ESTIMATION DE LA PRÉCISION D'UNE ENQUÊTE TÉLÉPHONIQUE : LE CAS DE L'ENQUÊTE VIRAGE

Nicolas RAZAFINDRATSIMA, Géraldine CHARRANCE, Gwennaëlle BRILHAULT

Institut National d'Etudes Démographiques (INED), Service des enquêtes et sondages

razafind@ined.fr; geraldine.charrance@ined.fr ; gwennaelle.brilhault@ined.fr

Mots-clés: estimation de variance, enquête téléphonique, procédures SAS

Résumé

L'INED a mené, en 2015/2016, une enquête sur les « violences et rapports de genre » (VIRAGE) auprès de femmes et d'hommes, âgés de 20 à 69 ans, habitant en France métropolitaine. Le volet principal de l'enquête a été réalisé par téléphone, et a permis de recueillir des informations sur les violences que les personnes ont subies dans différentes sphères (professionnelle, privée, familiale, etc.). Les données permettent d'estimer le nombre et la proportion de personnes victimes de violences, d'étudier les facteurs explicatifs et de construire des typologies (Hamel et al. 2016; Lebugle et l'équipe Virage 2017).

L'échantillon a été obtenu grâce à une génération aléatoire de numéros fixes et mobiles. Près de 200 000 numéros ont été générés, permettant, au final, d'interroger 27 268 personnes (20 971 auprès de lignes fixes, 6 297 par mobiles exclusifs). Le taux de réponse s'élève à 49,0%. Un calage sur marges a été réalisé sur les données du recensement de la population de 2012, avec les marges suivantes : le sexe, l'âge, le diplôme, le lieu de naissance, la catégorie socio-professionnelle, le type de ménage, et la région.

Notre objectif est de trouver une méthode d'estimation de la variance des résultats obtenus à la fois fiable et facilement utilisable par les utilisateurs des données, en général non statisticiens. Nous visons à proposer une méthode fondée sur les procédures SAS d'analyse d'enquêtes par sondage, avec une syntaxe simple et facile à mettre en œuvre, et donnant des résultats de variance proches de la « vraie » variance. Nous privilégions pour cela une optique « prudente », c'est-à-dire que nous simplifions la procédure d'estimation pour qu'elle soit facilement utilisable par tous les utilisateurs, quitte à surestimer légèrement la variance des résultats de l'enquête.

Notre démarche a consisté en deux étapes : Dans un premier temps, nous estimons empiriquement la « vraie » variance. Pour cela, nous approximons le plan de sondage de l'enquête, en faisant des hypothèses simplificatrices, comme celles utilisées pour l'enquête ACSF (Warszawski et al. 1997). Ensuite, nous prenons en compte le calage sur marges effectué, en utilisant un estimateur par régression (Ardilly 2006).

Abstract

This paper aims at proposing a simplified variance estimation for VIRAGE, a phone survey conducted by Ined in 2015 on a random sample of 27,268 people (15,556 women, 11,712 men) aged 20-69, living in ordinary households in metropolitan France, on experience of interpersonal violence in the 12 previous months and in the respondents'lifetimes in different life spaces (family, school,

workplace, current or former union, public space). We first estimate the 'true' variance, then propose a simplified syntax, and finally compare the results of the two procedures.

1. Présentation de l'enquête et de sa méthodologie

1.1. Présentation de l'enquête

L'enquête Violences et rapports de genre (dite VIRAGE) est une enquête de grande envergure réalisée auprès de 27 268 femmes et hommes, dont l'objectif est de mesurer l'ampleur des violences subies tant par les femmes que par les hommes.

Cette enquête est née de la volonté de chercheur(e)s de l'Ined de répondre aux besoin de renouvellement des connaissances sur les violences à l'encontre des femmes exprimé dès 2009 par la Mission d'évaluation de la politique de prévention et de lutte contre les violences faites aux femmes. Elle appelait à la réalisation d'une enquête permettant d'actualiser et approfondir les résultats issus de l'enquête pionnière sur ce thème, L'enquête nationale sur les violences envers les femmes en France (Enveff), réalisée en 2000 par le centre de recherche de l'Institut démographique de l'Université Paris 1 (Jaspard et l'équipe Enveff, 2003). Quinze ans après Enveff, l'enquête Virage entend compléter d'autres données comme les données ministérielles et les enquêtes de victimation (Cadre de vie et sécurité, Insee), en prenant en compte l'ensemble des situations où se produisent les violences (espaces publics, lieux d'étude, de travail, relation de couple, cadre familial et entourage proche) et la multiplicité des formes qu'elle peut prendre pour caractériser la situation des victimes. Elle permet de contextualiser et d'explorer les conséquences des violences sur les victimes, leur état de santé et leurs parcours scolaire, professionnel, familial et conjugal.

L'objectif premier de l'enquête est d'approfondir les connaissances sur les violences. Elle pose pour hypothèse que le degré de gravité est fonction de la nature des violences (verbales, psychologiques, physiques, sexuelles), de leur fréquence et cumul, de leur ancienneté, du contexte dans lequel elles se produisent, des liens existant entre l'auteur et la victime et des conséquences sur le devenir des personnes. Elle porte une attention particulière au sexe des victimes et des auteurs et replace les situations de violence dans le cadre plus global des inégalités de genre.

Pour répondre à ces ambitions, le projet Virage s'articule autour de quatre volets. Le volet «principal», dont il est question ici, est une enquête téléphonique sur un échantillon aléatoire représentatif de la population âgée de 20 à 69 ans, résidant en France métropolitaine, en ménages ordinaires et trois volets complémentaires réalisés sur Internet afin de permettre l'analyse sur des populations spécifiques de petite taille, mais particulièrement exposées au risque de subir des violences : population étudiante (Virage-université), population LGBT (Virage-LGBT) et population victime de violences ayant consulté des associations de soutien (Virage-Victimes).

1.2. Méthodologie de l'enquête

1.2.1. Constitution de l'échantillon

La constitution de l'échantillon et la collecte des données ont été réalisés par téléphone. Plusieurs contraintes ont guidé ce choix. En premier lieu, la volonté de constituer un échantillon représentatif de la population des 20-69 ans. La génération aléatoire de numéros de téléphone est un bon moyen de constituer une base sondage quasi exhaustive, compte tenu du taux d'équipement téléphonique des Français. De plus, l'interrogation par téléphone, utilisée pour la première fois en 1992 pour l'enquête ACSF (Attitude et Comportements Sexuels des Français) et reprise dans de nombreuses autres enquêtes depuis (Baromètre Santé, CSF, Fecond, Enveff) permet d'aborder plus facilement qu'en face-à-face, des sujets sensibles (Spira, Bajos & groupe ACSF, 1993). Les moindres coûts par

rapport à une enquête en face-à-face, surtout sur un échantillon de grande taille, est également un argument qui joue en faveur du recours au téléphone.

L'enquête repose sur un échantillonnage aléatoire à deux degrés (sélection d'un ménage, puis d'un individu).
Le premier degré d'échantillonnage consiste en la génération aléatoire de numéros de téléphones fixes et mobiles. Ces derniers sont ensuite filtrés sur les racines attribuées par l'Autorité pour la régulation des communications et des postes (ARCEP), puis confrontés à l'annuaire inversé pour retrouver des adresses et éliminer les entreprises, et finalement testés par un automate d'appel afin de ne retenir que les numéros attribués.
La seconde phase consiste à sélectionner aléatoirement un individu. Pour les lignes fixes, le tirage se fait parmi les adultes âgés de 20 à 69 ans résidant habituellement dans le foyer contacté. Pour les lignes mobiles, le tirage se fait parmi les utilisateurs réguliers de la ligne, âgés de 20 à 69 ans. Dans les deux cas, la sélection de la personne à interroger se fait après contrôle de l'éligibilité de la ligne et selon la méthode Kish.
Le contrôle de l'éligibilité de la ligne consiste, pour les numéros de fixes, à s'assurer qu'il s'agit d'un logement ordinaire occupé en qualité de résidence principale. Pour les mobiles, nous avons eu recours à la notion du « mobile exclusif » utilisée jusqu'alors dans les enquêtes aléatoires par téléphone : il s'agit de mobiles détenus par des personnes qui les utilisent pour leurs communications privées sans posséder de ligne fixe privée. Néanmoins, avec la généralisation d'internet, de nombreux ménages disposent d'un numéro de téléphone fixe fourni avec leur réseau sans que celui-ci soit lié à un appareil téléphonique. Parallèlement, les techniques de filtrage ont facilité un changement dans le comportement des utilisateurs : même si la ligne de téléphone fixe est utilisée pour passer des appels, l'affichage du numéro appelant leur permet de ne pas répondre, et cela de manière systématique, aux appels inconnus. La question utilisée jusqu'ici dans les enquêtes pour définir un mobile exclusif (et notamment lors de l'enquête pilote) semble ainsi écarter injustement un certain nombre des mobiles en tant qu'inéligibles, qui pourraient correspondre à certains types particuliers d'utilisateurs[1]. La masse des numéros non éligibles (toutes raisons confondues) trouvés lors du pilote (54% des mobiles, hors faux numéros) semblait corroborer cette analyse. Par ailleurs, d'autres enquêtes par questionnaire téléphonique ont récemment modifié leur échantillonnage des numéros de téléphone pour prendre en compte cette évolution[2]. Pour l'enquête réelle, nous avons considéré comme éligibles les lignes mobiles utilisées pour les raisons privées quand les utilisateurs de la ligne n'ont pas de téléphone fixe dans leur résidence principale ou lorsqu'ils disposent (d'au moins) un téléphone filaire mais auquel ils ne répondent pas quand ils ne connaissent pas le numéro de l'appelant.

1.2.2.Protocole de collecte

La procédure repose ensuite sur un protocole d'appels autorisant jusqu'à 40 tentatives de contact sur un numéro (20 avant la sélection de la personne à interroger, et 20 après), en variant les plages horaires et les jours d'appels, en semaine de 10h à 21h et le samedi de 10h à 16h. Les refus avant sélection de la personne dans le ménage sont rappelés deux fois et les refus des personnes sélectionnées le sont une fois. Même si la collecte s'est faite majoritairement par téléphone avec saisie informatique directe des réponses par des enquêteurs professionnels (méthode CATI) (n=26 634, soit 97,7% de l'échantillon), le protocole prévoyait de proposer aux personnes sélectionnées qui ne pouvaient ou ne souhaitaient pas répondre oralement (malentendants, personnes ayant des difficultés à s'isoler dans un cadre confidentiel, refus explicites…) de répondre au questionnaire en ligne.

[1] La question utilisée dans le cadre de l'enquête pilote de Virage était : « En plus de ce téléphone portable, disposez-vous dans votre résidence principale d'une ligne de téléphone fixe, y compris par Internet ? » La réponse « Oui » classait le numéro de mobile comme inéligible.
[2] En France, il s'agit notamment du Baromètre santé 2014 mené par l'INPES, même si la méthodologie finalement choisie diffère de notre approche.

1.2.3.Bilan de collecte

Le terrain de l'enquête VIRAGE a débuté le mardi 10 février 2015 et s'est terminé mi-novembre 2015. Il s'est donc déroulé sur près de 10 mois, avec une activité moindre durant l'été mais sans interruption.
Les numéros générés ont été exploités par lot, permettant de répartir l'exploitation dans le temps et d'assurer un effort identique pour tous les numéros. Chaque lot était composé de 70% de lignes fixes et 30% de lignes mobiles. La durée moyenne d'exploitation d'un lot était de 107 jours.

Une base initiale de 211 235 numéros de téléphone a été constituée permettant in fine de recueillir 27 268 questionnaires (tableau 1). Cette déperdition résulte des défauts de la base de sondage (numéros non attribués, numéros non éligibles à l'enquête) ainsi que des difficultés à joindre et interroger les individus.
En premier lieu, 29% (n=61 284) des numéros exploités étaient soit non attribués, soit des numéros de fax. Par ailleurs, 9,4% des numéros étaient injoignables lors des 20 appels prévus (n=19 900). Un contact a pu être établi auprès de 130 051 numéros, permettant d'identifier 24 424 numéros comme hors cible de niveau ménage (entreprises, collectivités, résidences secondaires ou numéros hors France métropolitaine), soit 18,8% des numéros contactés. Finalement, parmi les 105 626 ménages joints, 22 366 ont refusé de participer à l'enquête (soit un taux de refus des ménages de 21,2%), 3 624 se sont révélés impossibles à enquêter -non francophone, handicap, autre (soit 3,4%) et 14 328 n'ont pu être recontactés afin de procéder à la sélection (soit 13,6%). Parmi les ménages participants (n=65 309), c'est-à-dire acceptant de procéder à la phase de sélection, 22 520 (soit 34,5%) ont été classés hors-champ, soit en raison de l'absence d'individu dans la tranche d'âge ciblée, soit parce qu'ils ne vérifiaient pas la condition de mobiles exclusifs. La sélection d'un individu a donc été réalisée auprès de 42 789 numéros, parmi lesquels 6 104 personnes sélectionnées ont refusé de participé à l'enquête (soit un taux de refus au niveau individu de 14,3%), 2 061 se sont révélées impossibles à enquêter (problème de langue, handicap, absence durant la période de l'enquête), soit 4,8%, et 5 138 sont restés injoignables après des prises de rendez-vous, sans jamais avoir débuté le questionnaire (12,0%). Finalement, parmi les 29 486 personnes sélectionnées ayant débuté le questionnaire, 27 362 l'ont rempli intégralement, portant le taux d'abandon à 9,3%.

Tableau 1. Bilan de la collecte Virage

	Fixes	Fixes avec adresse	Fixes sans adresse	Mobiles	Total
1.0 Eligibles interrogés[3]	21 005	12 496	8 507	6 357	27 362
2.0 Eligibles non interrogés	10 381	6 334	4 047	2 922	13 303
2.1 Refus après sélection/Abandons	6 527	3 989	2 538	1 701	8 228
2.2 Personnes sélectionnées injoignables	3 662	2 213	1 449	1 476	5 138
2.3 Personnes sélectionnées impossibles à enquêter	1 672	1 109	563	389	2 061
3.0 Eligibilité inconnue	44 505	22 408	22 097	15 713	60 218
3.1 Refus avant sélection	17 082	10 890	6 192	5 284	22 366
3.2 Numéros ou ménages injoignables	24 655	9 677	14 978	9 573	34 228
3.3 Ménages impossibles	2 768	1 841	927	856	3 624
4.0 Non éligibles (Hors cible)	35 395	11 716	23 679	11 549	46 944
5.0 Faux numéros, fax…	36 653	9 756	26 897	24 631	61 284
Total	149 419	63 687	85 730	61 816	211 235

Source : Enquête VIRAGE 2015, Ined

[3] L'effectif finalement retenu pour l'analyse est de 27 268 questionnaires. La différence s'explique par une erreur sur une des questions filtre du champ de l'enquête et par l'exclusion de quelques questionnaires aux réponses non cohérentes.

Le taux de réponse, calculé selon les recommandations de l'AAPOR (American Association for Public Opinion Research), est de 49,0%. Par ailleurs, les 51% restants, correspondant aux non-répondants, se répartissent comme suit : 33,8% de refus/abandon, 10,4% de non-contacts et 6,8% d'impossibles à enquêter.

1.3.Calcul des pondérations

Les pondérations répondent à deux objectifs :

- Corriger les biais liés au plan de sondage. Lors du tirage de l'échantillon, les individus n'ont pas tous la même probabilité d'être sélectionné, le poids de sondage permet donc de corriger cet effet.
- Corriger les biais dus à la non-participation à l'enquête. Par exemple, les personnes n'ayant pas ou peu de diplôme ont moins souvent répondu à l'enquête. Il en résulte une surreprésentation dans l'échantillon de personnes ayant un diplôme de l'enseignement supérieur. On peut corriger cette distorsion en attribuant un poids plus important aux catégories sous-représentées dans l'échantillon.

Les deux phases de calcul des pondérations répondent successivement à ces deux objectifs.

1.3.1. Calcul des poids de sondage

Dans un premier temps, afin de corriger les biais liés au plan de sondage et de tenir compte des probabilités inégales de sélection, des poids de sondage ont été calculés. Ils correspondent au rapport du nombre d'éligibles dans le ménage sur le nombre de lignes fixes et mobiles permettant de joindre la personne sélectionnée (information collectée lors de la passation du questionnaire).

Une fois cette première opération effectuée, on observe une proportion d'individus ne disposant que d'un mobile est de 7,4% dans l'échantillon pondéré ; or, d'après le baromètre du numérique, enquête menée en juin 2015 par le CREDOC, la proportion est estimée à 12,5% au sein de la tranche d'âge de la population cible. Le poids de sondage est alors corrigé en conséquence.

A l'issue de cette première phase de calcul, nous pouvons mesurer les effets de la non-réponse sur la structure de notre échantillon en la confrontant à la distribution de la population cible obtenue via une source externe.

L'échantillon souffre des biais suivants (cf. tableau 2):

- Déficit de moins de 40 ans, excédent de 50 ans ou plus
- Déficit de non-diplômés, excédent de diplômés du supérieur du 2e et 3e cycle
- Déficit de personnes nées à l'étranger
- Déficit d'artisans, commerçants, d'ouvriers et d'inactifs / excédent de cadres, professions intermédiaires et retraités
- Déficit de familles monoparentales (chez les femmes) et déficit des autres types de ménage
- Déficit de personnes résidant en Ile-de-France

Tableau 2. Distribution de certaines variables dans le recensement et dans VIRAGE

En %	Distributions sur la population cible (RRP2012)		Distributions sur l'échantillon VIRAGE (pondéré par le poids de sondage corrigé)	
	Hommes	Femmes	Hommes	Femmes
Classe d'âge				
20-29 ans	19,1	18,8	15,8	12,7
30-39 ans	20,6	20,2	16,1	16,8
40-49 ans	22,3	22,0	21,9	23,0

50-59 ans	20,8	21,2	23,1	23,5
60-69 ans	17,1	17,8	23,2	24,0
Diplôme				
Aucun diplôme	18,3	19,8	7,9	9,0
BEPC/BEP/CAP	34,0	26,9	31,1	27,4
Baccalauréat	18,6	19,9	19,7	20,9
Dipl. du supérieur 1er cycle	12,9	17,2	15,8	15,3
Dipl. du supérieur 2e et 3e cycle	16,3	16,3	25,5	27,4
Lieu de naissance				
En France	86,1	85,7	91,1	91,8
A l'étranger	13,9	14,3	8,9	8,2
Catégorie socio-professionnelle				
Agriculteurs exploitants	1,7	0,6	1,2	0,6
Artisans, commerçants et chefs d'ent.	6,5	2,5	4,2	1,6
Cadres et prof. Intellectuelles sup.	14,5	9,3	18,2	10,8
Profesions Intermédiaires	17,7	19,1	18,5	20,9
Employés	9,9	31,6	11,5	29,2
Ouvriers	26,9	6,2	19,6	5,0
Retraités	15,8	15,8	20,8	20,2
Autres personnes sans activité prof.	7,0	15,4	6,1	11,8
Type de ménage				
Personnes seules	17,3	15,8	18,7	17,7
Couples sans enfants	26,6	27,5	30,4	29,0
Familles monoparentales	4,9	10,2	4,4	9,3
Couple avec enfants	43,1	39,2	41,2	39,6
Autres type de ménage	8,1	7,3	5,4	4,4
Région				
Île-de-France	19,1	19,6	16,6	16,6
Champagne-Ardenne	2,1	2,1	2,2	2,3
Picardie	3,1	3,0	2,8	2,7
Haute-Normandie	2,9	2,9	2,7	2,7
Centre	4,0	4,0	4,5	4,2
Basse-Normandie	2,3	2,3	2,8	2,6
Bourgogne	2,6	2,5	2,4	2,5
Nord-Pas-de-Calais	6,4	6,4	5,7	5,7
Lorraine	3,8	3,7	3,9	3,7
Alsace	3,0	3,0	3,3	3,2
Franche-Comté	1,9	1,8	1,9	1,9
Pays de la Loire	5,7	5,6	6,6	6,9
Bretagne	5,1	4,9	5,8	5,6
Poitou-Charentes	2,8	2,8	2,8	2,9
Aquitaine	5,2	5,2	5,0	5,3
Midi-Pyrénées	4,6	4,6	5,2	5,2
Limousin	1,2	1,1	1,3	1,1
Rhône-Alpes	10,0	10,0	10,7	10,6
Auvergne	2,2	2,1	2,5	2,7
Languedoc-Roussillon	4,2	4,3	4,4	4,4
Provence-Alpes-Côte d'Azur	7,5	7,8	6,6	6,8
Corse	0,5	0,5	0,3	0,4

Source : recensement de la population (RRP) de 2012, Insee ; Enquête VIRAGE 2015, Ined. Le total de chacune des distributions affichée dans le tableau vaut 100.

1.3.2. Calage sur marges

La seconde étape consiste à corriger les biais dus à la non-participation. On procède donc de façon classique à une post-stratification, dont le principe est d'imposer à l'échantillon la structure observée sur la population cible. Cette opération a également pour objectif de réduire la variance et le biais de l'échantillon.
Compte tenu de la thématique de l'enquête et du fait que les analyses seront souvent menées par sexe, nous avons opté pour une post-stratification par sexe, afin d'assurer une meilleure représentativité de chaque population (masculine et féminine).

Les marges de calage, calculées à partir du fichier du recensement 2012[4] (téléchargeable sur www.insee.fr), sont les suivantes : l'âge en 5 classes, le diplôme en 5 classes, le lieu de naissance en 2 classes, la catégorie socio-professionnelle en 8 classes, le type de ménage en 5 classes, et la région en 22 classes. Le calage sur marges a été réalisé grâce à la macro CALMAR (Sautory, 1993).

Tableau 3. Statistiques sur les poids

	Poids de sondage sans corr. Mob. Exclu	Poids de sondage av. corr. Mob. Exclu	Poids après calage sur marges (avt troncature)	Poids après calage sur marges (après troncature)
Coefficient de variation (%)	45,11	46,14	84,89	82,67
Max/Min	50	67,08	297,76	49,44
Min	0,14	0,13	55,53	144,57
1%	0,50	0,47	240,38	241,00
Médiane	1	1,74	1108,27	1113,20
99%	4	3,77	5752,28	5773,10
Max	7	8,72	16534,86	7147,65

Source : Enquête VIRAGE 2015, Ined

A l'issue de cette dernière étape de calcul, il convient de contrôler les statistiques de poids, et notamment le rapport du poids maximum sur le poids minimum (tableau 3). La présence d'écarts importants dans les pondérations conduit à une augmentation de la variance. Il est donc préférable de limiter ce rapport de poids.

Le poids de sondage présente un rapport de poids de 50 (7/0,14). Une fois la correction de la proportion des mobiles exclusifs dans l'échantillon, le rapport de poids passe à 67,8 (8,72/0,13). Le calage conduit, quant à lui, à multiplier le rapport de poids par 4,4, passant de 67,8 à 297.

Le rapport de poids étant jugé trop élevé, nous avons tronqué les poids extrêmes (les poids les plus bas et les plus élevés). Cette opération consiste à modifier le poids de 1% de l'échantillon (0,5% des poids les plus bas et 0,5% des poids les plus élevés). Après cette troncature, les poids sont ensuite normalisés sur les populations féminines et masculines (la somme des poids par sexe est égale à la taille de la population par sexe). Le rapport de poids final est de 49,44, résultat nettement plus acceptable.

[4] année de référence du dernier fichier au niveau détail diffusé par l'INSEE au moment des opérations

2. L'estimation de la variance

Dans cette section, nous présentons deux méthodes d'estimation de la variance. D'abord, nous estimons une variance tenant compte du plan de sondage réel de l'enquête puis du calage sur marges (qu'on peut appeler « vraie » variance). Ensuite, nous proposons une autre méthode d'estimation, plus simple, destiné aux utilisateurs des données de l'enquête.

2.1. Estimation de la « vraie » variance

Nous nous sommes largement inspirés des méthodes d'estimation élaborées lors de l'enquête ACSF de 1992 (Warszawski et al. 1997).

ACSF a été la première enquête téléphonique probabiliste menée en France sur le thème sensible des comportements sexuels. L'enquête visait à interroger les adultes de 18-69 ans résidant en France métropolitaine. Le plan de sondage de ACSF était en deux phases. En première phase, on a tiré aléatoirement des numéros de téléphones, en prenant comme base de sondage l'annuaire téléphonique (il s'agissait de numéros fixes uniquement, les portables étant encore inexistants), qui permettaient d'accéder à des ménages. Cette phase est assimilée à un tirage d'unités primaires (UP). On sélectionne ensuite aléatoirement une personne dans le ménage, assimilé à un tirage d'unités secondaires (US). Un questionnaire court est administré à cette personne sélectionnée, ce qui permet de classer les individus en 4 strates, selon le risque d'exposition au virus du VIH. En 2ème phase, on sélectionne des individus selon des probabilités inégales selon leur strate d'appartenance, en sur-représentant les populations « à risque » en matière de VIH puis on administre le questionnaire ACSF. 20 055 individus ont été interrogés.

La variance d'un tel plan de sondage est la somme de deux composantes, liées respectivement à la 1ère phase de tirage et à la 2e phase. La 1ère phase est très proche du plan de sondage de l'enquête VIRAGE, d'où notre idée de se baser sur les méthodes utilisées dans l'enquête ACSF pour les appliquer au cas de Virage.

L'estimateur de variance dans la première phase d'ACSF

La première phase du tirage dans ACSF est un sondage en deux degrés sans remise. Il existe une formule théorique pour la variance d'un tel plan (par exemple dans Ardilly, 2006). L'estimation est toutefois difficile du fait que nous n'avons tiré et enquêté qu'un seul individu par UP, ce qui rend impossible le calcul des dispersions intra-grappes.

Dans ACSF, les auteurs ont proposé d'utiliser, comme estimateur de variance de la première phase, celui proposé par Hansen et Hurwitz (1943) (appelé estimateur HH par la suite). Cet estimateur est valable en principe pour des sondages à probabilités inégales avec remise. Il a l'avantage de pouvoir être mis en œuvre simplement.

En notant :

- s l'échantillon, de taille n

- Y la variable d'intérêt, qui prend les valeurs Y_k (k=1 à n)

- P_k la probabilité de sélection de l'unité k, qui est constante d'un tirage à l'autre (le tirage étant avec remise)

- $\hat{t}_{p,Y}$ l'estimateur du total de la variable Y

On a :

$$\hat{t}_{p,Y} = \frac{1}{n} \frac{\sum_s y_k}{p_k} = -z \text{ , où } z_k = \frac{y_k}{p_k}$$

Et, l'estimateur de Hansen-Hurwitz de la variance :

$$\widehat{V}(\hat{t}_{p,Y}) = \frac{1}{n} \hat{S}_z^2 \text{ , où } \hat{S}_z^2 = \frac{1}{n-1} \sum_s (z_i - ¿ - z)^2 ¿$$

Prise en compte de la post-stratification

On note e_k le résidu de la régression de Y_k sur les variables de post-stratification X_k

$e_k = y_k - X'_k \hat{\beta}_s$, où $\hat{\beta}_s$ sont les coefficients estimés de la régression de Y sur X.

On pose $\hat{t}_e$ le total estimé de ces résidus. La variance de l'estimateur post-stratifié du total de Y vaut alors : $\widehat{V}(\hat{t}_e)$

Hypothèses

Afin de pouvoir utiliser l'estimateur de Hansen-Hurwitz, on fait l'hypothèse que les individus sont directement tirés de la population française adulte, avec des probabilités de sélection inégales, à savoir le produit de la probabilité de tirage au premier et au second degré. Ignorer le tirage en deux degrés pour le calcul de la variance peut se justifier pour les raisons suivantes :

- il n'y a pas d'effet de grappe : en effet, une seule unité secondaire est tirée de chaque unité primaire
- la probabilité de tirer deux individus d'un même ménage est négligeable si les individus ont effectivement été tirés directement avec remise à partir de la population mère. Cela est dû au faible taux de sondage au premier degré et à la taille relativement faible des ménages (en moyenne 2,1 personnes ménage, et moins encore si on considère uniquement les individus de la tranche d'âge éligible dans VIRAGE).

De plus, le faible taux de sondage nous permet de maintenir l'hypothèse que le tirage est effectué avec remise.

2.2. Mise en œuvre pratique

2.2.1. Calcul des probabilités de sélection

La probabilité de tirage a été estimée *a posteriori*. La difficulté, en effet, est que nous n'avions pas, au départ, le nombre de téléphones (fixes et mobiles exclusifs) éligibles, c'est-à-dire permettant de joindre une personne de population cible.

A partir de la pondération issue du calage sur marge, nous avons pu estimer le nombre de téléphones fixes et mobiles exclusifs éligibles à, respectivement 19 083 000 et 8 174 000.

La probabilité de sélection de chaque personne enquêtée est alors estimée à :

Pour les personnes jointes sur le téléphone fixe :
nombre de lignes fixes utilisées dans le ménage x taux de sondage estimé des téléphones fixes x (1/nombre d'éligibles du ménage)

Pour les personnes jointes sur téléphone mobile exclusif :
nombre de lignes mobiles utilisées dans le ménage x taux de sondage estimé des téléphones mobiles x (1/nombre d'utilisateurs de mobile du ménage)

2.2.2. La variance selon la méthode Hansen-Hurwitz

La variance selon la méthode de Hansen-Hurwitz est estimée en deux étapes, à l'aide du logiciel SAS (version 9.3).

- dans un premier temps, la procédure *surveymeans* est utilisée. Un sondage stratifié est spécifié (option *strata*), avec comme strates le type de téléphone ayant permis de sélectionner l'individu (fixe ou mobile exclusif). La pondération utilisée est l'inverse de la probabilité de tirage.
  Les coefficients de variation issus de cette procédure sont notés CV2 dans les résultats de la section 3

- dans un deuxième temps, on tient compte du calage sur marges. La variable d'intérêt est régressée sur les variables de calage (procédure *GLM*). On relance la procédure *surveymeans*, sur les résidus, avec les spécifications ci-dessus, et on obtient une nouvelle estimation de la variance
  Les coefficients de variation issus de cette procédure sont notés CV3 ci-dessous

2.3. Une syntaxe simplifiée pour les utilisateurs

La plupart des utilisateurs du fichier de VIRAGE ne sont pas statisticiens. Aussi nous avons souhaité mettre à disposition des utilisateurs une syntaxe simplifiée. Il s'agissait de se fonder sur les procédures habituelles du logiciel SAS, sans introduire de pondération supplémentaire à part la pondération de référence (poids issu du calage sur marges) déjà présente dans le fichier diffusé.

La syntaxe proposée se base sur la procédure *surveymeans* de SAS, en spécifiant une stratification (option *strata*), avec comme strates le type de téléphone ayant permis de sélectionner l'individu (fixe ou mobile exclusif). La pondération utilisée est la pondération issue du calage sur marges (cf section 1.3.2).

L'avantage de cette syntaxe est que, d'une part, l'estimation du total et de la moyenne produite est, en principe, sans biais, et identique à ce qu'on obtiendrait avec d'autres procédures de statistique descriptive. D'autre part, on obtient une variance (celle d'un sondage stratifié à un degré, avec remise, à probabilités inégales), que nous allons comparer avec les estimations HH exposées ci-dessus.

Les coefficients de correspondants sont notés CV1 dans les résultats de la section 3.

3. Les résultats

Nous vérifions empiriquement sur quelques variables d'intérêt que les variances produites avec la syntaxe simplifiée sont plus grandes que la « vraie » variance.

Pour cela, nous avons comparé les coefficients de variations (CV)[5] pour l'estimation du total et de la moyenne des quatre variables d'intérêt suivantes :

[5] Rapport entre l'écart type selon la méthode utilisée et l'estimation du total ou la moyenne (obtenue avec la pondération issue du calage sur marges)

- Avoir subi une ou des violences sexuelles sans frotté collé sur la vie entière
- Avoir subi une ou des violences sexuelles avec frotté collé sur la vie entière
- Avoir souffert d'un trouble anorexique au cours de la vie (terminé ou non) avec un IMC (indice de masse corporelle) inférieur à 18,5
- Souffrir actuellement d'un épisode dépressif majeur

Ces quatre variables sont toutes dichotomiques et mesurent des phénomènes rares (incidence à moins de 10%). Nous avons mené les estimations sur l'ensemble de l'échantillon, puis à l'intérieur de sous-groupes définis par une combinaison sexe x tranches d'âges (20-34, 35-49, 50-69), qui font office de « domaines » dans la syntaxe de SAS. Une partie des résultats est reproduite dans le tableau 4.

Tableau 4. Résultats des estimations de variance selon les trois méthodes

Ensemble

		En %						
	Total	**CV1**	**CV2**	**CV3**	**Proportion**	**CV1**	**CV2**	**CV3**
Violences sexuelles sans frotté collé 12 mois + vie entière	1 484 610	3,8	3,4	3,4	3,8	3,8	3,3	3,3
Violences sexuelles avec frotté collé 12 mois + vie entière	3 672 378	2,4	2,2	2,2	9,4	2,4	2,1	2,1
Trouble anorexique au cours de la vie (terminé ou non) avec un IMC inférieur à 18,5	742 474	5,3	5,1	5,0	1,9	5,3	4,8	4,8
Episode dépressif majeur actuel	3 925 687	2,4	2,0	2,0	10,0	2,4	1,9	1,9

Source : Enquête VIRAGE 2015, Ined

Femmes 20-34 ans

		En %						
	Total	**CV1**	**CV2**	**CV3**	**Proportion**	**CV1**	**CV2**	**CV3**
Violences sexuelles sans frotté collé 12 mois + vie entière	325 936	9,6	6,5	6,3	5,8	9,3	7,5	7,5
Violences sexuelles avec frotté collé 12 mois + vie entière	901 059	5,5	4,2	3,8	16,0	5,1	4,6	4,5
Trouble anorexique au cours de la vie (terminé ou non) avec un IMC inférieur à 18,5	245 691	10,3	8,1	7,9	4,4	10,1	9,3	9,3
Episode dépressif majeur actuel	591 262	6,9	5,1	4,9	10,5	6,6	5,8	5,8

Source : Enquête VIRAGE 2015, Ined

Hommes 20-34 ans

		En %						
	Total	**CV1**	**CV2**	**CV3**	**Proportion**	**CV1**	**CV2**	**CV3**
Violences sexuelles sans frotté collé 12 mois + vie entière	75 999	19,3	12,3	12,3	1,4	19,2	15,9	16,0
Violences sexuelles avec frotté collé 12 mois + vie entière	258 363	10,9	7,3	7,2	4,7	10,7	9,4	9,4
Trouble anorexique au cours de la vie (terminé ou non) avec un IMC inférieur à 18,5	9 007	46,1	38,6	39,1	0,2	46,1	50,3	50,9
Episode dépressif majeur actuel	391 805	8,3	6,1	5,9	7,1	8,1	7,7	7,6

Source : Enquête VIRAGE 2015, Ined.

On constate que la syntaxe simplifiée donne des CV (i.e des variances) supérieurs à ceux des méthodes HH (avant et après prise en compte du calage) sauf, dans le tableau 4, pour la proportion

d'hommes de 20-34 ans touchés par un « trouble anorexique au cours de la vie (terminé ou non) avec un IMC inférieur à 18,5 ».

Dans le tableau 5, nous avons synthétisé les résultats en comparant les coefficients de variations CV1 (syntaxe simplifiée) et CV3 (méthode HH tenant compte du calage) pour les différentes sous-populations.

Tableau 5. Synthèse des CV selon la méthode simplifiée vs la méthode HH avec prise en compte du calage

	Total		**Proportion**		**Nombre de sous-population Testées**
	CV1 >= CV3	**CV1 < CV3**	**CV1 >= CV3**	**CV1 < CV3**	
Violences sexuelles sans frotté collé 12 mois + vie entière	5	2	7	0	7
Violences sexuelles avec frotté collé 12 mois + vie entière	4	3	7	0	7
Trouble anorexique au cours de la vie (terminé ou non) avec un IMC inférieur à 18,5	5	2	5	2	7
Episode dépressif majeur actuel	7	0	7	0	7
Total	**21**	**7**	**26**	**2**	**28**

Source : Enquête VIRAGE 2015, Ined. Lecture : Pour la variable « violences sexuelles sans frotté collé 12 mois + vie entière », les CV de 7 totaux sont comparés (pour l'ensemble de la population, puis dans chacune des sous-populations définie par la combinaison sexe x âge en 3 postes). Dans 5 cas, CV1 est supérieur ou égal à CV3, tandis que dans 2 cas, CV1 est inférieur à CV3.

Pour le total, CV1 donne des variances plus élevées que CV3 dans 21 cas sur 28. Pour la proportion, CV1 donne des variances plus élevées que CV3 pour 26 cas sur 28. Ainsi, pour les variables étudiées ici, et dans la plupart des sous-groupes de populations, la syntaxe approchée paraît donner des résultats satisfaisants.

Conclusion

Dans cette étude, nous avons voulu proposer aux utilisateurs de l'enquête VIRAGE une procédure d'estimation de variance simple à mettre en œuvre, qui donne des intervalles de confiance proches, en tout cas en général plus larges que ceux de la vraie valeur.

Nous nous sommes largement inspirés des estimations menées dans l'enquête ACSF pour établir nos « vraies » variances.

Les premiers résultats montrent que, numériquement, la syntaxe simplifiée donne des résultats assez satisfaisants. Il reste à confirmer ces tests sur d'autres variables d'intérêt.

Bibliographie

Ardilly, Pascal. 2006. *Les techniques de sondage*. 2e édition. Paris: Technip.

Hamel, Christelle, Alice Debauche, Elisabeth Brown, Amandine Lebugle, Tania Lejbowicz, Magali Mazuy, Amélie Charruault, Sylvie Cromer, et Justine Dupuis. 2016. « Viols et agressions sexuelles en France : premiers résultats de l'enquête Virage ». *Population et Sociétés*, n° 538 (novembre).

Hansen, M.H and Hurwitz, W.N. 1943. "On the theory of sampling from finite populations", *Annals of Mathematical Statistics*, 14, p. 333-362

Jaspard, Maryse et l'équipe Enveff. 2003. *Les violences envers les femmes en France. Une enquête nationale*. Paris, la Documentation française, 370 pages.
Lebugle, Amandine, et l'équipe Virage. 2017. « Les violences dans les espaces publics touchent surtout les jeunes femmes des grandes villes ». *Population et Sociétés*, n° 550 (décembre).
Sautory, Olivier. 1993. *La macro CALMAR. Redressement d'un échantillon par calage sur marges*. Insee, document de travail n° F9310, https://insee.fr/fr/information/2021902 [consulté le 6 juin 2018]
Spira Alfred, Bajos Nathalie & le groupe ACSF. 1993. *Les comportements sexuels en France*, Paris, La Documentation française
Warszawski, Josiane, Antoine Messiah, Joseph Lellouch, Laurence Meyer, et Jean-Claude Deville. 1997. « Estimating Means and Percentages in a Complex Sampling Survey: Application to a French National Survey on Sexual Behaviour (ACSF) ». *Statistics in Medicine* 16: 397-423.