

00

Bases statistiques et produits généraux

338-0066

Rapport de méthodes

Enquête sur la situation sociale et économique des étudiant·e·s des hautes écoles suisses 2009

Cadre de sondage, plan d'échantillonnage et méthodes d'estimation

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Département fédéral de l'intérieur DFI
Office fédéral de la statistique OFS

Neuchâtel, 2012

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 Mobilité et transports

12 Monnaie, banques, assurances

13 Protection sociale

14 Santé

15 Education et science

16 Culture, médias, 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

Rapport de méthodes

Enquête sur la situation sociale et économique des étudiant·e·s des hautes écoles suisses 2009

Cadre de sondage, plan d'échantillonnage et méthodes d'estimation

Auteurs Beatriz Andrade, Paul-André Salamin

Editeur Office fédéral de la statistique (OFS)

Office fédéral de la statistique (OFS)
Neuchâtel, 2012

Editeur:	Office fédéral de la statistique (OFS)
Complément d'information:	Beatriz Andrade, tél. 032 713 68 19, e-mail: Beatriz.Andrade@bfs.admin.ch Paul-André Salamin, Haute Ecole Spécialisée de Suisse occidentale, Valais, e-mail: pandre.salamin@hevs.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 / e-mail: order@bfs.admin.ch
Internet:	http://www.statistique.admin.ch
Numéro de commande:	338-0066
Prix:	gratuit
Série:	Statistique de la Suisse
Domaine:	0 Bases statistiques et produits généraux
Langue du texte original:	Français
Page de couverture:	OFS; concept: Netthoevel & Gaberthüel, Bienne; photo: © NorthShoreSurfPhotos – Fotolia.com
Graphisme/Layout:	Section DIAM, Prepress / Print
Copyright:	OFS, Neuchâtel 2012 La reproduction est autorisée, sauf à des fins commerciales, si la source est mentionnée
ISBN:	978-3-303-00472-2

Table des matières

1 Introduction **7**

2 Cadre de tirage **7**
2.1 Cadre de tirage 2009 pour les hautes écoles universitaires 8
2.2 Cadre de tirage 2009 pour les hautes écoles spécialisées/pédagogiques 8

3 Plan d'échantillonnage **10**
3.1 Calcul de la variance d'une proportion 10
3.2 Taille d'échantillon 11
3.3 Allocation 13
3.4 Résultats de l'allocation 16

4 Tirage des échantillons **21**

5 Méthodes d'estimation **25**
5.1 Extrapolation 25
5.2 Exemples d'extrapolation 33

6 Conclusion **39**

A Annexes **40**
A.1 Préparation des données 40
A.2 Allocation pour une moyenne 41

Références **43**

Rapports de méthodes du Service de méthodes statistiques **45**

Préambule

La section BILD-S de l'Office fédéral de la statistique a demandé au Service de méthodes statistiques (METH) un soutien pour l'élaboration du plan d'échantillonnage, la pondération des données et les estimations de l'enquête sur la situation sociale et économique des étudiant-e-s. Ce rapport décrit en détails ces différentes étapes. Ce travail a été réalisé par Beatriz Andrade et Paul-André Salamin, METH.

Un grand merci à la section BILD-S, en particulier à Laurence Boegli, Sarah Gerhard et Martin Teichgräber pour les fructueuses discussions tout au long de ce travail. Merci également à Monique Graf (METH) pour la relecture attentive du rapport.

Résumé

Ce rapport documente le plan d'échantillonnage, l'allocation et le tirage de l'échantillon ainsi que les méthodes d'estimation utilisées pour la statistique sur la situation sociale et économique des étudiant-e-s. Cette enquête a été menée pour la première fois en 2005 et a été reconduite en 2009. Le plan d'échantillonnage en 2009 est défini de la même manière qu'en 2005. Il a été construit dans le but de satisfaire les objectifs de précision de cette enquête. La stratification correspond au croisement des hautes écoles (université, école polytechnique, haute école spécialisée et haute école pédagogique) avec le domaine d'études (groupe de domaines d'études pour les hautes écoles universitaires et domaine d'études pour les hautes écoles spécialisées et pédagogiques). Une allocation qui garantit la précision au niveau haute école et domaine d'études resp. groupe de domaines d'études et qui tient compte de la non-réponse est appliquée. Le nouvel échantillon contient 11'500 étudiants pour les hautes écoles universitaires et 13'014 pour les hautes écoles spécialisées et pédagogiques, y compris les augmentations de l'échantillon demandées par une haute école universitaire et une haute école spécialisée.

L'échantillon de l'Université de Genève représente un cas particulier, car cette université a effectué presque en même temps une enquête auprès de ses étudiant-e-s avec un questionnaire similaire au questionnaire de l'OFS. Pour minimiser le nombre des étudiant-e-s qui sont interrogés deux fois, on a utilisé l'algorithme de Kish & Scott qui permet de faire une coordination entre les deux échantillons.

L'extrapolation de l'échantillon à la population est expliquée à l'aide de différents exemples. Pour terminer, des comparaisons entre estimateurs de variance sont présentées pour valider l'estimateur de variance utilisé dans cette enquête.

1 Introduction

La Statistique sur la situation sociale et économique des étudiant-e-s fournit des informations sur les conditions d'études des étudiant-e-s pour l'ensemble de la Suisse, différenciées par haute école (Université, école polytechnique, haute école spécialisée et haute école pédagogique) ou domaine d'études (groupe de domaines d'études pour les hautes écoles universitaires et domaine d'études pour les hautes écoles spécialisées et pédagogiques). La date de l'exécution de l'enquête est le semestre de printemps. L'enquête a été menée pour la première fois en 2005 à la même période et a été reconduite en 2009.

La population d'intérêt est formée de tous les étudiant-e-s immatriculés dans une haute école suisse (Université, école polytechnique, haute école spécialisée et haute école pédagogique) pour y obtenir une licence/diplôme, un bachelor ou un master. Cette population de base est définie grâce aux fichiers du Système d'information universitaire suisse (SIUS).

L'enquête est basée sur des échantillons stratifiés par haute école et groupe de domaines d'études resp. domaine d'études. Une pondération initiale avec un facteur de correction pour la non-réponse suivie d'un calage par marges permettent l'estimation des résultats pour l'ensemble de la population d'étude. La variance a été estimée par un estimateur par calage. Une comparaison de quatre estimateurs de variance permet de valider cette approche.

Le présent rapport documente le plan d'échantillonnage, l'allocation et le tirage de l'échantillon ainsi que les méthodes d'estimation utilisées pour 2009.

2 Cadre de tirage

La population d'intérêt est formée de tous les étudiant-e-s immatriculés au semestre d'automne 2009 dans une haute école suisse (Université, école polytechnique, haute école spécialisée et haute école pédagogique) pour y obtenir une licence/diplôme, un bachelor ou un master. Cette population de base est définie grâce aux fichiers du Système d'Information Universitaire Suisse (SIUS). L'état du SIUS date de janvier 2009 pour les hautes écoles universitaires (HEU) et les hautes écoles spécialisées et pédagogiques (HES/HEP).

Le cadre initial des HEU est formé de 96'352 étudiant-e-s et celui des HES/HEP de 57'212 étudiant-e-s.

Deux plans d'échantillonnage ont été définis pour les étudiant-e-s des :

– Hautes écoles universitaires (HEU)
– Hautes écoles spécialisées et pédagogiques (HES/HEP).

Dans le plan d'échantillonnage, la stratification correspond au croisement du type des hautes écoles (HEU ou HES/HEP) avec le domaine d'études (groupe de domaines d'études pour HEU et domaine d'études pour HES/HEP). On calcule la taille de l'échantillon en se donnant une précision à atteindre pour l'estimation d'une proportion au niveau des hautes écoles et des groupes de domaines d'études resp. des domaines d'études (strates marginales). On augmente la taille de l'échantillon, éventuellement jusqu'à la taille de la strate, pour tenir compte des taux de réponse par HEU et HES/HEP.

Les taux de réponses utilisés sont ceux observés lors de l'enquête 2005 moins 5%. Ils varient entre 46.4% et 68.4% pour les HEU et entre 50% et 68.1% pour les HES/HEP. Pour la Kalaidos Fachhochschule on a supposé un taux de réponse de 50%, puisqu'il s'agit d'une nouvelle haute école spécialisée. Les strates avec moins de 100 étudiant-e-s sont exhaustives. La

taille d'échantillon est d'au moins 50 unités par strate et la taille maximale des deux échantillons (HEU et HES/HEP) est 10'000 étudiant-e-s chacun sans compter les augmentations des échantillons de deux hautes écoles.

Le tirage des échantillons stratifiés a été effectué avec la procédure `proc surveyselect` du logiciel SAS. Un échantillon aléatoire simple pour les HEU respectivement pour les HES/HEP a été tiré dans chaque strate avec une taille de l'échantillon donnée par le plan d'échantillonnage décrit ci-dessous.

2.1 Cadre de tirage 2009 pour les hautes écoles universitaires

La stratification correspond au croisement des HEU avec le groupe de domaines d'études. Il y a 7 groupes de domaines d'études dans le cas des HEU (voir Table 1).

TABLE 1 Groupes de domaines d'études des HEU

FBG1	Sciences humaines et sociales
FBG2	Sciences économiques
FBG3	Droit
FBG4	Sciences exactes et naturelles
FBG5	Médecine et pharmacie
FBG6	Sciences techniques
FBG7	Interdisciplinaire et autre

La taille de la population par strate (HEU x groupes de domaines d'études) est résumée dans la Table 2.

TABLE 2 Nombre d'étudiant-e-s par HEU et groupe de domaines d'études

HEU	FGB1	FBG2	FBG3	FBG4	FBG5	FBG6	FBG7	Total
Basel	3279	947	1120	1327	1518	.	961	9152
Bern	4116	1508	1748	1466	1646	.	670	11154
Fribourg	4525	1010	1422	843	298	.	74	8172
Genève	5281	1135	1209	853	1346	.	436	10260
Lausanne	3847	1705	1307	902	1363	.	367	9491
Luzern	929	1	1076	.	.	.	8	2014
Neuchâtel	1673	346	535	447	74	22	16	3113
St. Gallen	407	3995	521	.	.	.	.	4923
Zürich	10589	2608	2975	2066	2081	.	19	20338
Svizzera italiana	647	732	.	129	.	655	.	2163
Fernstudien Schweiz	127	114	147	.	.	.	.	388
Institut universitaire Kurt Bösch	.	.	.	.	.	.	18	18
EPFL	.	.	.	1746	.	3012	.	4758
ETHZ	29	.	.	3987	375	5420	587	10398
Total	35449	14101	12060	13766	8701	9109	3156	96342

2.2 Cadre de tirage 2009 pour les hautes écoles spécialisées/pédagogiques

La stratification correspond au croisement des HES/HEP avec le domaine d'études. Il y a 13 domaines d'études dans le cas des HES/HEP (voir Table 3). La liste des HES/HEP se trouve dans la Table 4.

TABLE 3 Liste des domaines d'études des HES/HEP

FBG1	Architecture, construction et planification
FBG2	Technique et IT
FBG3	Chimie et sciences de la vie
FBG4	Agronomie et économie forestière
FBG5	Economie et services
FBG6	Design
FBG7	Sport
FBG8	Musique, arts de la scène et autres arts
FBG9	Linguistique appliquée
FBG10	Travail social
FBG11	Psychologie appliquée
FBG12	Santé
FBG13	Formation des enseignants

TABLE 4 Liste des HES/HEP

Nomenclature 2009	
100	Berner Fachhochschule
300	Haute Ecole Spécialisée de Suisse occidentale
400	Fachhochschule Nordwestschweiz
800	Fachhochschule Zentralschweiz
900	Scuola Universitaria Professionale della Svizzera Italiana
1000	Fachhochschule Ostschweiz
1100	Zürcher Fachhochschule
1250	Kalaidos Fachhochschule
1500	Andere PH (nicht integriert) und Institutionen der Lehrkräfteausbildung

La taille de la population par strate (HES/HEP x domaine d'études) est résumée dans la Table 5.

TABLE 5 Nombre d'étudiant-e-s par HES/HEP et domaine d'études

HES/HEP	FGB1	FBG2	FBG3	FBG4	FBG5	FBG6	FBG7	FBG8	FBG9	FBG10	FBG11	FBG12	FBG13	Total
100	393	1065	73	293	811	184	131	612	.	360	.	542	.	4464
300	823	2465	464	44	4078	581	.	1491	.	2477	.	2204	87	14714
400	275	1007	378	.	1802	423	.	717	.	1084	238	.	1570	7494
800	264	722	.	.	1135	308	.	648	.	584	.	.	.	3661
900	232	448	.	.	824	209	.	226	.	237	.	274	.	2450
1000	456	1179	.	.	1275	.	.	.	.	445	.	123	.	3478
1100	466	1460	717	.	3782	554	.	1101	500	598	230	785	1713	11906
1250	.	.	.	.	513	.	.	.	.	.	.	22	.	535
1500	.	.	.	.	.	.	.	.	.	.	.	.	8510	8510
Total	2909	8346	1632	337	14220	2259	131	4795	500	5785	468	3950	11880	57212

3 Plan d'échantillonnage

3.1 Calcul de la variance d'une proportion

Ce paragraphe s'appuie sur la réference [1]. Par la suite on utilise dans ce paragraphe domaine d'études à la place de domaine d'études et groupe de domaines d'études.

La stratification a été définie aux tables 2 et 5.

TABLE 6 Notations

U	Population des étudiant-e-s
N	Nombre d'étudiant-e-s dans la population U
U_{ij}	Population des étudiant-e-s dans la haute école i et le domaine d'études j
N_{ij}	Nombre d'étudiant-e-s dans la strate U_{ij}
y_k	Variable indicatrice d'un domaine $A \subseteq U$
U_{i+} et U_{+j}	Strates marginales
N_{i+} et N_{+j}	Nombre d'étudiant-e-s dans les strates marginales U_{i+} et U_{+j}
V	Variance désirée
D_h	Variance de y dans la strate U_h

Les variables utilisées pour la stratification sont la haute école (HEU ou HES/HEP) et le domaine d'études. Les croisements entre haute école et domaine d'études déterminent les strates. Il s'agit donc d'une stratification selon deux critères :

$$U = \bigcup_{i,j} U_{ij}.$$

On veut estimer une proportion

$$p = \frac{|A|}{|U|}$$

où $A \subseteq U$ est un domaine d'intérêt. Si on définit la variable indicatrice

$$y_k = \begin{cases} 1 & \text{si } k \in A \\ 0 & \text{si } k \notin A \end{cases}$$

alors on peut aussi écrire

$$p = \frac{|A|}{|U|} = \frac{1}{N} \sum_{k \in U} y_k = \frac{Y}{N}.$$

Nous avons aussi

$$p = \frac{1}{N} \sum_{i,j} \sum_{k \in U_{ij}} y_k = \sum_{i,j} \frac{N_{ij}}{N} \frac{1}{N_{ij}} \sum_{k \in U_{ij}} y_k = \sum_{i,j} W_{ij} p_{ij} \text{ où } W_{ij} = \frac{N_{ij}}{N} \text{ et } p_{ij} = \frac{1}{N_{ij}} \sum_{k \in U_{ij}} y_k.$$

On tire un échantillon aléatoire simple S_{ij} de taille n_{ij} dans chaque strate U_{ij}. L'estimateur de Horvitz-Thompson de p est donné par

$$\widehat{p} = \sum_{i,j} W_{ij} \widehat{p}_{ij} = \sum_{i,j} W_{ij} \frac{1}{n_{ij}} \sum_{k \in S_{ij}} y_k.$$

La variance de $\widehat{p}$ est donnée par

$$\mathrm{Var}\left(\widehat{p}\right) = \sum_{i,j} W_{ij}^2 \mathrm{Var}\left(\widehat{p}_{ij}\right) = \sum_{i,j} W_{ij}^2 \left(1 - \frac{n_{ij}}{N_{ij}}\right) \frac{1}{n_{ij}} D_{ij}^2,$$

où D_{ij}^2 est la variance de y dans la strate U_{ij}. Comme y ne prend que les valeurs 0 ou 1, on trouve que

$$D_{ij}^2 = \frac{N_{ij}}{N_{ij} - 1} p_{ij} \left(1 - p_{ij}\right).$$

Avec l'approximation $N_{ij} / \left(N_{ij} - 1\right) \approx 1$, on obtient finalement

$$\mathrm{Var}\left(\widehat{p}\right) \approx \sum_{i,j} W_{ij}^2 \left(1 - \frac{n_{ij}}{N_{ij}}\right) \frac{1}{n_{ij}} p_{ij} \left(1 - p_{ij}\right).$$

3.2 Taille d'échantillon

Ce paragraphe s'appuie sur la réference [1].

On détermine la taille de l'échantillon en se donnant une précision minimale à atteindre pour l'estimation d'une proportion au niveau des strates marginales $U_{i+} = \bigcup_j U_{ij}$ et $U_{+j} = \bigcup_i U_{ij}$. On note $N_{i+} = \sum_j N_{ij}$ et $N_{+j} = \sum_i N_{ij}$ les tailles des strates marginales. Les paramètres à estimer sont donc

$$p_i = \frac{1}{N_{i+}} \sum_{k \in U_{i+}} y_k = \frac{1}{N_{i+}} \sum_j \sum_{k \in U_{ij}} y_k = \sum_j \frac{N_{ij}}{N_{i+}} \frac{1}{N_{ij}} \sum_{k \in U_{ij}} y_k = \sum_j \left(\frac{N_{ij}}{N_{i+}}\right) p_{ij}$$

et

$$p_j = \frac{1}{N_{j+}} \sum_{k \in U_{+j}} y_k = \frac{1}{N_{+j}} \sum_i \sum_{k \in U_{ij}} y_k = \sum_i \frac{N_{ij}}{N_{+j}} \frac{1}{N_{ij}} \sum_{k \in U_{ij}} y_k = \sum_i \left(\frac{N_{ij}}{N_{+j}}\right) p_{ij}.$$

Les variances de $\widehat{p}_i$ et de $\widehat{p}_j$ sont données par

$$\mathrm{Var}\left(\widehat{p}_i\right) \approx \sum_j \left(\frac{N_{ij}}{N_{i+}}\right)^2 \left(1 - \frac{n_{ij}}{N_{ij}}\right) \frac{1}{n_{ij}} p_{ij} \left(1 - p_{ij}\right), \tag{1}$$

$$\mathrm{Var}\left(\widehat{p}_j\right) \approx \sum_i \left(\frac{N_{ij}}{N_{+j}}\right)^2 \left(1 - \frac{n_{ij}}{N_{ij}}\right) \frac{1}{n_{ij}} p_{ij} \left(1 - p_{ij}\right). \tag{2}$$

On suppose que nous ne disposons pas d'information a priori sur les p_{ij}. On prend la borne supérieure de $p_{ij}(1 - p_{ij})$ qui est atteinte quand $p_{ij} = p = 1/2$. Pour une allocation proportionnelle, la taille d'échantillon par strate marginale, sans la correction de population finie, est alors donnée, si on veut atteindre une variance V, par

$$\frac{D^2}{V} = \frac{p\left(1 - p\right)}{V} \leq \frac{1}{4V} = n_0 \text{ (voir annexe A.2.3).}$$

Avec les corrections de population finie sur U_{i+} et U_{+j} on obtient les tailles d'échantillon

$$n_{i+} = \frac{n_0}{1 + n_0 / N_{i+}} \text{ et } n_{+j} = \frac{n_0}{1 + n_0 / N_{+j}}. \tag{3}$$

Nous avons ensuite deux possibilités de faire une allocation proportionnelle dans les strates U_{ij}

$$n_{ij}^{(1)} = n_{i+} \frac{N_{ij}}{N_{i+}} \text{ et } n_{ij}^{(2)} = n_{+j} \frac{N_{ij}}{N_{+j}}$$

On prend $n_{ij} = \max(n_{ij}^{(1)}, n_{ij}^{(2)})$ comme taille de l'échantillon dans la strate U_{ij}, ce qui garantit la précision V pour $\widehat{p}_i$ et $\widehat{p}_j$.

Remarque 1 *Les formules 3 se déduisent de la théorie de la section A.2.3. On peut aussi utiliser directement les formules (1) ou (2). Par exemple, si* $p_{ij} = p$ *et* $n_{ij} = n_{i+}\,(N_{ij}/N_{i+})$*, alors*

$$\begin{aligned}
\mathrm{Var}\,(\widehat{p}_i) &\leq p\,(1-p) \sum_j \left(\frac{N_{ij}}{N_{i+}}\right)^2 \left(1 - \frac{n_{i+}N_{ij}}{N_{i+}N_{ij}}\right) \frac{N_{i+}}{n_{i+}N_{ij}} \\
&= p\,(1-p) \sum_j \left(\frac{N_{ij}}{N_{i+}}\right)^2 \left(1 - \frac{n_{i+}}{N_{i+}}\right) \frac{N_{i+}}{n_{i+}N_{ij}} \\
&= p\,(1-p) \left(1 - \frac{n_{i+}}{N_{i+}}\right) \sum_j \left(\frac{N_{ij}}{N_{i+}}\right)^2 \frac{N_{i+}}{n_{i+}N_{ij}} \\
&= p\,(1-p)\,\frac{1}{n_{i+}} \left(1 - \frac{n_{i+}}{N_{i+}}\right) \underbrace{\sum_j \frac{N_{ij}}{N_{i+}}}_{1}
\end{aligned}$$

On trouve finalement

$$\mathrm{Var}\,(\widehat{p}_i) \approx p\,(1-p)\,\frac{1}{n_{i+}} \left(1 - \frac{n_{i+}}{N_{i+}}\right).$$

et ainsi on peut aussi obtenir les tailles d'échantillon. Soit alors V *la variance désirée de l'estimateur* $\widehat{p}_i$*. Alors de*

$$\left(1 - \frac{n_{i+}}{N_{i+}}\right) \frac{1}{n_{i+}} p\,(1-p) = V$$

on tire, avec p=0.5,

$$\left(1 - \frac{n_{i+}}{N_{i+}}\right) \frac{1}{n_{i+}} = 4V = \frac{1}{n_0}$$

Ce qui implique que

$$n_{i+} = \frac{n_0}{1 + n_0/N_{i+}}.$$

De la même manière on obtient n_{+j}*.*

3.3 Allocation

Par la suite on utilise dans ce paragraphe branche à la place de domaine d'études et groupe de domaines d'études.

L'allocation est effectuée en plusieurs étapes :

1. On calcule avec la formule

$$n_0 = \frac{p(1-p)}{V}$$

la taille d'échantillon par strate marginale, sans correction de population finie, en posant $p = \frac{1}{2}$ qui est une borne supérieure. V est un paramètre.

2. Ensuite pour obtenir les tailles n_{i+} et n_{+j} dans les strates marginales on applique la correction de population finie à la taille n_0. On utilise alors les formules suivantes :

$$n_{i+} = \frac{n_0}{1 + n_0/N_{i+}} \text{ et } n_{+j} = \frac{n_0}{1 + n_0/N_{+j}}.$$

3. On procède alors à une répartition des différentes tailles déterminées dans l'étape deux dans les croisements *hautes écoles x branches*, c'est-à-dire dans les strates de tirage $h = 1, ...H$. La répartition est une fois réalisée proportionnellement à la taille par rapport au total par haute école et une fois par rapport au total par branche. Pour obtenir les tailles de l'échantillon on utilise la formule suivante :

$$n_{ij}^{(1)} = n_{i+}\frac{N_{ij}}{N_{i+}} \text{ et } n_{ij}^{(2)} = n_{+j}\frac{N_{ij}}{N_{+j}}$$

4. Dans l'étape suivante, on choisit le maximum entre les deux valeurs déterminées dans chaque strate h à l'étape trois.

$$n_{ij} = \max(n_{ij}^{(1)}, n_{ij}^{(2)})$$

5. On augmente la taille de l'échantillon, éventuellement jusqu'à la taille de la strate, pour tenir compte des taux de réponse (θ_i) par haute école observés lors de l'enquête 2005 moins 5 % à l'exception de la Kalaidos Fachhochschule où le taux de réponse est supposé être 50% ; n_{ij} est alors divisé par θ_i.

6. Les strates avec moins de $N_{exh} = 100$ étudiant-e-s sont exhaustives. Pour tenir compte des strates exhaustives on procède de la manière suivante : Si

$$n_{ij} > N_{ij} \text{ ou } N_{ij} < 100$$

alors

$$n_{ij} = N_{ij}.$$

7. La taille d'échantillon est d'au moins $n_{min} = 50$ unités par strate ou autrement dit si $n_{ij} < 50$ alors $n_{ij} = 50$.

Les étapes 5-7 sont résumées dans la figure 1.

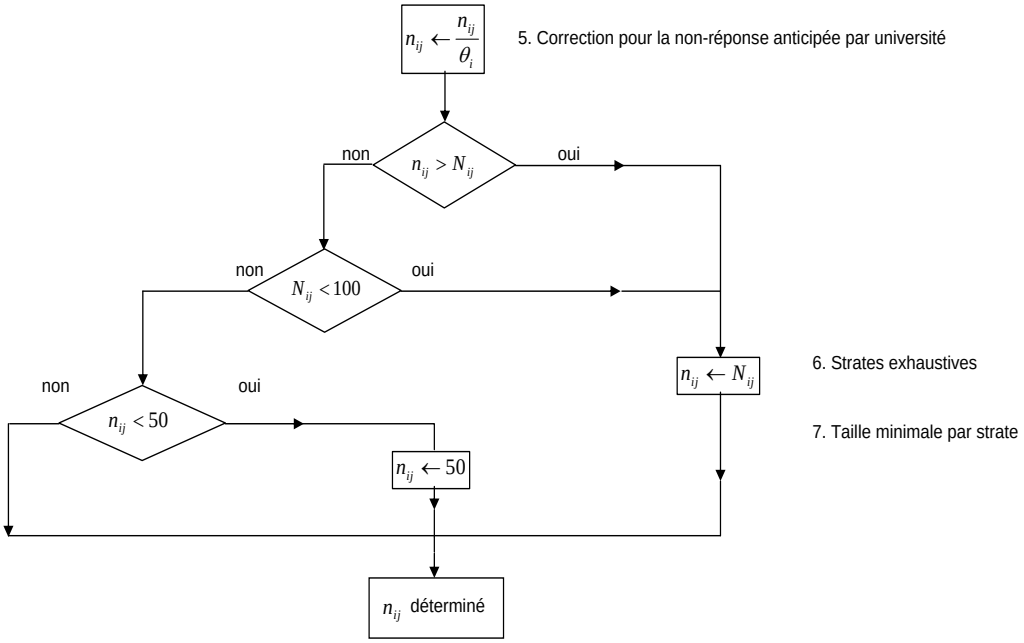


FIGURE 1 Résumé étapes 5-7

La taille de l'échantillon globale est alors $n = \sum_{i,j} n_{ij}$. Les étapes 1-7 définissent une fonction

$$V \rightarrow F(V) = n$$

$$n = F(V) = F(V, p, N_{ij}, \theta_i, N_{exh}, n_{min}).$$

On choisit $\sqrt{V}$ tel que $n \approx 10'000$. La précision par haute école est donnée dans la Table 7 et la fonction F(V) est représentée dans les Figures 2 et 3.

TABLE 7 Précision par haute école

	$\sqrt{V}$ (%)
HEU	2.39
HES/HEP	2.22

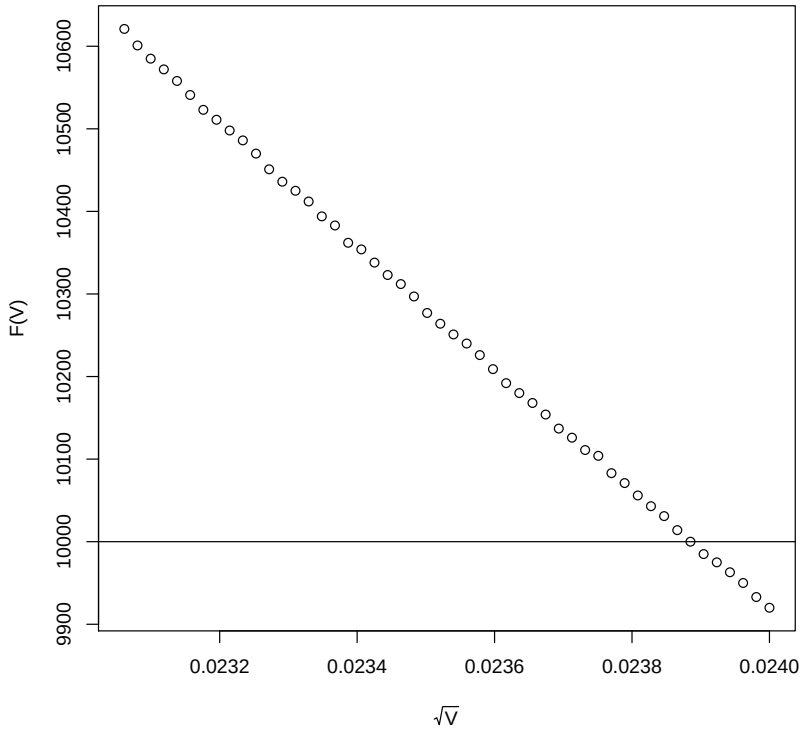


FIGURE 2 Fonction F(V) pour les HEU montre la relation entre la taille de l'échantillon et la précision $\sqrt{V}$. La ligne horizontale correspond à une taille d'échantillon de 10'000.

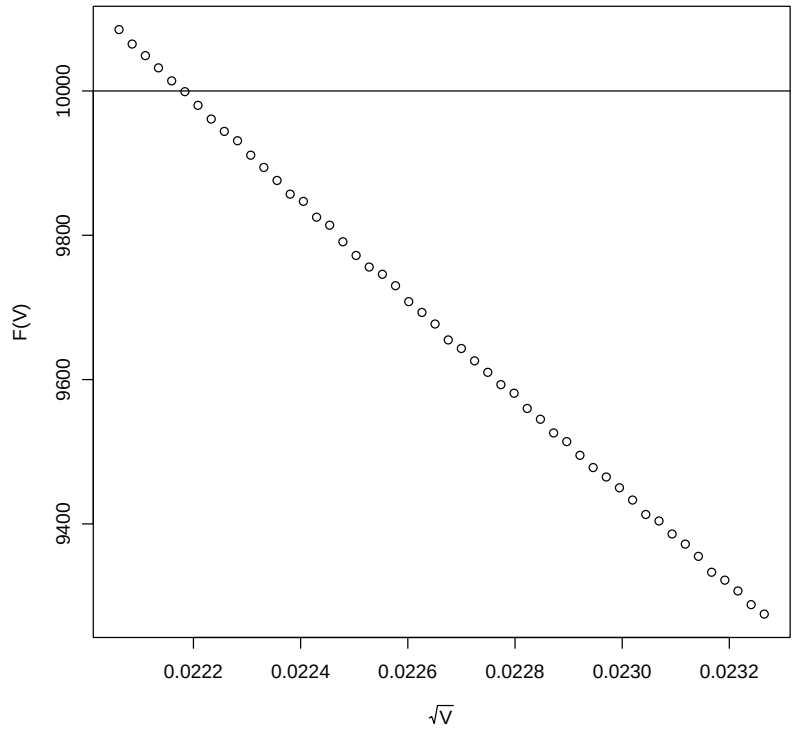


FIGURE 3 Fonction F(V) pour les HES/HEP montre la relation entre la taille de l'échantillon et la précision $\sqrt{V}$. La ligne horizontale correspond à une taille d'échantillon de 10'000.

3.4 Résultats de l'allocation

3.4.1 Détail de l'allocation des HEU

La taille de l'échantillon pour les HEU par strate marginale est donnée par

$$n_0 = \frac{p\,(1-p)}{V} = \frac{1}{4V} = \frac{1}{4 \cdot (0.023885)^2} = 438.22$$

(voir Table 7). Dans la Table 8 le nombre d'étudiant-e-s par HEU dans la population est donné par $N_{i+} = \sum_j N_{ij}$ et par groupe de domaines d'études par $N_{+j} = \sum_i N_{ij}$.

TABLE 8 Nombre d'étudiant-e-s par HEU et groupe de domaines d'études

HEU	FGB1	FBG2	FBG3	FBG4	FBG5	FBG6	FBG7	N_{i+}
Basel	3279	947	1120	1327	1518	.	961	9152
Bern	4116	1508	1748	1466	1646	.	670	11154
Fribourg	4525	1010	1422	843	298	.	74	8172
Genève	5281	1135	1209	853	1346	.	436	10260
Lausanne	3847	1705	1307	902	1363	.	367	9491
Luzern	929	1	1076	.	.	.	8	2014
Neuchâtel	1673	346	535	447	74	22	16	3113
St. Gallen	407	3995	521	.	.	.	.	4923
Zürich	10589	2608	2975	2066	2081	.	19	20338
Svizzera italiana	647	732	.	129	.	655	.	2163
Fernstudien Schweiz	127	114	147	.	.	.	.	388
Institut universitaire Kurt Bösch	.	.	.	.	.	.	18	18
EPFL	.	.	.	1746	.	3012	.	4758
ETHZ	29	.	.	3987	375	5420	587	10398
N_{+j}	35449	14101	12060	13766	8701	9109	3156	96342

Le n_{i+} (taille de l'échantillon par HEU) et le n_{+j} (taille de l'échantillon par groupe de domaines d'études) sont déterminés à partir du n_0 avec les formules (voir annexe A.2.3)

$$n_{i+} = \frac{n_0}{1 + n_0/N_{i+}} \text{ et } n_{+j} = \frac{n_0}{1 + n_0/N_{+j}}.$$

On trouve les n_{i+} dans la Table 9 et les n_{+j} dans la Table 10.

TABLE 9 Echantillon avec allocation proportionnelle par rapport au total par HEU

HEU	FGB1	FBG2	FBG3	FBG4	FBG5	FBG6	FBG7	n_{i+}
Basel	149.8	43.3	51.2	60.6	69.4	.	43.9	418
Bern	155.6	57.0	66.1	55.4	62.2	.	25.3	422
Fribourg	230.3	51.4	72.4	42.9	15.2	.	3.8	416
Genève	216.3	46.5	49.5	34.9	55.1	.	17.9	420
Lausanne	169.8	75.2	57.7	39.8	60.2	.	16.2	419
Luzern	166.0	0.2	192.3	.	.	.	1.4	360
Neuchâtel	206.4	42.7	66.0	55.2	9.1	2.7	2.0	384
St. Gallen	33.3	326.5	42.6	.	.	.	.	402
Zürich	223.3	55.0	62.7	43.6	43.9	.	0.4	429
Svizzera italiana	109.0	123.3	.	21.7	.	110.3	.	364
Fernstudien Schweiz	67.4	60.5	78.0	.	.	.	.	206
Institut universitaire Kurt Bösch	.	.	.	.	.	.	17.3	17
EPFL	.	.	.	147.2	.	254.0	.	401
ETHZ	1.2	.	.	161.2	15.2	219.2	23.7	420
Total	1728	882	738	663	330	586	152	5080 [1]

1. Les totaux sont arrondis

TABLE 10 Echantillon avec allocation proportionnelle par rapport au total par groupe de domaines d'études

HEU	FGB1	FBG2	FBG3	FBG4	FBG5	FBG6	FBG7	Total
Basel	40.0	28.5	39.3	40.9	72.8	.	117.2	339
Bern	50.3	45.5	61.3	45.2	78.9	.	81.7	363
Fribourg	55.3	30.4	49.9	26.0	14.3	.	9.0	185
Genève	64.5	34.2	42.4	26.3	64.5	.	53.2	285
Lausanne	47.0	51.4	45.8	27.8	65.4	.	44.7	282
Luzern	11.3	0.0	37.7	.	.	.	1.0	50
Neuchâtel	20.4	10.4	18.8	13.8	3.5	1.0	2.0	70
St. Gallen	5.0	120.4	18.3	.	.	.	.	144
Zürich	129.3	78.6	104.3	63.7	99.8	.	2.3	478
Svizzera italiana	7.9	22.1	.	4.0	.	30.1	.	64
Fernstudien Schweiz	1.6	3.4	5.2	.	.	.	.	10
Institut universitaire Kurt Bösch	.	.	.	.	.	.	2.2	2
EPFL	.	.	.	53.9	.	138.3	.	192
ETHZ	0.4	.	.	123.0	18.0	248.8	71.6	462
n_{+j}	433	425	423	425	417	418	385	2926[1]

A partir des n_{i+} et n_{+j} on calcule la taille $n_{ij} = \max(n_{ij}^{(1)}, n_{ij}^{(2)})$ qui est donnée dans la Table 11.

TABLE 11 Maximum de 9. et 10.

HEU	FGB1	FBG2	FBG3	FBG4	FBG5	FBG6	FBG7	Total
Basel	149.8	43.3	51.2	60.6	72.8	0	117.2	495
Bern	155.6	57.0	66.1	55.4	78.9	0	81.7	495
Fribourg	230.3	51.4	72.4	42.9	15.2	0	9.0	421
Genève	216.3	46.5	49.5	34.9	64.5	0	53.2	465
Lausanne	169.8	75.2	57.7	39.8	65.4	0	44.7	453
Luzern	166.0	0.2	192.3	0	0	0	1.4	360
Neuchâtel	206.4	42.7	66.0	55.2	9.1	2.7	2.0	384
St. Gallen	33.3	326.5	42.6	0	0	0	0	402
Zürich	223.3	78.6	104.3	63.7	99.8	0	2	572
Svizzera italiana	109.0	123.3	0	21.7	0	110.3	0	364
Fernstudien Schweiz	67.4	60.5	78.0	0	0	0	0	206
Institut universitaire Kurt Bösch	0	0	0	0	0	0	17.3	17
EPFL	0	0	0	147.2	0	254.0	0	401
ETHZ	1.2	0	0	161.2	18.0	248.8	71.6	501
Total	1728	905	780	683	424	616	400	5536[1]

On augmente la taille de l'échantillon, éventuellement jusqu'à la taille de la strate pour tenir compte des taux de réponse par HEU observés lors de l'enquête 2005. C'est-à-dire la taille de l'échantillon dans chaque strate est augmentée par le facteur $\frac{1}{\theta_i}$ où θ_i est donné dans la colonne TR de la Table 12. En plus on tient compte de la taille minimale de l'échantillon qui est fixée à 50 étudiant-e-s et la taille minimale par strate est fixée à 100 étudiant-e-s en dessous de cette limite la strate est prise de manière exhaustive.

Remarque 2 *Parmi les HEU, l'EPFL a demandé une augmentation de l'échantillon de 1500 étudiant-e-s. Cette augmentation est allouée proportionnellement au total de l'HEU. La taille de l'échantillon est calculée par*

$$\bar{n}_{ij} = \min\left(n_{ij} + \theta_i c_i \left(\frac{N_{ij}}{N_{i+}}\right), N_{ij}\right)$$

où c_i correspond à l'augmentation de la taille de l'échantillon et θ_i au taux de réponse par HEU. Le détail de l'allocation est résumé dans la Table 13.

TABLE 12 Ajustement de l'allocation en tenant compte du taux de réponse (TR) ainsi que des strates exhaustives et de la taille minimale de l'échantillon par cellule

HEU	FGB1	FBG2	FBG3	FBG4	FBG5	FBG6	FBG7	TR	Total
Basel	245	71	84	99	119	0	192	0.611	810
Bern	279	102	118	99	141	0	146	0.558	885
Fribourg	386	86	121	72	50	0	74	0.596	789
Genève	466	100	107	75	139	0	115	0.464	1002
Lausanne	295	131	100	69	114	0	78	0.575	787
Luzern	263	1	305	0	0	0	8	0.631	577
Neuchâtel	351	72	112	94	74	22	16	0.589	741
St. Gallen	60	589	77	0	0	0	0	0.554	726
Zürich	395	139	185	113	177	0	19	0.565	1028
Svizzera italiana	214	242	0	50	0	216	0	0.510	722
Fernstudien Schweiz	127	114	147	0	0	0	0	0.500	388
Institut universitaire Kurt Bösch	0	0	0	0	0	0	18	0.500	18
EPFL	0	0	0	273	0	470	0	0.540	743
ETHZ	29	0	0	236	50	364	105	0.684	784
Total	3110	1647	1356	1180	864	1072	771		10000

TABLE 13 Augmentation de l'échantillon de 1'500 étudiant-e-s pour l'EPFL

HEU	FGB1	FBG2	FBG3	FBG4	FBG5	FBG6	FBG7	Total
EPFL	0	0	0	823	0	1420	0	2243

3.4.2 Détail de l'allocation des HES/HEP

La taille de l'échantillon par strate marginale pour les HES/HEP est donnée par

$$n_0 = \frac{p\,(1-p)}{V} = \frac{1}{4V} = \frac{1}{4 \cdot (0.02216)^2} = 509.10$$

(voir Table 7). Dans la Table 14 le nombre d'étudiant-e-s par HES/HEP dans la population est donné par $N_{i+} = \sum_j N_{ij}$ et par domaine d'études par $N_{+j} = \sum_i N_{ij}$.

TABLE 14 Nombre d'étudiant-e-s par HES/HEP et domaine d'études

HES/HEP	FGB1	FBG2	FBG3	FBG4	FBG5	FBG6	FBG7	FBG8	FBG9	FBG10	FBG11	FBG12	FBG13	N_{i+}
100	393	1065	73	293	811	184	131	612	.	360	.	542	.	4464
300	823	2465	464	44	4078	581	.	1491	.	2477	.	2204	87	14714
400	275	1007	378	.	1802	423	.	717	.	1084	238	.	1570	7494
800	264	722	.	.	1135	308	.	648	.	584	.	.	.	3661
900	232	448	.	.	824	209	.	226	.	237	.	274	.	2450
1000	456	1179	.	.	1275	.	.	.	.	445	.	123	.	3478
1100	466	1460	717	.	3782	554	.	1101	500	598	230	785	1713	11906
1250	.	.	.	.	513	.	.	.	.	.	.	22	.	535
1500	.	.	.	.	.	.	.	.	.	.	.	.	8510	8510
N_{+j}	2909	8346	1632	337	14220	2259	131	4795	500	5785	468	3950	11880	57212

Le n_{i+} (taille de l'échantillon par HES/HEP) et le n_{+j} (taille de l'échantillon par domaine d'études) sont déterminés à partir du n_0 avec les formules (voir annexe A.2.3)

$$n_{i+} = \frac{n_0}{1 + n_0/N_{i+}} \text{ et } n_{+j} = \frac{n_0}{1 + n_0/N_{+j}}.$$

On trouve les n_{i+} dans la Table 15 et les n_{+j} dans la Table 16.

TABLE 15 Echantillon avec allocation proportionnelle par rapport au total par HES/HEP

HES/HEP	FGB1	FBG2	FBG3	FBG4	FBG5	FBG6	FBG7	FBG8	FBG9	FBG10	FBG11	FBG12	FBG13	n_{i+}
100	40.2	109.0	7.5	30.0	83.0	18.8	13.4	62.7	.	36.9	.	55.5	.	457
300	27.5	82.4	15.5	1.5	136.4	19.4	.	49.9	.	82.8	.	73.7	2.9	492
400	17.5	64.1	24.0	.	114.6	26.9	.	45.6	.	69.0	15.1	.	99.9	477
800	32.2	88.1	.	.	138.6	37.6	.	79.1	.	71.3	.	.	.	447
900	39.9	77.1	.	.	141.8	36.0	.	38.9	.	40.8	.	47.1	.	422
1000	58.2	150.5	.	.	162.8	.	.	.	.	56.8	.	15.7	.	444
1100	19.1	59.9	29.4	.	155.1	22.7	.	45.1	20.5	24.5	9.4	32.2	70.2	488
1250	.	.	.	.	250.1	.	.	.	.	.	.	10.7	.	261
1500	.	.	.	.	.	.	.	.	.	.	.	.	480.4	480
Total	235	631	76	31	1182	161	13	321	21	382	25	235	653	3968[3]

TABLE 16 Echantillon avec allocation proportionnelle par rapport au total par domaine d'études

HES/HEP	FGB1	FBG2	FBG3	FBG4	FBG5	FBG6	FBG7	FBG8	FBG9	FBG10	FBG11	FBG12	FBG13	Total
100	58.5	61.2	17.4	176.3	28.0	33.8	104.2	58.7	.	29.1	.	61.9	.	629
300	122.6	141.7	110.3	26.5	141.0	106.9	.	143.1	.	200.4	.	251.6	3.6	1248
400	41.0	57.9	89.9	.	62.3	77.8	.	68.8	.	87.7	124.0	.	64.5	674
800	39.3	41.5	.	.	39.2	56.6	.	62.2	.	47.2	.	.	.	286
900	34.6	25.8	.	.	28.5	38.4	.	21.7	.	19.2	.	31.3	.	199
1000	67.9	67.8	.	.	44.1	.	.	.	.	36.0	.	14.0	.	230
1100	69.4	83.9	170.5	.	130.7	101.9	.	105.7	252.3	48.4	119.8	89.6	70.4	1243
1250	.	.	.	.	17.7	.	.	.	.	.	.	2.5	.	20
1500	.	.	.	.	.	.	.	.	.	.	.	.	349.7	350
n_{+j}	433	480	388	203	492	415	104	460	252	468	244	451	488	4879[3]

A partir des n_{i+} et n_{+j} on calcule la taille $n_{ij} = \max(n_{ij}^{(1)}, n_{ij}^{(2)})$ qui est donnée dans la Table 17.

TABLE 17 Maximum de 15. et 16.

HES/HEP	FGB1	FBG2	FBG3	FBG4	FBG5	FBG6	FBG7	FBG8	FBG9	FBG10	FBG11	FBG12	FBG13	Total
100	59	109	17	176	83	34	104	63	0	37	0	62	0	744
300	123	142	110	26	141	107	0	143	0	200	0	252	4	1248
400	41	64	90	0	115	78	0	69	0	88	124	0	100	768
800	39	88	0	0	139	57	0	79	0	71	0	0	0	473
900	40	77	0	0	142	38	0	39	0	41	0	47	0	424
1000	68	151	0	0	163	0	0	0	0	57	0	16	0	454
1100	69	84	170	0	155	102	0	106	252	48	120	90	70	1267
1250	0	0	0	0	250	0	0	0	0	0	0	11	0	261
1500	0	0	0	0	0	0	0	0	0	0	0	0	480	480
Total	439	715	388	203	1187	415	104	498	252	542	244	477	654	6118[3]

On augmente la taille de l'échantillon, éventuellement jusqu'à la taille de la strate pour tenir compte des taux de réponse par HES/HEP observés lors de l'enquête 2005. C'est-à-dire la taille de l'échantillon dans chaque strate est augmentée par le facteur $\frac{1}{\theta_i}$ où θ_i est donné dans la colonne TR de la Table 18. En plus on tient compte de la taille minimale de l'échantillon qui est fixée à 50 étudiant-e-s et la taille minimale par strate est fixée à 100 étudiant-e-s en dessous de cette limite la strate est prise de manière exhaustive.

TABLE 18 Ajustement de l'allocation en tenant compte du taux de réponse (TR) ainsi que des strates exhaustives et de la taille minimale de l'échantillon par cellule

HES/HEP	FGB1	FBG2	FBG3	FBG4	FBG5	FBG6	FBG7	FBG8	FBG9	FBG10	FBG11	FBG12	FBG13	TR	Total
100	94	174	73	282	133	54	131	100	0	59	0	99	0	0.625	1199
300	206	239	186	44	237	180	0	241	0	337	0	424	87	0.594	2181
400	66	103	145	0	185	125	0	111	0	141	200	0	161	0.621	1237
800	61	136	0	0	214	88	0	122	0	110	0	0	0	0.647	731
900	66	128	0	0	235	64	0	64	0	68	0	78	0	0.603	703
1000	100	221	0	0	239	0	0	0	0	83	0	50	0	0.681	693
1100	110	133	270	0	246	161	0	167	400	77	190	142	112	0.631	2008
1250	0	0	0	0	500	0	0	0	0	0	0	22	0	0.500	522
1500	0	0	0	0	0	0	0	0	0	0	0	0	740	0.649	740
Total	703	1134	674	326	1989	672	131	805	400	875	390	815	1100		10014

Remarque 3 *Parmi les HES/HEP, l'Haute Ecole Spécialisée de Suisse occidentale a demandé une augmentation de l'échantillon de 3000 étudiant-e-s. Cette augmentation est allouée proportionnellement au total de l'HES. La taille de l'échantillon est calculée par*

$$\bar{n}_{ij} = \min\left(n_{ij} + \theta_i c_i \left(\frac{N_{ij}}{N_{i+}}\right), N_{ij}\right)$$

où c_i correspond à l'augmentation de la taille de l'échantillon et θ_i au taux de réponse par HES/HEP. Le détail de l'allocation est résumé dans la Table 19

TABLE 19 Augmentation de l'échantillon de 3'000 étudiant-e-s pour l'HES-SO

HES/HEP	FGB1	FBG2	FBG3	FBG4	FBG5	FBG6	FBG7	FBG8	FBG9	FBG10	FBG11	FBG12	FBG13	Total
HES-SO	375	746	281	44	1076	300	0	548	0	847	0	877	87	5181

4 Tirage des échantillons

Le tirage des échantillons stratifiés a été effectué avec la procédure `proc surveyselect` du logiciel SAS. Un échantillon aléatoire simple pour les HEU, respectivement pour les HES/HEP, a été tiré dans chaque strate avec une taille de l'échantillon donnée par le plan d'échantillonnage décrit dans la Section 3.

L'Observatoire de la vie étudiante (OVE) de l'université de Genève (UniGE) représente un cas particulier, puisqu'il mène une étude longitudinale auprès de ses étudiant-e-s depuis 2006. En 2009, elle a eu lieu presque en même temps que celle de l'OFS. Les deux questionnaires sont proches et les deux sont longs. L'université de Genève a contacté l'OFS pour proposer une collaboration entre l'OFS et l'OVE pour éviter que des étudiant-e-s soient sélectionnés pour répondre aux deux enquêtes.

Pour minimiser l'intersection entre les deux échantillons OFS et OVE, on a eu recours à la méthode de Kish & Scott [7] qui permet de coordonner deux échantillons stratifiés. Cette méthode garantit les propriétés d'un échantillon aléatoire stratifié pour l'échantillon résultant.

L'idée est la suivante : L'échantillon S_1 qui correspond à l'échantillon de l'OVE est donné. Pour l'échantillon de l'OFS on effectue dans un premier temps un tirage S_2 indépendant de S_1.
Il peut donc y avoir des étudiant-e-s qui sont à la fois dans les deux échantillons. Pour faire de sorte qu'un minimum d'étudiant-e-s soient tirés deux fois on utilise la méthode de Kish & Scott qui permet de minimiser le nombre d'étudiant-e-s dans l'intersection des deux échantillons. Premièrement on garde les étudiant-e-s qui sont uniquement dans l'échantillon $S_2 \setminus S_1$. Ensuite pour pouvoir minimiser le nombre d'étudiant-e-s qui se trouvent dans l'intersection des deux échantillons, on tire un échantillon S_3 dans le reste de la population U, c'est-à-dire parmi ceux qui ne sont ni dans S_1 ni dans S_2. Ceci peut être fait seulement si le nombre d'étudiant-e-s qu'on doit remplacer est plus petit que le reste des étudiants dans la population U (sans S_1 et S_2). Dans le cas contraire on prend toutes les étudiant-e-s du reste de la population U, c'est-à-dire parmi ceux qui n'ont pas encore été sélectionnés et on tire les étudiant-e-s qui manquent encore dans l'intersection des deux échantillons S_1 et S_2.

Cette procédure est aussi visualisée dans les Figures 4 et 5.

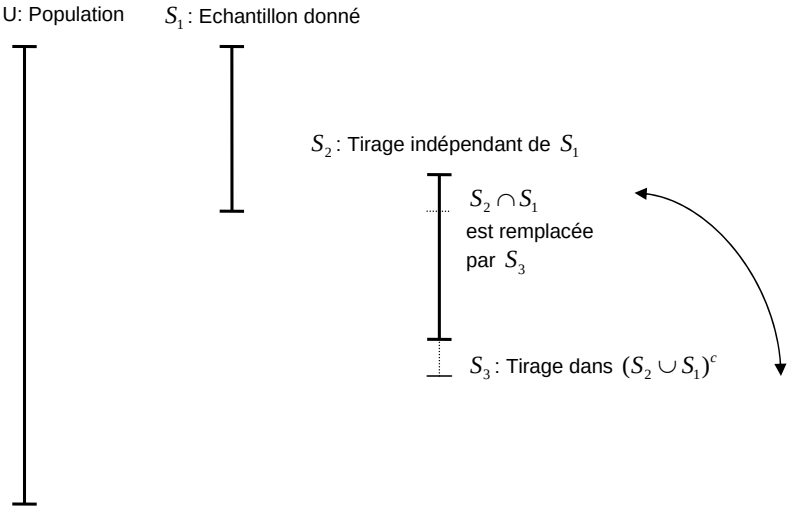


FIGURE 4 Remplacement de l'intersection $S_1 \cap S_2$ par un échantillon aléatoire S_3 tiré dans $(S_1 \cup S_2)^c$ dans le cas où $|S_1 \cap S_2| \leq |S_1 \cup S_2|^c$

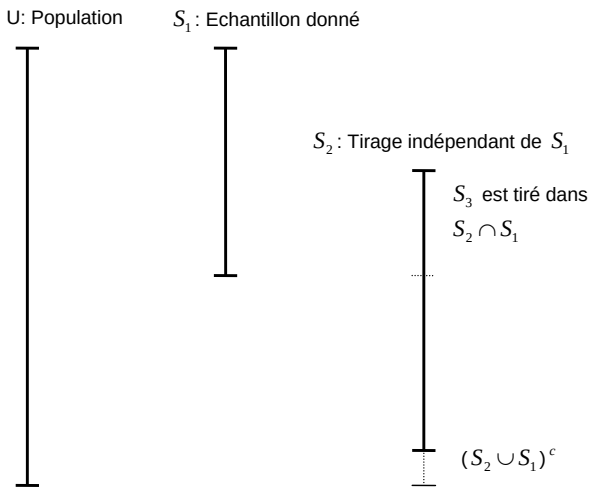


FIGURE 5 Remplacement de l'intersection $S_1 \cap S_2$ par le complément $S_1 \cap S_2$ et un échantillon aléatoire S_3 tiré dans l'intersection $S_1 \cap S_2$ dans le cas où $|S_1 \cap S_2| > |S_1 \cup S_2|^c$

En résumant, on considère une population $U = \{1, \ldots, k, \ldots, N\}$ avec les stratifications $U = \bigcup_{h \in H} U_h$ et $U = \bigcup_{g \in G} U_g$ dans laquelle nous tirons deux échantillons aléatoires stratifiés S_1 et S_2 indépendants. Dans le cas d'une coordination négative l'algorithme de Kish & Scott consiste à modifier l'échantillon S_2 afin d'obtenir un recouvrement minimal avec l'échantillon S_1. Pour ce faire on effectue dans chaque croisement de strates $U_{hg} = U_h \cap U_g$ les opérations suivantes :

1. Si $k \in S_2 \setminus S_1$ on garde l'unité k.
2. Si $|S_1 \cap S_2| \leq |(S_1 \cup S_2)^c|$, on tire S_3 de taille $|S_1 \cap S_2|$ dans $(S_1 \cup S_2)^c$. L'échantillon final dans le croisement de strates $U_{hg} = U_h \cap U_g$ est alors $\tilde{S}_2 = (S_2 \setminus S_1) \cup S_3$.
3. Si $|S_1 \cap S_2| > |(S_1 \cup S_2)^c|$, on tire S_3 de taille $|S_1 \cap S_2| - |(S_1 \cup S_2)^c|$ dans $S_1 \cap S_2$. L'échantillon final dans le croisement de strates $U_{hg} = U_h \cap U_g$ est alors $\tilde{S}_2 = (S_2 \setminus S_1) \cup (S_1 \cup S_2)^c \cup S_3$.

Tirage échantillon de l'OVE

L'étude longitudinale de l'OVE se base sur un panel initial qui a commencé en 2006. Les étudiant-e-s inscrits pour le semestre d'automne 2005 à l'UNIGE ont été pris en compte. Les années suivantes (2007, 2008 et 2009) il y a eu chaque fois un renouvellement. Les strates de l'OVE sont alors données par les années d'entrée 2006, 2007, 2008 et 2009. L'échantillon S_1 correspond à l'échantillon de l'OVE avec les étudiants des années 2006, 2007 et 2008. La population et l'échantillon brut dans la base de sondage de l'OFS sont représentés dans la Table 20. L'échantillon brut 2009 est vide, puisque l'échantillon OVE n'a pas encore été tiré.

TABLE 20 Population et échantillon brut dans les strates OVE[2] et dans la base de sondage OFS

Année	Population	Echantillon brut
2006	3499	841
2007	1647	550
2008	2111	792
2009	3003	0
Total	10260	2183

Tirage échantillon de l'OFS

Pour l'échantillon de l'OFS on a procédé à un tirage S_2 indépendamment du tirage de l'OVE. Les groupes de domaines d'études définissent les strates de l'OFS, plus de détails se trouvent dans la Section 3.4. Le nombre d'étudiant-e-s par groupe de domaines d'études provient du système universitaire suisse et est résumé dans la Table 21. Un calcul de l'intersection des deux échantillons S_1 et S_2 a montré que l'intersection n'est pas vide, mais correspond à 217 étudiant-e-s. Pour minimiser le nombre d'étudiant-e-s qui sont présent dans les deux échantillons on a par la suite appliqué la méthode de Kish & Scott.

TABLE 21 Groupes de domaines d'études, population d'intérêt, échantillon brut OFS

		Population	Echantillon brut
FBG1	Sciences humaines et sociales	5281	466
FBG2	Sciences économiques	1135	100
FBG3	Droit	1209	107
FBG4	Sciences exactes et naturelles	853	75
FBG5	Médecine et pharmacie	1346	139
FBG7	Interdisciplinaire et autre	436	115
Total		10260	1002

Application Kish & Scott

Avant d'appliquer Kish & Scott $S_1 \cap S_2$ contenait 217 étudiant-e-s. Pour remplacer les éléments qui se trouvent dans l'intersection des deux échantillons on procède à un tirage d'un échantillon S_3 dans $(S_1 \cup S_2)^c$ et dans les nouvelles strates qui correspondent au croisement des strates de l'échantillon S_1 et ceux de l'échantillon S_2. Les nouvelles strates (voir Table 22) sont donc obtenues après croisement des variables $fachl1$ (strates OFS) et $year$ (strates OVE). La strate 6, qui correspond aux sciences techniques, est vide. Donc il y a 24 nouvelles strates non vides. Tous les éléments de l'intersection ont été remplacés par l'échantillon S_3. On constate qu'après l'application de Kish & Scott cette intersection devient vide, puisqu'on avait $|S_1 \cap S_2| < |(S_1 \cup S_2)^c|$. Les résultats de l'application de l'algorithme de Kish & Scott sont donnés dans la Table 23.

2. Livraison de la population et de l'échantillon brut par l'OVE

TABLE 22 Nouvelles strates sans le fachl1=6

Strates	$fachl1$	$year$
1	1	2006
2	2	2006
3	3	2006
4	4	2006
5	5	2006
7	7	2006
8	1	2007
9	2	2007
10	3	2007
11	4	2007
12	5	2007
14	7	2007
15	1	2008
16	2	2008
17	3	2008
18	4	2008
19	5	2008
21	7	2008
22	1	2009
23	2	2009
24	3	2009
25	4	2009
26	5	2009
28	7	2009

TABLE 23 Recouvrement des échantillons OFS et OVE avant et après avoir appliqué la méthode Kish & Scott

Strate	Taille strate dans SIUS	éléments que dans S_1 $(S_1 \setminus S_2)$	éléments dans recouvrement $(S_1 \cap S_2)$	élements que dans S_2 $(S_2 \setminus S_1)$	élements dans S_1 ou S_2 $(S_1 \cup S_2)$	éléments nouveau du cadre SIUS $(S_1 \cup S_2)^c$	éléments maintenus de l'intersection
1	1786	410	33	118	561	33	0
2	316	54	5	19	78	5	0
3	418	89	9	34	132	9	0
4	323	66	5	24	95	5	0
5	547	125	14	33	172	14	0
7	109	19	12	22	53	12	0
8	866	267	19	54	340	19	0
9	197	56	5	18	79	5	0
10	204	59	5	10	74	5	0
11	126	46	6	7	59	6	0
12	191	59	6	10	75	6	0
14	63	17	5	8	30	5	0
15	1064	355	45	66	466	45	0
16	253	88	7	16	111	7	0
17	255	84	12	16	112	12	0
18	170	52	6	6	64	6	0
19	240	84	13	17	114	13	0
21	129	36	10	23	69	10	0
22	1565	0	0	131	131	0	0
23	369	0	0	30	30	0	0
24	332	0	0	21	21	0	0
25	234	0	0	21	21	0	0
26	368	0	0	46	46	0	0
28	135	0	0	35	35	0	0
Total	10260	1966	217	785	2968	217	0

L'OVE a par la suite utilisé la même méthode pour tirer leur échantillon.

5 Méthodes d'estimation

5.1 Extrapolation

Une pondération des données récoltées a été effectuée sur la base des probabilités d'inclusion d'un individu dans une strate et d'un facteur de correction pour la non-réponse. De plus, un calage a été effectué sur des caractéristiques connues sur l'ensemble de la population (sexe, classes d'âge, domicile avant le début des études, niveau d'études). La repondération par calage permet de tenir compte de l'information auxiliaire disponible sur un certain nombre de variables, appelées variables de calage. Le redressement consiste à remplacer les pondérations initiales, qui sont en général les poids de sondage des individus (égaux aux inverses des probabilités d'inclusion), par des poids de calage aussi proches que possible des pondérations initiales. Cette pondération permet l'estimation des résultats pour l'ensemble de la population d'étude.

5.1.1 Répondants

L'échantillon net est défini à partir des personnes qui ont répondu à au moins deux des trois questions clés. Les trois questions clés sont résumées dans la Table 24 ci-dessous.

TABLE 24 Variables d'enquête

antfor	*Menez-vous actuellement des études ?* 1 = Oui 2 = Oui, mais j'ai pris un semestre/une année de congé 3 = Non, j'ai interrompu mes études, mais je souhaite les reprendre ultérieurement 4 = Non, j'ai arrêté mes études 5 = Non, j'ai déjà terminé mes études
erwja	*Au cours des douze derniers mois, avez-vous exercé une (ou plusieurs) activité(s) rémunérée(s) ?* 1 = Oui, à la fois durant les semestres et les périodes sans cours (vacances semestrielles) 2 = Oui, durant les semestres uniquement 3 = Oui, durant les périodes sans cours (vacances semestrielles) uniquement 4 = Non 5 = Je n'étais pas encore inscrit-e à la haute école lorsque j'exerçais une activité rémunérée -9 = Missing normal, ne pas répondu à la question -1 = N'ont jamais vu la question à cause de filtre -5 = Réponse n'est pas plausible
logwo	*Où habitez-vous pendant ce semestre d'études ?* 1 = Chez mes parents 2 = Dans un foyer pour étudiant-e-s 3 = Dans une chambre chez des particuliers 4 = Seul-e dans un studio, un appartement, une maison 5 = Dans un appartement, une maison avec mon partenaire/mes enfants 6 = En colocation dans un appartement 7 = Autre -9 = Missing normal, ne pas répondu à la question -1 = N'ont jamais vu la question à cause de filtre -5 = Réponse n'est pas plausible

Les tables suivantes indiquent le nombre d'étudiant-e-s des HEU et HES/HEP dans la population après traitement des doublons (voir annexe A.1) et dans l'échantillon brut et net. Pour les HEU le nombre d'étudiant-e-s dans la population est résumé dans la Table 25 et l'échantillon brut et net correspondants dans les Tables 26 et 27.

TABLE 25 Nombre d'étudiant-e-s par HEU et groupe de domaines d'études sans doublons

HEU	FGB1	FBG2	FBG3	FBG4	FBG5	FBG6	FBG7	Total
Basel	3278	947	1120	1327	1518	.	960	9150
Bern	4116	1507	1747	1466	1646	.	670	11152
Fribourg	4522	1010	1422	842	298	.	74	8168
Genève	5281	1135	1208	851	1346	.	436	10257
Lausanne	3850	1705	1307	902	1363	.	367	9494
Luzern	929	1	1076	.	.	.	8	2014
Neuchâtel	1670	346	535	446	74	22	17	3110
St. Gallen	406	3993	520	.	.	.	.	4919
Zürich	10593	2607	2975	2066	2081	.	19	20341
Svizzera italiana	646	731	.	129	.	655	.	2161
Fernstudien Schweiz	127	114	146	.	.	.	.	387
Institut universitaire Kurt Bösch	.	.	.	.	.	.	18	18
EPFL	.	.	.	1743	.	3007	.	4750
ETHZ	29	.	.	3979	375	5416	587	10386
Total	35447	14096	12056	13751	8701	9100	3156	96307

TABLE 26 Echantillon brut selon haute école et groupe de domaines d'études pour les HEU

HEU	FGB1	FBG2	FBG3	FBG4	FBG5	FBG6	FBG7	Total
Basel	245	71	84	99	119	0	192	810
Bern	279	102	118	99	141	0	146	885
Fribourg	386	86	121	72	50	0	74	789
Genève	466	100	107	75	139	0	115	1002
Lausanne	295	131	100	69	114	0	78	787
Luzern	263	1	305	0	0	0	8	577
Neuchâtel	351	72	112	94	74	22	16	741
St. Gallen	60	589	77	0	0	0	0	726
Zürich	395	139	185	113	177	0	19	1028
Svizzera italiana	214	242	0	50	0	216	0	722
Fernstudien Schweiz	127	114	147	0	0	0	0	388
Institut universitaire Kurt Bösch	0	0	0	0	0	0	18	18
EPFL	0	0	0	823	0	1420	0	2243
ETHZ	29	0	0	236	50	364	105	784
Total	3110	1647	1356	1730	864	2022	771	11500

TABLE 27 Echantillon net selon haute école et groupe de domaines d'études pour les HEU

HEU	FGB1	FBG2	FBG3	FBG4	FBG5	FBG6	FBG7	Total
Basel	153	41	48	64	77	.	120	503
Bern	172	68	71	71	98	.	92	572
Fribourg	248	61	69	52	37	.	50	517
Genève	271	53	66	43	70	.	65	568
Lausanne	193	94	65	44	67	.	49	512
Luzern	172	0	218	.	.	.	6	396
Neuchâtel	200	46	67	60	48	10	8	439
St. Gallen	35	317	37	.	.	.	.	389
Zürich	243	71	114	78	117	.	12	635
Svizzera italiana	118	147	.	23	.	79	.	367
Fernstudien Schweiz	80	64	84	.	.	.	.	228
Institut universitaire Kurt Bösch	.	.	.	.	.	.	14	14
EPFL	.	.	.	530	.	827	.	1357
ETHZ	19	.	.	153	35	257	68	532
Total	1904	962	839	1118	549	1173	484	7029

Pour les HES/HEP le nombre d'étudiant-e-s dans la population après traitement des doublons est résumé dans la Table 28 et l'échantillon brut et net correspondants dans les Tables 29 et 30.

TABLE 28 Nombre d'étudiant-e-s par HES/HEP et domaine d'études sans doublons

HES/HEP	FGB1	FBG2	FBG3	FBG4	FBG5	FBG6	FBG7	FBG8	FBG9	FBG10	FBG11	FBG12	FBG13	Total
100	393	1065	73	293	809	183	131	610	.	360	.	542	.	4459
300	823	2463	464	44	4068	580	.	1486	.	2476	.	2204	87	14695
400	274	1006	378	.	1801	423	.	715	.	1083	238	.	1570	7488
800	264	722	.	.	1135	308	.	648	.	584	.	.	.	3661
900	232	448	.	.	822	209	.	226	.	237	.	274	.	2448
1000	456	1179	.	.	1275	.	.	.	.	445	.	123	.	3478
1100	466	1459	717	.	3778	554	.	1101	500	598	230	785	1710	11898
1250	.	.	.	.	510	.	.	.	.	.	.	22	.	532
1500	.	.	.	.	.	.	.	.	.	.	.	.	8501	8501
Total	2908	8342	1632	337	14198	2257	131	4786	500	5783	468	3950	11868	57160

TABLE 29 Echantillon brut selon haute école et domaine d'études pour les HES/HEP

HES/HEP	FGB1	FBG2	FBG3	FBG4	FBG5	FBG6	FBG7	FBG8	FBG9	FBG10	FBG11	FBG12	FBG13	Total
100	94	174	73	282	133	54	131	100	0	59	0	99	0	1199
300	375	746	281	44	1076	300	0	548	0	847	0	877	87	5181
400	66	103	145	0	185	125	0	111	0	141	200	0	161	1237
800	61	136	0	0	214	88	0	122	0	110	0	0	0	731
900	66	128	0	0	235	64	0	64	0	68	0	78	0	703
1000	100	221	0	0	239	0	0	0	0	83	0	50	0	693
1100	110	133	270	0	246	161	0	167	400	77	190	142	112	2008
1250	0	0	0	0	500	0	0	0	0	0	0	22	0	522
1500	0	0	0	0	0	0	0	0	0	0	0	0	740	740
Total	872	1641	769	326	2828	792	131	1112	400	1385	390	1268	1100	13014

TABLE 30 Echantillon net selon haute école et domaine d'études pour les HES/HEP

HES/HEP	FGB1	FBG2	FBG3	FBG4	FBG5	FBG6	FBG7	FBG8	FBG9	FBG10	FBG11	FBG12	FBG13	Total
100	64	107	45	173	90	33	64	53	.	39	.	69	.	737
300	205	438	155	20	569	145	.	229	.	533	.	505	44	2843
400	42	66	97	.	117	74	.	46	.	103	134	.	111	790
800	48	83	.	.	148	59	.	76	.	78	.	.	.	492
900	37	72	.	.	156	42	.	21	.	46	.	57	.	431
1000	70	147	.	.	163	.	.	.	.	60	.	40	.	480
1100	68	86	169	.	167	92	.	93	308	57	146	98	66	1350
1250	.	.	.	.	302	.	.	.	.	.	.	16	.	318
1500	.	.	.	.	.	.	.	.	.	.	.	.	506	506
Total	534	999	466	193	1712	445	64	518	308	916	280	785	727	7947

5.1.2 Pondération initiale

Le poids d'extrapolation d_k d'un individu k dans la strate $h = (i, j)$ (qu'on notera par la suite h) tient compte de la probabilité d'inclusion (π_k) de l'individu k et du taux de réponse (q_h) dans la strate et est calculé, pour $k \in U_h$,

$$d_k = \pi_k^{-1} \times q_h^{-1} = \frac{N_h}{n_h} \times \frac{n_h}{r_h} = \frac{N_h}{r_h}$$

où r_h correspond aux répondants dans la strate h.

5.1.3 Modèle logistique

Les variables auxiliaires à utiliser pour le calage ont été déterminées par un modèle logistique. Ces variables auxiliaires sont définies dans la Table 31. La Table 32 des HEU indique que les cinq effets `uni, fachl1, sexe, wo` et `clage` entrent en considération dans le modèle (p-value < 0.05). La variable `stufe` n'est pas significative, mais puisqu'elle est importante pour les analyses, on l'utilise aussi pour le calage. Ainsi ces six variables sont par la suite utilisées pour faire le calage des HEU. La Table 33 des HES/HEP indique que les six effets `fhs, gan1, sexe, wo, clage, stufe` entrent en considération dans le modèle (p-value < 0.05). Donc pour le calage des HES/HEP ces six variables ont été utilisées.

TABLE 31 Variables auxiliaires

Variable	Définition
uni	HEU
fachl1	Groupe de domaines d'études des HEU
fhs	HES/HEP
gan1	Domaine d'études des HES/HEP
sexe	Sexe
	sexe = 1 : homme
	sexe = 2 : femme
wo	Domicile avant le début des études
	wo = 1 : en Suisse
	wo = 2 : à l'étranger
annais	Date de naissance (yyyymm)
age	Age
	age = 2009 - annais
clage	Classes d'âge
	clage = 25 si age $\leq$ 25
	clage = 30 si 26 $\leq$ age $\leq$ 30
	clage = 99 si 31 $\leq$ age
stufe	stufe=10 : licence, diplôme et autres
	stufe=15 : Bachelor
	stufe=25 : Master

TABLE 32 HEU

Effect	DF	Wald Chi-Square	Pr > ChiSq
uni	13	69.1846	<.0001
fachl1	6	16.6096	0.0108
sexe	1	13.2908	0.0003
wo	1	81.5947	<.0001
clage	2	40.6249	<.0001
stufe	2	1.1725	0.5564

TABLE 33 HES

Effect	DF	Wald Chi-Square	Pr > ChiSq
fhs	8	67.4837	<.0001
gan1	12	102.1473	<.0001
sexe	1	32.5221	<.0001
wo	1	97.8061	<.0001
clage	2	9.4230	0.0090
stufe	2	58.4235	<.0001

5.1.4 Calage et estimation d'un total

Ce paragraphe s'appuie sur la référence [6].

Pour déterminer les poids finaux on a utilisé la macro CALMAR (`calmar2.sas`) de l'INSEE avec la méthode linéaire (voir [5]). Avec le calage l'information des variables auxiliaires est utilisée dans les estimations afin de pouvoir faire des ajustements de non-réponse et de respecter les totaux de référence. Le fichier input correspond à l'échantillon net. On indique avec `poids` le poids initial. Les marges sont alors définies dans la variable `marmen`. Avec $m = 1$ (méthode linéaire) on indique la méthode qui a été utilisée. Dans notre cas la méthode de la distance quadratique a été utilisée, ce qui correspond à la fonction G donnée par la formule 5 (voir plus bas). L'identifiant des étudiant-e-s qui correspond au numéro matricule est indiqué avec la variable `ident`. Après l'application de la macro de `calmar` on obtient un fichier output qui contient les poids après calage.

Soit $S \subset U$ l'échantillon dans la population $U = \bigcup_{h \in H} U_h$ finie contenant H strates et soit d_k le poids de sondage. On connaît y_k et x_k pour les éléments $k \in S$. On note x_k un vecteur de p variables auxiliaires. De plus les totaux $x_U = \sum_{k \in U} x_k$ sont connus sur la population. On veut estimer le total

$$Y = \sum_{k \in U} y_k.$$

L'objectif est de changer les poids de sondage initiaux d_k qui apparaissent dans l'estimateur de Horvitz-Thompson, qui est donné par

$$\widehat{Y}_{HT} = \sum_{k \in S} d_k y_k.$$

On cherche alors des poids w_k qui permettent de satisfaire les équations de calage

$$\sum_{k \in S} w_k x_k = x_U$$

qui soient proches des poids initiaux. On est alors amené à un problème d'optimisation :

$$\min_{w_k, k \in S} \sum_{k \in S} G(w_k, d_k) \tag{4}$$

sous les contraintes

$$\sum_{k \in S} w_k x_k = x_U.$$

Dans le cas d'un estimateur par régression, qui est utilisé dans cette enquête, la fonction de distance est définie par

$$G(w, d) = \frac{1}{2} \frac{(w-d)^2}{d} = d \ \frac{1}{2} \Big(\frac{w}{d} - 1 \Big)^2 = d\Psi(\frac{w}{d}), \tag{5}$$

où $\Psi(x) = \frac{1}{2}(x-1)^2$.

Le problème d'optimisation (4) devient alors :

$$\min_{w_k, k \in S} \sum_{k \in S} d_k \Psi \Big(\frac{w_k}{d_k} \Big)$$

sous les contraintes

$$\sum_{k \in S} w_k x_k = x_U$$

Les équations de Lagrange sont données par :

$$\psi \Big(\frac{w_k}{d_k} \Big) - x_k^t \lambda = 0,$$

où $\lambda = (\lambda_1, ..., \lambda_p)^t$ et $\psi(x) = \Psi'(x) = x - 1$.

Donc

$$\psi \Big(\frac{w_k}{d_k} \Big) = x_k^t \lambda \Longrightarrow \frac{w_k}{d_k} = \psi^{-1}(x_k^t \lambda) = F(x_k^t \lambda),$$

où $F(x) = 1 + x$ est l'inverse de la fonction ψ.

On obtient alors les poids de calage :

$$w_k = d_k F(x_k^t \lambda) = d_k g_k$$

où $g_k = F(x_k^t \lambda)$ sont appelés les poids g (g-weights). Comme dans notre cas $F(u) = 1 + u$ on peut aussi écrire :

$$w_k = d_k (1 + x_k^t \lambda). \tag{6}$$

Les équations de calage qui sont utilisées pour calculer le vecteur des multiplicateurs de Lagrange sont données par

$$\sum_{k \in S} w_k x_k = \sum_{k \in S} d_k F(x_k^t \lambda) x_k = x_U.$$

On peut résoudre les équations de calage explicitement. On obtient

$$\sum_{k\in S} d_k x_k (1 + x_k^t \lambda) = x_U \Longleftrightarrow \sum_{k\in S} d_k x_k + \Big\{\sum_{k\in S} d_k x_k x_k^t\Big\}\lambda = x_U.$$

Puisque

$$\sum_{k\in S} d_k x_k x_k^t = X^t D X$$

on peut aussi écrire les équations de calage comme

$$\widehat{x}_{HT} + (X^t D X)\ \lambda = x_U,$$

où X est une matrice $n \times p$ et D est une matrice $n \times n$.

Si la matrice $X^t D X$ est de rang plein alors la solution est donnée par

$$\lambda = (X^t D X)^{-1}(x_U - \widehat{x}_{HT}),$$

ou

$$\lambda = \Big(\sum_{k\in S} d_k x_k x_k^t\Big)^{-1}\Big(\sum_{k\in U} x_k - \sum_{k\in S} d_k x_k\Big).$$

En remplaçant ce λ dans l'équation (6) on obtient les poids de calage qui sont donnés par :

$$w_k = d_k F(\lambda^t x_k) = d_k\Big\{1 + (x_U - \widehat{x}_{HT})^t (X^t D X)^{-1} x_k\Big\}.$$

Les poids de calage peuvent aussi être écrits comme $w_k = d_k g_k$ où les poids g_k dépendant de l'échantillon S sont donnés par

$$g_k = 1 + (x_U - \widehat{x}_{HT})^t (X^t D X)^{-1} x_k.$$

L'estimateur linéaire de calage du total $Y = \sum_{k\in U} y_k$ est

$$\widehat{Y} = \sum_{k\in S} d_k g_k y_k.$$

5.1.5 Estimation de la variance après le calage

L'estimateur de régression généralisé d'un total

$$Y = \sum_{k\in U} y_k$$

est donné par

$$\widehat{Y} = \widehat{Y}_{HT} + (x_U - \widehat{x}_{HT})^t \widehat{\beta}_{HT}$$

où

$$\widehat{Y}_{HT} = \sum_{k\in S} d_k y_k, \quad \widehat{x}_{HT} = \sum_{k\in S} d_k x_k \quad \text{et} \quad \widehat{\beta}_{HT} = \Big(\sum_{k\in S} d_k x_k x_k^t\Big)^{-1}\Big(\sum_{k\in S} d_k x_k y_k\Big).$$

Par un développement en série de Taylor au premier ordre de $\widehat{Y}$, on trouve

$$\widehat{Y} = \widehat{Y}_{HT} + (x_U - \widehat{x}_{HT})^t \widehat{\beta}_{HT} \approx \widehat{Y}_{HT} + (x_U - \widehat{x}_{HT})^t \widehat{\beta},$$

où

$$\widehat{\beta} = \Big(\sum_{k\in U} x_k x_k^t\Big)^{-1}\Big(\sum_{k\in U} x_k y_k\Big)$$

est l'estimateur obtenu par la méthode des moindres carrés. Cette approximation de l'estimateur de régression généralisé peut aussi être écrite sous la forme

$$\widehat{Y} \approx x_U^t\widehat{\beta} + \sum_{k\in S} d_k\Big(y_k - x_k^t\widehat{\beta}\Big) = \sum_{k\in U} x_k^t\widehat{\beta} + \sum_{k\in S} d_k e_k$$

où

$$e_k = y_k - x_k^t\widehat{\beta}$$

sont les résidus sur la population.

Au premier ordre l'estimateur de régression généralisé est donc non biaisé,

$$E(\widehat{Y}) \approx E(\widehat{Y}_{HT}) + E(x_U - \widehat{x}_{HT})^t\widehat{\beta} = Y,$$

et sa variance est donnée approximativement par

$$\begin{aligned}\mathrm{Var}(\widehat{Y}) &\approx \mathrm{Var}\Big(\sum_{k\in S} d_k e_k\Big) = \sum_{k,k'\in U}\Delta_{kk'}d_k d_{k'}e_k e_{k'}\\ &= \sum_{k,k'\in U}\Delta_{kk'}d_k\Big(y_k - x_k^t\widehat{\beta}\Big)d_{k'}(y_{k'} - x_{k'}^t\widehat{\beta}\Big)\end{aligned} \tag{7}$$

où $\Delta_{kk'} = \pi_{kk'} - \pi_k\pi_{k'}$ avec $\pi_{kk'} = P(S \ni k, k')$.

Un estimateur de cette variance peut être calculé avec la formule suivante :

$$\begin{aligned}\widehat{\mathrm{Var}}\Big(\widehat{Y}\Big) &= \sum_{k,k'\in S}\frac{\Delta_{kk'}}{\pi_{kk'}}d_k\Big(y_k - x_k^t\widehat{\beta}_{HT}\Big)d_k\Big(y_{k'} - x_{k'}^t\widehat{\beta}_{HT}\Big)\\ &= \sum_{k,k'\in S}\frac{\Delta_{kk'}}{\pi_{kk'}}d_k d_{k'}\widehat{e}_k\widehat{e}_{k'}\end{aligned} \tag{8}$$

ou bien

$$\begin{aligned}\widehat{\mathrm{Var}}\Big(\widehat{Y}\Big) &= \sum_{k,k'\in S}\frac{\Delta_{kk'}}{\pi_{kk'}}w_k\Big(y_k - x_k^t\widehat{\beta}_{HT}\Big)w_k\Big(y_{k'} - x_k^t\widehat{\beta}_{HT}\Big)\\ &= \sum_{k,k'\in S}\frac{\Delta_{kk'}}{\pi_{kk'}}w_k w_{k'}\widehat{e}_k\widehat{e}_{k'}\end{aligned} \tag{9}$$

Comme $w_k = d_k g_k$ la formule (9) peut aussi être écrite sous la forme suivante :

$$\widehat{\mathrm{Var}}\Big(\widehat{Y}\Big) = \sum_{k,k'\in S}\frac{\Delta_{kk'}}{\pi_{kk'}}d_k d_{k'}(g_k\widehat{e}_k)(g_{k'}\widehat{e}_{k'}).$$

5.2 Exemples d'extrapolation

Les estimations des totaux des variables d'intérêt et de leur variabilité sont effectuées à l'aide du logiciel SAS. En particulier la procédure PROC SURVEYMEANS a été utilisée.

5.2.1 Estimation d'un total

Analyse globale sur un item pour les HEU

La Table 34 montre l'estimation du total d'étudiant-e-s qui oui ou non ont exercés une activité rémunérée au cours des douze derniers mois (`erwja`) cf. Table 24 ainsi que les estimations de la précision obtenues en utilisant le poids après calage `wcal`. La colonne `Sum` de la Table 34 correspond à l'estimation du total de la variable `erwja`. La colonne `Std Dev` correspond à l'écart-type.

Pour faire ces estimations on part du fichier data=`net_uni` qui contient tous les répondants. Avec l'instruction total=`total_uni` on tient compte de la correction de population finie, puisque `total_uni` correspond au nombre d'unités par strate. La définition des strates est indiquée avec la commande `strata` (les variables `uni` et `fachL1` sont définies dans la Table 31). Le poids après calage `wcal` est utilisé. Avec les deux codes suivants de la procédure PROC SURVEYMEANS de SAS on obtient les mêmes résultats pour l'estimation des totaux et l'estimation des écart-types.

– Dans ce premier code on a crée une variable réponse (`resp=1` pour toutes les unités dans le fichier `net_uni`) et on utilise en même temps l'instruction `domain` qui permet que la variable `erwja` soit reconnue en tant que variable catégorielle.

```
proc surveymeans data=net_uni total=total_uni sum missing;
  domain erwja;
  strata uni fachL1;
  var resp;
  weight wcal;
run;
```

– Dans ce deuxième code on utilise les instructions `class` et `var` ce qui permet que la variable `erwja` soit reconnue en tant que variable catégorielle.

```
proc surveymeans data=net_uni total=total_uni sum missing;
  class erwja;
  strata uni fachL1;
  var erwja;
  weight wcal;
run;
```

TABLE 34 HEU : Estimation du total des activités rémunérées exercées au cours des douze derniers mois (erwja) ainsi que de la précision correspondante

erwja	Sum	Std Dev
-9	48	26.1
-5	503	98.6
-1	6472	376.7
1	53984	667.3
2	4177	268.6
3	9082	367.1
4	18487	520.3
5	3556	247.2

Analyse globale d'un item par groupe de domaines d'études pour les HEU

La Table 35 montre les résultats de l'estimation du total des étudiant-e-s par groupe de domaines d'études et par activité rémunérée pour les HEU. Avec les trois codes suivants de la procédure PROC SURVEYMEANS de SAS on obtient les mêmes résultats.

– Ce premier code est en effet le même que celui de l'exemple précédent sauf que l'on rajoute `by fachL1` pour obtenir les estimations de la variable `erwja` par `fachL1`.

```
proc surveymeans data=net_uni total=total_uni sum missing;
  domain erwja;
  by fachl1;
  strata uni fachL1;
  var resp;
  weight wcal;
run;
```

– Ce code est le même comme le deuxième code de l'exemple précédent sauf que dans ce cas on rajoute aussi `by fachL1` pour obtenir les estimations de la variable `erwja` par `fachL1`.

```
proc surveymeans data=net_uni total=total_uni sum missing;
  class erwja ;
  by fachl1;
  strata uni fachL1;
  var erwja ;
  weight wcal;
run;
```

– Dans ce troisième code on utilise les instructions `class` et `var` qui permettent que la variable `erwja` soit reconnue en tant que variable catégorielle. L'instruction `domain` est alors utilisée pour faire les estimations de la variable `erwja` par `fachL1`.

```
proc surveymeans data=net_uni_erwja_dom total=total_uni  sum  missing;
  domain fachL1;
  class erwja;
  strata uni fachL1;
  var erwja;
  weight wcal;
run;
```

Remarque 4 *On peut aussi procéder de la manière suivante : On crée une nouvelle variable* `erwja_dom` *qui définit un domaine pour les modalités -1, -9 et -5 de la variable* `erwja` *et un autre domaine pour le reste des modalités. On utilise les instructions* `class` *et* `var` *qui permettent que la variable* `erwja` *soit reconnue en tant que variable catégorielle. L'instruction* `domain` *est alors utilisée pour faire les estimations de la variable* `erwja` *par* `fachL1`*. Le code est alors le suivant :*

```
data net_uni_erwja_dom; set net_uni;
  if erwja eq -1 or erwja eq -9 or erwja eq -5 then erwja_dom=0;
  else erwja_dom=1;
run;
proc surveymeans data=net_uni_erwja_dom total=total_uni  sum  missing;
  domain erwja_dom*fachL1;
  class erwja;
  strata uni fachL1;
  var erwja;
  weight wcal;
run;
```

TABLE 35 HEU : Estimation du total des étudiant-e-s par groupe de domaines d'études (fachl1) et par activité rémunérée (erwja)

	erwja								All
fachL1	-9	-5	-1	1	2	3	4	5	
1	18	194	2211	24323	1232	2169	4350	951	35447
2	.	92	856	6559	659	1777	3350	803	14096
3	19	63	1323	7096	359	796	2080	320	12056
4	.	39	1056	6113	813	1408	3679	642	13751
5	.	38	236	4546	398	1172	2045	266	8701
6	11	58	587	3217	6622	1519	2609	478	9100
7	.	18	202	2131	94	241	374	96	3156
Total	48	503	6472	53984	4177	9082	18487	3556	96307

Pour les HES/HEP on a procédé de la même manière que pour les HEU.

5.2.2 Comparaison de différents estimateurs de variance

Pour faire les estimations on prend toujours le poids de calage et on prend les variances sans calculer les résidus. Pour pouvoir comparer l'estimation de variance sans les résidus et avec les résidus, on a aussi estimé la variance avec les formules 8 et 9. Le but de cette comparaison est de valider l'approche choisie pour cette enquête, c'est-à-dire de montrer que l'on peut effectivement employer l'estimateur naïf pour faire le calcul de variance. Cet estimateur naïf est facile à utiliser, mais pas tout à fait juste, car dans le calcul de g_k tout l'échantillon intervient et non pas seulement l'unité k. Donc c'est d'autant plus juste si x_{HT} est proche de x_U. Pour des raisons pratiques on a renoncé de calculer les résidus pour chaque estimateur et on a alors calculé les précisions avec les codes donnés au chapitre 5.2.1 sans calculer les résidus.

On veut estimer un total $Y = \sum_{k\in U} y_k$. On considère l'estimateur de Horvitz-Thompson, donné par

$$\widehat{Y}_{HT} = \sum_{k\in S} d_k y_k,$$

et l'estimateur par calage donné par

$$\widehat{Y} = \sum_{k\in S} w_k y_k.$$

Dans le cas d'un sondage stratifié, la variance (7) prend la forme

$$\mathrm{Var}(\widehat{Y}) \approx \sum_{h\in H} N_h^2 \Big(1 - \frac{n_h}{N_h}\Big) \frac{1}{n_h} S^2_{\widehat{e},U_h}$$

où $S^2_{\widehat{e},U_h} = \frac{1}{N_h - 1} \sum_{k\in U_h} \Big(\widehat{e}_k - \frac{1}{N_h} \sum_{k\in U_h} \widehat{e}_k\Big)^2$ et $\widehat{e}_k = y_k - x_k^t \widehat{\beta}_{HT}$.

Dans le cas d'un sondage stratifié (strate h et poids a quelconque) la procédure `proc surveymeans` calcule un estimateur de la variance donné par

$$\widehat{\mathrm{Var}}(\widehat{Y}) = \sum_{h\in H} n_h (1 - f_h) \frac{1}{n_h - 1} \sum_{k\in S_h} \big(a_k y_k - \frac{1}{n_h} \sum_{k\in S_h} a_k y_k\big)^2. \tag{10}$$

où $f_h = n_h/N_h$ est le taux de sondage par strate.

Si a_k est le poids de sondage, c'est-à-dire $a_k = d_k$ où $d_k = N_h/n_h$ pour $k \in S_h$, alors (10) est donnée par la forme standard

$$\widehat{\mathrm{Var}}(\widehat{Y}) = \sum_{h\in H} N_h^2 (1 - f_h) \frac{1}{n_h} S^2_{y,S_h}$$

où

$$S^2_{y,S_h} = \frac{1}{n_h - 1} \sum_{k\in S_h} \big(y_k - \frac{1}{n_h} \sum_{k\in S_h} y_k\big)^2.$$

Comme le poids de calage est $w_k = g_k d_k$ les estimateurs de variance (u est défini dans la Table 36) sont donnés par

$$\widehat{\mathrm{Var}}(\widehat{Y}) = \sum_{h\in H} N_h^2 (1 - f_h) \frac{1}{n_h} S^2_{u,S_h}. \tag{11}$$

où

$$S^2_{u,S_h} = \frac{1}{n_h - 1} \sum_{k \in S_h} \left(u_k - \frac{1}{n_h} \sum_{k \in S_h} u_k\right)^2.$$

Ces quatre estimateurs de variance sont résumés dans la Table 36.

TABLE 36 Définition des quatre estimateurs de variance

Estimateur de variance	Proc surveymeans		Formule 11		Remarque
	Poids	Variable	Poids	Variable u_k	
Horvitz-Thompson est. 1	d_k	y_k	d_k	y_k	avec poids d_k dans `surveymeans`
cal_est. 2	w_k	y_k	d_k	$g_k y_k$	avec poids w_k dans `surveymeans`
est. 3	d_k	$\widehat{e}_k$	d_k	$\widehat{e}_k$	cf. formule 8 avec poids d_k dans `surveymeans` tient compte de la variabilité du poids de calage
est. 4	w_k	$\widehat{e}_k$	d_k	$g_k \widehat{e}_k$	cf. formule 9 avec poids w_k dans `surveymeans` tient compte de la variabilité du poids de calage

Remarquons que c'est équivalent de prendre le poids w_k et la variable y_k resp. $\hat{e}_k$ ou le poids d_k et la variable $g_k y_k$ resp. $g_k \hat{e}_k$.

On a pris comme exemple les trois variables clés, c'est-à-dire les variables `antfor`, `erwja` et `logwo` pour les HEU, cf. Table 24. Pour chaque modalité des trois variables clés on a estimé le total avec ces quatre estimateurs. Les modalités 6 resp. 8 de la variable `erwja` resp. `logwo` regroupent les modalités -1, -5 et -9 qui sont définies dans la Table 24.

Les écart-types std1, std2, std3 et std4 du total ainsi que les coefficients de variation sont résumés dans la Table 37. La dernière colonne dans la Table 37 correspond à la variation relative de std2 par rapport à std3. On évalue cette quantité, puisqu'on veut montrer que l'on peut prendre l'estimateur 2 pour l'estimation de variance au lieu de l'estimateur 3 qui est obtenu par linéarisation du calage et qui découle directement de la théorie, car l'estimateur 2 est plus simple à utiliser.

Rappelons que les coefficients de variation de $\widehat{Y}$ sont donnés par le quotient :

$$\widehat{CV}(\widehat{Y}) = \frac{\sqrt{\widehat{\mathrm{Var}}(\widehat{Y})}}{\widehat{Y}}.$$

avec l'estimateur du total $\widehat{Y}$ et l'estimateur de variance $\widehat{\mathrm{Var}}(\widehat{Y})$. Il est souvent exprimé en %.

Globalement les estimateurs de précision sont proches. Il n'y a pas de grande différence entre les quatre estimateurs de précision.

– Comparaison de l'approche naïve versus calage : Si on compare std2 avec std3, on remarque que std2 est pour les cas examinés toujours plus grand que std3, mais les différences absolues entre ces deux écart-types sont petites.
– Comparaison estimateur de Horvitz-Thompson versus calage : Si on compare std1 et std3 cela montre l'effet de calage. On constate que std3 n'est pas toujours plus petit que std1 et que le gain en utilisant std3 n'est pas très élevé. Ceci est souvent le cas si on fait un calage sur marges.
– Comparaison résidus versus g-weighted : On peut aussi comparer std3 et std4. On constate qu'aucun n'est systématiquement plus petit que l'autre.

Dans le graphique 6 on a représenté les variations des coefficients de variation relatives pour les différents estimateurs.

TABLE 37 Estimation des variables `erwja`, `logwo` et `antfor`

VarName	n	HT	cal_est	std1	std2	std3	std4	cv1	cv2	cv3	cv4	(std2-std3)/std3
erwja_1	3668	54044	53984	657.9	667.3	654.9	652.6	1.2%	1.2%	1.2%	1.2%	1.9%
erwja_2	324	4210	4177	272.2	268.6	267.6	271.2	6.5%	6.4%	6.4%	6.5%	0.4%
erwja_3	811	9385	9082	381.5	367.1	363.3	376.8	4.1%	4.0%	4.0%	4.1%	1.0%
erwja_4	1488	18414	18487	515.2	520.3	508.9	505.7	2.8%	2.8%	2.8%	2.7%	2.2%
erwja_5	286	3628	3556	252.3	247.2	245.8	250.8	7.0%	7.0%	6.9%	7.1%	0.6%
erwja_6	452	6625	7022	366.4	388.4	369.1	350.2	5.5%	5.5%	5.3%	5.0%	5.2%
logwo_1	2444	33994	32569	643.2	616.2	584.5	600.1	1.9%	1.9%	1.8%	1.8%	5.4%
logwo_2	376	4125	4282	265.3	274.5	268.9	260.9	6.4%	6.4%	6.3%	6.1%	2.1%
logwo_3	115	1452	1490	161.7	165.3	164.6	161.1	11.1%	11.1%	11.0%	10.8%	0.4%
logwo_4	818	10644	10992	428.7	450.3	441.9	422.9	4.0%	4.1%	4.0%	3.8%	1.9%
logwo_5	849	11724	12581	467.3	505.6	463.5	437.4	4.0%	4.0%	3.7%	3.5%	9.1%
logwo_6	1885	26656	26224	609.8	601.3	595.7	601.9	2.3%	2.3%	2.3%	2.3%	0.9%
logwo_7	15	121	139	35.8	44.2	44.0	35.7	29.6%	31.8%	31.7%	25.7%	0.5%
logwo_8	527	7590	8030	387.2	411.8	391.1	370.0	5.1%	5.1%	4.9%	4.6%	5.3%
antfor_1	6435	87575	87156	414.0	433.2	417.5	399.5	0.5%	0.5%	0.5%	0.5%	3.8%
antfor_2	98	1404	1423	167.7	171.8	171.5	167.4	11.9%	12.1%	12.1%	11.8%	0.2%
antfor_3	87	1259	1257	164.6	164.4	164.2	164.3	13.1%	13.1%	13.1%	13.1%	0.1%
antfor_4	59	757	741	122.2	118.9	118.7	121.8	16.1%	16.0%	16.0%	16.4%	0.2%
antfor_5	350	5311	5731	333.8	360.1	336.8	313.0	6.3%	6.3%	5.9%	5.5%	6.9%

- n = nombre d'observations
- HT = Estimateur Horvitz-Thompson pour le total
- cal_est = Estimateur du total en utilisant les poids calés
- std1 = Ecart-type obtenu avec l'estimateur de Horvitz-Thompson qui correspond à l'estimateur 1 de la Table 36
- std2 = Ecart-type obtenu avec l'estimateur qui utilise les poids calés qui correspond à l'estimateur 2 de la Table 36
- std3 = Ecart-type obtenu avec l'estimateur cf. formule 8 qui correspond à l'estimateur 3 de la Table 36
- std4 = Ecart-type obtenu avec l'estimateur cf. formule 9 qui correspond à l'estimateur 4 de la Table 36
- cv1, cv2, cv3, cv4 = coefficients de variation correspondants
- (std2-std3)/std3 = variation relative de std2 par rapport à std3

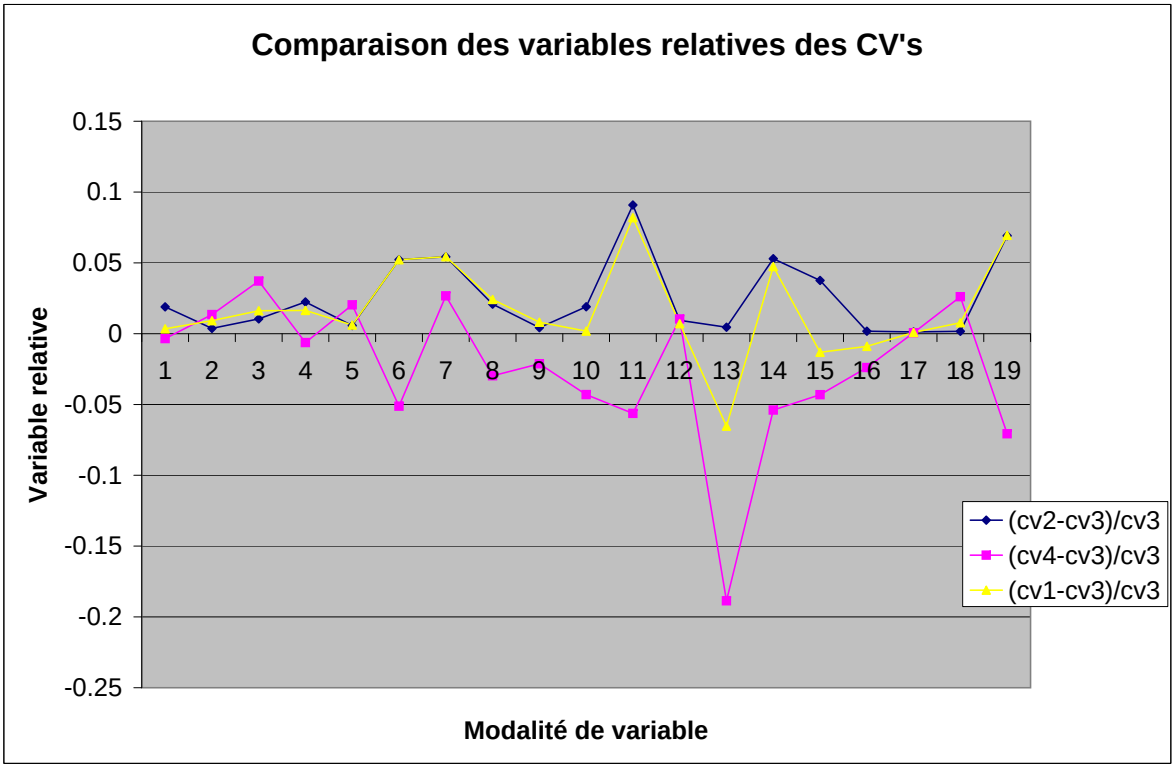


FIGURE 6 Comparaison des coefficients de variation obtenus avec les différents estimateurs

Les résultats ont montrés qu'on peut utiliser l'estimateur 2, qui tient compte des poids calés, car les résultats obtenus sont très semblables aux résultats obtenus avec l'estimateur 3 qui découle directement de la théorie. L'avantage de l'estimateur 2 est son implémentation qui est facile.

6 Conclusion

Le but de ce rapport est de décrire les travaux méthodologiques qui ont conduit aux résultats définitifs de l'enquête sur la situation sociale et économique des étudiant-e-s 2009. Ce document est partagé en 4 parties :

– Résumé sur les cadres de tirage.
– Description des plans d'échantillonnage.
– Tirage des échantillons et résumé de la méthode Kish & Scott.
– Extrapolation, exemples d'extrapolation et comparaison de quatre estimateurs de variance.

Le plan de sondage est un plan stratifié par haute école et par groupe de domaine d'études resp. domaine d'études. L'unité d'échantillonnage est l'étudiant. La taille totale de l'échantillon est fixée de telle façon à obtenir une précision suffisante pour l'estimation. En plus, une HEU et une HES/HEP ont demandé une augmentation de leurs échantillons pour améliorer la précision dans leur cas.

La pondération est basée sur les cadres nettoyés (sans les doublons) et un calage sur les marges a également été effectué. Pour l'estimation de variance, l'estimateur par calage a été employé qui tient compte de la variabilité des poids. Cette approche a été validée avec une comparaison de quatre estimateurs de variance. On a pu constater que les différences entre les estimateurs qu'on a comparés sont petites sur les variables clés de l'enquête. On peut tolérer l'estimateur par calage même si la théorie nous pousserait à utiliser l'estimateur basé sur les résidus.

Des programmes SPLUS et SAS ont été utilisé pour faire ces calculs. Des extraits de ces programmes sont donnés dans l'annexe de ce rapport.

A Annexes

A.1 Préparation des données

La plus grande partie de la préparation des données a été effectuée par la section BILD-S elle-même, selon leurs règles de plausibilisation. Certains recodifications ont eu lieu pendant le travail de la pondération ainsi que le traitement des doublons du cadre de sondage qui ont été identifiés après le tirage de l'échantillon. On distingue les doublons où les deux immatriculations d'une même personne sont dans l'échantillon (type 11), les doublons où une immatriculation est dans l'échantillon et l'autre ne l'est pas (type 10) et finalement les doublons où les deux immatriculations d'une même personne ne sont pas dans l'échantillon (type 00), cf. la Table 38 ci-dessous.

TABLE 38 Doublons

Doublons	Type	Nombre HEU	Nombre HES/HEP
a	11	1	5
b	10	13	19
c	00	31	28
Total		**45**	**52**

Une liste détaillée de ces doublons a été établie, mais ne fait pas partie de ce rapport pour des raisons de protection de données.

Le traitement des doublons dans le cas des étudiant-e-s des HEU a été fait à partir du cadre initial qui est formé de 96'352[3] étudiant-e-s. Le nombre d'étudiant-e-s des HEU resp. des HES/HEP après traitement des doublons est donné dans les Tables 39 resp. 40.

TABLE 39 Nombre d'étudiant-e-s par HEU et groupe de domaines d'études sans doublons

HEU	FGB1	FBG2	FBG3	FBG4	FBG5	FBG6	FBG7	Total
Basel	3278	947	1120	1327	1518	.	960	9150
Bern	4116	1507	1747	1466	1646	.	670	11152
Fribourg	4522	1010	1422	842	298	.	74	8168
Genève	5281	1135	1208	851	1346	.	436	10257
Lausanne	3850	1705	1307	902	1363	.	367	9494
Luzern	929	1	1076	.	.	.	8	2014
Neuchâtel	1670	346	535	446	74	22	17	3110
St. Gallen	406	3993	520	.	.	.	.	4919
Zürich	10593	2607	2975	2066	2081	.	19	20341
Svizzera italiana	646	731	.	129	.	655	.	2161
Fernstudien Schweiz	127	114	146	.	.	.	.	387
Institut universitaire Kurt Bösch	.	.	.	.	.	.	18	18
EPFL	.	.	.	1743	.	3007	.	4750
ETHZ	29	.	.	3979	375	5416	587	10386
Total	35447	14096	12056	13751	8701	9100	3156	96307

3. La différence de 10 étudiant-e-s par rapport à la Table 2 vient des ajustements faits après la définition de l'allocation.

TABLE 40 Nombre d'étudiant-e-s par HES/HEP et domaine d'études sans doublons

HES/HEP	FGB1	FBG2	FBG3	FBG4	FBG5	FBG6	FBG7	FBG8	FBG9	FBG10	FBG11	FBG12	FBG13	Total
100	393	1065	73	293	809	183	131	610	.	360	.	542	.	4459
300	823	2463	464	44	4068	580	.	1486	.	2476	.	2204	87	14695
400	274	1006	378	.	1801	423	.	715	.	1083	238	.	1570	7488
800	264	722	.	.	1135	308	.	648	.	584	.	.	.	3661
900	232	448	.	.	822	209	.	226	.	237	.	274	.	2448
1000	456	1179	.	.	1275	.	.	.	.	445	.	123	.	3478
1100	466	1459	717	.	3778	554	.	1101	500	598	230	785	1710	11898
1250	.	.	.	.	510	.	.	.	.	.	.	22	.	532
1500	.	.	.	.	.	.	.	.	.	.	.	.	8501	8501
Total	2908	8342	1632	337	14198	2257	131	4786	500	5783	468	3950	11868	57160

A.2 Allocation pour une moyenne

Cette annexe s'appuie sur la référence [1].

A.2.1 Estimation d'une moyenne

TABLE 41 Notations

N	nombre d'étudiant-e-s
N_h	nombre d'étudiant-e-s dans la strate h
U_h	univers des étudiant-e-s de la strate h
S_h	échantillon d'étudiant-e-s de la strate h
n_h	taille de S_h
$W_h = \frac{N_h}{N}$	taille relative de U_h
D_h	écart-type de y dans la strate U_h

Soit U la population des étudiant-e-s décomposée en H strates U_h, tel que $U = \bigcup_h U_h$. Soit n_h et N_h les tailles d'échantillon et de population de la strate h, respectivement. La moyenne des étudiant-e-s $\overline{Y}$ au niveau global vaut :

$$\overline{Y} = \frac{Y}{N} = \frac{1}{N} \sum_{k \in U} y_k.$$

Nous avons aussi

$$\overline{Y} = \frac{1}{N} \sum_h \sum_{k \in U_h} y_k = \sum_h \frac{N_h}{N} \frac{1}{N_h} \sum_{k \in U_h} y_k = \sum_h W_h \overline{Y}_h.$$

Donc si on veut estimer la moyenne générale $\overline{Y}$, ceci revient à estimer les moyennes $\overline{Y}_h$ de chacun des H groupes.

On tire un échantillon aléatoire simple S_h de taille n_h dans chaque strate U_h. L'estimateur de Horvitz-Thompson de $\overline{Y}$ est donné par

$$\widehat{\overline{Y}} = \sum_h W_h \widehat{\overline{Y}}_h = \sum_h W_h \frac{1}{n_h} \sum_{k \in S_h} y_k.$$

A.2.2 Calcul de la variance et de l'allocation

La variance de $\widehat{\overline{Y}}$ est donnée par

$$\mathrm{Var}\left(\widehat{\overline{Y}}\right) = \sum_h W_h^2 \mathrm{Var}\left(\widehat{\overline{Y}}_h\right) = \sum_h W_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{1}{n_h} D_h^2,$$

où D_h^2 est la variance de y dans la strate U_h. Nous avons

$$\sum_h W_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{1}{n_h} D_h^2 = \sum_h \left\{ \frac{W_h^2 D_h^2}{n_h} - \frac{1}{N} W_h D_h^2 \right\}.$$

Soit $A_h = W_h D_h$, $B_h = W_h D_h^2$ et $A = \sum_h A_h, B = \sum_h B_h$. Alors

$$\mathrm{Var}\left(\widehat{\overline{Y}}\right) = \sum_h \frac{A_h^2}{n_h} - \frac{1}{N} \sum_h B_h = \sum_h \frac{A_h^2}{n_h} - \frac{B}{N}.$$

L'allocation optimale est donnée par

$$n_h = n \frac{A_h}{A} = n \frac{W_h D_h}{\sum_h W_h D_h} = n \frac{N_h D_h}{\sum_h N_h D_h}.$$

Sous l'allocation optimale,

$$\mathrm{Var}\left(\widehat{\overline{Y}}\right) = \frac{A}{n} \sum_h A_h - \frac{B}{N} = \frac{1}{n} A^2 - \frac{B}{N}.$$

Si $D_h = D$ pour tout $h \in H$, alors on retrouve l'allocation proportionnelle

$$n_h = n \frac{N_h D}{\sum_h N_h D} = n \frac{N_h}{\sum_h N_h} = n \frac{N_h}{N}.$$

$$\begin{aligned} A &= \sum_h A_h = D \sum_h W_h = D \\ \frac{B}{N} &= \frac{1}{N} D^2 \sum_h W_h = \frac{D^2}{N} \end{aligned}$$

et

$$\mathrm{Var}\left(\widehat{\overline{Y}}\right) = \frac{1}{n} D^2 - \frac{D^2}{N} = \left(1 - \frac{n}{N}\right) \frac{1}{n} D^2.$$

A.2.3 Taille d'échantillon par strate

Soit V la variance désirée de l'estimateur de $\overline{Y}$. Alors de

$$V = \frac{1}{n} A^2 - \frac{B}{N}$$

on tire

$$n = \frac{A^2}{V + B/N} = \frac{\left(\sum_h W_h D_h\right)^2}{V + \left(\sum_h W_h D_h^2\right)/N}.$$

Si $D_h = D$ pour tout $h \in H$, alors

$$n = \frac{D^2}{V + D^2/N} = \frac{D^2/V}{1 + \left(D^2/V\right)/N} = \frac{n_0}{1 + n_0/N}$$

où $n_0 = D^2/V$.

Références

[1] Salamin, P.-A. (2005). Plan de sondage auprès des universités pour l'enquête sur la situation sociale des étudiants 2004, Document de travail, OFS, Neuchâtel, Suisse.

[2] Salamin, P.-A. (2005). Plan de sondage auprès des hautes écoles spécialisées pour l'enquête sur la situation sociale des étudiants 2004, Document de travail, OFS, Neuchâtel, Suisse.

[3] Salamin, P.-A. (2008) Notes on sampling theory. Document interne, OFS, Neuchâtel, Suisse.

[4] Boegli, L., Inversion L., Müller P., Teichgräber M. (2007). Conditions de vie et d'études dans les hautes écoles suisses : Publication principale de l'enquête sur la situation sociale des étudiant-e-s 2005. Numéro de commande : 255-0501, Office Fédéral de la statistique (OFS), Neuchâtel. URL http://www.bfs.admin.ch/bfs/portal/fr/index/news/publikationen.html?publicationID=2578.

[5] Institut national de la statistique et de études économiques, Paris. La macro CALMAR.

[6] Paul-André Salamin (2007). Calibration : Advanced Methods of Survey Sampling, incl. Case study : Social situation of students survey 2004.

[7] Kish L. and Scott A. (1971). Journal of the American Statistical Association, Volume 66, Number 35 : Retaining Units after Changing Strata and Probabilities.

[8] Nedyalkova, D., Pea, J. & Tillé, Y. (2008), Sampling Procedures for Coordinating Stratified Samples : Methods Based on Microstrata,International Statistical Review, 76, 368-386.

[9] Site de l'enquête sur la situation sociale des étudiant-e-s http://www.bfs.admin.ch/bfs/portal/fr/index/infothek/erhebungen__quellen/blank/blank/soz_lag/00.html/portal/fr/index/infothek/erhebungen__quellen.html

Methodenberichte des Dienstes Statistische Methoden des BFS
Rapports de méthodes du Service de méthodes statistiques de l'OFS
Methodology reports published by the SFSO's Statistical Methods Unit

Andrade, B., Salamin P.-A. (2012). Enquête sur la situation sociale et économique des étudiant-e-s des hautes écoles suisses 2009. Cadre de sondage, plan d'échantillonnage et méthodes d'estimation. Numéro de commande : 338-0066

Potterat, J., Panchard, C., Kilchmann, D. (2012). Umweltschutzausgaben der Unternehmen 2009 (UWSA2009). Stichprobenplan, Einsetzungen, Gewichtung und Schätzverfahren. Bestellnummer : 338-0065

Potterat, J. (2012). Benutzung der Umsteigeschlüssel NOGA 2002-2008. Bestellnummer : 338-0064

Potterat, J. (2011). Kosten und Nutzen der Berufsbildung aus Sicht der Betriebe im Jahr 2009 (KNBB09). Stichprobenplan, Gewichtung und Schätzverfahren. Bestellnummer : 338-0063

Kilchmann, D., Potterat, J., Genoud, S. (2011). Gütertransporterhebung 2008. Stichprobenplan, Datenaufbereitung, Gewichtung und Schätzverfahren. Bestellnummer : 338-0062

Graf, E. (2010). Enquête suisse sur la santé 2007. Plan d'échantillonnage, pondérations et analyses pondérées des données. Numéro de commande : 338-0061

Eichenberger P., Hulliger B., Potterat J. (2010). Describing the Anticipated Accuracy of the Swiss Population Survey. Order number : 338-0060

Graf, E. (2010). Étude empirique de l'attrition du Panel Suisse de Ménages : vers une caractérisation du profil du non-répondant. Numéro de commande : 338-0059

Graf, E. (2009). Weightings of the Swiss Household Panel : SHP_I wave 9, SHP_II wave 4, SHP_I et SHP_II combined. Order number : 338-0058

Graf, E. (2009). Pondérations du Panel Suisse de Ménages : PSM_I vague 9, PSM_II vague 4, PSM_I et PSM_II combinés. Numéro de commande : 338-0057-05

Qualité, L., Tillé, Y. (2009). Estimation de la précision d'évolutions dans l'enquête sur la valeur ajoutée. Numéro de commande : 338-0056

Renaud, A., Panchard, C. et Potterat, J. (2008). Statistique de l'emploi. Révision 2007 : méthodes d'estimation. Numéro de commande : 338-0055

Graf, E. (2008). Pondérations du PSM. PSM_I vague 8, PSM_II vague 3, PSM_I et PSM_II combinés. Numéro de commande : 338-0054

Andrade, B., Graf, M. (2008). Enquête suisse sur la structure des salaires 2006. Aspects méthodologiques du modèle des salaires "Salarium". Numéro de commande : 338-0053

Renaud, A. (2008). Statistique de l'emploi. Révision 2007 : cadre de sondage et échantillonage. Numéro de commande : 338-0052

Graf, E. (2008). Pondérations du SILC pilote. SILC_I vague 2, SILC_II vague 1, SILC_I et SILC_II combinés. Numéro de commande : 338-0051

Kilchmann, D. (2008). Statistik der sozialmedizinischen Institutionen 1999-2004 und Krankenhausstatistik 1999-2002. Einsetzungen für fehlende Daten. Bestellnummer : 338-0050

Renaud, A. (2008). Technologies de l'information et de la communication. Estimations sur la base de la statistique de la valeur ajoutée. Numéro de commande : 338-0049

Assoulin, D. (2007). Wertschöpfungsstatistik. Einsetzungsversuche für fehlende Antworten grosser Unternehmen. Bestellnummer : 338-0048

Kilchmann, D. (2007). Beherbergungsstatistik Campingplätze. Stichprobenrahmen und Schätzverfahren 2005/06. Bestellnummer : 338-0047

Gabler, S., Häder, S. (2007). Haushalts- und Personenerhebungen. Machbarkeit von Random Digit Dialing in der Schweiz. Bestellnummer : 338-0046

Ferrez, J., Graf, M. (2007). Enquête suisse sur la structure des salaires. Programmes R pour l'intervalle de confiance de la médiane. Numéro de commande : 338-0045

Renaud, A. (2007). Harmonisation de la scolarité obligatoire en Suisse (HarmoS). Design général de l'enquête et échantillon des écoles. Numéro de commande : 338-0044

Potterat, J. (2007). Betriebszählung 2005. Statistische Methoden zur Schätzung der provisorischen Ergebnisse. Bestellnummer : 338-0043

Hulliger, B. (2006). Umweltschutzausgaben der Unternehmen 2003, Stichprobenplan, Datenaufbereitung und Schätzverfahren. Bestellnummer : 338-0042

Renfer, J.-P. (2006). Enquête sur les chiffres d'affaires du commerce de détail. Plan d'échantillonnage et méthodes d'estimation. Numéro de commande : 338-0041

Salamin, P.-A. (2006). Statistique de l'aide sociale dans le domaine de l'asile. Plan de sondage et extrapolations pour l'enquête pilote 2005. Numéro de commande : 338-0040

Renaud, A. (2006). Statistique suisse des bénéficiaires de l'aide sociale. Pondération des communes 2004. Numéro de commande : 338-0039

Graf, M. (2006). Swiss Earnings Structure Survey 2002-2004. Compositional data in a stratified two-stage sample : Analysis and precision assessment of wage components. Order number : 338-0038

Potterat, J. (2006). Pensionskassenstatistik 2004. Statistische Methoden zur Schätzung der provisorischen Ergebnisse. Bestellnummer : 338-0037

Potterat, J. (2006). Kosten und Nutzen der Berufsbildung aus Sicht der Betriebe im Jahr 2004. Stichprobenplan, Gewichtung und Schätzverfahren. Bestellnummer : 338-0036

Kilchmann, D. (2006). Vierteljährliche Wohnbaustatistik. Stichprobenplan, statistische Datenaufarbeitung und Schätzverfahren 2005. Bestellnummer : 338-0035

Kilchmann, D. (2006). Erhebung über Forschung und Entwicklung in der schweizerischen Privatwirtschaft 2004. Bereinigung der Stichprobe, Ersatz fehlender Werte und Schätzverfahren. Bestellnummer : 338-0034

Kilchmann, D., Eichenberger, P., Potterat, J. (2005). Volkszählung 2000. Statistische Einsetzungsverfahren Band 2. Bestellnummer : 338-0033

Kilchmann, D., Eichenberger, P., Potterat, J. (2005). Volkszählung 2000. Statistische Einsetzungsverfahren Band 1. Bestellnummer : 338-0032

Graf, M., Matei, A. (2005). Enquête suisse sur la structure des salaires 2002. La précision du salaire brut standardisé médian. Numéro de commande : 338-0031

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). Bereinigung der Stichprobe, Ersatz fehlender Werte und Schätzverfahren. Erhebung über F+E in der schweizerischen Privatwirtschaft 2000. 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 é lèves (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.

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:

Moyen de diffusion	*Contact*
Service de renseignements individuels	032 713 60 11 info@bfs.admin.ch
L'OFS sur Internet	www.statistique.admin.ch
Communiqués de presse: information rapide concernant les résultats les plus récents	www.news-stat.admin.ch
Publications: information approfondie (certaines sont disponibles sur disquette/CD-Rom)	032 713 60 60 order@bfs.admin.ch
Données interactives (banques de données, accessibles en ligne)	www.statdb.bfs.admin.ch

Informations sur les divers moyens de diffusion sur Internet à l'adresse www.statistique.admin.ch → Services → Les publications de Statistique suisse

Rapports de méthodes du Service de méthodes statistiques

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.

Ce rapport documente le plan d'échantillonnage, l'allocation et le tirage de l'échantillon ainsi que les méthodes d'estimation utilisées pour la statistique sur la situation sociale et économique des étudiant·e·s. Cette enquête a été menée pour la première fois en 2005 et a été reconduite en 2009. Le plan d'échantillonnage en 2009 est défini de la même manière qu'en 2005. Il a été construit dans le but de satisfaire les objectifs de précision de cette enquête. La stratification correspond au croisement des hautes écoles (université, école polytechnique, haute école spécialisée et haute école pédagogique) avec le domaine d'études (groupe de domaines d'études pour les hautes écoles universitaires et domaine d'études pour les hautes écoles spécialisées et pédagogiques). Une allocation qui garantit la précision au niveau haute école et domaine d'études resp. groupe de domaines d'études et qui tient compte de la non-réponse est appliquée. Le nouvel échantillon contient 11'500 étudiants pour les hautes écoles universitaires et 13'014 pour les hautes écoles spécialisées et pédagogiques, y compris les augmentations de l'échantillon demandées par une haute école universitaire et une haute école spécialisée.

L'échantillon de l'Université de Genève représente un cas particulier, car cette université a effectué presque en même temps une enquête auprès de ses étudiant·e·s avec un questionnaire similaire au questionnaire de l'OFS. Pour minimiser le nombre des étudiant·e·s qui sont interrogés deux fois, on a utilisé l'algorithme de Kish & Scott qui permet de faire une coordination entre les deux échantillons.

L'extrapolation de l'échantillon à la population est expliquée à l'aide de différents exemples. Pour terminer, des comparaisons entre estimateurs de variance sont présentées pour valider l'estimateur de variance utilisé dans cette enquête.

N° de commande
338-0066

Commandes
Tél.: 032 713 60 60
Fax: 032 713 60 61
E-mail: order@bfs.admin.ch

Prix
gratuit

ISBN 978-3-303-00472-2