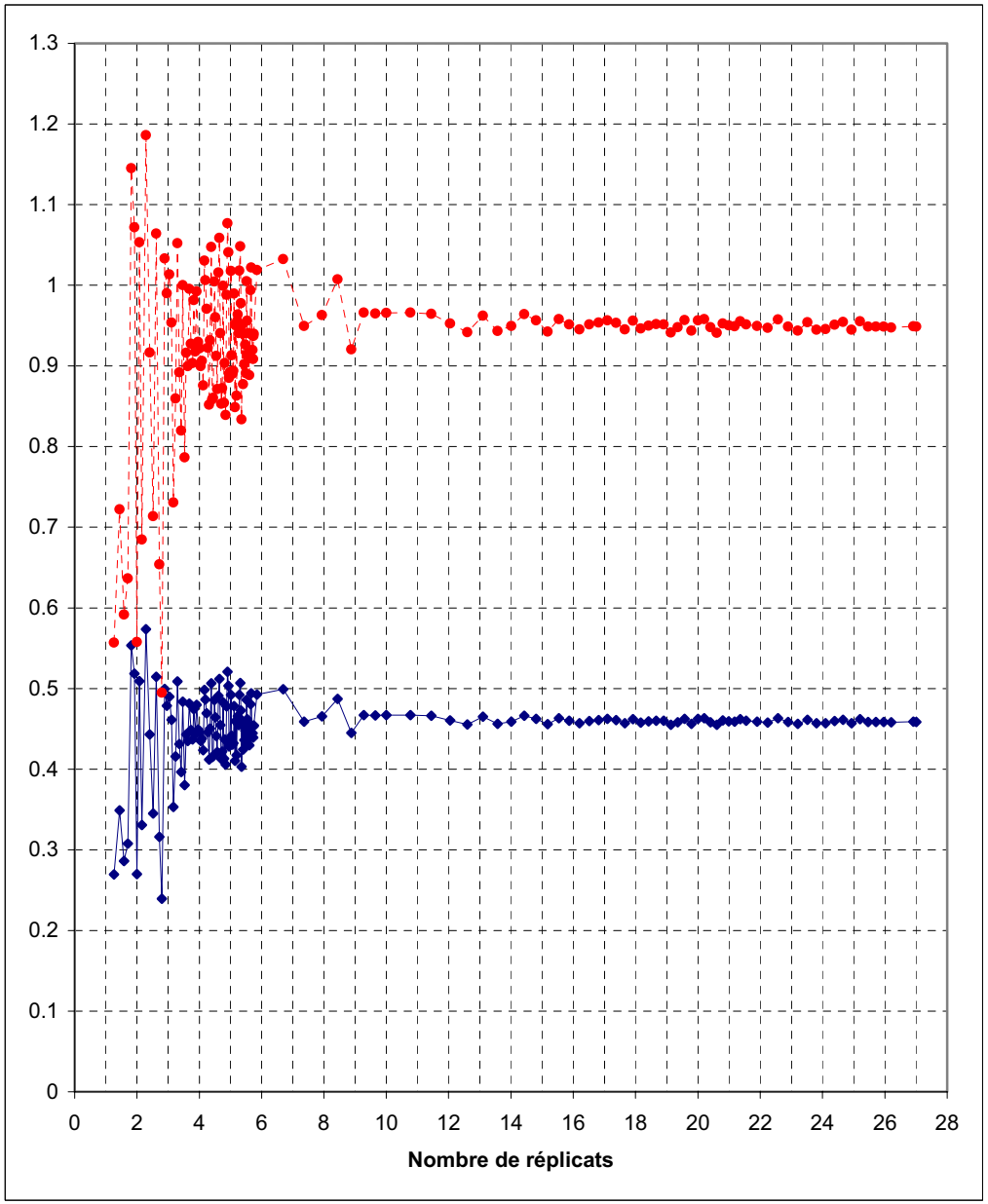


G2 : Sans poids, sans respecter les strates, échelle racine cubique, sex=1



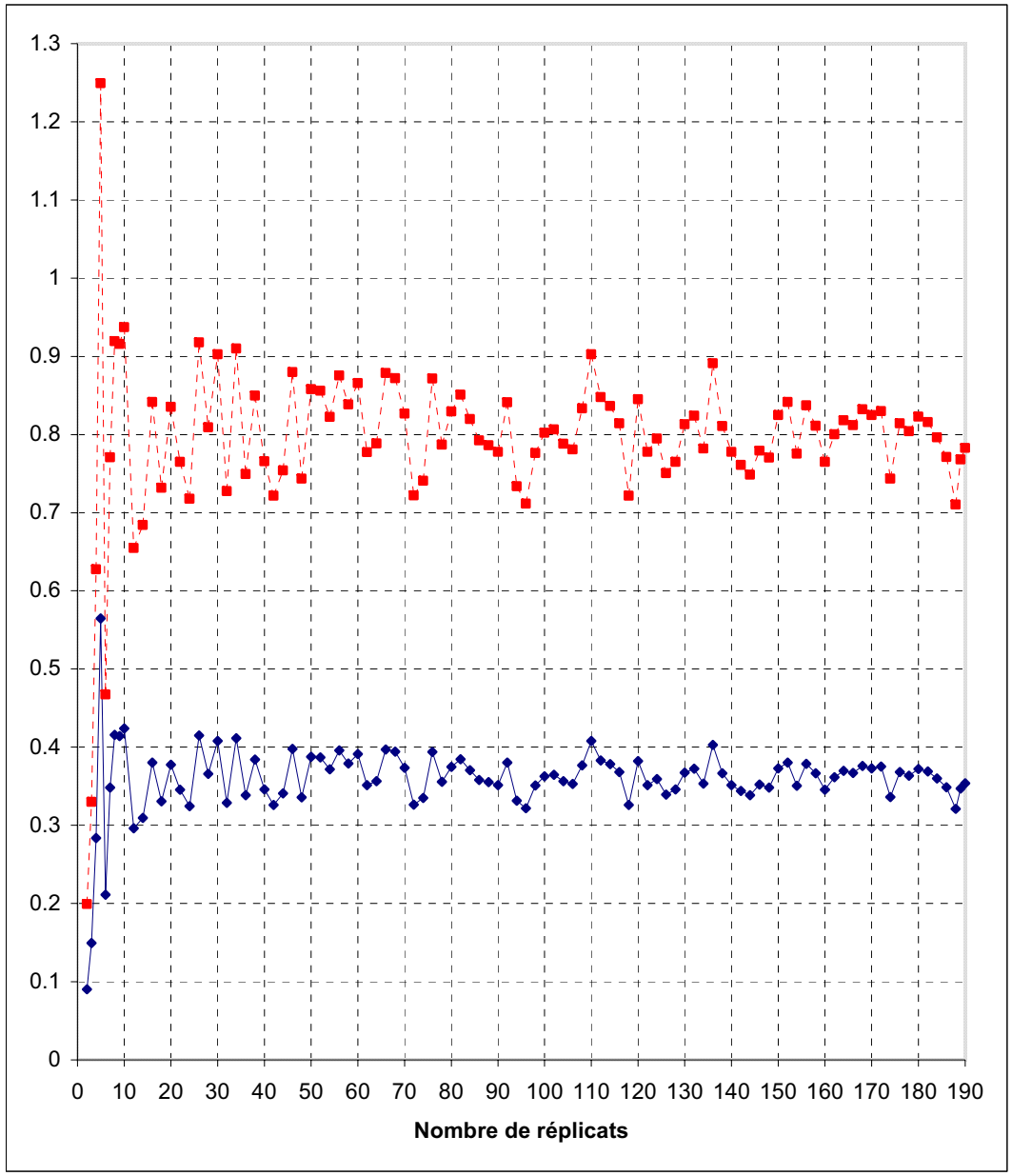


G3 : Avec poids w^{FIN}, sans respecter les strates, échelle conventionnelle, sex=1

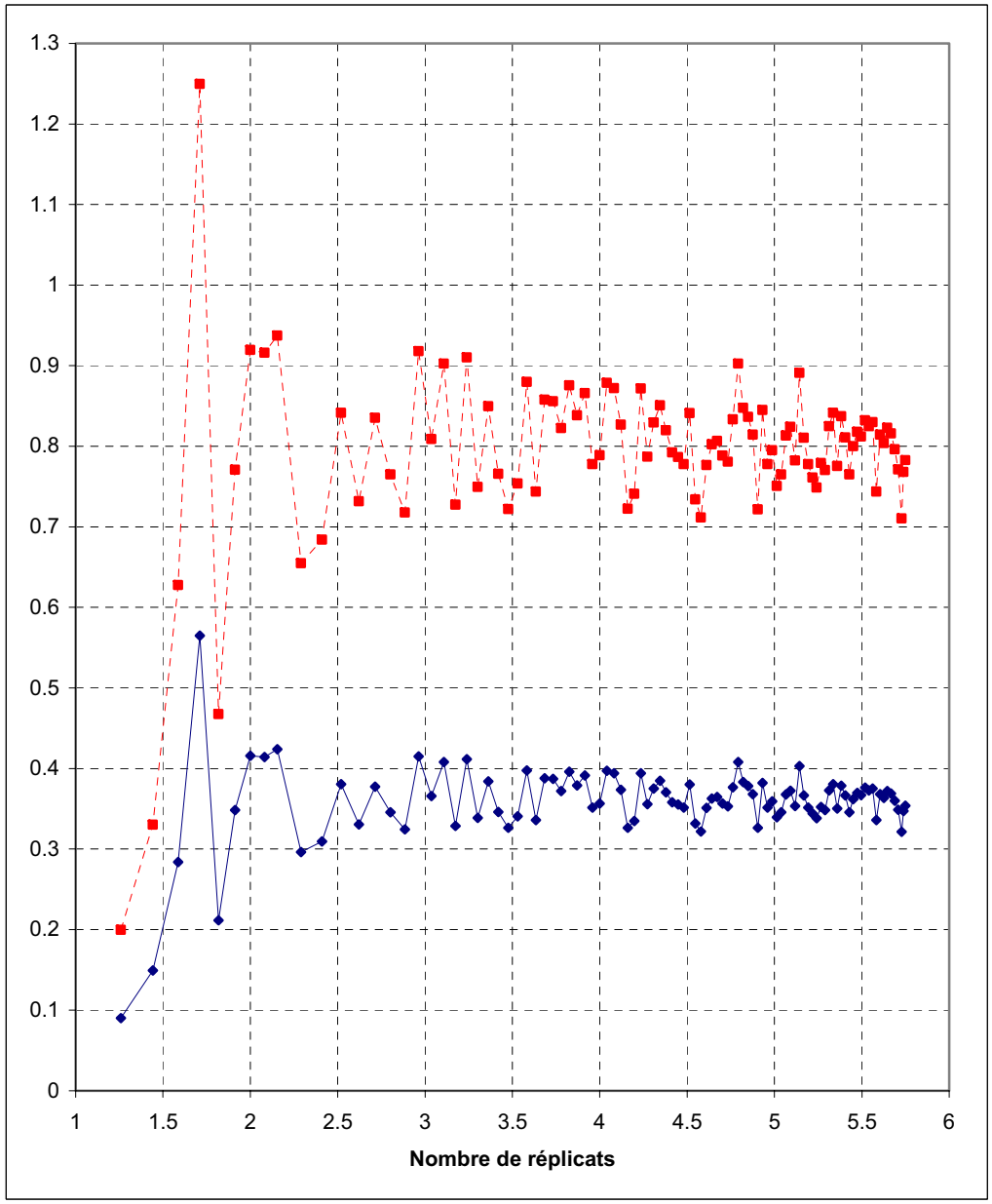


G4 : Avec poids w^{FIN}, sans respecter les strates, échelle racine cubique, sex=1



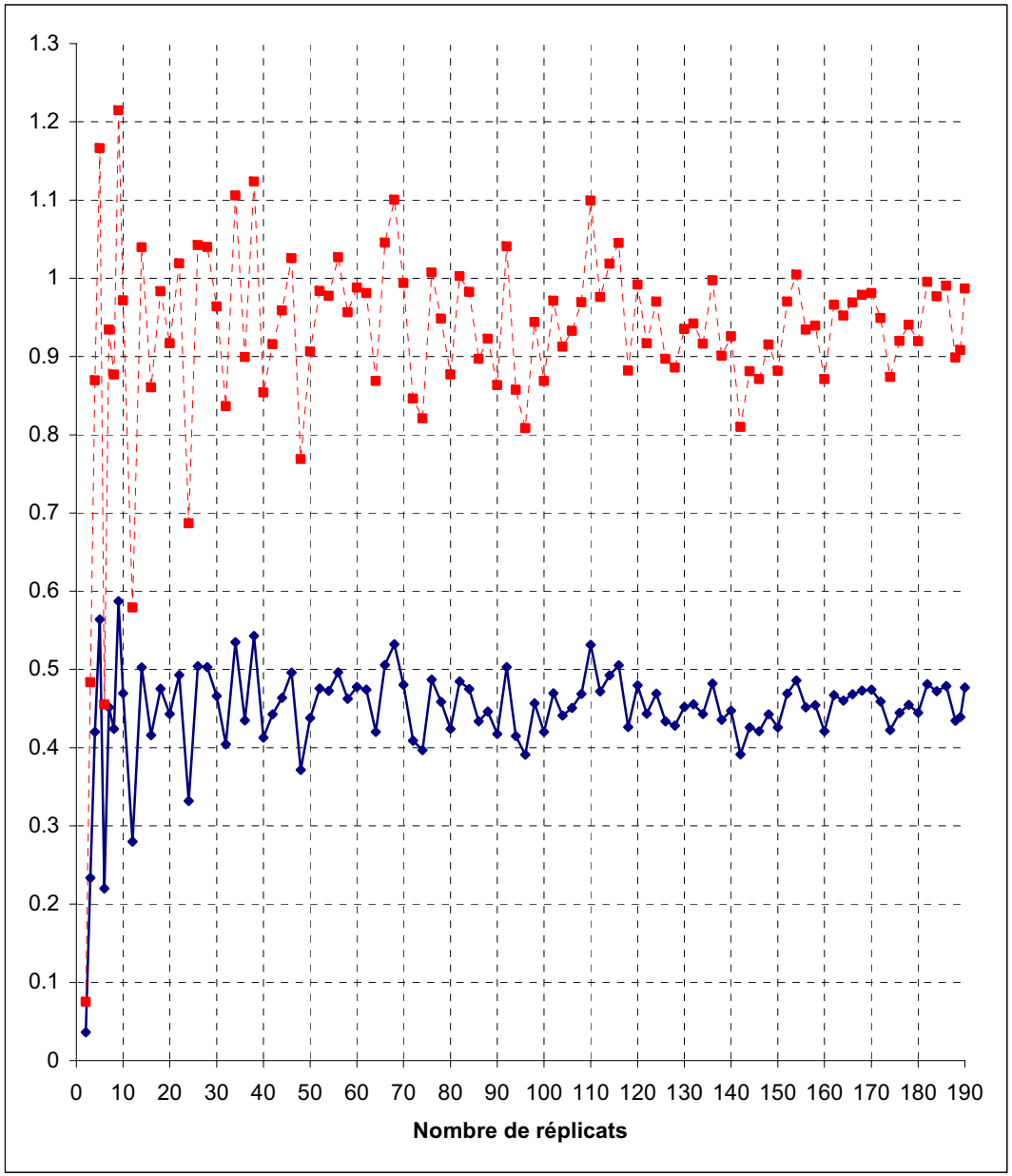


G5 : Sans poids, en respectant les strates, échelle conventionnelle, sex=1

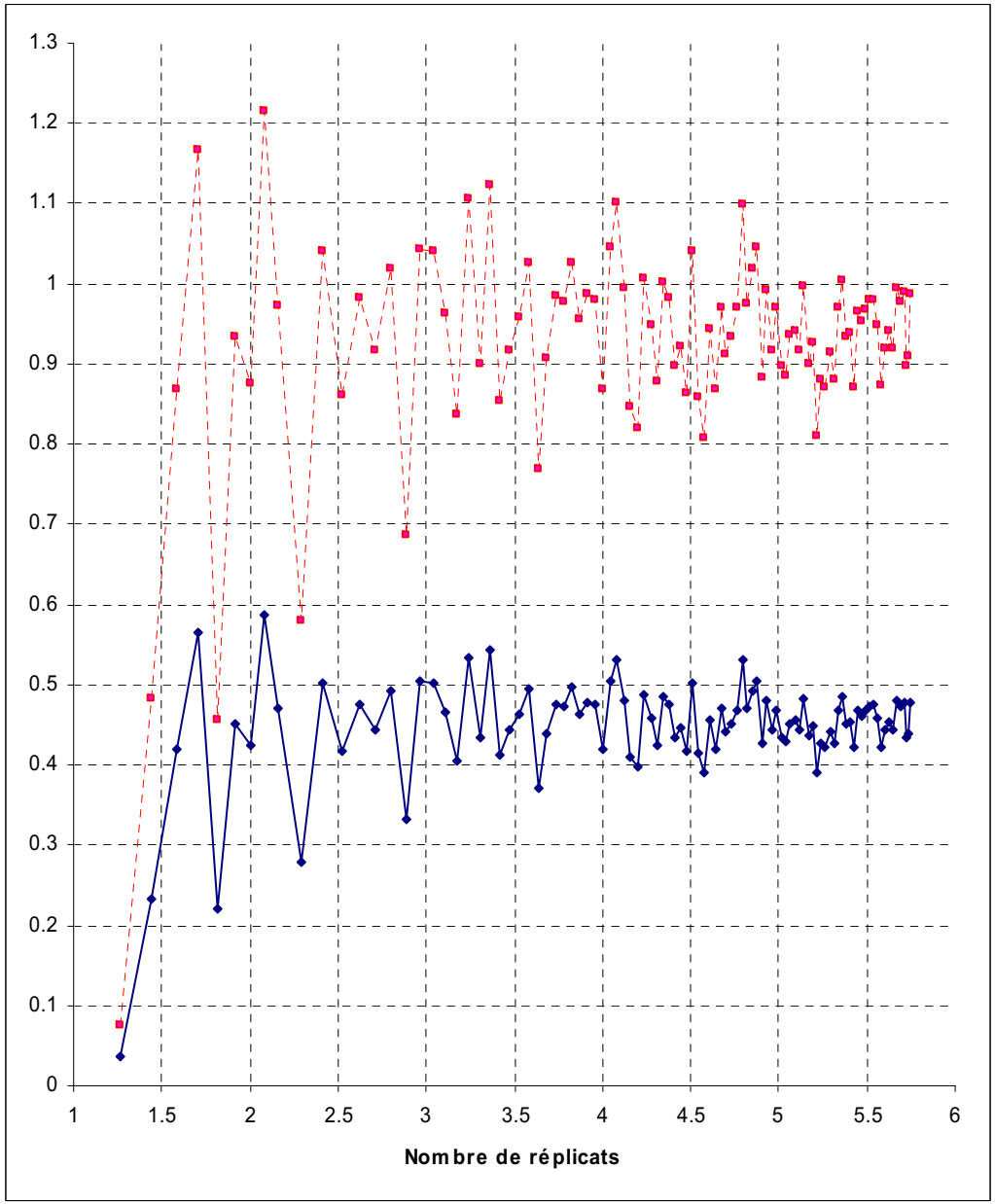


G6 : Sans poids, en respectant les strates, échelle racine cubique, sex=1



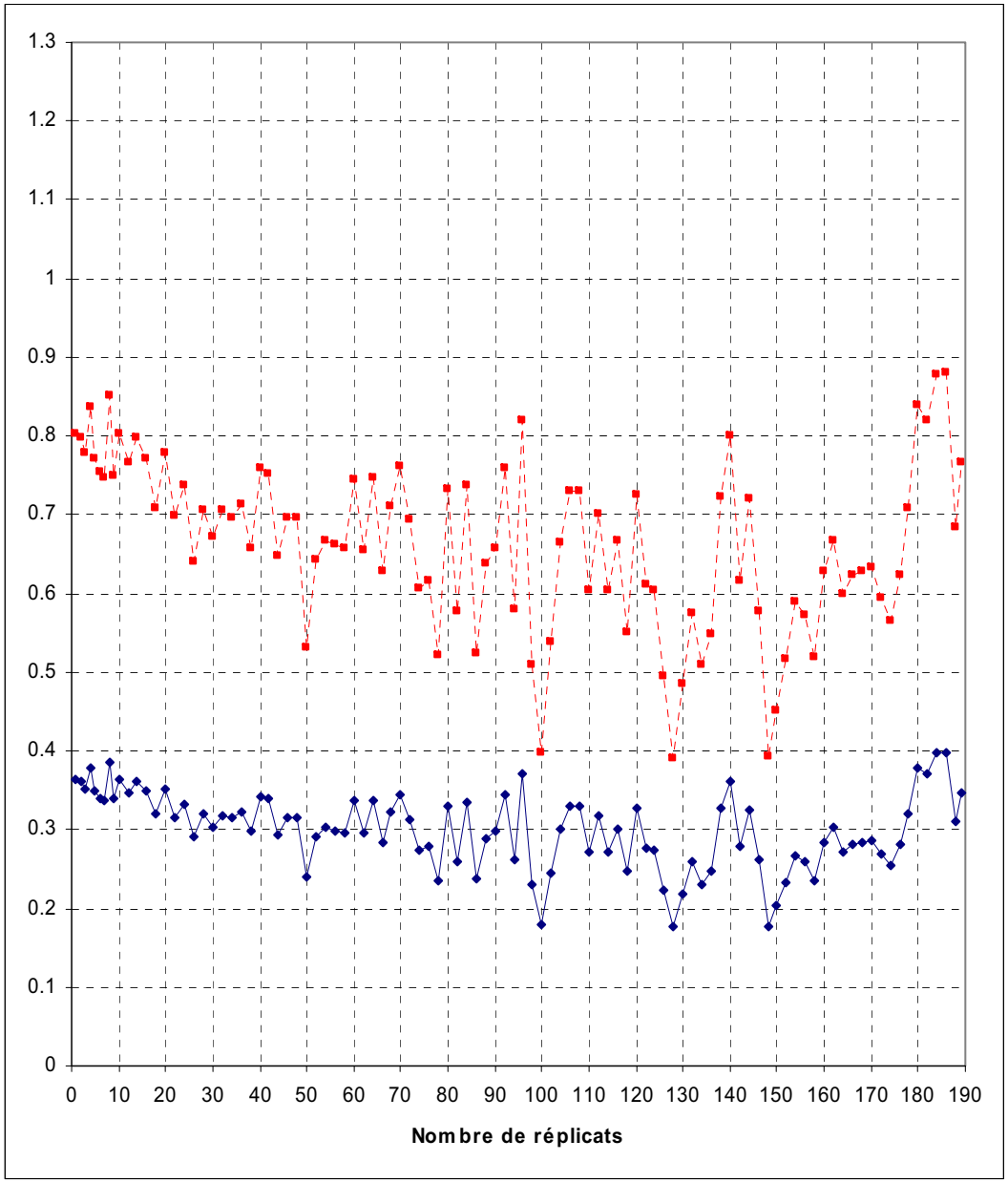


G7 : Avec poids w^{FIN}, en respectant les strates, échelle conventionnelle, sex=1

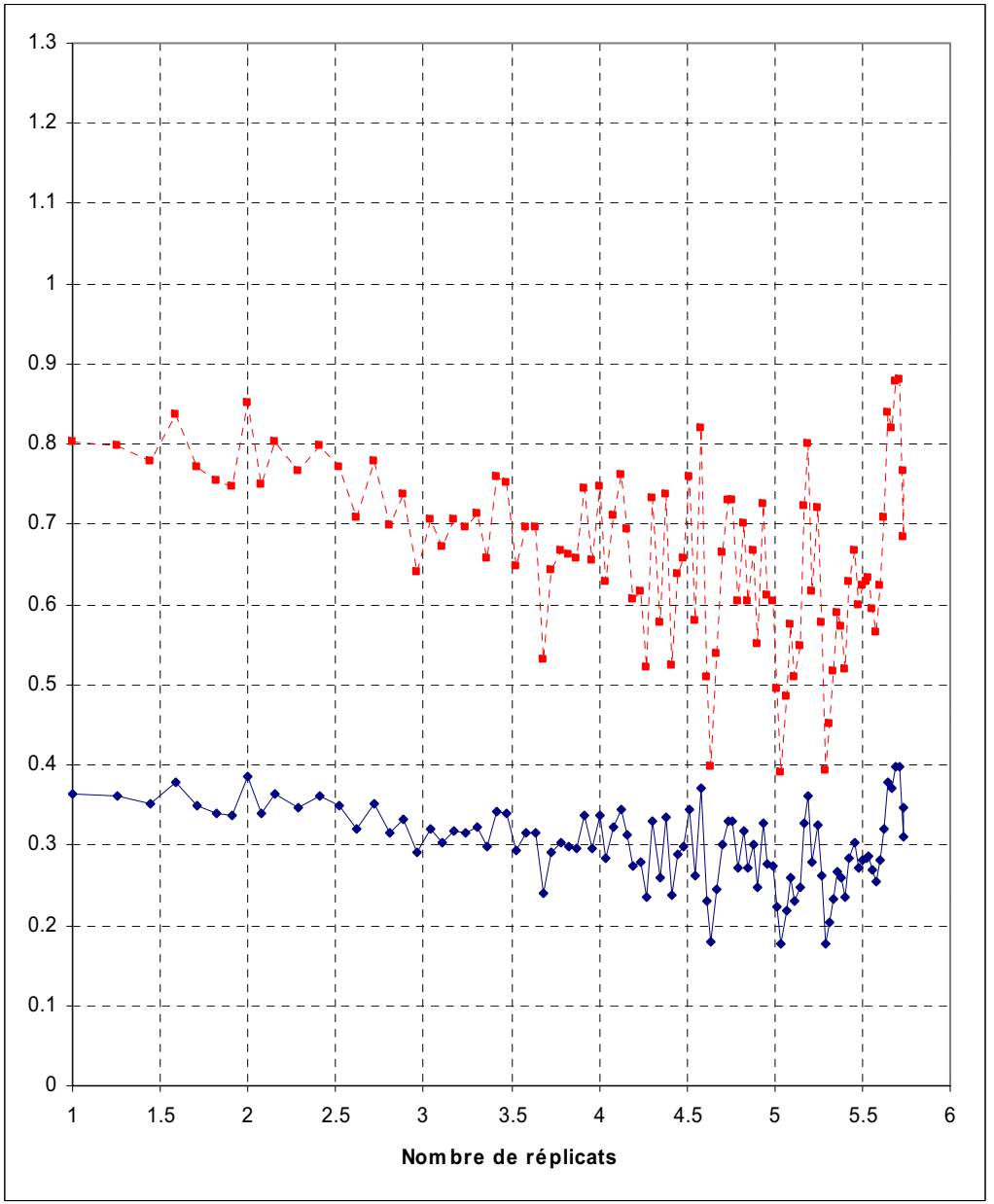


G8 : Avec poids w^{FIN}, en respectant les strates, échelle racine cubique, sex=1

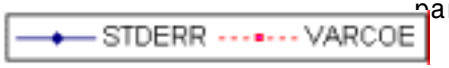


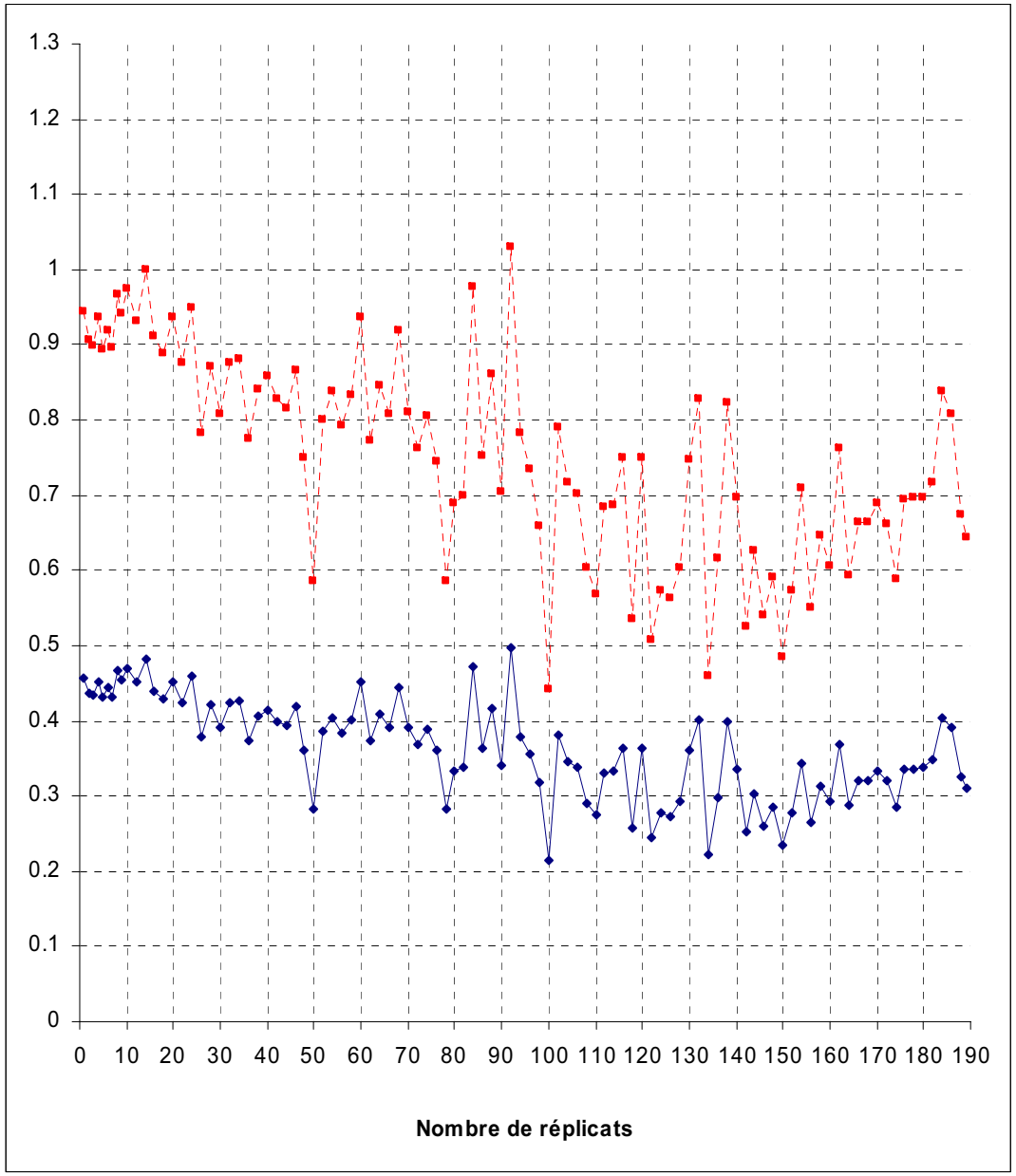


G9 : Sans poids, en enlevant un nbr. donné d'obs. par strates, échelle conventionnelle, sex=1

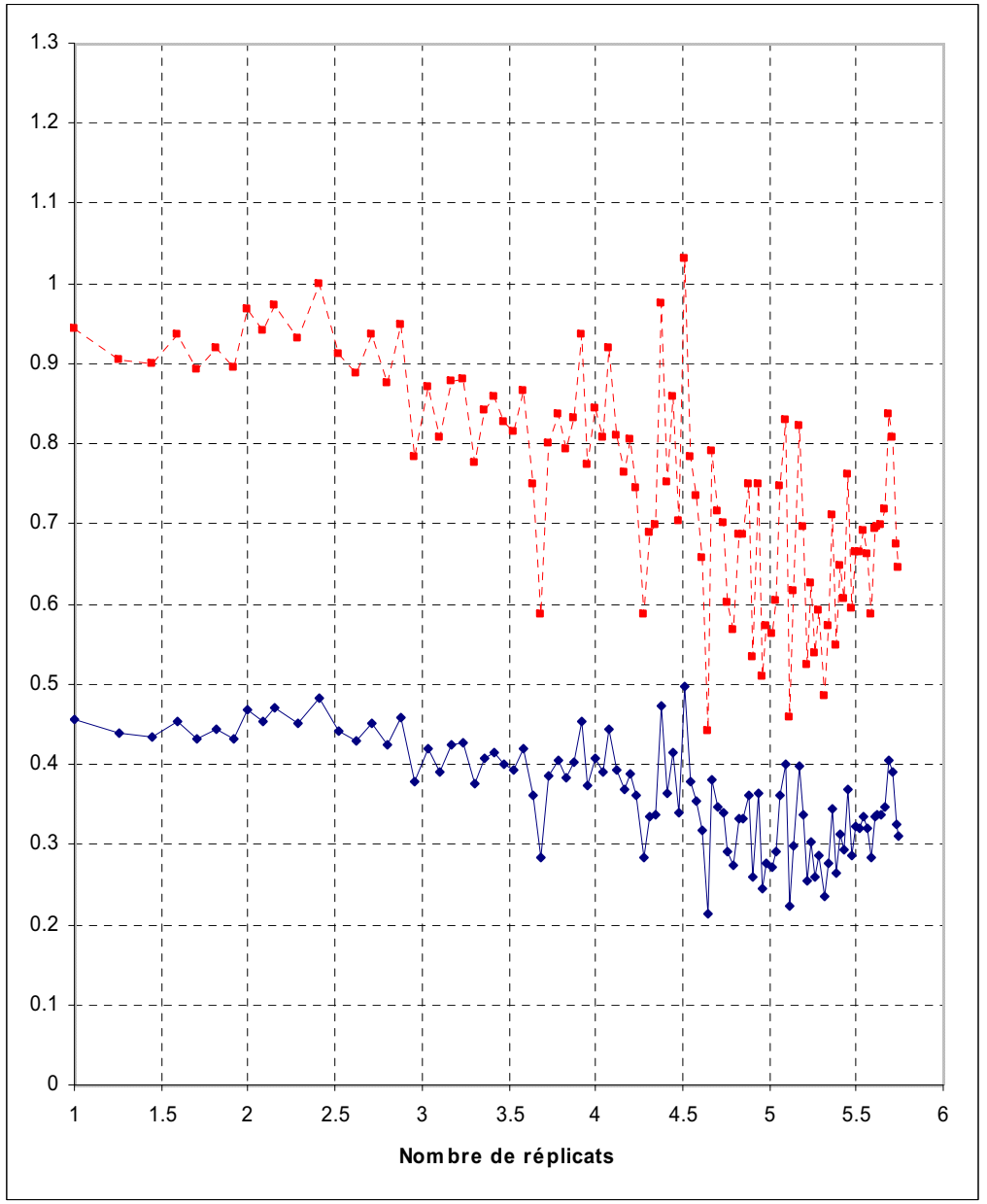


G10 : Sans poids, en enlevant un nbr. donné d'obs. par strates, échelle racine cubique, sex=1





G11 : Avec poids w^{FIN}, en enlevant un nbr. donné d'obs. par strates, échelle conventionnelle, sex=1



G12 : Avec poids w^{FIN}, en enlevant un nbr. donné d'obs. par strates, échelle racine cubique, sex=1



Références

Canty, A. J. and Davison, A. C. (1998). Variance estimation for two complex surveys in Switzerland. Rapport non publié.

Canty, A. J. and Davison, A. C. (1999). Resampling-based variance estimation for labour force surveys. *The Statistician*, **48** :379-391.

Cochran, W.G. (1977). *Sampling techniques*. John Wiley & Sons, New York.

Deville, J.-C. and Särndal, C.-E. (1992). Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association*, **87**(418) :376-382.

Deville, J.-C., Särndal, C.-E. and Sautory, O. (1993). Generalized Raking Procedures in Survey Sampling. *Journal of the American Statistical Association*, **88**(423) :1013-1020.

OFS (1991). Section de la santé. Conception générale de l'enquête suisse sur la santé 1992/93. OFS, Berne.

OFS (1994). Comment, T. *SAKE-News* 2/94. Enquête suisse sur la population active. Intervalles de confiance. OFS, Berne.

OFS (1994). Vonlanthen, C. Pondération de l'enquête suisse sur la santé 1992/93. OFS, Berne.

OFS (1997). Vonlanthen, C. Méthodologie statistique de l'enquête suisse sur la santé 1992/93. OFS, Berne.

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

Renaud, A. (2004). Analyse de données d'enquêtes. Quelques méthodes et illustrations avec les données de l'OFS. Rapport de méthodes 338-0024, OFS.

Särndal, C. E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer Series in Statistics.

Sautory, O. (1993). La macro CALMAR. Redressement d'un échantillon par calage sur marges. Document de travail n° F 9310 de la direction générale de l'Institut National de la Statistique et des études économiques (INSEE).

Vonlanthen, C. (1997). Méthodes statistiques de l'enquête suisse sur la santé 1997. Rapport non publié. OFS, Neuchâtel.

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

Graf, E., Renfer J.-P. (2005). Enquête suisse sur la santé 2002. Plan d'échantillonnage, pondération et estimation de la précision. Numéro de commande 338-0030

Potterat, J. (2005). Mietpreis-Strukturerhebung 2003. Gewichtung und Schätzverfahren. Bestellnummer: 338-0029

Potterat, J. (2005). Landwirtschaftliche Betriebszählung 2003. Schätzverfahren für die Zusatzerhebung. Bestellnummer: 338-0028

Renaud, A. (2004). Coverage estimation for the Swiss population census 2000. Estimation methodology and results. Order number: 338-0027

Kilchmann, D. (2004). Revision des Schweizerischen Lohnindex. Schätzmethoden der Lohnindices und deren Varianzschätzer. Bestellnummer: 338-0026

Graf, M. (2004). Enquête suisse sur la structure des salaires 2002. Plan d'échantillonnage et extrapolation pour le secteur privé. Numéro de commande: 338-0025

Renaud, A. (2004). Analyse de données d'enquêtes. Quelques méthodes et illustration avec des données de l'OFS. Numéro de commande 338-0024

Renaud, A., Potterat, J. (2004). Estimation de la couverture du recensement de la population de l'an 2000. Echantillon pour l'estimation de la sous-couverture (P-sample) et qualité du cadre de sondage des bâtiments. Numéro de commande: 338-0023

Graf, M. (2004). Fusion de données. Etude de faisabilité. Numéro de commande: 338-0022

Potterat, J. (2003). Mietpreis-Strukturerhebung 2003. Entwicklung des Stichprobenplans und Ziehung der Stichprobe. Bestellnummer: 338-0021

Potterat, J. (2003). Landwirtschaftliche Betriebszählung 2003. Stichprobenplan der Zusatzerhebung. Bestellnummer: 338-0020

Renaud, A. (2003). Estimation de la couverture du recensement de la population de l'an 2000. Echantillon pour l'estimation de la sur-couverture (E-sample). Numéro de commande: 338-0019

Hulliger, B. (2003). Erhebung über Forschung und Entwicklung in der schweizerischen Privatwirtschaft 2000. Bereinigung der Stichprobe, Ersatz fehlender Werte und Schätzverfahren. Bestellnummer: 338-0018

Renfer, J.-P. (2003). Enquête 2000 sur la recherche et le développement dans l'économie privée en Suisse. Plan d'échantillonnage. Numéro de commande: 338-0017

Potterat, J. (2003). Kosten und Nutzen der Berufsbildung aus Sicht der Betriebe. Schätzverfahren. Bestellnummer: 338-0016

Graf, M., Matei, A. (2003). Stratégie de choix des modèles de désaisonnalisation. Application aux séries de l'emploi total. Numéro de commande : 338-0015

Potterat, J., Salamin, P.A. (2002). Betriebszählung 2001. Methoden für die Datenbereinigung. Bestellnummer: 338-0014

Renaud, A. (2002). Programme international pour le suivi des acquis des élèves (PISA). Plans d'échantillonnage pour PISA 2000 en Suisse. Numéro de commande: 338-0013

Renfer, J.-P. (2002). Enquête 2001 sur les coûts et l'utilité de la formation des apprentis du point de vue des établissements. Plan d'échantillonnage. Numéro de commande: 338-0012

Potterat, J., Salamin, P.A. (2002). Betriebszählung 2001. Stichprobenplan und Schätzverfahren für die provisorischen Ergebnisse. Bestellnummer: 338-0011

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

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

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

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

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

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

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

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

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

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

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

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

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

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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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 info@bfs.admin.ch	Service de renseignements individuels
Das BFS im Internet	www.statistik.admin.ch	L'OFS sur Internet
Medienmitteilungen zur raschen Information der Öffentlichkeit über die neusten Ergebnisse	www.news-stat.admin.ch	Communiqués de presse: information rapide concernant les résultats les plus récents
Publikationen zur vertieften Information (zum Teil auch als Diskette/CD-Rom)	032 713 60 60 order@bfs.admin.ch	Publications: information approfondie (certaines sont disponibles sur disquette/CD-Rom)
Online-Datenbank	032 713 60 86 www.statweb.admin.ch	Banque de données (accessible en ligne)

Nähere Angaben zu den verschiedenen Diffusionsmitteln liefert das laufend nachgeführte Publikationsverzeichnis im Internet unter der Adresse www.statistik.admin.ch>>News>>Neuerscheinungen.

La Liste des publications mise à jour régulièrement, donne davantage de détails sur les divers moyens de diffusion. Elle se trouve sur Internet à l'adresse www.statistique.admin.ch >>Actualités>>Nouvelles publications.

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.

Die Methodenberichte sind auf der Internetseite des BFS in elektronischer Form verfügbar.

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.

Les rapports de méthodes sont disponibles sous forme électronique sur le site internet de l'OFS.

L'ESS02 a pour objectif principal de fournir, à partir d'un échantillon de près de 20'000 personnes résidant en Suisse et âgées de 15 ans et plus, des résultats sur l'état de santé de la population, les comportements en matière de santé, l'assurance-maladie et le recours aux prestations du système de santé.

Ce rapport décrit les méthodes utilisées pour établir le plan d'échantillonnage en abordant les questions relatives à la délimitation de la population, au cadre de sondage et aux unités d'échantillonnage. Le tirage, la gestion et l'érosion des échantillons font également partie du chapitre consacré à l'échantillonnage.

L'ESS se compose en fait de deux enquêtes: la première est une enquête téléphonique alors que la seconde, sous forme écrite, ne s'adresse qu'aux personnes ayant répondu à la première. La pondération relative à chacune de ces enquêtes y est décrite dans les deux chapitres suivants.

Enfin les méthodes pour calculer la précision des résultats obtenus sont décrites et des exemples de calculs d'intervalles de confiance sont fournis.

N° de commande: 338-0030
Commandes: 032 713 60 60
Fax: 032 713 60 61
E-Mail: order@bfs.admin.ch
Prix: gratuit
ISBN 3-303-00314-9