Rapport de méthodes

Analyse de données d'enquêtes

0

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

Quelques méthodes et illustration avec des données de l'OFS

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

OFS BFS UST

Neuchâtel, 2004

Statistik der Schweiz

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

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

Statistique de la Suisse

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

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

Statistik der Schweiz

Methodenbericht

Statistique de la Suisse

Rapport de méthodes

Analyse de données d'enquêtes

Quelques méthodes et illustration avec des données de l'OFS

Auteur **Anne Renaud**
Office fédéral de la statistique

Editeur **Office fédéral de la statistique**

OFS BFS UST

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

Neuchâtel, 2004

Préambule

Diverses discussions internes à l'OFS ont montré un besoin de conseils pour l'analyse de données d'enquêtes. Le Service de méthodes statistique (METH) a donc décidé de faire un résumé des connaissances dans le domaine et de fournir des exemples sur des jeux de données de l'OFS. Le présent rapport, réalisé par Anne Renaud de METH, a pour but de poursuivre les discussions et l'information sur le sujet. Il servira de base à la rédaction d'une recommandation pour l'analyse des données d'enquêtes à l'OFS.

Le rapport a été réalisé grâce à la participation de plusieurs personnes. Un grand merci à Beat Hulliger et Paul-André Salamin de METH pour les discussions méthodologiques et la relecture du rapport. Un grand merci également à Anne Cornali de la Section des revenus, de la consommation et des conditions de vie (EKL), à Beat Schmid de la Section des analyses socio-économiques (LEB) et à Jacques Méry de la Section sur les salaires et les conditions de travail (LOHN) pour leurs conseils dans l'utilisation de leurs données et les remarques sur le rapport.

Résumé

L'analyse de données d'enquêtes par échantillon requiert l'utilisation de méthodes spécifiques qui tiennent compte du plan d'échantillonnage et de la pondération des observations.

Une introduction théorique et des exemples d'application sont présentés dans le cadre de trois types d'analyses courantes: les statistiques descriptives (moyennes et médianes), l'estimation de coefficients dans des modèles de régression linéaire et les tests d'indépendance entre variables catégorielles (chi-carré).

Les exemples basés sur les données de l'Enquête sur les revenus et la consommation de 1998, en particulier le module «Conditions de vie», et celles de l'Enquête suisse sur la structure des salaires de 1998 permettent de voir l'effet de l'utilisation de méthodes adéquates en comparaison avec des méthodes classiques. Ces dernières produisent en effet des résultats biaisés qui peuvent conduire à des conclusions erronées.

Mots-clé

Rapport de méthodes; analyse de données d'enquêtes; estimation; pondération; biais; variance; effet de sondage; design effect; exemples; statistique publique

Complément d'information: Anne Renaud, tél. 032 713 62 65
Anne.Renaud@bfs.admin.ch

Réalisation: Service de méthodes statistiques, OFS

Diffusion: Office fédéral de la statistique
CH-2010 Neuchâtel
Tél. 032 713 60 60 / Fax 032 713 60 61
Order@bfs.admin.ch

Internet: http://www.statistik.admin.ch

Numéro de commande: 338-0024

Prix: gratuit

Série: Statistique de la Suisse

Domaine: 0 Bases statistiques et produits généraux

Langue du texte original: Français

Graphisme/Layout: OFS

Copyright: OFS, Neuchâtel 2004
La reproduction est autorisée, sauf à des fins commerciales,
si la source est mentionnée.

ISBN: 3-303-00295-9

Table des matières

1 Introduction **5**

2 Estimation sur la base d'un échantillon aléatoire **7**

2.1 Échantillon et estimation 7

2.2 Biais, variance et MSE d'un estimateur 8

2.3 Estimateur de Horwitz-Thompson 9

2.4 L'effet de sondage 9

2.5 Estimation dans des domaines 10

3 Choix auxquels est confronté l'analyste **11**

3.1 Approches design-based et model-based de la modélisation 11

3.2 Cas de la statistique publique 13

3.3 Choix des poids dans les analyses 14

4 Démarche dans le présent rapport **15**

4.1 Méthodes comparées 15

4.2 Mesure des effets des poids et du plan 16

5 Données pour les illustrations **17**

5.1 Enquête sur les revenus et la consommation 1998 et module "Conditions de vie" 17

5.2 Enquête suisse sur la structure des salaires 1998 18

6 Moyennes et médianes **20**

6.1 Estimateurs classiques 20

6.2 Estimateurs avec poids et plan 21

6.3 Estimateurs incomplets 22

6.4 Littérature 23

6.5 Logiciels 23
6.6 Exemple sur les données de l'ERC98 25
6.7 Exemple sur les données de la LSE98 28

7 Modèles de régression linéaire **36**
7.1 Estimateurs classiques 36
7.2 Estimateurs avec poids et plan 37
7.3 Littérature 38
7.4 Logiciels 38
7.5 Exemple sur les données de la LSE98 39

8 Tests d'indépendance entre variables catégorielles **46**
8.1 Chi-carré classique 46
8.2 Chi-carré avec poids et plan 47
8.3 Logiciels 50
8.4 Exemple sur les données de l'ERC98 50
8.5 Exemple sur les données de la LSE98 54

9 Conclusion **57**

A Trucs dans SAS **59**
A.1 Moyenne et variance de la moyenne pour l'ERC98 59
A.2 Variance de la médiane non pondérée sans plan pour l'ERC98 60
A.3 Variance de la médiane pondérée sans plan 61
A.4 Moyenne et variance de la moyenne pour la LSE98 62
A.5 Coefficients de régression et prévisions avec IML pour la LSE98 63
A.6 Calcul de χ^2_c et χ^2_p pour l'ERC98 64

B Trucs dans Wesvar **67**

Bibliographie **70**

Chapitre 1

Introduction

Les données d'enquêtes ne satisfont pas aux hypothèses de la statistique classique[1] car les unités sont tirées dans une population finie avec des probabilités inégales et les observations ne sont pas indépendantes les unes des autres.

Dans le cas habituel de la statistique publique, l'échantillon est basé sur un plan stratifié ou en grappe afin de satisfaire les objectifs de l'enquête tout en minimisant la taille de l'échantillon. L'algorithme de tirage, sans remise et avec des probabilités inégales dans la population finie, amène à des poids d'échantillonnage inégaux. Ces poids sont ensuite adaptés afin de tenir compte de la non-réponse, d'un éventuel calage sur des données auxiliaires ou des valeurs extrêmes dommageables aux analyses. Les poids finaux peuvent donc fortement varier d'une unité à l'autre.

Les méthodes classiques appliquées à des échantillons de la statistique publique donnent des résultats biaisés, avec une variabilité souvent sous estimée, car elles ignorent la complexité de l'échantillon. Des résultats raisonnables peuvent être espérés uniquement dans le cas d'une enquête basée sur un échantillon aléatoire simple tiré avec un faible taux de sondage et pour laquelle le taux de réponse est homogène dans tout l'échantillon. Ce type d'enquête est rare dans le cadre de la statistique publique.

Les méthodes spécifiques pour les données d'enquêtes ont été développées depuis les années 1960. Elles prennent en compte les poids et la structure du plan de sondage (stratification, grappes). Ces méthodes sont proposées dans les logiciels de statistique courants tel que SAS et SPSS, notamment pour des statistiques descriptives tels que des moyennes et des totaux ou encore des modèles de régression linéaire.

Toute une série de logiciels spécialisés offrent également un choix de méthodes plus important[2]. Si les logiciels à disposition ne proposent pas de procédure adéquate, il est possible d'appliquer des méthodes approximatives qui prennent en compte une partie de l'effet des poids et du plan dans les résultats (par ex. utilisation d'une mesure de l'effet de sondage). Ces méthodes de substitution sont cependant moins fiables que les méthodes spécifiques.

[1] Hypothèses de base de la statistique classique : observations indépendantes et identiquement distribuées (iid) sélectionnées dans une population infinie.

[2] Liste de logiciels : `http ://www.fas.harvard.edu/~stats/survey-soft/survey-soft.html`.

Après une introduction sur l'estimation en général et sur la question des poids et du plan dans les analyses, ce rapport aborde trois types d'analyses courantes : les statistiques descriptives (moyennes et médianes), l'estimation de coefficients dans des modèles de régression linéaire et les tests d'indépendance entre variables catégorielles.

Pour chaque type d'analyse, on présente une brève introduction théorique puis des exemples basés sur deux enquêtes de l'OFS. Des conseils et recommandations pratiques sont également proposés pour les applications avec les logiciels SAS v8 et Wesvar v4.

Chapitre 2

Estimation sur la base d'un échantillon aléatoire

Dans ce chapitre, nous introduisons quelques éléments de la théorie de l'échantillonnage utilisés dans le présent rapport. Pour une description plus complète, voir par exemple Cochran (1977), Särndal et al. (1992) ou encore Ardilly (1994) et Tillé (2001) en français.

2.1 Échantillon et estimation

Dans la théorie de l'échantillonnage, on désire estimer un paramètre θ, fonction d'une variable d'intérêt y, dans une population $U = \{1, .., i, .., N\}$ de taille finie N.

La caractéristique y est par exemple le salaire ou un ensemble de variables telles que le sexe, la taille et l'âge. Le paramètre θ est par exemple la moyenne $\theta = \bar{y}$ ou le total $\theta = Y = \sum_1^N y_i$. Il peut également être un coefficient d'un modèle de régression ou une statistique plus complexe.

Une enquête sur l'ensemble de la population U n'étant que rarement réalisable (relevé exhaustif ou recensement), on passe souvent par des échantillons pour avoir une information sur la population U.

Un échantillon s_{brut} de taille n_{brut} est tiré dans U selon un plan de sondage noté $p(s_{brut})$. Selon ce plan, chaque unité i de la population U a une probabilité connue π_i de faire partie de l'échantillon et l'inverse de cette probabilité $1/\pi_i$ définit le poids d'échantillonnage, $i = 1, .., N$.

Dans le cas le plus simple, un échantillon de taille n_{brut} est tiré dans U de façon à ce que chaque unité a la même probabilité de faire partie de l'échantillon. On a alors $\pi_i = n_{brut}/N$ ou encore $p(s_{brut}) = 1/\binom{N}{n_{brut}}$ (plan aléatoire simple).

En pratique, l'échantillon est basé sur un plan plus complexe que celui du tirage aléatoire simple. Les plans sont en effet stratifiés et/ou en grappes afin de satisfaire les objectifs de l'enquête tout en minimisant la taille de l'échantillon. On utilise aussi des tirages successifs (plusieurs niveaux) pour sélectionner des groupes d'unités puis un sous-échantillon de ces derniers (ex. écoles puis élèves).

L'enquête vise à recueillir les données y_i sur l'échantillon $s_{brut} \in U$ entier mais, en pratique,

une partie des données ne sont pas relevées suite à des refus par exemple. Il reste donc l'ensemble $s \subset s_{brut}$ de taille n pour l'estimation $\widehat{\theta}_s$ de θ. Les poids d'échantillonnage $1/\pi_i$ sont alors adaptés afin d'obtenir les poids finaux w_i pour $i \in s$. Les poids finaux w_i prennent en compte une correction pour la non-réponse, un éventuel calage sur des valeurs auxiliaires et une éventuelle adaptation pour éviter des valeurs extrêmes.

L'estimateur $\widehat{\theta}_s$ de θ doit être bien choisi afin d'avoir un biais et une variabilité faibles. De plus, il s'agit de définir un estimateur de sa variance ; lui-même si possible sans biais et de faible variabilité.

2.2 Biais, variance et MSE d'un estimateur

Le biais, la variance, l'écart-type, le coefficient de variation et l'erreur quadratique moyenne donnent une information sur les erreurs d'échantillonnage.

Le *biais* de l'estimateur $\widehat{\theta}_s$ est défini par :

$$B(\widehat{\theta}_s) = E(\widehat{\theta}_s) - \theta \tag{2.1}$$

avec $E(\widehat{\theta}_s) = \sum_s p(s)\widehat{\theta}_s$, l'espérance de $\widehat{\theta}_s$.

La *variance* de l'estimateur $\widehat{\theta}_s$ est définie par la moyenne des carrés des écarts à la moyenne :

$$V(\widehat{\theta}_s) = \sum_s p(s)[\widehat{\theta}_s - E(\widehat{\theta}_s)]^2 \tag{2.2}$$

et son *écart-type* par :

$$Std(\widehat{\theta}_s) = \sqrt{V(\widehat{\theta}_s)}. \tag{2.3}$$

On définit également le *coefficient de variation* de $\widehat{\theta}_s$:

$$cv(\widehat{\theta}_s) = Std(\widehat{\theta}_s)/E(\widehat{\theta}_s) \tag{2.4}$$

L'*erreur quadratique moyenne* (mean square error) de l'estimateur $\widehat{\theta}_s$ est définie par :

$$MSE(\widehat{\theta}_s) = E(\widehat{\theta}_s) - \theta)^2 = V(\widehat{\theta}_s) + [B(\widehat{\theta}_s)]^2 \tag{2.5}$$

L'estimateur est non biaisé si $B(\widehat{\theta}_s) = 0$. Dans ce cas, $MSE(\widehat{\theta}_s) = V(\widehat{\theta}_s)$.

Si le paramètre $\widehat{\theta}_s$ suit (au moins approximativement) une loi normale, on peut définir un intervalle de confiance à $(1-\alpha)\%$ comme suit : $[\widehat{\theta}_s - z_{\alpha/2}\, Std(\widehat{\theta}_s), \widehat{\theta}_s + z_{\alpha/2}\, Std(\widehat{\theta}_s)]$, avec $z_{\alpha/2}$ le $\alpha/2$-quantile de la loi normale. Si le risque de première espèce est de 5% ($\alpha = 5\%$), on retrouve l'intervalle $[\widehat{\theta}_s - 1.96\, Std(\widehat{\theta}_s), \widehat{\theta}_s + 1.96\, Std(\widehat{\theta}_s)]$.

Un facteur de correction pour la population finie est intégré aux estimateurs de la variance afin de tenir compte du fait que la population est de taille finie N. Ce facteur réduit la variance lorsque la taille n de l'échantillon se rapproche de la taille N de la population. Dans le cas d'un échantillon aléatoire simple, il est égal à $(1 - n/N)$.

Le but est habituellement de trouver le couple (plan, estimateur) de sorte que l'estimateur soit sans biais avec la variance minimale. Il est également possible de chercher la combinaison permettant d'avoir une erreur quadratique moyenne minimale, sans exiger que l'estimateur soit sans biais.

Le biais tout comme la variance ne sont pas connus exactement. La variance peut être estimée sur la base des données disponibles. Le biais ne peut être estimé que lorsque des données auxiliaires de référence sont disponibles.

2.3 Estimateur de Horwitz-Thompson

L'estimateur de Horwitz-Thompson ou π-*estimateur* du total $\theta = Y = \sum_{i=1}^{N} y_i$ pour un plan d'échantillonnage $p(s)$ est défini par :

$$\widehat{\theta}_s = \widehat{Y} = \sum_{i \in s} \frac{y_i}{\pi_i}. \tag{2.6}$$

On note que $\widehat{Y}$ estime Y sans biais si les π_i sont strictement positifs pour $i = 1, .., N$. L'expression de sa variance est donnée par exemple dans Särndal et al. (1992).

L'estimateur de Horwitz-Thompson est très utilisé en pratique pour des totaux mais aussi pour toutes les fonctions de totaux telles que les moyennes, les proportions, les quotients entre totaux ou encore les coefficients d'un modèle de régression.

2.4 L'effet de sondage

L'*effet de sondage* (ou design effect), popularisé par L. Kish dès le milieu des années 1960, est une mesure de l'effet du plan d'échantillonnage sur la variabilité. Il mesure le gain ou la perte de précision obtenue en choisissant le plan $p(s)$ par rapport à un plan aléatoire simple de même taille n.

L'effet de sondage $\mathit{deff}_p(\widehat{\theta}_s)$ du plan d'échantillonnage $p(s)$ pour un estimateur non biaisé $\widehat{\theta}_s$ du paramètre θ est défini par :

$$\mathit{deff}_p(\widehat{\theta}_s) = \frac{V_p(\widehat{\theta}_s)}{V_{SI}(\widehat{\theta}_{s,SI})} \tag{2.7}$$

avec $\widehat{\theta}_{s,SI}$ l'estimateur de θ sous le plan aléatoire simple sans remise SI, $V_{SI}(\widehat{\theta}_{s,SI})$ sa variance sous le plan SI et $V_p(\widehat{\theta}_s)$ la variance de $\widehat{\theta}_s$ sous le plan $p(s)$; voir par exemple Särndal et al. (1992).

L'effet de sondage *deff* est supérieur à 1 en cas de précision moindre du plan $p(s)$ par rapport à un plan aléatoire simple SI. Il est inférieur à 1 en cas de gain en précision. Si *deff* est supérieur à 1, un intervalle de confiance basé sur le plan $p(s)$ est plus large que celui basé sur un échantillon SI de même taille.

Pour une précision donnée, la taille de l'échantillon augmente en fonction de l'effet de sondage. La *taille effective* $n_{eff} = n/\textit{deff}$ est définie comme la taille d'un échantillon SI nécessaire pour obtenir la même précision que l'échantillon de taille n sous le plan $p(s)$.

En pratique, on utile souvent une définition modifiée de *deff*. Il s'agit de l'effet de sondage *deft* défini par :

$$\textit{deft}_p(\widehat{\theta}_s) = \sqrt{\frac{V_p(\widehat{\theta}_s)}{V_{SIR}(\widehat{\theta}_{s,SIR})}} \tag{2.8}$$

avec $\widehat{\theta}_{s,SIR}$ l'estimateur de θ sous le plan aléatoire simple avec remise SIR et la variance correspondante $V_{SIR}(\widehat{\theta}_{s,SIR})$. Le calcul de la variance de référence est plus simple car il ne doit plus tenir compte du tirage sans remise ; tirage qui nécessite notamment la prise en compte du facteur de correction de la population finie dans les formules. De plus *deft* peut être utilisé directement pour l'accroissement d'un intervalle de confiance par exemple.

L'estimation de l'effet de sondage ne pose pas de difficulté majeure dans le cas de moyennes ou de totaux dans des plans simples. Elle est plus difficile pour des statistiques et des enquêtes plus complexes.

L'effet de sondage est utilisé comme outil dans le développement d'un plan d'échantillonnage ou pour mesurer l'effet d'un plan sur les estimations et les analyses, voir par ex. Sections 6 et 7. Il est également utilisé pour modifier des statistiques d'inférence lorsque l'estimation de la variance $V_p(\widehat{\theta}_s)$ est difficile ; voir Section 8.

L'utilisation d'un effet de sondage estimé $\widehat{\textit{deft}}$ pour estimer une variance par une formule du type $\widehat{V}_p(\widehat{\theta}_s) = \widehat{\textit{deft}}^2 \widehat{V}_{SIR}(\widehat{\theta}_{s,SIR})$ est souvent peu fiable. L'effet de sondage peut en effet fortement varier d'un paramètre à l'autre et d'une enquête à l'autre. Il est lié notamment à la variabilité des poids, ainsi qu'à la corrélation entre les poids et le paramètre étudié ; voir par exemple Park and Lee (2001, 2002). Des informations générales peuvent cependant être citées (Särndal et al, 1992) : (1) l'effet de sondage est en principe supérieur à 1 dans le cas de plans avec grappes ; (2) l'effet de sondage de statistiques combinant plusieurs variables est en général inférieur aux effets de sondage des moyennes des variables combinées (par ex. l'effet sur le quotient entre le total de x et le total de y est inférieur à l'effet sur la moyenne de x) ; (3) l'effet de sondage de statistiques combinant plusieurs variables tend à croître si les effets de sondage des moyennes des variables utilisées augmentent (par ex. l'effet x/y augmente si l'effet sur x augmente).

2.5 Estimation dans des domaines

Dans la pratique des statistiques d'enquêtes, les estimations ne sont pas seulement désirées pour la population U dans son ensemble mais également pour des sous-ensembless de la population $U_d \subset U$ de taille $N_d < N$. On parle alors de domaines.

Dans de nombreux cas, la taille N_d de la sous-population est inconnue. De plus le nombre n_d d'éléments dans $s_d = s \cap U_d$ est aléatoire si la définition de la sous-population n'est pas liée au plan d'échantillonnage. Dans le cas d'une moyenne, la variance de l'estimateur $\sum_{s_d} y_i/n_d$ dépendra alors non seulement de la variance du total $\sum_{s_d} y_i$ mais également de celle de n_d ; voir par exemple Särndal et al. (1992).

Chapitre 3

Choix auxquels est confronté l'analyste

Dans l'analyse de données d'enquêtes, le statisticien se trouve face à trois questions :

1. Dois-je prendre en compte les poids dans mon analyse ? Si oui, quels poids et comment ?
2. Dois-je prendre en compte l'information sur le plan d'échantillonnage (strates et grappes) ? Si oui, comment ?
3. Quels sont les risques encourus si je ne prends pas en compte les poids et le plan ?

L'unanimité existe sur la nécessité d'utiliser les poids pour l'estimation de statistiques descriptives telles que des moyennes ou des totaux. Dans le cas contraire, le biais peut être très important.

Les opinions divergent sur l'utilisation des poids dans des analyses telles que l'estimation de coefficients dans des modèles de régression. Les deux approches principales sont l'approche *design-based* du statisticien d'enquête qui inclus des poids dans toutes les analyses et l'approche *model-based* du modélisateur qui ne prend pas en compte les poids.

Il est important de noter que, quelle que soit l'approche choisie pour l'estimation des coefficients d'un modèle (model-based ou design-based), le plan d'échantillonnage, et en particulier la dépendance provenant du tirage de grappes, doit être pris en compte dans l'inférence (par ex. calcul de variance, tests d'hypothèse) (Lohr, 1999). Dans le cas contraire, les conclusions peuvent être erronées car basées sur des résultats biaisés.

La question reste donc uniquement au niveau de l'utilisation ou non des poids pour l'estimation des paramètres d'un modèle.

3.1 Approches design-based et model-based de la modélisation

Les notions de design-based et model-based sont brièvement décrites dans cette section. Des informations complémentaires peuvent être trouvées par exemple dans Skinner et al. (1989), Chambers et Skinner (2003) ou encore Little (2004). La notation utilisée dans cette section est reprise de ces références.

Soient une population U de taille finie N et la matrice des variables d'enquête $Y_U = (y_1, .., y_N)$, avec y_i le vecteur des variables relevées pour l'unité i. Soit également le vecteur des variables d'inclusion $I_U = (I_1, .., I_N)$, avec $I_i = 1$ si l'unité i est dans l'échantillon s et $I_i = 0$ sinon.

On note $f(i_U)$ le méchanisme d'échantillonnage, avec i_U une réalisation du vecteur aléatoire I_U. Si la population U est considérée comme étant une réalisation d'un tirage avec remise parmi une *superpopulation* (fictive) de taille infinie, on note $f(y_U; \theta)$ la distribution de la réalisation y_U de la matrice aléatoire Y_U, avec θ le vecteur des paramètres définissant le modèle de superpopulation. Dans le cas général de l'inférence, on définit alors (y_U, i_U) la réalisation de la matrice aléatoire (Y_U, I_U).

On dit qu'un plan d'échantillonnage est *informatif* si le méchanisme d'échantillonnage dépend de y_U. La distribution conjointe de Y_U et I_U est donc de la forme $f(i_U|Y_U = y_U)f(Y_U = y_U)$. Si le méchanisme d'échantillonnage ne dépend pas de y_U, le plan est *non-informatif* et la distribution conjointe est de la forme $f(i_U)f(Y_U = y_U)$. En pratique, i_U dépend souvent de y_U par le biais des variables z_U définissant le plan d'échantillonnage (ex. strates). On parle alors de plan non-informatif conditionnellement à z_U si on a une distribution de la forme $f(i_U|Z_U = z_U)$.

Dans l'approche *design-based*, la seule source aléatoire provient du vecteur i_U. La matrice des valeurs y_U est considérée fixe et il n'est donc pas nécessaire de définir un modèle pour générer y_U. Le but est de trouver un estimateur $\widehat{\theta}$ de θ qui soit approximativement sans biais et qui ait une faible variabilité. Il est construit en intégrant les poids d'échantillonnage et sa variabilité provient de i_U. Les approximations normales liées aux grands échantillons sont en général utilisées pour l'inférence. Dans certains cas, des modèles sont utilisés pour motiver le choix des estimateurs. On parle alors d'une approche *model-assisted* ; voir Särndal et al (1992).

Dans l'approche *model-based* classique, la seule source de variabilité provient du modèle qui a généré y_U. De son côté, i_U est considéré comme fixe. Comme dans l'approche design-based, il est possible de chercher un estimateur $\widehat{\theta}$ de θ, par exemple par le maximum de vraisemblance, avec des tests basés sur le ratio du maximum de vraisemblance et une approximation normale. Cette approche est justifiée dans certains cas malgré l'ignorance de la distribution induite par le plan d'échantillonnage et des poids. Les conditions d'*ignorabilité* du plan d'échantillonnage sont par exemple présentes quand la vraisemblance ne dépend plus de i_U. La notion d'ignorabilité du plan est cependant complexe car elle ne dépend pas uniquement du plan et de l'information disponible sur le plan, mais également du modèle et du paramètre d'intérêt.

Plusieurs auteurs, voir résumé dans Lohr (1999), proposent l'utilisation de l'approche design-based afin d'avoir des résultats robustes par rapport à un éventuel mauvais choix du modèle (misspecification). Les pionniers de l'analyse des données complexes tels que Kish and Frankel (1974), Nathan and Holt (1980) ou encore Holt et al. (1980) ont cette approche. D'autres auteurs pensent que l'approche model-based, avec prise en compte du plan pour l'inférence, est préférable. Les variables liées à l'échantillonnage telles que les strates ou les grappes peuvent également être intégrées dans le modèle ; voir par exemple la disaggregated approach de Skinner et al. (1989) ou l'introduction sur les mixed models intégrant les grappes dans Lohr (1999).

La tendance est depuis quelques années à trouver un bon compromis entre les deux approches. Kish (1992) propose par exemple de choisir entre pondération ou pas sur la base de l'erreur quadratique moyenne (MSE=biais^2 + variance). Le biais d'un estimateur est proportionnel à la corrélation entre la variable d'intérêt et le poids. Il varie peu avec la taille de l'échantillon. De son

coté, la variance est supérieure pour le cas pondéré mais diminue avec la taille de l'échantillon. Dans le cas d'un grand échantillon, l'estimateur pondéré a donc souvent une MSE inférieure à l'estimateur non pondéré. Pour les petits échantillons, l'inverse peut être observé.

Un consensus semble s'être établi sur le rôle essentiel des poids dans le cas d'un plan non-ignorable (Pfeffermann, 1993). On estime que les poids préviennent contre le biais dû à un plan d'échantillonnage non-ignorable et contre les spécifications erronées du modèle de la population.

En guise de résumé, Lohr (1999) recommande l'approche model-based si l'échantillon n'est pas aléatoire, si le modèle théorique est connu et a été validé par des études empiriques ou encore si l'échantillon est petit car l'approche design-based demande de grands échantillons pour l'inférence. Elle propose de prendre en compte les poids et l'approche design-based dans le cas de statistiques publiques, notamment pour garder une cohérence avec les résultats obtenus pour des statistiques descriptives pondérées telles des moyennes globales ou par domaines.

3.2 Cas de la statistique publique

Les échantillons de la statistique publique sont construits pour atteindre des objectifs de précision concernant des statistiques descriptives telles que des moyennes, des médianes ou des totaux. Ils sont souvent mal adaptés pour la modélisation car le plan est complexe et les poids parfois très variables. Heureusement, les tailles d'échantillons sont habituellement grandes (plusieurs milliers), ce qui permet de limiter les risques liés aux petits échantillons.

Dans la pratique, ni l'ignorabilité du plan d'échantillonnage ni la spécification exacte du modèle ne peuvent être assurées. On recommande donc l'utilisation des poids dans la modélisation afin de rester cohérent avec les autres statistiques telles des moyennes, pour réduire l'effet d'une misspécification du modèle et pour éviter le biais dû aux plans non-ignorables. L'inférence sur les coefficients, de son côté, prend en compte la pondération et le plan d'échantillonnage dès que les poids varient et que les observations sont tirées selon un plan différent de l'aléatoire simple. Dans le cas de petits échantillons ou de petits domaines, il est par contre possible de choisir l'approche model-based ; en prenant des précautions liées à la petite taille et à l'ignorance de la structure de l'échantillon.

Notons qu'une comparaison des résultats avec et sans poids peut être utile par exemple lors du choix d'un modèle. Si le modèle décrit bien le mécanisme générant les données, les résultats obtenus avec et sans poids devraient être similaires. Une grande différence entre valeurs pondérées ou pas pourrait indiquer un mauvais fit du modèle pour une partie de la population. Notons que, si les valeurs des paramètres varient peu avec et sans poids, la variance des paramètres varie toujours selon l'approche choisie. En général, la variance design-based est plus élevée que la variance model-based.

3.3 Choix des poids dans les analyses

Dans la statistique publique, on détermine en général un unique jeu de poids finaux par échantillon pour les analyses. Ce jeu de poids est basé sur les variables et les estimateurs principaux de l'enquête et tient compte de l'échantillonnage avec des corrections telles que non-réponse et calage.

Dans certains cas, les résultats obtenus pour des variables non prioritaires dans les analyses, et pas considérées dans le calage, sont moins plausibles que ceux obtenus sur la base des poids d'échantillonnage uniquement. Certains statisticiens recommandent donc d'utiliser les poids avant correction pour la non-réponse, ou avant un calage sur des données auxiliaires ; voir par exemple Gabler et al. (1994).

Ce type de réflexion aborde la question de la difficulté du choix du modèle de non-réponse et du modèle de calage. En cas de besoins spécifiques, il est possible de créer un nouveau jeu de poids. La modification des poids fournis avec les données doit cependant être justifiée et documentée. Avec ou sans correction, les poids d'échantillonnages doivent dans tous les cas être pris en compte dans les estimations.

En pratique, on ignore souvent l'effet explicite du calage, de la correction de la non-réponse (en général non-ignorable) ou encore de l'imputation (perturbation des relations entre variables) sur la variabilité des estimateurs. Seuls les poids finaux et le plan d'échantillonnage sont pris en compte. Diverses méthodes, parfois complexes, existent cependant pour prendre en compte ces effets. Elles ne sont pas abordées dans ce rapport.

Chapitre 4

Démarche dans le présent rapport

Dans ce rapport, les données sont abordées dans une approche descriptive et design-based. C'est-à-dire que l'on mesure une caractéristique sur la population finie, à un certain moment. L'estimation est faite à partir des données relevées sur un échantillon tiré dans cette population, sans choix de modèle théorique.

On fait l'hypothèse qu'il n'y a pas de non-réponse partielle (item non-response) et que les erreurs "hors échantillonnage" (ex. erreur de mesure) sont négligeables. De plus, on considère que les poids finaux peuvent être utilisés tels quels dans les estimations.

On parle d'un *estimateur ponctuel* (point estimator) dans le cas d'une moyenne, une médiane ou encore d'un coefficient d'un modèle, et d'un *estimateur de variabilité* (variance estimator) pour la variance d'un estimateur ponctuel.

4.1 Méthodes comparées

Dans les analyses présentées dans les exemples, on compare les estimateurs ponctuels et de variabilité dans les trois cas suivants :

1. Estimation simple classique sans tenir compte des poids, du plan d'échantillonnage et du fait que la population est de taille finie.
2. Estimation pondérée, mais sans tenir compte du plan d'échantillonnage ni de la population finie.
3. Estimation pondérée, tenant compte du plan d'échantillonnage et de la population finie.

Le premier cas correspond à l'approche naïve d'un statisticien qui considère les données comme provenant d'un échantillonnage aléatoire simple dans une population de taille infinie (statistique classique). Les estimateurs ponctuels $\widehat{\theta}_{nonpond}$ et de variabilité $\widehat{V}_{SIR}(\widehat{\theta}_{nonpond})$ correspondant sont vraisemblablement biaisés.

Dans le deuxième cas, le statisticien prend en compte les poids afin d'obtenir des estimateurs ponctuels $\hat{\theta}_{pond}$ non biaisés mais il néglige de prendre en compte l'effet de l'échantillonnage et de la population finie pour les estimateurs de variabilité $\widehat{V}_{SIR}(\hat{\theta}_{pond})$. Il ne prend donc pas en compte les dépendances présentes dans l'échantillon ni l'effet bénéfique d'une éventuelle stratification.

Dans le troisième cas, le statisticien prend en compte les poids et le plan afin d'obtenir les estimateurs ponctuel $\widehat{\theta}_{pond}$ et de variabilité $\widehat{V}_p(\widehat{\theta}_{pond})$ adéquats. De plus, il tient compte du fait que la population est de taille finie. Il utilise une méthode correcte dans le cas de données provenant d'une enquête par échantillonnage de la statistique publique.

Les trois approches sont comparées à l'aide de statistiques construites pour l'estimateur ponctuel et l'estimateur de variabilité.

4.2 Mesure des effets des poids et du plan

On mesure l'effet de la pondération sur l'estimateur ponctuel par le *biais technique* $BT(\hat{\theta})$:

$$BT(\widehat{\theta}) = \widehat{\theta}_{nonpond} - \widehat{\theta}_{pond} \tag{4.1}$$

avec $\hat{\theta}_{pond}$ l'estimateur pondéré et $\hat{\theta}_{nonpond}$ l'estimateur non pondéré.

Le biais technique relatif est défini par $BTR(\widehat{\theta}) = BT(\widehat{\theta})/\widehat{\theta}_{pond}$.

L'effet des poids et du plan sur l'estimateur de variabilité est mesuré par le *weight effect* (*weft*), le *design effect* (*deft*) et le *complete effect* (*ceft*) :

$$weft(\widehat{\theta}) = \sqrt{\frac{\widehat{V}_{SIR}(\widehat{\theta}_{pond})}{\widehat{V}_{SIR}(\widehat{\theta}_{nonpond})}} = \frac{\widehat{Std}_{SIR}(\widehat{\theta}_{pond})}{\widehat{Std}_{SIR}(\widehat{\theta}_{nonpond})} \tag{4.2}$$

$$deft(\widehat{\theta}) = \sqrt{\frac{\widehat{V}_{p}(\widehat{\theta}_{pond})}{\widehat{V}_{SIR}(\hat{\theta}_{pond})}} = \frac{\widehat{Std}_{p}(\widehat{\theta}_{pond})}{\widehat{Std}_{SIR}(\hat{\theta}_{pond})} \tag{4.3}$$

$$ceft(\hat{\theta}) = weft(\widehat{\theta})\, deft(\hat{\theta}) = \sqrt{\frac{\widehat{V}_{p}(\widehat{\theta}_{pond})}{\widehat{V}_{SIR}(\hat{\theta}_{nonpond})}} = \frac{\widehat{Std}_{p}(\widehat{\theta}_{pond})}{\widehat{Std}_{SIR}(\hat{\theta}_{nonpond})} \tag{4.4}$$

On définit également l'*erreur quadratique moyenne technique* : (1) $MSET_{nonpond,SIR}(\hat{\theta})$ sans poids ni plan, (2) $MSET_{pond,SIR}(\hat{\theta})$ avec poids sans plan, et (3) $MSET_{pond,p}(\hat{\theta})$ avec poids et plan :

$$MSET_{nonpond,SIR}(\hat{\theta}) = BT(\hat{\theta})^2 + \widehat{V}_{SIR}(\widehat{\theta}_{nonpond}) \tag{4.5}$$

$$MSET_{pond,SIR}(\hat{\theta}) = \widehat{V}_{SIR}(\widehat{\theta}_{pond}) \tag{4.6}$$

$$MSET_{pond,p}(\hat{\theta}) = \widehat{V}_{p}(\widehat{\theta}_{pond}) \tag{4.7}$$

Dans les exemples, on mettra en évidence les risques encourus lorsque qu'une méthode inadéquate est appliquée, notamment sur les conclusions erronées qui peuvent en découler.

Chapitre 5

Données pour les illustrations

Deux jeux de données de l'OFS sont utilisés pour illustrer les méthodes d'analyse de données d'enquêtes. Le premier provient d'une enquête auprès des ménages avec strates mais pas de grappes. Le deuxième provient d'une enquête auprès d'entreprises, avec stratification et grappes.

Les résultats numériques présentés dans ce rapport ont pour but d'illustrer les concepts décrits. Ce ne sont pas des résultats officiels des enquêtes concernées.

5.1 Enquête sur les revenus et la consommation 1998 et module "Conditions de vie"

L'Enquête sur les revenus et la consommation de 1998 (ERC98) est une enquête auprès des ménages en Suisse organisée par la Section des prix et de la consommation de l'OFS (PREIS)[1]. La base de sondage est le registre pour les enquêtes auprès des ménages, c'est-à-dire une liste de numéros de téléphone ("Stichprobenregister für Haushalterhebungen", SRH). Le plan d'échantillonnage est stratifié par grandes régions (7 régions NUTS). Voir OFS (1999 et 2000a) pour une description de la méthodes d'enquête et des résultats.

Deux modules étaient rattachés à l'ERC98 : "Comportement en matière de voyages" et "Conditions de vie". Ces modules sont respectivement sous la responsabilité de la Section du tourisme (TOUR)[2] et de la Section de la culture, de la politique et des conditions de vie (KULT)[3]. Les ménages ayant participé à l'ERC98 et qui acceptent de répondre à des questions supplémentaires participent aux modules. Chaque ménage est interrogé sur un seul module, avec une probabilité de 50% pour chaque module. Les questions sont posées à une personne cible sélectionnée aléatoirement parmi les personnes du ménage ayant 15 ans ou plus.

Les données du module "Conditions de vie" et quelques variables au niveau des ménages provenant de l'ERC98 sont utilisées dans ce rapport. Pour plus de détails, voir les analyses regroupées dans Branger et al. (2002).

[1] Depuis le 1er janvier 2004 : section "Revenus, consommation et conditions de vie" (EKL).
[2] Depuis le 1er janvier 2004 : partiellement intégré dans d'autres sections.
[3] Depuis le 1er janvier 2004 : section "Analyses socio-économiques" (LEB).

Notons qu'un effet de grappe existe car des ménages (grappes d'individus) sont tirés avant de sélectionner une personne cible. En pratique, on ne tient pas compte de ce tirage à deux niveaux dans les estimations de la variance. La variance intra-ménage n'est pas connue et on considère que les personnes cibles ont été tirées dans un plan à un niveau, avec des probabilités inégales. Notons que cette simplification ne pourrait être faite si l'on sélectionnait plus de 1 personne par ménage.

Les données du module "Conditions de vie" comprennent 4'369 personnes pour les analyses. Le taux de réponse est de l'ordre de 35% pour l'ERC98 et de 94% pour le module.

La pondération des ménages de l'ERC98 prend en compte le plan d'échantillonnage et un modèle de non-réponse en deux parties : taux fixe et modèle logistique (Cornali, 2000). La pondération des personnes du module "Conditions de vie" reprend la pondération des ménages de l'ERC98 et l'adapte pour tenir compte de la sélection aléatoire du module (50%), de la non-réponse au module, et du choix de la personne cible dans le ménage. Les poids sont ensuite corrigés par un calage sur des données de population de référence et winzorisés pour éviter les valeurs extrêmes.

5.2 Enquête suisse sur la structure des salaires 1998

L'Enquête suisse sur la structure des salaires de 1998 (LSE98) est une enquête bisannuelle auprès des entreprises organisée par la Section sur la structure des salaire de l'OFS (LOHN)[4]. La base de sondage pour le secteur privé est le Registre des entreprises et des établissements (REE et REE-AGR). Le secteur public est relevé de façon indépendante. Le plan d'échantillonnage comporte deux niveaux (entreprises et employés) avec des tirages aléatoires simples sans remise aux deux niveaux. Au premier niveau, les entreprises sont stratifiées selon leur taille et leur activité économique (Peters et Hulliger, 1996). Les probabilités d'inclusion au niveau des employés sont inégales.

En 1998, quelque 7'100 entreprises représentant 516'000 salariés sont exploitables. Le taux de réponse net des entreprises s'élève à 83%. La pondération finale de la LSE98 est construite de façon à pouvoir être utilisée dans des analyses telles que l'estimation des salaires médians pour différentes classes économiques, pour les secteurs public et privé ou pour les hommes et les femmes. Elle tient compte du plan d'échantillonnage, de la non-réponse, de l'imputation de valeurs aberrantes et d'un calage sur des données auxiliaires au niveau des emplois (Peters, 1997).

L'OFS publie les résultats de la LSE, voir par exemple OFS (2000b), en prenant en compte la pondération. Une estimation (non publiée) de la variance des salaires moyens ou médians est calculée avec une approche simplifiée de l'estimateur linéarisé (Peters et Hulliger, 1996).

Une analyse des données de la LSE94 a permis de montrer l'effet non négligeable du plan d'échantillonnage sur les variances. On note par exemple un effet de sondage *deff* du salaire moyen prenant des valeurs entre 1 et 23 selon l'activité économique (Peters et Hulliger, 1996) et entre 5.75 et 12.3 pour des paramètres d'un modèle de régression (information de B.Hulliger). L'effet de la variabilité des poids des employés sur le salaire moyen est également ressorti de

[4]Depuis le 1er janvier 2004, nouvelle appellation : Section sur les salaires et les conditions de travail (LOHN).

l'étude de Canty et Davison (1998). Cette étude propose d'éviter l'utilisation de l'effet de sondage, vu son manque de stabilité, et recommande le bootstrap pour les estimations de la variance. Les estimateurs de la variance des moyennes et médianes ont été revus pour la LSE2000 (Graf, 2002).

Les données de l'Enquête suisse sur la structure des salaires sont utilisées par divers chercheurs externes à l'OFS, notamment pour analyser le lien entre le salaire et diverses variables relevées dans l'enquête (ex. sexe, niveau de qualification, années d'expérience).

Les données du secteur privé de la LSE98 telles que fournies aux chercheurs externes à l'OFS sont utilisées dans ce rapport. Elles sont donc anonymisées et ne contiennent qu'une partie des variables disponibles en interne.

Chapitre 6

Moyennes et médianes

La moyenne et la médiane sont des mesures de la tendance centrale très utilisées dans les statistiques publiques.

Nous donnons des résultats pour les moyennes et médianes comme base des statistiques descriptives. Les résultats pour les proportions et les totaux peuvent être directement déduit de ceux sur les moyennes. Ceux pour les autres quantiles peuvent être déterminés de façon similaire à ceux de la médiane.

6.1 Estimateurs classiques

Dans la statistique classique, on considère que les n données sont des réalisations indépendantes $y_1, .., y_n$ d'une variable aléatoire y suivant une loi de moyenne μ et de variance σ^2.

Pour faire le lien avec la théorie de l'échantillonnage, l'ensemble des n réalisations peut être considéré comme un échantillon de taille n tiré selon un plan aléatoire simple avec remise dans une population de taille infinie.

6.1.1 Moyenne

L'estimateur classique $\widehat{\overline{y}}_{iid}$ de la moyenne est donné par :

$$\widehat{\overline{y}}_{iid} = \frac{\sum_{i=1}^{n} y_i}{n} = \widehat{\mu} \tag{6.1}$$

L'estimation de la variance de la moyenne est donnée par :

$$\widehat{V}_{iid}(\widehat{\overline{y}}_{iid}) = \frac{s^2}{n} \tag{6.2}$$

avec s^2 l'estimateur de la variance de y :

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \widehat{\overline{y}}_{iid})^2}{n-1} = \widehat{\sigma}^2 \tag{6.3}$$

6.1.2 Médiane

L'estimateur classique de la médiane $\widehat{med}_{iid}$ est la valeur centrale des données. Elle est supérieure à la moitié des données et inférieure à l'autre moitié. Si n est paire, la médiane est définie par la moyenne entre les deux valeurs centrales ou par la valeur centrale supérieure.

La médiane est également définie par $\widehat{med}_{iid} = F^{-1}(0.5)$, l'inverse de la fonction de répartition empirique $F(.)$ exprimée en 0.5. Pour un jeu de n données, on a $F(y) = |A_y|/n$, avec $|A_y|$ la taille de l'ensemble $A_y = \{k|y_k \leq y\}$. La fonction $F(.)$ est en escalier. Par conséquent, la valeurs de $F^{-1}(0.5)$ doit parfois être choisie parmi plusieurs valeurs ; voir par exemple Särndal et al. (1992).

L'estimation de la variance d'une médiane est plus complexe que celle d'une moyenne.

Selon une approximation normale, la variance de la moyenne $V(\overline{y})$ et la variance de la médiane $V(med)$ sont liées par la formule $V(med) = \frac{\pi}{2}V(\overline{y})$. Cette formule est cependant parfois très loin de la vraie relation entre les deux variances.

Si les données sont indépendantes mais ne suivent pas une loi normale, la variance de la médiane peut être estimée par une méthode non-paramétrique telle que celle de Woodruff, voir par exemple Särndal et al. (1992).

L'idée de la méthode de Woodruf est d'estimer la variance de $F(\widehat{med}_{iid})$ puis un intervalle de confiance $[l, u]$ autour de $F(\widehat{med}_{iid})$ en utilisant des hypothèses de normalité, valable en cas de gros échantillons. On revient ensuite dans les valeurs de base en déterminant l'intervalle $[F^{-1}(l), F^{-1}(u)]$ autour de $\widehat{med}_{iid}$. L'intervalle ainsi déterminé est souvent asymétrique autour de la médiane mais il permet tout de même d'estimer un écart-type et une variance. L'écart-type est par exemple choisi sous la forme $Std = x/z_{\alpha/2}$ avec $z_{\alpha/2}$ la valeur de la loi normale au niveau α et $x =$ longueur de l'intervalle/2 ou, plus conservateur, $x = \max[(\widehat{med}_{iid} - F^{-1}(l)), (F^{-1}(u) - \widehat{med}_{iid})]$. Dans le deuxième cas, l'intervalle $\widehat{med}_{iid} \pm 2\ Std$ englobe les bornes de $[F^{-1}(l), F^{-1}(u)]$ ($\alpha = 5\%$). La variance estimée est donnée par $\widehat{V}_{iid}(\widehat{med}_{iid}) = Std^2$.

6.2 Estimateurs avec poids et plan

Dans le cas d'une enquête, un poids w_i est associé à chaque observation $i = 1, .., n$ de l'échantillon. De plus, les n éléments ont été tirés dans une population finie de taille N selon un plan d'échantillonnage $p(s)$ différent du plan aléatoire simple.

L'estimateur pondéré $\widehat{\overline{y}}_{pond}$ de la moyenne est défini par :

$$\widehat{\overline{y}}_{pond} = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i} \tag{6.4}$$

L'estimateur de la variance $\widehat{V}_p(\widehat{\overline{y}}_{pond})$ tenant compte des poids et du plan a une forme qui dépend de la stratification, de l'allocation, des niveaux d'échantillonnage et des algorithmes de tirage ; voir par exemple Särndal et al (1992), Cochran (1977) ou, plus théorique, Wolter (1985) et Skinner (1989).

La variance estimée de la moyenne peut être calculée de manière explicite dans la plupart des cas présents dans la statistique publique ; pour l'échantillon entier ou des sous-échantillons particuliers (domaines). Des simplifications sont souvent appliquées en pratique. Parmi ces dernières on note la non prise en compte du facteur de correction de la population finie ou encore la prise en compte uniquement du niveau supérieur d'échantillonnage sans la décomposition de la variance inter-grappes et intra-grappes. Dans le cas de statistiques ou de plans complexes, des méthodes de linéarisation ou de rééchantillonnage prennent le relais.

L'estimateur pondéré de la médiane $\widehat{med}_{pond}$ peut prendre diverses formes. Il peut être défini de manière similaire à la médiane simple, en remplaçant les y_i par les $w_i y_i$, $i = 1, .., n$, ou encore être défini comme la valeur correspondant à la valeur 0.5 de la fonction de répartition empirique, avec approximation linéaire si nécessaire.

On utilisera par défaut la forme suivante :

$$\widehat{med}_{pond} = \begin{cases} \frac{(y_k + y_{k+1})}{2} & \text{si } \sum_{i=1}^{k} w_i = \frac{W}{2} \\ y_{k+1} & \text{si } \sum_{i=1}^{k} w_i < \frac{W}{2} < \sum_{i=1}^{k+1} w_i \end{cases} \tag{6.5}$$

avec $W = \sum_{i=1}^{n} w_i$, la somme des poids.

La variance de la médiane est estimée en général sur la base de la méthode de Woodruff, voir Särndal et al. (1992) pour une description générale et Graf (2002) pour le cas d'un plan stratifié à deux niveaux, avec tirage aléatoire simple aux deux niveaux. La variance de $F(med)$ peut être estimée selon une méthode de linéarisation (ex. Graf, 2002) ou de rééchantillonnage (ex. WesVar). Dans tous les cas, il s'agit de choisir l'estimateur de l'écart-type une fois que l'intervalle de confiance de la médiane est déterminé ; voir Section 6.1. Notons que les méthodes de rééchantillonnage ne sont pas recommandées pour une estimation directe de la variance d'un percentile (sans passer par la fonction de répartition). Les résultats sont souvent fortement biaisés. Le Jackknife ne marche par exemple que dans certains cas avec de grands échantillons et des grappes de taille moyenne (WesStat, 2000).

6.3 Estimateurs incomplets

Dans le cas de données d'enquêtes, on utilisera les notations suivantes pour l'estimateur de la moyenne et de la médiane sans prise en compte des poids : $\widehat{\overline{y}}_{nonpond} = \widehat{\overline{y}}_{iid}$ et $\widehat{med}_{nonpond} = \widehat{med}_{iid}$. De même, les estimations de leur variance sans tenir compte des poids et du plan sont notées $\widehat{V}_{SIR}(\widehat{\overline{y}}_{nonpond}) = \widehat{V}_{iid}(\widehat{\overline{y}}_{iid})$ et $\widehat{V}_{SIR}(\widehat{med}_{nonpond}) = \widehat{V}_{iid}(\widehat{med}_{iid})$.

On définit également l'estimateur incomplet $\widehat{V}_{SIR}(\widehat{\overline{y}}_{pond})$ de la variance de $\widehat{\overline{y}}_{pond}$, tenant compte uniquement des poids mais pas du plan et de la population finie :

$$\widehat{V}_{SIR}(\widehat{\overline{y}}_{pond}) = \frac{s^2_{pond}}{n} \tag{6.6}$$

avec s^2_{pond} l'estimateur de la variance dans la population :

$$s^2_{pond} = \frac{\sum_{i=1}^{n} w_i (y_i - \widehat{\overline{y}}_{pond})^2}{df} \tag{6.7}$$

et $df = \sum_i w_i$. Notons que df peut prendre la valeur $\sum_i w_i - 1$, mais que la valeur $n - 1$ se serait pas correcte.

6.4 Littérature

De nombreuses publications présentent des comparaisons entre moyennes classiques et pondérées, avec calcul de la variance selon la méthode classique ou tenant compte du plan d'échantillonnage. La différence entre les résultats dépend notamment du jeu de données, de la variabilité du poids et de la corrélation entre la variable d'intérêt et le poids.

Brogan (1998) présente par exemple une comparaison entre les résultats obtenus par différents logiciels sur les données de l'enquête "Behavioral Risk Factor Surveillance System". Il montre ainsi l'importance de l'emploi des méthodes adéquates.

On citera aussi ici les analyses faites sur la première et la deuxième National Heath and Nutrition Examination Survey (NHNES) aux États-Unis. L'échantillon de cette enquête a été tiré selon un plan d'échantillonnage stratifié avec des grappes ; voir par exemple DHEW (1975) ou DHHS (1982) pour la description de l'enquête et des analyses. Lee et al. (1989) observe peu de différence entre les moyennes simples et pondérées d'une série de variable mais des *deff* variant entre 0.89 et 21 pour ces mêmes estimations.

Il semble que peu de comparaisons aient été réalisées sur l'influence des poids et du plan sur les médianes (ou d'autres quantiles).

6.5 Logiciels

Deux logiciels de statistique sont utilisés pour les analyses présentées ci-dessous : SAS v8 et WesVar v4. Dans cette section, on décrit brièvement les outils concernant les moyennes et médianes. Des exemples d'application sont présentés dans les Sections 6.6 et 6.7 et dans l'annexe.

PROC UNIVARIATE de SAS permet de calculer les moyennes et les médianes, simples et pondérées, ainsi que la variance de la moyenne, avec et sans poids mais sans tenir compte du plan d'échantillonnage. Une estimation de la variance de la médiane sans poids ni plan est fournie indirectement avec l'option CIPCTLDF.

PROC SURVEYMEANS de SAS permet de tenir compte des informations sur l'échantillonnage (poids, strates et grappes). Elle calcule les estimateurs pondérés ou non de la moyenne et leur variance selon la méthode de linéarisation de Taylor. Dans le cas de plans d'échantillonnage en grappe, seul le premier niveau est considéré dans l'estimation. SAS fait donc l'hypothèse que les grappes ont été tirées avec remise ou que le taux de sondage au premier niveau est petit. L'option DOMAIN permet de faire des analyses pour des domaines et l'option BY effectue des analyses séparées. Si le nombre d'éléments d'un sous-groupe n'est pas connu exactement dans la population on choisit l'option DOMAIN car l'option BY ne donne pas des résultats corrects.

SAS n'offre pas la possibilité de calculer la variance de la médiane en prenant compte des poids et du plan.

WesVar permet de calculer des estimateurs simples et pondérés des moyennes et médianes. Il produit un estimateur de la variabilité par la méthode du Jackknife pour la moyenne et une méthode basée sur la méthode de Woodruff pour la médiane (la variance de la valeur dans la fonction de répartition empirique est estimée par Jackknife). Ces estimations tiennent compte des poids, des strates et des grappes. Comme SAS, seul le premier niveau d'échantillonnage est considéré dans l'estimation de la variance et on peut faire des estimations par domaines.

SAS et Wesvar utilisent une méthode légèrement différente pour l'estimation de la médiane pondérée. SAS choisi la valeur qui vient juste après le milieu des données (avec poids) et WesVar donne une valeur entre les deux observations se trouvant au centre des données. Les résultats peuvent donc être sensiblement différents si les valeurs centrales sont éloignées les unes des autres.

Monique Graf (METH, OFS) a programmé deux macros SAS permettant d'estimer une moyenne, respectivement une médiane, et leur variance dans le cas d'un plan d'échantillonnage stratifié à deux niveaux. Les deux macros traitent correctement les domaines et donnent une variance estimée qui tient compte de la correction pour la population finie et des deux niveaux d'échantillonnage. La macro pour la médiane se base sur la méthode de Woodruff avec linéarisation (Graf,2002). L'écart-type de la médiane est obtenu par le quotient entre le plus grand demi-intervalle de confiance et 1.96.

Le traitement des strates avec un unique élément varie selon les logiciels. C'est un élément important car l'estimateur de la variance est peu stable lorsque les strates contiennent peu d'éléments. Par défaut, le logiciel SAS regroupe les strates avec une unique unité. Si plusieurs strates sont dans ce cas, elles sont regroupées dans une nouvelle strate. Si une seule strate est dans ce cas, elle est combinée avec la strate suivante. L'option NOCOLLAPSE permet d'éviter le regroupement et de fixer la variance à 0 dans ces strates (moins conservateur). Wesvar a un masque qui permet d'enregistrer la liste des strates avec un unique élément. Elles sont alors considérées comme "self-representing" de variance nulle. Wesvar ne propose pas de regroupement automatique. Au besoin, il peut être fait en dehors du logiciel. Dans les macros de M. Graf, la variance des strates avec un unique élément est imputée sur la base de la variance d'autres strates. Pour ce rapport, la macro est utilisée avec une imputation globale, sans traitement différent selon des groupes de strates.

6.6 Exemple sur les données de l'ERC98

L'effet des poids et du plan sur des moyennes et des médianes est illustré sur les données du module "Conditions de vie" de l'ERC98.

6.6.1 But et données

On désire estimer la moyenne et la médiane du revenu mensuel total du ménage des personnes de 15 ans ou plus en Suisse. L'unité d'analyse est donc la personne de 15 ans ou plus, et la variable d'intérêt est le revenu mensuel total de son ménage (variable EREVTOT).

Les informations sur l'échantillonnage sont les strates (variable XSCHICHT) et les poids finaux (variable XGEWZP). Il n'y a pas de grappes.

L'échantillon comporte $n = 4'369$ personnes avec un poids w_i =XGEWZP prenant des valeurs entre 210 et 4'510 et un revenu mensuel du ménage y =EREVTOT de moyenne 8521, voir Table 6.1.

TAB. 6.1 – Quelques statistiques sur le module "Conditions de vie" de l'ERC98 : variables XGEWZP et EREVTOT par strate XSCHICHT.

		XGEWZP				
XSCHICHT	n	mean	std	min	median	max
1.Esp. Mittelland	1036	1312.9	657.1	373.0	1135.9	4510.0
2.Région lémanique	702	1500.2	712.7	421.5	1378.9	4510.0
3.Zürich	776	1291.4	653.0	403.6	1160.9	4510.0
4.Nordwestschweiz	605	1351.1	715.3	424.0	1123.5	4510.0
5.Ostschweiz	554	1511.6	790.5	494.1	1284.4	4510.0
6.Zentral-CH	378	1429.5	707.0	460.8	1254.7	4497.2
7.Ticino	318	834.1	508.8	210.1	680.2	3581.6
Total	4369	1344.9	706.4	210.1	1170.1	4510.0
		EREVTOT				
XSCHICHT	n	mean	std	median		
1.Esp. Mittelland	1036	8121.7	4680.3	7357.5		
2.Région lémanique	702	8461.7	6023.6	7305.3		
3.Zürich	776	9342.0	7623.7	7697.8		
4.Nordwestschweiz	605	9047.6	6036.6	7978.3		
5.Ostschweiz	554	7701.2	3831.1	7029.2		
6.Zentral-CH	378	8957.4	7280.4	7616.9		
7.Ticino	318	7859.8	4817.3	7077.7		
Total	4369	8521.2	5900.1	7469.5		

6.6.2 Résultats pour la moyenne

La moyenne de la variable y =EVETOT varie nettement entre la valeur non pondérée $\widehat{\overline{y}}_{nonpond} = 8521.2$ et pondérée $\widehat{\overline{y}}_{nonpond} = 8835.8$; voir Table 6.2. L'estimation non pondérée est sous-estimée par rapport à l'estimation pondérée.

La variabilité de la moyenne augmente dès que l'on tient compte des poids et surtout du plan d'échantillonnage. L'écart-type passe de 89.3 à 93.7 si on intègre les poids, puis augmente à 109.0 en prenant en compte le plan. Des intervalles de confiance basés sur l'écart-type sans

poids ni plan sont donc sous-estimés par rapport à ceux basés sur l'estimation complète avec poids et plan.

TAB. 6.2 – Moyennes sans pondération $\widehat{\overline{y}}_{nonpond}$ et avec pondération $\widehat{\overline{y}}_{pond}$, avec l'écart-type classique sans pondération $\widehat{Std}_{SIR}(\widehat{\overline{y}}_{nonpond})$, avec pondération uniquement $\widehat{Std}_{SIR}(\widehat{\overline{y}}_{pond})$ et avec pondération et plan $\widehat{Std}_{p}(\widehat{\overline{y}}_{pond})$.

$\widehat{\overline{y}}_{nonpond}$	$\widehat{\overline{y}}_{pond}$	$\widehat{Std}_{SIR}(\widehat{\overline{y}}_{iid})$	$\widehat{Std}_{SIR}(\widehat{\overline{y}}_{pond})$	$\widehat{Std}_{p}(\widehat{\overline{y}}_{pond})$
8521.2	8835.8	89.3	93.7	109.0

PROC UNIVARIATE est appliquée pour l'estimation de $\widehat{Std}_{SIR}(\widehat{\overline{y}}_{nonpond})$ et $\widehat{Std}_{SIR}(\widehat{\overline{y}}_{pond})$; voir les calculs dans l'Annexe A.1.

PROC SURVEYMEANS est appliquée pour l'estimation de $\widehat{Std}_{p}(\widehat{\overline{y}}_{pond})$ sans l'option TOTAL car le facteur de correction pour la population finie est négligeable dans notre cas.

L'estimation obtenue par Wesvar (Jacknife) est égale à 109.1. Elle est donc très proche de la valeur 109.0 de SAS (linéarisation) ; voir les informations générales sur Wesvar dans l'Annexe B.

L'approche naïve, sans poids ni plan, donne un estimateur biaisé de la moyenne ($BTR(\widehat{\overline{y}}) = -3.6\%$) avec une erreur quadratique technique nettement moins bonne que l'approche complète ; voir Table 6.3. L'approche incomplète, avec les poids mais pas le plan, donne une MSET inférieure à celle qui tient compte des poids et du plan mais sa composante de variance est biaisée. L'effet du plan *deft* = 1.16 domine l'effet des poids *weft* = 1.05 sur la variabilité. Un intervalle de confiance construit avec la méthode complète est 22% plus grand que celui construit avec la méthode naïve (*ceft* $= 1.22$).

TAB. 6.3 – Effet des poids et du plan sur l'estimation de la moyenne.

Biais technique	$BT(\widehat{\overline{y}})$ = - 314.6
	$BTR(\widehat{\overline{y}})$ = -3.6%
Effet sur la variance	*weft* = 1.05
	deft = 1.16
	ceft = 1.22
MSE technique	$MSET_{nonpond,SIR}(\widehat{\overline{y}})$ = 106'941
	$MSET_{pond,SIR}(\widehat{\overline{y}})$ = 8'778
	$MSET_{pond,p}(\widehat{\overline{y}})$ = 11'888

6.6.3 Résultats pour la médiane

La médiane de y =EVETOT est également influencée par la pondération et le plan d'échantillonnage ; voir Table 6.4.

La variabilité de l'estimateur augmente de 84.4 à 96.2 dès que l'on prend en compte les poids puis diminue à 87.0 lorsque les poids et le plan sont intégrés. Il faut cependant noter que les intervalles de confiance déterminés pour $\widehat{Std}_{SIR}(\widehat{med}_{nonpond})$ et $\widehat{Std}_{SIR}(\widehat{med}_{pond})$ sont asymétriques ; voir calculs dans les Annexes A.2 et A.3. Les écarts-types de la Table 6.4, définis comme le plus grand demi intervalle divisé par 1.96, sont donc plutôt conservateurs et pas très fiables.

TAB. 6.4 – Médianes sans pondération $\widehat{med}_{nonpond}$ et avec pondération $\widehat{med}_{pond}$, avec l'écart-type classique sans pondération $\widehat{Std}_{SIR}(\widehat{med}_{nonpond})$, avec pondération mais pas le plan $\widehat{Std}_{SIR}(\widehat{med}_{pond})$ et avec pondération et plan $\widehat{Std}_{p}(\widehat{med}_{pond})$ (dans Wesvar).

$\widehat{med}_{nonpond}$	$\widehat{med}_{pond}$	$\widehat{Std}_{SIR}(\widehat{med}_{nonpond})$	$\widehat{Std}_{SIR}(\widehat{med}_{pond})$	$\widehat{Std}_{p}(\widehat{med}_{pond})$
7469.5	7720.6	84.4	96.2	87.0

L'approche complète présente ici la plus petite erreur quadratique technique ; voir Table 6.5. Les valeurs de *weft* et *deft*, et par conséquent de *ceft*, sont plutôt conservatrices (intervalles asymétriques). L'effet du plan *deft* est inférieur à 1. Cela est probablement dû à l'effet bénéfique de la stratification, en combinaison avec d'autres effets liés aux poids et au plan. L'effet global *ceft* reste inférieur à celui sur la moyenne.

TAB. 6.5 – Effet des poids et du plan sur l'estimation de la médiane.

Biais technique	$BT(\widehat{med})$ = - 251.1
	$BTR(\widehat{med})$ = -3.3%
Effet sur la variance	*weft* = 1.14
	deft = 0.90
	ceft = 1.03
MSE technique	$MSET_{nonpond,SIR}(\widehat{med})$ = 70'201
	$MSET_{pond,SIR}(\widehat{med})$ = 9'256
	$MSET_{pond,p}(\widehat{med})$ = 7'574

6.6.4 Interprétation des résultats

Le biais dans les estimations ponctuelles et la sous-estimation de la variance de la méthode naïve peuvent conduire à des interprétations erronées dans les analyses. Une différence peut par exemple être significative selon la méthode naïve alors qu'elle ne l'est pas si on prend en compte tous les éléments (poids et plan).

Comme illustration, on peut noter que les intervalles de confiance du revenu mensuel des ménage des hommes et des femmes ne se recoupent pas dans le cas non pondéré et sans tenir compte du plan ; voir IC_{SIR} dans la Table 6.6. Par contre, ils se recoupent lorsque l'on prend

en compte les poids et le plan ; voir IC_p. Notons cependant que l'écart-type de la différence n'a pas été calculé. Seul ce dernier, dépendant aussi de la covariance entre les deux estimateurs, pourrait permettre de conclure en une différence significative ou pas entre hommes et femmes.

TAB. 6.6 – Moyennes sans pondération $\widehat{\overline{y}}_{nonpond}$ et avec pondération $\widehat{\overline{y}}_{pond}$, avec l'écart-type classique sans pondération $\widehat{Std}_{SIR}(\widehat{\overline{y}}_{nonpond})$ et avec pondération et plan $\widehat{Std}_p(\widehat{\overline{y}}_{pond})$ (dans SAS avec l'option DOMAIN). On note également les intervalles de confiance estimés IC_{SIR} et IC_p, respectivement.

Sexe	n	$\widehat{\overline{y}}_{nonpond}$	$\widehat{Std}_{SIR}(\widehat{\overline{y}}_{nonpond})$	IC_{SIR}
Hommes	2009	8809.5	122.5	[8564 ; 9055]
Femmes	2360	8275.8	128.0	[8020 ; 8532]
Sexe	n	$\widehat{\overline{y}}_{pond}$	$\widehat{Std}_p(\widehat{\overline{y}}_{pond})$	IC_p
Hommes	2009	9100.6	142.9	[8820 ; 9381]
Femmes	2360	8589.2	163.5	[8269 ; 8910]

6.7 Exemple sur les données de la LSE98

L'effet des poids et du plan d'échantillonnage sur les moyennes, les médianes et leur variance est également illustré sur les données de la LSE98.

Notons que les résultats de la LSE sont diffusés sous forme de médianes et non pas de moyennes. Les calculs sur les moyennes présentés ci-dessous permettent d'illustrer les méthodes. Sur le plan thématique, les résultats numériques sur les moyennes sont cependant moins pertinents que ceux sur les médianes.

6.7.1 But et données

On désire estimer la moyenne et la médiane du salaire mensuel brut standardisé des postes du secteur privé (variable `MBLS`).

Les informations sur l'échantillonnage sont les grappes (entreprises, variable `BURNR_N`), et les strates (variable `STRA`). Le poids final de chaque emploi est donné dans la variable `GEWICHT`.

Dans le cas de la LSE, on est intéressé par le salaire `MBLS` ajusté par le taux d'occupation standardisé `IBGRS`. Les estimations présentées dans le présent rapport prennent donc en compte la standardisation `IBGRS` pour les résultats non pondérés et le poids `GEWIBGRS=GEWICHT*IBGRS` pour les estimateurs pondérés. La variable `IBGRS` n'étant pas disponible dans les données reçues de la section LOHN, elle est construite comme suit : `IBGRS=GEWIBGRS/GEWICHT`.

On retire du jeu de données les 893 emplois avec un poids nul car ils sont dans des entreprises avec un unique employé (hors population selon les données collectées). On retire également les 23'550 emplois de la strate `STRA`=-9 et des strates du secteur primaire (`NOGA2 in ('01',`

'02')). Les emplois dans la strate STRA=-9 sont par ex. dans des entreprises de télécommunication publique au moment du tirage et privée pour exploitation.

Le jeu de données final contient 415'787 emplois dans 7006 entreprises (grappes) et 200 strates du secteur privé. Le salaire standardisé MBLS a un coefficient de variation de std/mean=51%, le poids de standardisation IBGRS a des valeurs entre 0 et 2.15 et le poids complet GEWIBGRS entre 0.01 et 96.82 ; voir Table 6.7.

TAB. 6.7 – Quelques statistiques sur les données de la LSE98.

Variable	n	mean	std	min	median	max
MBLS	415787	5866.8	3005.4		5206	
IBGRS	415787	0.93	0.24	0.005	1.02	2.15
GEWIBGRS	415787	4.97	9.97	0.01	1.33	96.82

Le nombre BURNH d'entreprises par strate, dans le cadre de sondage, a été fourni dans un fichier séparé car il n'est pas disponible dans le fichier de base pour les analyses.

Wesvar n'est pas utilisé ici car le fichier, et notamment le nombre de grappes, est trop gros pour faire des estimations sans préparatifs préliminaires (choix des replicates, etc.).

Les résultats numériques présentés ci-dessous sont données pour la région lémanique. Cette région est définie par l'ensemble des cantons de Vaud, Genève et Valais (critère : ARBKTO in ('VD','GE','VS')).

6.7.2 Premiers résultats

Le salaire moyen des hommes et des femmes travaillant dans la région lémanique varie sensiblement entre valeurs non pondérées et pondérées ; voir Table 6.8.

La moyenne pour les hommes diminue lorsque l'on prend en compte les poids. Le comportement est inverse pour les femmes. La différence entre les salaires moyens des hommes et des femmes diminue donc de 1979 à 1525 (-23%) lorsque l'on passe d'une estimation non pondérée à une estimation pondérée.

Le biais technique des hommes (384, *i.e.* 6%) est supérieur à celui des femmes (-97, *i.e.* -2%). Cette différence est partiellement expliquée par une corrélation négative entre le poids et la variable MBLS ; corrélation légèrement plus forte pour les hommes que les femmes.

Il est possible d'obtenir différentes estimations de la variabilité de l'estimation de la moyenne des salaires. Les six estimations de la Table 6.8 sont définies comme suit, de la moins complète à la plus complète :

– $\widehat{Std}(1) = \widehat{Std}_{SIR}(\widehat{\overline{y}}_{nonpond})$: estimation naïve avec PROC UNIVARIATE sans tenir compte des poids d'échantillonnage et du plan ;
– $\widehat{Std}(2) = \widehat{Std}_{SIR}(\widehat{\overline{y}}_{pond})$: estimation avec PROC UNIVARIATE en tenant compte des poids mais pas du plan ;
– $\widehat{Std}(3)$: estimation avec PROC SURVEYMEANS en tenant compte des poids et du plan mais sans tenir compte de la correction pour la population finie et en considérant les sous-

TAB. 6.8 – Moyennes sans pondération $\widehat{\overline{y}}_{nonpond}$ et avec pondération $\widehat{\overline{y}}_{pond}$ ainsi que 6 estimations de l'écart-type. La valeur de $\widehat{Std}(1) = \widehat{Std}_{SIR}(\widehat{\overline{y}}_{nonpond})$ est la moins complète. Celles de $\widehat{Std}(5) = \widehat{Std1}_p(\widehat{\overline{y}}_{pond})$ et $\widehat{Std}(6) = \widehat{Std2}_p(\widehat{\overline{y}}_{pond})$ sont les plus complètes, voir texte.

	n	$\widehat{\overline{y}}_{nonpond}$	$\widehat{\overline{y}}_{pond}$	$\widehat{Std}(1)$	$\widehat{Std}(2)$	$\widehat{Std}(3)$	$\widehat{Std}(4)$	$\widehat{Std}(5)$	$\widehat{Std}(6)$
Hommes	41674	6796.8	6412.5	19.1	19.3	124.3	133.6	100.1	101.2
Femmes	31148	4790.6	4887.6	11.5	12.2	70.3	76.7	60.7	61.6

groupes "hommes région lémanique" et "femmes région lémanique" comme des échantillons indépendants (option BY) ;

– $\widehat{Std}(4)$: estimation identique à $\widehat{Std}(3)$ mais en considérant les sous-groupes comme des domaines (option DOMAIN) ;
– $\widehat{Std}(5) = \widehat{Std1}_p(\widehat{\overline{y}}_{pond})$: estimation complète avec PROC SURVEYMEANS tenant compte des poids et du plan, ainsi que de la correction pour la population finie et du fait que les sous-groupes sont des domaines ; voir Annexe A.4.
– $\widehat{Std}(6) = \widehat{Std2}_p(\widehat{\overline{y}}_{pond})$: estimation complète avec la macro de M. Graf qui prend en compte les deux niveaux d'échantillonnage, les poids, le plan, la correction pour la population finie et le fait que les sous-groupes sont des domaines.

L'effet du passage de l'écart-type sans poids ni plan $\widehat{Std}(1) = \widehat{Std}_{SIR}(\widehat{\overline{y}}_{iid})$ à celui avec les poids mais pas de plan $\widehat{Std}(2) = \widehat{Std}_{SIR}(\widehat{\overline{y}}_{pond})$ est faible. Le facteur *weft* est égal à 1.01 pour les hommes et 1.06 pour les femmes.

Le plus grand effet est dans la prise en compte du plan d'échantillonnage, *i.e.* des grappes et de la variabilité des poids, avec des *deft* égaux à 5.18 pour les hommes et 4.98 pour les femmes, pour le passage de $\widehat{Std}(2) = \widehat{Std}_{SIR}(\widehat{\overline{y}}_{pond})$ à $\widehat{Std}(5) = \widehat{Std1}_p(\widehat{\overline{y}}_{pond})$.

On note aussi la croissance sensible de la variabilité lors du passage de l'estimation des sous-groupes comme des échantillons indépendants $\widehat{Std}(3)$ (option BY) à l'estimation en tenant compte du fait que ce sont des domaines $\widehat{Std}(4)$ (option DOMAIN) (facteur 1.07 pour les hommes, respectivement 1.09 pour les femmes).

La prise en compte du facteur de correction de population finie a un apport positif sur la variance ; facteur 0.75, respectivement 0.79 entre $\widehat{Std}(4)$ et $\widehat{Std}(5) = \widehat{Std1}_p(\widehat{\overline{y}}_{pond})$. Nous sommes en effet en présence d'un échantillon avec un taux de sondage égal ou proche de 1 dans plusieurs strates.

Le passage de la méthode naïve $\widehat{Std}(1) = \widehat{Std}_{SIR}(\widehat{\overline{y}}_{nonpond})$ à la méthode complète $\widehat{Std}(5) = \widehat{Std1}_p(\widehat{\overline{y}}_{pond})$ conduit à une augmentation sensible de la variabilité de l'estimateur. Les intervalles de confiance sont multipliés par un facteur supérieur à 5 dû au tirage en grappe et à la variabilité des poids (*ceft* $= 5.23$ pour les hommes et *ceft* $= 5.30$ pour les femmes).

Les deux estimations complètes $\widehat{Std}(5) = \widehat{Std1}_p(\widehat{\overline{y}}_{pond})$ et $\widehat{Std}(6) = \widehat{Std2}_p(\widehat{\overline{y}}_{pond})$ sont proches. L'approximation faite dans SAS est donc satisfaisante même si elle ne prend pas en compte les deux niveaux de l'échantillonnage en grappe.

La valeur de la médiane est également modifiée par les poids, voir la Table 6.9. Comme pour les moyennes, la valeur diminue pour les hommes et augmente pour les femmes. Le biais technique, 10% pour les hommes et -4% pour les femmes, est plus important que dans le cas de la moyenne.

TAB. 6.9 – Médianes sans pondération $\widehat{med}_{nonpond}$ et avec pondération $\widehat{med}_{pond}$ ainsi que les estimations de l'écart-type naïve $\widehat{Std}_{iid}(\widehat{med}_{nonpond})$, avec les poids mais pas le plan $\widehat{Std}_{SIR}(\widehat{med}_{pond})$ et complète $\widehat{Std}_{p}(\widehat{med}_{pond})$ avec la macro, voir texte.

	$\widehat{med}_{nonpond}$	$\widehat{med}_{pond}$	$\widehat{Std}_{SIR}(\widehat{med}_{nonpond})$	$\widehat{Std}_{SIR}(\widehat{med}_{pond})$	$\widehat{Std}_{p}(med_{pond})$
Hommes	5911	5375	14.8	28.1	71.4
Femmes	4224	4411	13.3	33.2	64.8

La procédure UNIVARIATE ne permet pas d'obtenir un intervalle de confiance estimé de la médiane en présence de poids. Nous utilisons une macro SAS pour estimer $\widehat{Std}_{SIR}(\widehat{med}_{nonpond})$ avec le poids `IBGRS` et $\widehat{Std}_{SIR}(\widehat{med}_{pond})$ avec le poids `GEWIRBGRS` ; voir Annexe A.3.

Le passage de la méthode naïve à la méthode complète conduit à une augmentation de l'écart-type de la médiane par un facteur *ceft* $= 4.83$ pour les hommes et *ceft* $= 4.88$ pour les femmes. L'effet des poids et celui du plan sont importants (*weft* $= 1.90$ pour les hommes et 2.5 pour les femmes ; *deft* $= 2.55$ pour les hommes et 1.95 pour les femmes). Il faut cependant noter que l'estimation de l'écart-type avec poids (en particulier sans plan) est plutôt conservatrice car les intervalles de Woodruff sont asymétriques.

L'erreur quadratique moyenne pour la méthode naïve ($MSET_{nonpond,SIR}(\widehat{med}) = 285'715$ pour les hommes, 35'145 pour les femmes) est très supérieure à celle de la méthode complète ($MSET_{pond,p}(\widehat{med}) = 5'102$ pour les hommes et 4'199 pour les femmes). Il est donc important de choisir la méthode complète.

La figure 6.1 résume l'effet des poids et du plan sur les estimations de la moyenne et de la médiane.

6.7.3 Résultats sur différents domaines

Le salaire moyen et médian des hommes et des femmes travaillant dans la région lémanique est étudié par classes NOGA afin de voir l'effet de la prise en compte des poids et du plan dans différents domaines.

Les groupes de classes NOGA sont repris de la construction du plan d'échantillonnage ; voir Peters (1997). Ci-dessous, le code `NOGA`=23 comporte les classes 23 et 24 et le code `NOGA`=29 comporte les classes 29, 34 et 35.

Les résultats sont comparés dans les sous-ensembles (domaines = NOGA) ayant plus de 6 entreprises et dont la différence entre le nombre d'emplois et le nombre d'entreprises est supérieur à 54[1]. En plus de limiter les problèmes de protection des données, elles permettent d'éviter les

[1]Règles de publication utilisées dans le cadre de la LSE pour les résultats au niveau Suisse dans le secteur

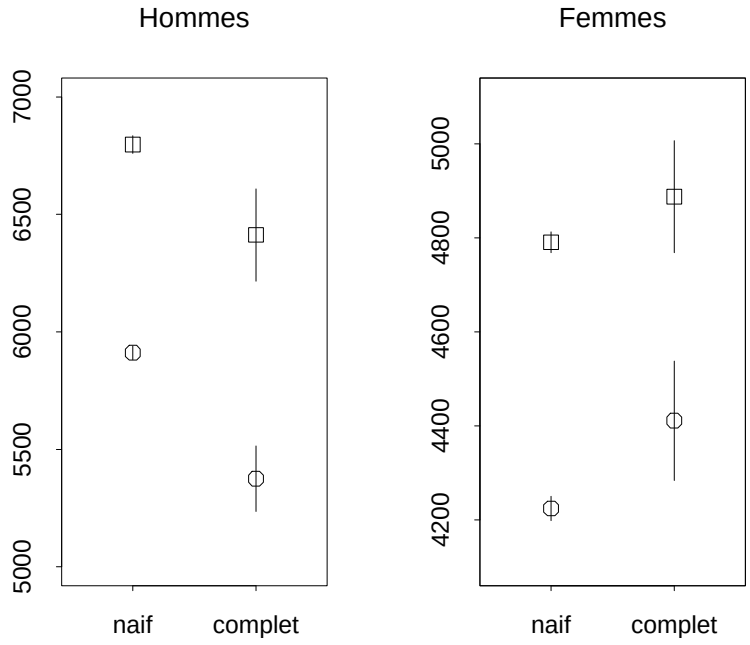


FIG. 6.1 – Salaire moyen (carré) et médian (rond) pour les hommes (gauche) et les femmes (droite). Valeur non pondérée avec intervalle basé sur la méthode naïve et valeur pondérée avec intervalle de confiance basé sur une méthode complète.

strates avec peu d'entreprises dans les calculs de variance. On exclu donc les classes NOGA $\in \{10, 16-21, 25, 26, 36, 61, 64, 75, 90\}$ pour atteindre 27 domaines chez les femmes et les hommes. Il est à noter que certains de ces domaines restent petits et que les estimations peuvent donc être peu fiables.

Les différences entre les moyennes, respectivement les médianes, non pondérées et pondérées sont sensibles ; voir Figure 6.2. Le biais technique relatif des moyennes varie de -18 à 49% pour les hommes, et de -10 à 18% pour les femmes. Les valeurs correspondantes pour la médiane sont -13% à 67% pour les hommes et -13% à 24% pour les femmes.

Il est intéressant de noter que plus la corrélation ρ entre MBLS et GEWICHT est grande, plus le biais relatif *BTR* est petit. La corrélation entre ρ et *BTR* est égale à -0.8. Dans le cas des hommes de la classe NOGA=15, les biais relatifs de 46% pour la moyenne et de 67% pour la médiane correspondent à $\rho = -0.2$. Dans le cas des hommes de la NOGA=65, les biais relatifs de -18% pour la moyenne et -13% pour la médiane correspondent à $\rho = 0.2$. Un modèle permettant d'estimer le biais à partir de la relation entre la variable d'intérêt et le poids est cependant difficile à déterminer car les valeurs observées sont très variables.

L'effet des poids et du plan sur l'estimateur de la variance (complet avec macro) de la moyenne varie fortement entre les différents domaines, ; voir Table 6.10. En moyenne, l'effet des poids reste assez faible (moyenne : *weft* $= 1.02$), mais l'effet du plan est important (moyenne : *deft* $=$

privé : au moins 5 entreprises et au moins 60 personnes.

3, maximum 8). Deux *ceft* sont légèrement inférieurs à 1 (0.92). Il s'agit des classes NOGA 67 et 93 pour lesquels *weft* sont égaux à 0.92, resp. 0.8 et les *deft* sont supérieurs à 1.

TAB. 6.10 – Effet des poids et du plan sur la variance de la moyenne dans les 54 domaines NOGA2 x sexe de la région lémanique.

	n	mean	cv	min	median	max
weft	54	1.02	17.51	0.62	0.98	1.62
deft	54	3.03	50.03	0.99	2.54	8.11
ceft	54	3.18	59.13	0.92	2.72	9.76

Il est intéressant de noter que, dans 18 domaines sur 54, l'erreur quadratique moyenne technique est meilleure pour l'estimateur sans poids ni plan que pour l'estimateur avec poids et plan. Il s'agit de cas avec un biais technique faible et un fort effet du plan d'échantillonnage. Par exemple, dans la classe NOGA=23, $MSET_{nonpond,SIR}(\widehat{\overline{y}}) = 35'333$ et $MSET_{pond,p}(\widehat{\overline{y}}) = 133'450$ avec $BT(\widehat{\overline{y}}) = -184$ (-2.4%) et $deft = 8.1$ ($ceft = 9.76$).

Les résultats obtenus pour l'écart-type de la moyenne avec la procédure SURVEYMEANS de SAS (un niveau) sont en moyenne 4% inférieures à ceux obtenus avec la macro de M. Graf (deux niveaux) ; voir graphique gauche de la Figure 6.3. Dans 8 cas sur 54 la différence relative est supérieure à 10%.

La différence extrême de la classe NOGA 29 pour les femmes (point nettement sous la droite 1-1) est un cas difficile à traiter car il n'y a que 101 employés dans 8 entreprises et 4 strates. Le choix du traitement des strates avec une seule entreprise (collapse dans SURVEYMEANS, imputation d'une variance dans la macro) ou encore l'utilisation d'un ou de deux niveaux ont alors des influences importantes sur les résultats. La procédure SURVEYMEANS, qui ne considère que le premier niveau pour les estimations donne ici très probablement une valeur sous-estimée.

L'écart-type de la moyenne et celui de la médiane ne sont pas bien corrélés, voir graphique droite de la Figure 6.3. Le lien théorique $Std(med) = \sqrt{(\pi/2)}Std(moy)$ n'est manifestement pas utilisable pour estimer la variabilité de la médiane à partir de celle de la moyenne.

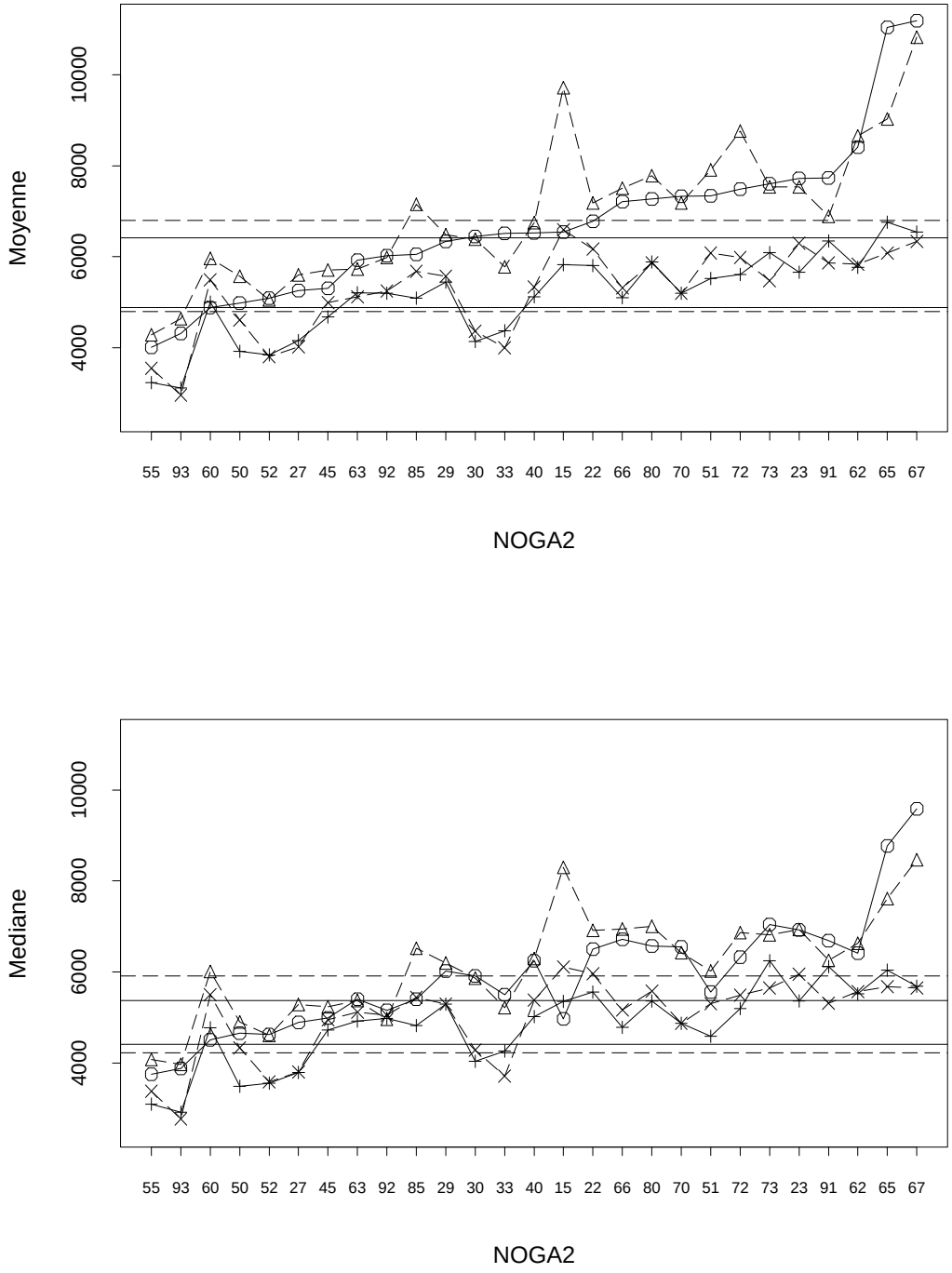


FIG. 6.2 – Pour 27 classes NOGA triées selon le salaire moyen pondéré des hommes : moyennes (graphe du haut) et médianes (graphe du bas) non pondérées (lignes traitillées) et pondérées (lignes continues). Valeurs pour les hommes (○ et △) et pour les femmes (+ et ×). Les lignes horizontales indiquent les valeurs globales (les deux valeurs pour les hommes sont toujours au dessus de celles des femmes).

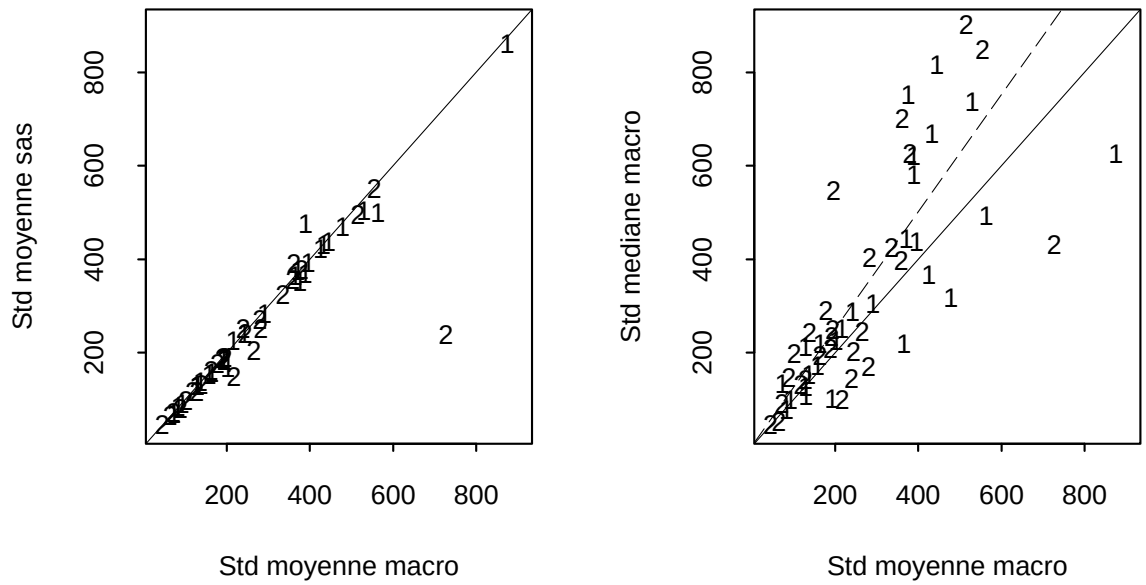


FIG. 6.3 – Relation, pour 27 classes NOGA, entre l'écart-type de la moyenne obtenu avec la macro et avec la procédure SURVEYMEANS (gauche) et entre l'écart-type de la moyenne et de la médiane avec la macro ; pour les hommes (1) et les femmes (2). La droite au dessus de la diagonale du graphique de droite indique la relation théorique sous normalité $std(med) = \sqrt{(\pi/2)}std(moy)$.

Chapitre 7

Modèles de régression linéaire

Les modèles de régression linéaire sont couramment utilisés pour étudier la relation entre une variable dépendante continue et des variables explicatives continues relevées sur les mêmes observations.

Les autres méthodes multivariées telles que la régression logistique ou même l'analyse en composante principale ne sont pas abordées ici. Quelques informations peuvent être trouvées par exemple dans Lohr (1999) ou des articles comme Skinner et al. (1986).

7.1 Estimateurs classiques

Soit un ensemble de n observations pour lesquelles on a relevé une variable dépendante y et une série de variables explicatives $x_1, .., x_p$.

Dans un modèle de régression linéaire, la variable y est exprimée comme une combinaison linéaire des p variables $x_1, .., x_p$:

$$\mathbf{y} = \mathbf{X}\beta + \epsilon \tag{7.1}$$

avec $\mathbf{y}$ le vecteur de taille n des observations relatives à la variable y, $\mathbf{X}$ la matrice $n \times (p+1)$ des observations relatives aux variables $x_1, .., x_p$ avec une première colonne $\mathbf{1}$ pour la constante du modèle (intercept), et ϵ le vecteur de taille n des erreurs aléatoires.

La valeur du vecteur $\beta = (b_0, b_1, .., b_p)'$ de taille $p+1$ est estimée, sans biais, par la méthode des moindres-carrés :

$$\widehat{\beta}_{iid} = (\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{y} \tag{7.2}$$

$$= (\sum_{i=1}^{n} \mathbf{x_i}\mathbf{x_i'})^{-1} \sum_{i=1}^{n} \mathbf{x_i} y_i \tag{7.3}$$

avec $\mathbf{x_i} = (1, x_{1i}, .., x_{pi})'$ le vecteur de taille $p+1$ des valeurs pour l'observation $i = 1, .., n$.

Si les résidus ϵ sont identiquement et indépendamment distribuées selon la loi $N(0, \sigma^2)$, les coefficients formant le vecteur $\widehat{\beta}_{iid}$ suivent une loi normale de variance $V_{iid}(\widehat{\beta}_{iid}) = (\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\sigma^2$. Le niveau de signification des coefficients b_r est donc testé à l'aide de $t = b_r / Std(b_r)$, $r =$

$1, .., p$, qui suit une loi t de Student avec $n-p-1$ degrés de libertés. L'ensemble des coefficients est testé avec une loi $F_{p,n-p-1}$.

L'homoscédasticité des résidus (variance constante) est contrôlée à l'aide de tests spécifiques ou empiriquement par la représentation graphique du résidu en fonction des variables explicatives et de la variable réponse. La non-corrélation des résidus peut de son côté être contrôlée à l'aide du test de Durbin-Watson.

Durant la procédure menant au choix du modèle, on évite de conserver des coefficients non significatifs, notamment pour limiter la variabilité des prédictions faites avec le modèle.

Divers critères permettent de comparer la qualité des modèles. On retiendra par exemple la racine de l'erreur quadratique moyenne (ci-dessous : root MSE) ou le coefficient de corrélation (ci-dessous : R-square), ajusté si le nombre de variables est important et le nombre d'observations petit.

Une fois que le modèle est choisi, que la normalité des résidus est contrôlée et que les coefficients ont été déterminés, le modèle peut être utilisé pour prédire une valeur y_0 pour une nouvelle observation $\mathbf{x_0}$.

La prévision est donnée par $y_0 = \mathbf{x_0'}\widehat{\beta}_{iid}$, avec une variance $V(y_0) = \sigma^2\mathbf{x_0'}(\mathbf{X'X})^{-\mathbf{1}}\mathbf{x_0}$ pour la valeur moyenne $E(y|\mathbf{x} = \mathbf{x_0})$ et $V(y_0) = \sigma^2(1 + \mathbf{x_0'}(\mathbf{X'X})^{-\mathbf{1}}\mathbf{x_0})$ pour une valeur unique y. La théorie générale sur la régression linéaire peut par exemple être lue dans Saporta (1990). Une approche très pratique se trouve également dans Rubinfeld (2000).

7.2 Estimateurs avec poids et plan

Les poids w_i associés aux éléments $i = 1, .., n$ de l'échantillon s et le plan de sondage $p(s)$ utilisé pour le tirage ont une influence non seulement sur les statistiques descriptives telles que les moyennes et les médianes, mais également sur l'analyse des données telle que la régression linéaire.

Les approches divergent dans l'utilisation ou pas des poids dans l'estimation des coefficients d'un modèle de régression, voir Section 3.1. Dans le cas de la statistique publique, on considère que les poids et le plan doivent être pris en compte ; ce choix pouvant éventuellement être remis en question dans le cas de très petits échantillons.

On définit β_{pond}, le coefficient pondéré, dont l'estimateur est :

$$\widehat{\beta}_{pond} = (\mathbf{X'WX})^{-\mathbf{1}}\mathbf{X'Wy} \tag{7.4}$$

$$= (\sum_{i=1}^{n} \mathbf{x_i}\mathbf{x_i}'w_i)^{-1} \sum_{i=1}^{n} \mathbf{x_i}y_i w_i \tag{7.5}$$

avec W la matrice diagonale des poids w_i, $i = 1, .., n$.

Comme dans le cas des statistiques descriptives, l'estimateur de la variance de $\widehat{\beta}_{pond}$ doit tenir compte des poids et du plan.

L'estimateur $\widehat{\beta}_{pond}$ est une fonction non linéaire de totaux. Sa variance peut être estimée par une méthode de linéarisation ou de rééchantillonnage, voir par exemple Särndal et al. (1992) pour

le cas général et Lohr (1999) pour un développement dans le cas d'un unique coefficient. Une fois que la variance des coefficients est disponible, on peut déterminer leur niveau de signification avec la loi de Student. Des méthodes alternatives telles que l'utilisation de l'inégalité de Bonferroni sont également disponibles pour les tests mais elles sont moins pratiques ; voir par exemple Korn and Graubard (1990).

La prévision d'une valeur pour un nouveau x_0 est donnée par $y_0 = \mathbf{x_0'}\widehat{\beta}_{pond}$, avec la variance $V(y_0) = \mathbf{x_0'}\widehat{\mathbf{\Sigma}}\mathbf{x_0}$ pour la valeur moyenne $E(y|\mathbf{x} = \mathbf{x_0})$. La matrice variance-covariance $\widehat{\mathbf{\Sigma}}$ est estimée sur la base des données, avec les poids et le plan.

En général, l'effet du plan sur la variance des coefficients de régression est inférieur à celui sur la variance des moyennes car la variabilité est en partie expliquée dans la variable y. L'estimation de la variance $\widehat{V}_p(\widehat{\beta}_{pond}) = \widehat{V}_{SIR}(\widehat{\beta}_{pond})\widehat{\textit{deft}}$, avec $\widehat{\textit{deft}}$ l'effet du sondage estimé pour la moyenne de la variable concernée, devrait donc fournir en principe une estimation plutôt conservatrice de la variance (Skinner et al.,1989).

Notons qu'une régression avec des poids telle que décrite ci-dessus ne correspond pas à une régression selon les moindres-carrés pondérés qui tient compte d'une variance variable. Cette dernière prend en compte la variabilité des données et non pas celle des poids (Lohr,1999).

Dans les comparaisons, nous noterons comme suit les coefficients non pondérés et sans tenir compte du plan : $\widehat{\beta}_{nonpond} = \widehat{\beta}_{iid}$ avec la variance $V_{SIR}(\widehat{\beta}_{nonpond}) = V_{iid}(\widehat{\beta}_{iid})$.

7.3 Littérature

Divers auteurs ont comparé les estimations de coefficients de modèles de régression avec et sans poids. Holt et al. (1980) et Nathan and Holt (1980) notent par exemple que la méthode des moindres-carrés ordinaires produit des résultats biaisés dès que l'on a des poids inégaux. Lee et al. (1989) observe peu de différences entre les coefficients simples et pondérées d'un modèle de régression linéaire liant le body mass index avec diverses variables de la NHNES. Les effets de sondage varient entre 0.8 et 2.18. Ils sont donc inférieurs aux valeurs 0.8 à 21 observées sur les moyennes des variables concernées.

7.4 Logiciels

PROC REG de SAS permet de calculer les coefficients simples et pondérés à l'aide des moindre-carrés ordinaires ainsi que leur variabilité. Les estimations de variance ne tiennent pas compte du plan d'échantillonnage.

PROC SURVEYREG de SAS permet de tenir compte des informations sur l'échantillonnage (poids, strates et grappes). Elle calcule les estimateurs pondérés ou non des coefficients et les variances selon la méthode de linéarisation de Taylor. Dans le cas d'un plan d'échantillonnages en grappe, seul le premier niveau est considéré dans l'estimation. Il n'y a pas d'option DOMAIN telle que dans SURVEYMEANS. Seul l'option BY est disponible pour des estimations indépendantes pour des sous-groupes. La correction pour la population finie est prise en compte si les totaux par strates sont donnés dans l'option TOTAL. Le nombre de degrés de libertés pour

les tests est égal par défaut au nombre de grappes auquel on soustrait le nombre de strates (après éventuel regroupement). L'estimateur de variabilité n'est pas donné si l'on intègre l'option TOTAL en combinaison avec l'option BY. Le facteur de correction de la population finie n'a en effet pas de sens si l'on travaille sur des sous-groupes indépendants (de taille inconnue).

Le logiciel WesVar permet également de calculer des estimateurs simples et pondérés des coefficients, avec une variabilité basée par la méthode du Jackknife en tenant compte du plan d'échantillonnage (poids, strates et grappes).

Le traitement des strates avec un unique élément est effectué comme dans le cas des moyennes et médianes, voir Section 6.5.

7.5 Exemple sur les données de la LSE98

Comme illustration de la régression linéaire, nous étudions le lien entre le logarithme du salaire standardisé et différentes variables relevées dans la LSE98.

Notons que les modèles de base utilisés dans cette section sont repris du rapport de Brandenberg (1997). Ce type de modèle est en effet couramment utilisé pour étudier des différences salariales entre groupes de personnes en fonction de divers critères tels que le sexe, les années d'expérience et la formation.

7.5.1 But et données

On désire estimer les coefficients de modèles de régression liant le logarithme du salaire standardisé $\texttt{LNMBLS} = \log(\texttt{MBLS})$ avec le nombre d'années de formation EDU, le nombre d'années d'expérience EXP, le nombre d'années de service dans l'entreprise actuelle TEN et l'expérience professionnelle avant l'entrée dans la présente entreprise EXPR ; voir Table 7.5.1.

Les deux modèles complets suivants servent de base aux analyses présentées ci-dessous :

$$\texttt{LNMBLS} = b_0 + b_1\texttt{EDU1} + b_2\texttt{EDUEXP} + b_3\texttt{EDUEXP2} + \epsilon \qquad \text{(Modèle A)}$$

et

$$\begin{aligned}\texttt{LNMBLS} =& b_0 + b_1\texttt{EDU1} + b_2\texttt{EDUEXPR} + b_3\texttt{EDUEXPR2} + \\ & b_4\texttt{EDUTEN} + b_5\texttt{EDUTEN2} + b_6\texttt{EDUEXPRTEN} + \epsilon\end{aligned} \qquad \text{(Modèle B)}$$

On ne présente ci-dessous que des modèles avec des résidus raisonnables et dont tous les coefficients sont significativement différents de 0 avec un risque de 5%. De plus, on suppose que les variables sont continues.

Les données de base sont les mêmes que celles utilisées pour les estimations des moyennes et médianes ; voir Section 6.7.1 et Table 7.2. On introduit également la standardisation IBGRS dans les estimations qui ne tiennent pas compte des poids d'échantillonnage. Pour les illustrations, on exclu les observations dont les valeurs explicatives sont manquantes.

TAB. 7.1 – Définition des variables de base pour les modèles de régression sur les données de la LSE98. Les années de service DIENSTJA et l'âge ALTER sont relevés dans l'enquête.

Variable	Construction
TEN	$=$ DIENSTJA
EXP	$= \max(0, (\texttt{ALTER} - \texttt{EDU} - 7))$
EXPR	$= \max(0, (\texttt{ALTER} - \texttt{EDU} - \texttt{TEN} - 7))$
EDU	`= .  if AUSBILD=-9/*valeurs manquantes*/`
	`= 17.5  if AUSBILD=1 /*université, haute école*/`
	`= 16  if AUSBILD=2 /*école prof. sup. / HES */`
	`= 15  if AUSBILD=3 /*formation prof. sup. */`
	`= 14  if AUSBILD=4 /*brevet d'enseignement */`
	`= 13  if AUSBILD=5 /*maturité */`
	`= 12  if AUSBILD=6 /*apprentissage complet */`
	`= 9.5  if AUSBILD=7 /*formation en entreprise*/`
	`= 8  if AUSBILD=8 /*scolarité obligatoire */`
	`= .  if AUSBILD=9 /*autres forma. complètes*/`

Variable	Construction	Variable	Construction
EDU1	$=$ EDU$/10$	EDUEXP	$=$ EDU1 $\cdot$ EXP1
EXP1	$=$ EXP$/10$	EDUEXP2	$=$ EDU1 $\cdot$ EXP1^2
TEN1	$=$ TEN$/10$	EDUEXPR	$=$ EDU1 $\cdot$ EXPR1
EXPR1	$=$ EXPR$/10$	EDUTEN	$=$ EDU1 $\cdot$ TEN1
		EDUTEN2	$=$ EDU1 $\cdot$ TEN1^2
		EDUEXPRTEN	$=$ EDU1 $\cdot$ EXPR1 $\cdot$ TEN1

TAB. 7.2 – Quelques statistiques sur les données de la LSE98 utilisées dans les modèles de régression.

Variable	n	nmiss	mean	std	min	median	max
EDU1	415787	92317	1.18	0.25	0.8	1.2	1.75
EXP1	415787	345	1.65	1.35	0.0	1.5	6.30
TEN1	415787	111403	1.07	0.93	0.1	0.8	5.50
EXPR1	415787	24655	1.12	0.95	0.0	0.9	5.80

7.5.2 Premiers résultats

Nous définissons le sous-groupe des femmes de la région lémanique "Lac_femmes" (critère : `GESCHLE=2 and ARBKTO in ('VD','GE','VS')`) et appliquons le modèle A sur les données afin d'estimer le logarithme du salaire `LNMBLS` en fonction du nombre d'années de formation `EDU` et d'expérience `EXP`. Après exclusion des observations avec des valeurs manquantes, "Lac_femmes" comporte 20'607 observations.

Tous les coefficients du modèle A sont significativement différents de 0 (p-value<0.0001)[1] avec la méthode naïve (sans poids ni plan), avec la méthode incomplète (avec poids mais pas le plan) et avec la méthode complète (poids et plan). Le fit du modèle est similaire pour les trois méthodes, avec un root MSE un peu moins bon pour la méthode incomplète : root MSE=0.260 et R-square=0.388 pour la méthode naïve, root MSE=0.674 et R-square=0.401 pour la méthode incomplète et root MSE=0.289 et R-square=0.401 pour la méthode complète.

Comme dans le cas des moyennes et médianes, on observe une modification des estimations ponctuelles suite à l'introduction des poids. Dans le cas du modèle A, les coefficients sont modifiés avec un biais technique relatif prenant des valeurs entre 0.4% et 21.4% ; voir Table 7.3. La variabilité des estimateurs est également modifiée par les poids et le plan d'échantillonnage. L'effet des poids est faible (*weft* $\simeq 1$) mais celui du plan d'échantillonnage est important (*deft* entre 2.4 et 3.3).

TAB. 7.3 – Coefficients et statistiques de test du modèle A pour le jeu de données "Lac_femmes", avec les coefficients sans et avec pondération $\widehat{\beta}_{nonpond}$ et $\widehat{\beta}_{pond}$ et les écarts-types sans poids ni plan $Std_{SIR}(\widehat{\beta}_{nonpond})$, avec poids mais sans plan $Std_{SIR}(pond)$, avec poids et plan avec SAS $Std1_p(\widehat{\beta}_{pond})$, et avec poids et plan avec Wesvar $Std2_p(\widehat{\beta}_{pond})$. Statistiques *weft*, *deft* et *ceft* et t-value sans poids ni plan $t_{SIR}(\widehat{\beta}_{nonpond})$, pondérée sans plan $t_{SIR}(\widehat{\beta}_{pond})$ et avec poids et plan dans SAS $t1_p(\widehat{\beta}_{pond})$.

Variables	$\widehat{\beta}_{nonpond}$	$\widehat{\beta}_{pond}$	$Std_{SIR}(\widehat{\beta}_{nonpond})$	$Std_{SIR}(\widehat{\beta}_{pond})$	$Std1_p(\widehat{\beta}_{pond})$	$Std2_p(\widehat{\beta}_{pond})$
`Intercept`	7.428	7.399	0.0093	0.0095	0.0327	0.0329
`EDU1`	0.636	0.647	0.0087	0.0088	0.0286	0.0288
`EDUEXP`	0.216	0.199	0.0058	0.0058	0.0137	0.0137
`EDUEXP2`	-0.034	-0.028	0.0014	0.0014	0.0034	0.0034

Variables	*weft*	*deft*	*ceft*	$t_{SIR}(\widehat{\beta}_{nonpond})$	$t_{SIR}(\widehat{\beta}_{pond})$	$t1_p(\widehat{\beta}_{pond})$
`Intercept`	1.02	3.45	3.50	795.4	780.0	226.5
`EDU1`	1.00	3.27	3.27	72.8	74.03	22.6
`EDUEXP`	1.00	2.36	2.35	37.1	34.25	14.5
`EDUEXP2`	1.01	2.43	2.45	-25.0	-20.2	-8.3

Les écart-types obtenus par les méthodes complètes linéarisées $Std1_p(\widehat{\beta}_{pond})$ de SAS et Jackknife $Std2_p(\widehat{\beta}_{pond})$ de Wesvar sont presque identiques. Dans les deux cas, nous considérons les données comme des sous-ensembles indépendants et non pas des domaines ; ce qui tend à sous-estimer la variance. Cependant, la correction pour la population finie n'est pas prise en

[1] La valeur absolue de la t-value pour un test de Student à 5% est égale à 1.96 si le nombre de degrés de liberté est très grand.

compte ; ce qui tend à sur-estimer la variance. Nous ne pouvons donc pas dire si l'estimation de la variance est plutôt sur ou sous-estimée.

Dans cet exemple, le fait de regrouper ou non les strates avec un unique élément (option NO-COLLAPSE) a une influence négligeable, inférieure à 1%, sur les estimations des écart-types.

Dans le cas présenté ci-dessus, le grand nombre d'observations dans le jeu de données amène à des coefficients fortement significatifs pour toutes les méthodes (naïve, incomplète et complète).

7.5.3 Influence du plan sur le choix du modèle

On montre dans cette section que le choix de la méthode peut avoir une influence sur le choix des variables conservées dans le modèle.

Nous définissons les sous-groupes "Lac_hommes_3040" composé des 9'847 hommes de la région lémanique entre 30 et 40 ans (critère : `30<=ALTER<40`) et "Lac_hommes_3050" composé des 17'975 hommes de la même région entre 30 et 50 ans.

Pour les deux jeux de données, on commence par choisir un modèle comportant un ensemble de variables de base et on recherche un modèle final avec un sous-ensemble de ces variables. On applique la méthode de sélection itérative nommée "backward" qui retire à chaque étape la variable non significative avec la plus petite p-value jusqu'à obtenir un ensemble de variables ayant toutes des coefficients significatifs.

Avec le modèle B et le jeu de données, "Lac_hommes_3040", la procédure de sélection aboutit au retrait des variables `EDUEXPR2` puis `EDUTEN2` en se basant sur l'estimation naïve tout comme sur l'estimation complète. L'effet des poids et du plan est cependant toujours sensible, surtout au niveau de la variabilité des estimateurs des coefficients ; voir Table 7.4. Toutes les p-values de la table sont inférieures à 0.0001, excepté le coefficient de `EDUEXPRTEN` dans l'estimation avec les poids et le plan (p-value=0.024). La statistique de fit R-square égale environ 0.39 pour les deux estimations.

TAB. 7.4 – Modèle B sur les données des hommes entre 30 et 40 ans dans la région lémanique. Voir les tableaux précédents pour la notation.

Variables	$\widehat{\beta}_{nonpond}$	$\widehat{\beta}_{pond}$	$t_{SIR}(\widehat{\beta}_{nonpond})$	$t_{SIR}(\widehat{\beta}_{pond})$	$t1_p(\widehat{\beta}_{pond})$	*weft*	*deft*	*ceft*
`Intercept`	7.537	7.384	419.6	385.3	107.3	1.07	3.59	3.83
`EDU1`	0.764	0.838	68.3	67.7	16.7	1.11	4.05	4.48
`EDUEXPR`	0.067	0.094	7.0	9.3	4.3	1.05	2.19	2.31
`EDUTEN`	0.131	0.135	15.6	13.5	6.3	1.18	2.14	2.54
`EDUEXPRTEN`	0.096	0.077	9.1	6.2	2.3	1.17	2.74	3.22

Avec le modèle A et le jeu de données "Lac_hommes_3050", la procédure de sélection aboutit à des modèles différents. Avec l'estimation naïve aucune variable n'est retirée car les coefficients sont tous significativement différents de zéro. Avec poids et plan, on retire la variable `EDUEXP2` car sa p-value=0.07 est supérieure à notre limite de 0.05. Le modèle obtenu est donc différent,

voir Table 7.5. Les p-values finales de tous les coefficients sont < 0.0001. Les statistiques de fit des modèles finaux sont similaires : root MSE=0.30 et R-square=0.38 pour la méthode sans poids ni plan, et root MSE=0.32 et R-square=0.39 pour la méthode complète.

TAB. 7.5 – Modèles sur les données des hommes entre 30 et 50 ans dans la région lémanique. Voir les tableaux précédents pour la notation.

Variables	$\widehat{\beta}_{nonpond}$	$Std_{SIR}(\widehat{\beta}_{nonpond})$	$t_{SIR}(\widehat{\beta}_{nonpond})$	$\widehat{\beta}_{pond}$	$Std1_p(\widehat{\beta}_{pond})$	$t1_p(\widehat{\beta}_{pond})$
`Intercept`	7.579	0.0115	661.9	7.439	0.0438	169.8
`EDU1`	0.707	0.0151	46.94	0.838	0.0396	21.2
`EDUEXP`	0.171	0.0155	11.04	0.101	0.0081	12.5
`EDUEXP2`	-0.017	0.0040	-4.28			

Le choix de la méthode d'estimation n'a donc pas uniquement une conséquence sur la valeur des paramètres et de leur variabilité mais également sur le choix du modèle.

7.5.4 Effet de sondage des moyennes et des coefficients

Divers auteurs ont essayé de trouver une relation entre l'effet de sondage d'un coefficient d'un modèle de régression et celui de la moyenne des variables utilisées dans le modèle ; voir résumé dans Särndal et al. (1992). Une telle relation est en effet pratique si l'estimation complète (avec poids et plan) de la variance de la moyenne est possible mais pas celle des coefficients.

Kish et Frankel (1974) ont notamment proposé d'utiliser l'estimation conservatrice :

$$\textit{deft}(\widehat{\theta}_s) = \textit{deft}(\widehat{\overline{x}}_s) \tag{7.6}$$

ou un modèle du type suivant (si $\textit{deft}(\widehat{\overline{x}}_s) > 1$) :

$$\textit{deft}(\widehat{\theta}_s)^2 = 1 + f_{\hat{\theta}_s}[\textit{deft}(\widehat{\overline{x}}_s)^2 - 1] \tag{7.7}$$

avec le facteur $f_{\hat{\theta}_s}$ unique pour la statistique $\widehat{\theta}_s$, et x une variable utilisée dans le calcul de $\widehat{\theta}_s$.

Dans les modèles analysés dans les sections précédentes, on note que les effets de sondage $\textit{deft}(\widehat{\theta}_s)$ des coefficients varient entre 2 et 4 ; voir Table 7.6. Ils sont inférieurs à ceux des moyennes dans tous les cas sauf le dernier du tableau. La première proposition de Kish et Frankel n'est donc pas toujours conservatrice.

Le facteur $\widehat{f}_{\hat{\theta}_s} = [\textit{deft}(\widehat{\theta}_s)^2 - 1]/[\textit{deft}(\widehat{\overline{x}}_s)^2 - 1]$ varie entre modèles et jeux de données. Il est donc difficile de choisir une valeur $\widehat{f}_{\hat{\theta}_s}$ pour estimer l'effet de sondage du coefficient $\widehat{\theta}_s$ d'un autre modèle.

Il est recommandé d'appliquer une méthode d'estimation complète (avec poids et plan) afin d'éviter le passage par le *deft* d'une variable auxiliaire. Cette étape intermédiaire ne devrait être utilisée que lorsque cela n'est pas possible autrement.

TAB. 7.6 – Effet de sondage des coefficients du modèle de régression *deft*$(\widehat{\theta}_s)$ et de la moyenne des variables correspondantes *deft*$(\widehat{\overline{x}}_s)$. Facteur $\widehat{f}_{\hat{\theta}_s}$ décrit dans le texte.

	x	*deft*$(\widehat{\theta}_s)$	*deft*$(\widehat{\overline{x}}_s)$	$f_{\hat{\theta}_s}$
"Lac_femmes"	EDU1	3.27	4.26	0.57
	EDUEXP	2.36	3.09	0.54
	EDUEXP2	2.43	2.85	0.69
"Lac_hommes3040"	EDU1	4.05	4.26	0.90
	EDUEXPR	2.19	2.88	0.52
	EDUTEN	2.14	2.96	0.46
	EDUEXPRTEN	2.74	1.43	6.23

7.5.5 Prévisions

Le sous-groupe "Lac_hommes_3040" et les résultats de la Table 7.4 sont repris pour illustrer l'effet des poids et du plan sur les prévisions du salaire standardisé $y = \texttt{MBLS} = e^{\texttt{LNMBLS}}$ en fonction des variables de base EDU, EXPR et TEN.

Des prévisions du salaire moyen, avec intervalles de confiance, sont effectuées pour 180 combinaisons entre $\texttt{EDU} = \{8, 9.5, 12, 15, 16\}$, $\texttt{EXPR} = \{0, 4, 8, 12, 16, 20\}$ et $\texttt{TEN} = \{1, 6, 11, 16, 21, 25\}$ selon le modèle déterminé sans les poids et le plan, respectivement selon le modèle complet.

Première constatation : des différences sensibles sont observées entre les prévisions sans pondération $\widehat{y}_{nonpond}$ et les prévisions avec pondération $\widehat{y}_{pond}$. En moyenne, la valeur pondérée est inférieur à celle non pondérée. Les différences les plus grandes sont observées pour des valeurs de TEN élevées telles que 21 ou 25. On note cependant que la prévision est peu fiable dans ce secteur car le jeu de données utilisé pour l'estimation des coefficients ne contient que peu de ces valeurs.

Lorsque l'on prend en compte la variabilité des estimateurs, on note que les intervalles de confiance se recoupent dans 137 cas sur 180. Dans les 43 cas restants, la valeur pondérée $\widehat{y}_{pond}$ est inférieure à la valeur non pondérée $\widehat{y}_{nonpond}$ et les intervalles ne se recoupent pas. Tous ces cas se trouvent dans les niveaux de formation $\texttt{EDU} = \{8, 9.5, 12\}$.

La Figure 7.1 illustre l'effet des poids et du plan sur la prévision pour $\texttt{TEN} = 6$ (6 années de service), $\texttt{EDU} = 8$ (scolarité obligatoire. à gauche) et $\texttt{EDU} = 12$ (apprentissage complet, à droite) et EXPR prenant les valeurs $\{0, 4, 8, 12, 16, 20\}$. Dans le graphique de gauche, toutes les prévisions sont significativement différentes. Dans le graphique de droite, seule une partie des différences sont significatives. Notons que les intervalles sont donnés pour la prévision d'une valeur moyenne. Ceux pour une valeur donnée y_0 seraient plus larges.

Les calculs de prévision ont été fait dans la procédure IML de SAS. Une estimation de la matrice de variance-covariance des coefficients est disponible comme output des procédure PROC REG et PROC SURVEYREG (COVB) ; voir les calculs dans l'Annexe A.5.

L'interprétation des différences entre les résultats pondérés et non pondérés n'est pas triviale.

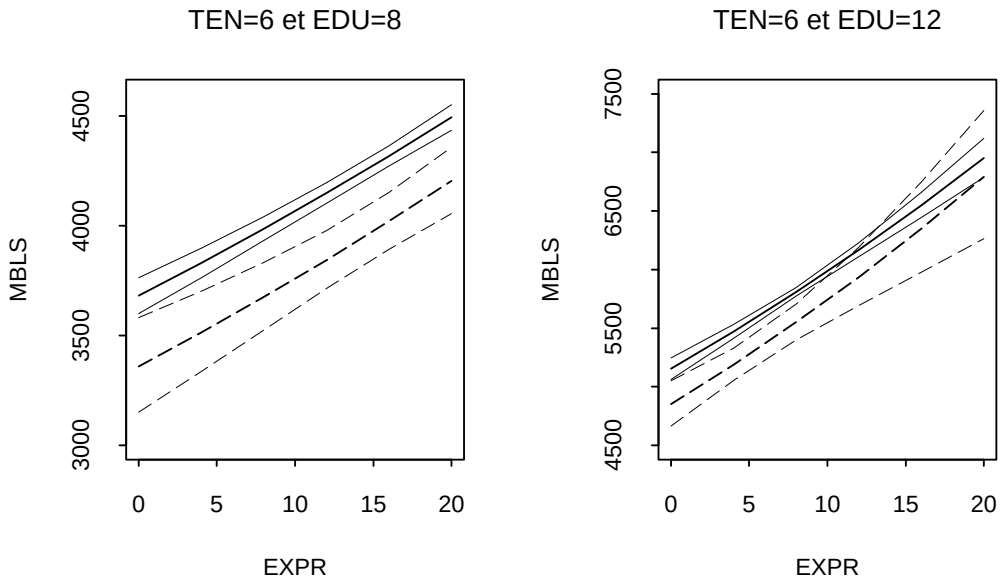


FIG. 7.1 – Prévision $\widehat{y}_{nonpond}$ (ligne continue en gras) avec intervalle de confiance (lignes continues), et $\widehat{y}_{pond}$ (ligne traitillée en gras) avec intervalle de confiance (lignes traitillées). Valeur pour $TEN = 6$ et $EDU = 8$ à gauche et pour $TEN = 6$ et $EDU = 12$ à droite

Ces différences peuvent être dues à un réel effet des poids mais aussi peut-être au fait que le modèle n'est pas bien spécifié.

Si on compare les prévisions non pondérées et pondérées en prenant, pour toutes les deux, l'intervalle de confiance complet qui prend en compte les poids et le plan, il ne reste que peu de différences significatives (9 sur les 43). Les valeurs pondérées et non pondérées sont effectivement différentes mais cette différence est relativisée dès que l'on considère la variabilité réelle due au plan d'échantillonnage. Il ne semble donc pas y avoir de gros problème de spécification du modèle dans notre exemple.

On retrouve ici le point suivant : le plus important dans l'estimation de coefficients d'un modèle de régression n'est pas l'utilisation des poids ou pas. C'est la prise en compte de la variabilité due au fait que l'échantillon n'a pas été tiré selon un plan aléatoire simple.

Chapitre 8

Tests d'indépendance entre variables catégorielles

Une méthode courante d'analyse de données catégorielles est le test d'indépendance entre variables. L'analyse passe en général par la représentation en tableaux croisés et le test du chi-carré.

Seul les tableaux croisés entre deux variables sont abordés dans ce rapport. Les méthodes peuvent être étendues aux tableaux avec plus de deux variables ou encore aux modèles log-linéaires avec interactions croisées entre les variables ; voir par exemple Lohr (1999).

8.1 Chi-carré classique

Soient n observations indépendantes classifiées selon deux facteurs : facteur R avec r catégories et facteur C avec c catégories.

Chaque observation a une probabilité p_{ij} de se trouver dans la case (i, j), avec $i = 1, .., r$ et $j = 1, .., c$. On définit $p_{i+} = \sum_{j=1}^{c} p_{ij}$ la probabilité de tomber dans la catégorie i et $p_{+j} = \sum_{i=1}^{r} p_{ij}$ celle de tomber dans la catégorie j. La somme $p_{++} = \sum_{i=1}^{r} \sum_{j=1}^{c} p_{ij}$ est donc égale à 1.

Le nombre d'unités observées dans la case (i, j) est noté x_{ij}.

Si toutes les observations de l'échantillon de taille n sont indépendantes, les x_{ij} suivent une loi multinomiale avec rc catégories. C'est le cas en présence d'un échantillon SIR ou, approximativement, en présence d'un échantillon SI si la taille de l'échantillon est petite par rapport à celle de la population.

L'hypothèse nulle d'indépendance est : $H_0 : p_{ij} = p_{i+}p_{+j}$ pour $i = 1, .., r$ et $j = 1, .., c$.

Sous H_0, le nombre espéré $m_{ij} = np_{ij}$ d'observations dans la case (i, j) peut être estimé par

$$\widehat{m}_{ij} = n\widehat{p}_{i+}\widehat{p}_{+j} = n\frac{x_{i+}}{n}\frac{x_{+j}}{n} = \frac{x_{i+}x_{+j}}{n} \tag{8.1}$$

La statistique du chi-carré de Pearson est définie par :

$$X_c^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(x_{ij}-\widehat{m}_{ij})^2}{\widehat{m}_{ij}} = n\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(\widehat{p}_{ij}-\widehat{p}_{i+}\widehat{p}_{+j})^2}{\widehat{p}_{i+}\widehat{p}_{+j}} \tag{8.2}$$

La statistique pour le test du ratio de vraisemblance (likelihood ratio test, chi-deux de vraisemblance) est donnée par :

$$G_c^2 = 2\sum_{i=1}^{r}\sum_{j=1}^{c}x_{ij}\ln(\frac{x_{ij}}{\widehat{m}_{ij}}) = 2n\sum_{i=1}^{r}\sum_{j=1}^{c}\widehat{p}_{ij}\ln(\frac{\widehat{p}_{ij}}{\widehat{p}_{i+}\widehat{p}_{+j}}) \tag{8.3}$$

Dans le cas d'un échantillon de grande taille, X^2 et G^2 sont distribués approximativement selon une loi χ^2 avec $(r-1)(c-1)$ degrés de liberté sous H_0. En pratique, on estime en général qu'il faut au moins une observation dans chaque case et $n \geq 5rc$ pour utiliser le test.

Les statistiques X_c^2 et G_c^2 peuvent aussi être utilisées pour mesurer l'homogénéité de proportions entre différentes populations. On a dans ce cas r populations classifiées dans c catégories et l'hypothèse nulle est $H_0 = p_{1j}/p_{1+} = ... = p_{rj}/p_{r+} = p_{+j}$ pour tout $j = 1,..,c$. Ces statistiques peuvent également être utilisées pour un test d'ajustement ("goodness of fit"), avec $H_0 : p_i = p_i^{(0)}$ pour $i = 1,..,k$ catégories, et $p_i^{(0)}$ une valeur pre-spécifiée ou estimée sur la base des données.

Les statistiques X_c^2 et G_c^2 dépendent fortement de la taille de l'échantillon n. Le test est donc sensible à une augmentation artificielle de la taille comme c'est le cas dans un échantillonnage avec effet de grappe (taille réelle supérieure à la taille effective de l'échantillon).

8.2 Chi-carré avec poids et plan

Dans le cas d'une enquête, un poids w_k est associé à chaque observation $k = 1,..,n$ de l'échantillon. De plus, les n éléments ont été tirés dans une population finie de taille N selon un plan d'échantillonnage $p(s)$ différent du plan aléatoire simple.

Comme dans le cas des statistiques descriptives et des modèles de régression, les poids et le plan doivent être pris en compte dans les tests d'indépendance entre variables catégorielles.

8.2.1 Statistique pondérée

Dans le cas d'échantillonnage à probabilités inégales, les nombres observés x_{ij} ne reflètent pas les fréquences relatives. L'estimateur pondéré $\widehat{p}_{ij}$ de la proportion dans chaque cellule est alors défini par :

$$\widehat{p}_{ij,pond} = \frac{\sum_{k=1}^{n} w_k\, y_{kij}}{\sum_{k=1}^{n} w_k} \tag{8.4}$$

avec la variable indicatrice $y_{kij} = 1$ si l'observation k est dans la cellule (i,j) et $y_{kij} = 0$ sinon.

La statistique classique du chi-carré de l'équation (8.2) est adaptée pour tenir compte de la pondération (forme similaire pour G_p^2) :

$$X_p^2 = n \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(\widehat{p}_{ij,pond} - \widehat{p}_{i+,pond}\,\widehat{p}_{+j,pond})^2}{\widehat{p}_{i+,pond}\,\widehat{p}_{+j,pond}} \tag{8.5}$$

8.2.2 Inférence

Dans le cas d'une enquête par échantillonnage, les probabilités ne suivent plus des lois multinomiales. Par conséquent, X_p^2 et G_p^2 ne suivent plus une distribution approximative du χ^2. Dans le cas d'un échantillon avec effet de grappe, la vraie *p*-value sera par exemple souvent supérieure à celle fournie par la théorie classique. Un test sans tenir compte du plan pourrait ainsi conclure à l'indépendance alors qu'elle n'est en fait pas significative ; et vice-versa.

Dans le cas d'un échantillonnage SIR, la statistique X_p^2 peut être directement comparée avec une distribution $\chi^2_{(r-1)(c-1)}$. Si l'échantillonnage est sans remise, on corrige la valeur de X_p^2 par le facteur $(1 - n/N)$ avec n la taille de l'échantillon et N la taille de la population. La valeur de $X_p^2/(1 - n/N)$ suit une loi $\chi^2_{(r-1)(c-1)}$; voir Skinner et al. (1989).

Dans le cas d'un échantillonnage stratifié sans grappes et si les lignes du tableau de contingence correspondent aux strates, le test du chi-carré peut également être appliqué sur la valeur fournie par l'équation (8.5). On est en effet dans un cas équivalent - si le facteur de population finie est négligeable - à une comparaison de proportions entre r populations indépendantes.

Dans les autres cas, la précision de l'estimateur de $p_{ij,pond}$ dépend du plan d'échantillonnage.

Divers travaux théoriques ont été réalisés depuis les années 70 pour comprendre l'effet du plan d'échantillonnage sur les tests d'indépendance et proposer des méthodes permettant de le corriger. Un résumé des méthodes les plus courantes est disponible dans Lohr (1999) ou Skinner et al. (1989). Thomas et al. (1996) regroupe les différentes méthodes dans les quatre types suivants : (1) statistique de Wald, (2) approche de Rao-Scott qui corrige les statistiques classique X_p^2 et G_p^2, (3) approche Jackknife des tests classiques X_p^2 et G_p^2, et (4) inégalité de Bonferroni.

Dans le cas d'enquêtes complexes, les tests d'indépendance et d'homogénéité ne sont plus toujours identiques. On se restreint ci-dessous aux tests d'indépendance.

Sous $H_0 : \theta = 0$ avec $\theta = [\theta_{11}\theta_{12}..\theta_{r-1,c-1}]^T$ et $\theta_{ij} = p_{ij} - p_{i+}p_{+j}$, la statistique de Wald $X_W^2 = \widehat{\theta}^T \widehat{V}(\widehat{\theta})^{-1}\widehat{\theta}$, avec $\widehat{V}(\widehat{\theta})$ la matrice de covariance estimée de $\widehat{\theta}$, suit approximativement une loi $\chi^2_{(r-1)(c-1)}$. Contrairement à celle de Pearson, elle peut être généralisée à des plans complexes si on dispose d'un estimateur consistent de la matrice de covariance des estimateurs dans les différentes cellules. Cette méthode n'est pas recommandée si on a un trop petit nombre d'unités d'échantillonnage ou si le nombre de cases dans le tableau est grand (Lohr, 1999). Des modifications ont été présentées par Thomas et al (1996).

Rao et Scott (1981, 1984, 1987) décomposent la statistique classique X_p^2 en une somme pondérée $\delta_1 W_1 + .. + \delta_{(r-1)(c-1)} W_{(r-1)(c-1)}$ de distributions χ_1^2 indépendantes W_i, $i = 1, .., (r-1)(c-1)$. Les poids δ_i sont les valeurs propres de la matrice des effets de sondage. Cette matrice est égale au produit $\Sigma\Sigma_0^{-1}$ entre la matrice de covariance des estimateurs des proportions $p_{ij,pond}$ et l'inverse de la matrice de covariance sous H_0 (forme symétrique : $\Sigma_0^{-1/2}\Sigma\,\Sigma_0^{-1/2}$). Dans le cas

d'une loi multinomiale (échantillon SI), les δ_i sont tous égaux à 1. Dans les autres cas, la statistique X_p^2 peut fortement dévier d'une distribution χ^2. Rao et Scott proposent des statistiques corrigées X_p^2/δ qui suivent asymptotiquement une loi du $\chi^2_{(r-1)(c-1)-1}$; voir exemple plus bas.

Fay (1985) reprend la forme de Rao et Scott et propose d'utiliser la méthode du Jackknife pour estimer le paramètre de correction δ. Cette méthode permet notamment de mieux aborder le problème en cas de grande variabilité des δ_i mais l'échantillon doit être de taille suffisante. Pour le Jackknife, le jeu de données résultat complet doit être disponible.

Un autre type de méthode, basé sur l'inégalité de Bonferroni, est proposé dans Thomas et al. (1996) et cité également dans Lohr (1999). Cette méthode fourni un test conservateur qui semble bien fonctionner en pratique. Elle teste les $\theta_{ij} = p_{ij} - p_{i+}p_{+j}$ les uns après les autres et les compare avec une loi t. La méthode est relativement simple mais nécessite tout de même une estimation de la variance de chaque $\widehat{\theta}_{ij}$, par exemple par une méthodes de reéchantillonnage.

D'autres méthodes reposent sur l'approche model-based. Celles présentées dans Lohr (1999) font l'hypothèse que les effets de sondage sont constants pour chaque cellule et marge. Ce type de méthode n'est pas abordé ci-dessous.

Les publications citées dans cette section comparent les résultats obtenus par différentes méthodes et sur différents jeux de données. Cependant, il est difficile de choisir une méthode passe-partout pouvant être utilisée dans tous les cas de façon raisonnable. Les points les plus influents sont la variabilité des δ_i, la taille de l'échantillon et le nombre de cases dans le tableau. Il est à noter que les grands échantillons de la statistique publiques présentent l'avantage de se rapprocher des conditions asymptotiques. De petites tailles peuvent cependant apparaître dans certaines cases et il faut donc rester vigilant.

8.2.3 Corrections du type Rao-Scott

La correction du premier ordre du type de Rao-Scott est définie par $\delta = E[X_p^2]/(r-1)(c-1)$. Elle a pour but d'adapter la moyenne de la statistique de test pour obtenir la moyenne d'une $\chi^2_{(r-1)(c-1)}$, i.e. $(r-1)(c-1)$.

L'estimateur corrigé est donc :

$$X_F^2 = \frac{(r-1)(c-1)X_p^2}{E[X_p^2]} \tag{8.6}$$

où

$$E[X_p^2] \approx \sum_{i=1}^{r}\sum_{j=1}^{c}(1-p_{ij})d_{ij} - \sum_{i=1}^{r}(1-p_{i+})d_i^R - \sum_{j=1}^{c}(1-p_{+j})d_j^C \tag{8.7}$$

avec d_{ij} l'effet de sondage (*deff*) pour l'estimation de p_{ij}, d_i^R le *deff* pour l'estimation de p_{i+} et d_j^C le *deff* pour l'estimation de p_{+j}. On utilise également la notation $E[X_p^2] \approx D_{++} - D_C - D_R$ avec les sommes sur les cases D_{++}, les lignes D_R et les colonnes D_C.

Pour une estimation de $E[X_p^2]$, on reprend les probabilités de l'équation (8.4) et on estime les *deff* sur la base des données. Notons que l'approximation du premier ordre de $E[X_p^2]$ peut s'annuler ou même devenir négative dans certains cas particuliers. Exemple : tableau 2 x 2. Si

$d_{ij} = d$, $d_i^R = d_j^C = d_m$, $\forall i, j$ et $d_m/d \geq 1.5$, alors $E[X_p^2] \leq 0$. Lorsque l'approximation est inférieure ou égale à 0, la correction n'est évidemment pas applicable.

La correction du premier ordre a l'avantage d'utiliser uniquement les effets de sondage et non pas toutes les données. Elle est cependant partielle car seule la moyenne est ajustée et les queues de la distribution ne le sont pas. De plus, la correction du premier ordre risque d'aboutir à une sous-estimation de la p-value si les effets de sondage varient beaucoup (Lohr, 1999).

Une correction du deuxième degré est possible mais demande également l'estimation de la matrice de variances-covariances complète pour les $\widehat{p}_{ij}$. Si les effets de sondage sont similaires, les corrections du premier et du deuxième ordre devraient être similaires. D'autres corrections sont décrites par exemple dans Särndal et al. (1992) et Chambers et Skinner (2003).

8.3 Logiciels

PROC FREQ de SAS fourni une valeur, pondérée ou pas, pour la statistique du chi-carré. L'utilisation de l'option WEIGHT est cependant faite dans le sens de fréquences et non pas de poids d'échantillonnage. La valeur du chi-carré fournie dans le cas pondéré doit être adaptée par le facteur n/N ; voir Annexe A.6.

SAS v8 ne propose pas d'estimations du chi-carré tenant compte du plan d'échantillonnage. Une correction du premier ordre peut cependant être effectuée "à la main" dans SAS. Dans ce cas, les effets de sondage des proportions dans les cases et des marges du tableau sont estimés avec SURVEYMEANS selon la méthode décrite dans la Section 6. La procédure SURVEYFREQ prévue dans SAS v9 permettra de faire les estimations en tenant compte du plan d'échantillonnage.

Wesvar 4 calcule les chi-carrés non pondérés et pondérés, mais ne corrige pas la valeur obtenue pour tenir compte du plan d'échantillonnage. Aucune correction n'est donc effectuée.

8.4 Exemple sur les données de l'ERC98

Pour illustrer les méthodes d'analyse de tableaux croisés, nous prenons quelques variables catégorielles du module "Conditions de vie" de l'ERC98.

8.4.1 But et données

Les variables catégorielles sélectionnées pour l'étude de dépendances sont le sexe (`SEX`), le niveau de formation (`BILD3`), le statut d'occupation du logement (`WOHNTYP2`) et l'insatisfaction dans la vie (`LZUFR2`); voir Table 8.1. Le jeu de données comporte 4'369 personnes, voir Section 6.6.

On étudie le lien entre l'insatisfaction et l'occupation du logement (tableau "`ZUFxWOH`"), ou le sexe (tableau "`ZUFxSEX`"), et entre la formation et l'occupation du logement (tableau "`BILDxWOH`").

TAB. 8.1 – Variables catégorielles utilisées. Les proportions indiquées sont les valeurs dans l'échantillon sans pondération.

Nom	Catégories	Notation	Proportions
Sexe	hommes	SEX=1	46.0%
	femmes	SEX=2	54.0%
Formation	école obligatoire	BILD3=1	20.2%
	degré secondaire II	BILD3=2	58.9%
	degré tertiaire	BILD3=3	20.9%
Occupation	locataire	WOHNTYP2=1	64.2%
	propriétaire	WOHNTYP2=2	35.8%
Insatisfaction	satisfaite	LZUFR2=0	95.7%
	insatisfaite	LZUFR2=1	4.3%

Le modèle log-linéaire correspondant par exemple au lien entre l'insatisfaction et l'occupation du logement est :

$$ln(p_{ij}) = \mu + \alpha_i + \beta_j + \gamma_{ij} \tag{8.8}$$

avec p_{ij} la probability d'être dans la case (i, j), α_i les effets de la variable LZUFR2, $i = 1, 2$, β_j les effets de la variable WOHNTYP2, $j = 1, 2$ et γ_{ij} les effets combinés des deux variables. Le test d'indépendance entre les deux variables correspond alors au test de l'hypothèse : H_0 : $\gamma_{ij} = 0$ pour $i = 1$ et $j = 1$.

8.4.2 Résultats

L'observation des tableaux croisés montre l'effet des poids sur le nombre d'éléments dans les différentes cases ; voir la Table 8.2 et l'Annexe A.6 pour l'explication sur les poids corrigés.

Dans le tableau ZUFxWOH, on note que la pondération ne modifie pas la répartition marginale entre satisfaits et insatisfaits (95.7% et 4.3%) mais légèrement celle entre locataires et propriétaires (les propriétaires passent de 35.8 à 38.1%).

Dans le tableau ZUFxSEX, la répartition entre hommes et femmes est modifiée pour les satisfaits (passage de 46 à 48% d'hommes).

Dans le tableau BILDxWOH, le plus grand changement a lieu dans les niveaux de formation avec un passage de par exemple 20% à 25% pour l'école obligatoire.

TAB. 8.2 – Tableaux croisés ZUFxWOH, ZUFxSEX et BILDxWOH avec les fréquences, la proportion par rapport au total, la proportion par rang et la proportion par colonne (similaire à la sortie de SAS).

ZUFxWOH sans poids.

	locataire	propriétaire	Total
satisfaite	2657	1526	4183
	60.8%	34.9%	95.7%
	63.5%	36.5%	
	94.8%	97.5%	
insatisfaite	147	39	186
	3.4%	0.9%	4.3%
	79.0%	21.0%	
	5.2%	2.5%	
Total	2804	1565	4369
	64.2%	35.8%	100%

ZUFxWOH avec poids corrigés.

	locataire	propriétaire	Total
satisfaite	2558.3	1623.8	4182.1
	58.6%	37.2%	95.7%
	61.2%	38.8%	
	94.6%	97.5%	
insatisfaite	145.4	41.5	186.9
	3.3%	1.0%	4.3%
	77.8%	22.2%	
	5.4%	2.5%	
Total	2703.7	1665.3	4369
	61.9%	38.1%	100%

ZUFxSEX sans poids.

	hommes	femmes	Total
satisfaite	1931	2252	4183
	44.2%	51.5%	95.7%
	46.2%	53.8%	
	96.1%	95.4%	
insatisfaite	78	108	186
	1.8%	2.5%	4.3%
	41.9%	58.1%	
	3.9%	4.6%	
Total	2009	2360	4369
	46.0%	54.0%	100%

ZUFxSEX avec poids corrigés.

	hommes	femmes	Total
satisfaite	2015.2	2166.9	4182.1
	46.1%	49.6%	95.7%
	48.2%	51.8%	
	95.7%	95.8%	
insatisfaite	91.5	95.4	186.9
	2.1%	2.2%	4.3%
	48.9%	51.1%	
	4.3%	4.2%	
Total	2106.7	2262.3	4369
	48.2%	51.8%	100%

BILDxWOH sans poids.

	locataire	propriétaire	Total
école obl	578	304	882
	13.2%	7.0%	20.2%
	65.5%	34.5%	
	20.6%	19.4%	
sec II	1662	913	2575
	38.0%	20.9%	58.9%
	64.5%	35.5%	
	59.2%	58.3%	
tertiaire	564	348	912
	12.9%	8.0%	20.9%
	61.8%	38.2%	
	20.1%	22.2%	
Total	2804	1565	4369
	64.1%	35.8%	100%

BILDxWOH avec poids corrigés.

	locataire	propriétaire	Total
école obl	701.4	391.1	1092.5
	16.1%	9.0%	25.0%
	64.2%	35.8%	
	25.9%	23.5%	
sec II	1502.3	920.8	2423.1
	34.40%	21.1%	55.5%
	62.0%	38.0%	
	55.6%	55.3%	
tertiaire	499.9	353.5	853.4
	11.4%	8.1%	19.5%
	58.6%	41.4%	
	18.5%	21.2%	
Total	2703.7	1665.3	4369
	61.9%	38.1%	100%

La statistique du chi-carré est modifiée par la pondération ; voir Table 8.3. La correction du premier ordre amène également une correction selon un facteur $\widehat{\delta}$ de l'ordre de 0.9 à 1.3. Selon les cas, la statistique reste presque constante, augmente ou diminue dès que l'on prend en compte les poids et le plan.

Pour le tableau ZUFxWOH, les valeurs non pondérées et pondérées - avec et sans plan - indiquent un refus clair de l'hypothèse d'indépendance ; risque inférieur à 0.1%, voir Table 8.4. Le fait d'être insatisfait semble donc lié au type d'occupation du logement.

Dans le cas du tableau ZUFxSEX, les valeurs du chi-carré indiquent toutes que l'hypothèse ne peut être refusée qu'avec un risque de première espèce α supérieur à 20%. Il ne parait donc pas y avoir de lien entre le sexe et le fait d'être insatisfait dans la vie.

Le troisième tableau est plus particulier. Le risque de première espèce passe d'une valeur supérieure à 20% pour X_c^2, à une valeur proche de 10% pour X_p^2. La correction du premier ordre amène alors le risque vers une valeur proche de 5%. D'une indépendance clairement refusée, on passe donc à une indépendance presque acceptée, avec le risque de 5%, entre la formation et le type d'occupation du logement.

TAB. 8.3 – Statistique du chi-carré X_c^2 sans et X_p^2 avec prise en compte des poids, et degrés de liberté $df = (r-1)(c-1)$. Correction du premier ordre effectué dans SAS : $D_{..}$, D_R, D_C, $E[X_p^2]$, $\widehat{\delta} = E[X_p^2]/df$ et résultat $X_F^2 = X_p^2/\widehat{\delta}$.

Tableau	X_c^2	X_p^2	df	D_{++}	D_R	D_C	$E[X_p^2]$	$\widehat{\delta}$	X_F^2
ZUFxWOH	18.64	20.94	1	3.94	1.33	1.28	1.33	1.33	15.74
ZUFxSEX	1.28	0.04	1	3.94	1.33	1.28	1.34	1.34	0.03
BILDxWOH	3.02	6.45	3	6.60	2.64	1.28	2.68	0.89	7.22

TAB. 8.4 – Valeurs de $\chi_{df,\alpha}$ pour différentes combinaisons (df,α). Extrait de la table du chi-carré.

$\chi_{df,\alpha}$		α				
		0.1%	1%	5%	10%	20%
df	1	10.83	6.64	3.84	2.71	1.64
	3	16.27	11.35	7.82	6.25	4.64

Comme pour les statistiques descriptives et des modèles, on note que l'utilisation de méthodes sans poids et/ou sans plan, peut amener à des conclusions erronées dans le cas de tests d'indépendance.

8.5 Exemple sur les données de la LSE98

Différentes dépendances entre variables catégorielles sont testées également sur la base des données de la LSE98.

8.5.1 But et données

On désire étudier le lien entre le fait de travailler à temps partiel et le sexe, la classe d'âge ou encore l'état civil des salariés.

Pour les besoins de l'illustration, nous choisissons de restreindre l'information sur le taux d'activité à 2 catégories (temps partiel oui ou non). De même, seule la classe d'âge, et non pas l'âge lui-même, est retenue. Les 3 types d'état civil sont également regroupés en deux catégories ; voir Table 8.5.

TAB. 8.5 – Variables catégorielles utilisées.

Nom	Catégories	Notation	Critère
sexe	hommes	`GESCHLE=1`	`GESCHLE=1`
	femmes	`GESCHLE=2`	`GESCHLE=2`
classe d'âge	age1	`CLAG=1`	`15<=ALTER<30`
	age2	`CLAG=2`	`30<=ALTER<40`
	age3	`CLAG=3`	`40<=ALTER<50`
	age4	`CLAG=4`	`50<=ALTER`
état civil groupé	célibataire	`ZIV=1`	`ZIVSTA=1`
	marié, divorcé, veuf	`ZIV=2`	`ZIVSTA in (2,3)`
occupation	partiel	`PART=1`	`IBGR<0.9`
	plein	`PART=0`	`IBGR>=0.9`

Les données sont celles utilisées dans les exemples des Sections 6.7 et 7.5. L'analyse concerne les emplois sans la variable du salaire `MBLS`. Il n'est donc pas nécessaire d'introduire la standardisation `IBGRS`. Les valeurs non pondérées n'utilisent donc aucun ajustement. Les valeurs pondérées reprennent la variable `GEWICHT`. Les sous-ensembles sur lesquels sont appliquées les méthodes d'estimation du chi-carré ne contiennent pas de valeurs manquantes ; voir Table 8.6.

8.5.2 Résultats

Les tableaux croisés avec et sans poids montrent des différences parfois importantes. Dans l'exemple du tableau *S* (sexe x partiel), la pondération amène à une augmentation de la proportion des temps partiels de 7.4 à 8.9% chez les hommes. Cette proportion reste autour de 41% pour les femmes ; voir la Table 8.7. Les proportions marginales indiquent une tendance à moins de femmes et plus de temps partiels. Dans le tableau *ZH2*, l'effet de la pondération est aussi net, notamment sur la proportion de temps partiel chez les célibataires.

TAB. 8.6 – Tableaux croisés utilisés dans les analyses de la LSE98.

Tableaux	Variables	Restriction	$r \times c$	n
S	GESCHLE x PART	-	2×2	72'822
CH	CLAG x PART	GESCHLE=1	4×2	41'639
CF	CLAG x PART	GESCHLE=2	4×2	31'109
ZH2	ZIV x PART	GESCHLE=1 et CLAG=2	2×2	13'283
ZH4	ZIV x PART	GESCHLE=1 et CLAG=4	2×2	10'293

TAB. 8.7 – Tableaux croisés *S* (sexe x partiel) et *ZH2* (état civil x partiel), avec les fréquences et les proportions par rapport au total, par rang et par colonne.

S sans poids.

	plein	partiel	Total
hommes	38579	3095	41674
	53.0%	4.3%	57.2%
	92.6%	7.4%	
	67.7%	19.6%	
femmes	18415	12733	31148
	25.3%	17.5%	42.8%
	59.1%	40.9%	
	32.3%	80.5%	
Total	56994	15828	72822
	78.3%	21.7%	100%

S avec poids corrigés.

	plein	partiel	Total
hommes	38724	3768.4	42493
	53.2%	5.2%	58.4%
	91.1%	8.9%	
	68.3%	23.3%	
femmes	17937	12393	30329
	24.6%	17.0%	41.7%
	59.1%	40.9%	
	31.7%	76.7%	
Total	56660.6	16161.4	72822
	77.8%	22.2%	100%

ZH2 sans poids.

	plein	partiel	Total
célibataire	3010	295	3305
	22.7%	2.2%	24.9%
	91.1%	8.9%	
	24.1%	36.8%	
autres	9472	506	9978
	71.3%	3.8%	75.1%
	94.9%	5.1%	
	75.9%	63.2%	
Total	12482	801	13283
	94.0%	6.0%	100%

ZH2 avec poids corrigés.

	plein	partiel	Total
célibataire	3044.2	454.5	3498.7
	22.9%	3.4%	26.3%
	87.0%	13.0%	
	25.2%	38.1%	
autres	9045.9	738.46	9784.3
	68.1%	5.6%	73.7%
	92.5%	7.6%	
	74.8%	61.9%	
Total	12090.1	1192.9	13283
	91.0%	9.0%	100%

La valeur de la statistique du chi-carré dépend fortement de la prise en compte ou non des poids ; voir Table 8.8. Dans les tableaux *S*, *CH*, et *ZH4*, on observe une diminution suivant un facteur variant entre 1 et 33. Dans les tableaux *CF* et *ZH2*, par contre, la statistique augmente de 66%, resp. 43%.

TAB. 8.8 – Statistique du chi-carré X_c^2 sans et X_p^2 avec prise en compte des poids et degrés de liberté $df = (r-1)(c-1)$. Correction du premier ordre effectué dans SAS : $D_{..}$, D_R, D_C, $E[X_p^2]$, $\widehat{\delta} = E[X_p^2]/df$ et résultat $X_F^2 = X_p^2/\widehat{\delta}$.

Tableau	X_c^2	X_p^2	*df*	D_{++}	D_R	D_C	$E[X_p^2]$	$\widehat{\delta}$	X_F^2
S	11726.16	10490.13	1	107.97	38.96	66.73	2.29	2.29	4584.80
CH	370.69	96.11	3	83.69	22.59	39.35	21.75	7.25	13.26
CF	442.34	734.33	3	67.40	19.41	30.04	17.95	5.98	122.71
ZH2	65.10	93.38	1	36.44	5.74	23.20	7.50	7.50	12.46
ZH4	17.11	0.52	1	17.33	4.63	9.79	2.91	2.91	0.18

On note la différence de comportement entre *CH* et *CF* : diminution de la statistique pour l'un et augmentation pour l'autre. Chez les hommes (*CH*), l'effet des poids amène à une réduction de la différences entre proportions de partiels des jeunes et plus âgés. Le lien entre temps partiel et âge est donc moins fort. Chez les femmes *CF*, l'effet de la pondération est moins clair. On n'observe pas clairement une modification de la relation entre les deux variables. La tendance semble cependant aller vers un déséquilibre car le lien entre l'âge et le temps partiel est plus fort en prenant en compte la pondération ($X_p^2 > X_c^2$).

Les statistiques X_c^2 et X_p^2 indiquent un lien fortement significatif pour les tableaux *S*, *CH*, *CF* et *ZH2*, avec un risque inférieur à 0.1%.

Dans le cas du tableau *ZH4*, une interprétation basée sur X_c^2 indique un lien significatif entre l'état civil et le temps partiel (risque $\leq$ 0.1%). L'interprétation basée sur X_p^2 indique cependant le contraire (risque > 10%). On retrouve le risque d'interprétation erronée encouru lorsque les poids et le plan ne sont pas pris en compte.

La correction du premier ordre a un effet important sur X_p^2. Les valeurs sont réduites par le facteur $\widehat{\delta} = df/E[X^2]$ prenant des valeurs entre 2.3 et 7.5. Les valeurs corrigées X_F^2 indiquent toujours une forte dépendance pour les tableaux *S* et *CF* et le manque de dépendance pour le tableau *ZH4*. Au contraire des femmes, le lien entre la classe d'âge et le temps partiel devient moins clair pour les hommes (tableau *CH*). La p-value tombe à une valeur proche de 0.01. Elle n'est donc plus inférieure à 0.001. De même, le lien entre l'état civil et le temps partiel devient moins clair pour les plus jeunes (tableau *ZH2*). Il reste cependant significatif avec un risque inférieur à 0.1%.

Comme dans le cas des moyennes, des médianes et des coefficients de régression, on note que l'influence du plan d'échantillonnage est plus important dans le cas de la LSE98, échantillon en grappe, que dans le cas de l'ERC98, échantillon stratifié. On note également que les valeurs de corrections varient fortement de cas en cas. Il est donc difficile d'utiliser une correction estimée sur un tableau pour corriger la valeur d'un autre tableau. Ces valeurs dépendent en effet des variables, des poids, des informations sur l'échantillonnage, et des corrélation entre eux.

Chapitre 9

Conclusion

Une analyse basées sur des données d'enquêtes et ayant pour but d'obtenir des résultats pour une population finie doit prendre en compte les poids et le plan d'échantillonnage. Les poids permettent d'obtenir des estimateurs ponctuels non biaisés. Le plan, combiné avec les poids, permet d'avoir des estimateurs de variabilité non biaisés et donc adéquats pour l'inférence.

L'interprétation des résultats peut clairement être faussée si les effets de la pondération et de l'échantillonnage sont partiellement ou complètement négligés. Ces erreurs, conséquence d'une inférence effectuée sur des valeurs biaisées, se retrouvent dans toutes les méthodes d'analyse (différence entre des moyennes de groupes, régression, test d'indépendance entre variables catégorielles, etc.).

On n'observe aucune sur-estimation ou sous-estimation systématique des estimateurs ponctuels non pondérés par rapport aux estimateurs pondérés de l'ERC98 et de la LSE98 (biais technique). Par contre, l'effet du plan sur la variance (effet de sondage) est nettement supérieur pour la LSE98 car cette enquête est basée sur un échantillonnage en grappe.

Un soin particulier doit être mis dans le choix des estimateurs. Les estimateurs de la variance peuvent notamment être basés sur une méthode de linéarisation ou de rééchantillonnage, avec prise en compte ou pas du facteur de population finie, avec considération de sous-groupes comme des domaines ou pas, avec regroupement ou pas des strates avec un unique élément, etc. Il est également important de faire attention aux résultats basés sur de petits sous-ensembles de l'échantillon.

Il est peu probant d'utiliser des résultats obtenus avec un estimateur pour en déduire des résultats pour un autre estimateur (ex. sur la base d'une moyenne pour une médiane ou un coefficient d'un modèle de régression). L'effet des poids et du plan varie entre les enquêtes, entre les estimateurs et entre les sous-groupes considérés. Il dépend de nombreux éléments, dont les corrélations entre les variables utilisées dans l'analyse et celles définissant le plan et la pondération.

Dans la pratique, il est donc important d'appliquer, pour chaque estimateur, une méthode prenant en compte les poids et le plan. Toute analyse de la statistique publique devrait passer par une méthode complète dès que cela est possible. Une simplification pour cause de manque d'information ou l'utilisation de résultats obtenus dans une autre enquête, dans un autre domaine ou pour un autre estimateur, ne doit être faite qu'en dernier recourt et avec un grand esprit critique.

Les logiciels de statistique courants offrent des méthodes prenant en compte les poids et le plan

d’échantillonnage. SAS v8 est pratique et puissant pour les moyennes et totaux (PROC SURVEYMEANS) et les modèles de régression (PROC SURVEYREG). SAS v9 offrira des procédures pour les régressions logistiques (PROC SURVEYLOGISTIC) et pour les tests du chi-carré (PROC SURVEYFREQ). De son côté, Wesvar 4 offre plus de méthodes d’analyses (médiane, régression logistique) mais est moins pratique avec de gros jeux de données. Différents logiciels spécialisés et des macros peuvent être appliqués pour satisfaire à d’autres besoins.

Annexe A

Trucs dans SAS

A.1 Moyenne et variance de la moyenne pour l'ERC98

Dans PROC UNIVARIATE avec les poids, la valeur de $\sqrt{\widehat{V}_{SIR}(\widehat{\overline{y}}_{pond})}$ est obtenue par l'une des manières suivantes :

1. stdmean $\times\sqrt{(n-1)/n}$, si VARDEF=DF (appliqué dans l'exemple) ;
2. std/$\sqrt{\sum_i w_i}$, si VARDEF=N ;
3. std/$\sqrt{n}$, si VARDEF=WEIGHT.

En effet, la valeur stdmean=std/$\sqrt{\sum_i w_i}$ est donnée uniquement si VARDEF=DF et celle de std est construite comme suit :

$$\text{std} = \sqrt{\frac{\sum_{i=1}^{n} w_i (y_i - \widehat{\overline{y}}_{pond})^2}{d}} \tag{A.1}$$

avec $d = n - 1$ si VARDEF=DF (default), $d = n$ si VARDEF=N et $d = \sum_i w_i$ si VARDEF=WEIGHT.

Notons que PROC SURVEYMEANS utilise un développement de Taylor pour l'estimation de la variance d'une moyenne. Les résultats de cette procédure sans indication de strates et de poids ne sont pas identiques à ceux obtenus avec PROC UNIVARIATE sans poids ni plan.

Input SAS :

```
* base sans pondération ni plan;
proc univariate data=evelb;
 var EREVTOT;
 output out=inter n=n mean=mean stdmean=stdmean;
run; proc print noobs;run;

* pondération mais pas plan;
proc univariate data=evelb vardef=df;
 var EREVTOT;
 weight XGEWZP;
 output out=inter n=n mean=mean stdmean=stdmean;
run; proc print noobs;run;
```

```
* pondération et plan;
proc surveymeans data=evelb mean;
 var EREVTOT;
 weight XGEWZP;
 strata XSCHICHT;
 ods output Statistics=inter;
run; proc print data=inter noobs;run;
```

Output SAS :

```
* base sans pond ni plan;
                 n       mean       stdmean
               4369    8521.19     89.2630

* pondération mais pas plan;
                 n       mean       stdmean
              4369    8835.77      93.7125
 [Note, avec la correction sqrt[(n-1)/n]: std(mean)= 93.70]

* pondération et plan;

                      The SURVEYMEANS Procedure

                             Data Summary

                 Number of Strata                   7
                 Number of Observations          4369
                 Sum of Weights            5875912.98

                             Statistics
                                                        Std Error
Variable   Label                                Mean     of Mean
-----------------------------------------------------------------------
erevtot    Totaleinkommen der Haushalte    8835.769861 109.031353
-----------------------------------------------------------------------
```

A.2 Variance de la médiane non pondérée sans plan pour l'ERC98

Les options CIPCTLDF et CIPCTLNORMAL dans la procédure UNIVARIATE permettent d'avoir une estimation de l'intervalle de confiance de différents percentiles. L'option CIPCTLDF, qui utilise une méthode distribution-free, est sélectionnée pour nos besoins. Cette option n'est pas disponible avec une pondération.

Input SAS :

```
proc univariate data=evelb CIPCTLDF;
 var EREVTOT;
 output out=inter n=n median=median;
run;
```

Output SAS :

```
                      The UNIVARIATE Procedure
          Variable:  erevtot  (Totaleinkommen der Haushalte)

                                Moments

   N                         4369    Sum Weights               4369
```

```
   Mean               8521.18829    Sum Observations    37229071.6
...
                      Basic Statistical Measures

            Location                    Variability

        Mean     8521.188     Std Deviation              5900
        Median   7469.450     Variance                34811672
...
                       Quantiles (Definition 5)

               95\% Confidence Limits    -------Order Statistics-------
Quantile         Distribution Free      LCL Rank  UCL Rank
Coverage ...
50% Median      7331.65      7634.96        2121      2251       95.07
...
```

```
[Note: l'IC estimé pour la médiane 7469.450 est donné par
[7331.65;7634.96]. On estime l'écart-type comme suit:
max(7469.450-7331.65, 7634.96-7469.450)/1.96 = 84.44.]
```

A.3 Variance de la médiane pondérée sans plan

La macro variaMed, se basant sur la méthode de Woodruff, est utilisée pour estimer la variabilité d'une médiane pondérée sans tenir compte du plan. Elle est fournie dans ce rapport comme complément pour la compréhension de la méthode de Woodruf et non pas comme recommandation pour l'application. La méthode complète, avec poids et plan, est recommandée (ex. macro de M. Graf).

```
%macro variaMed (dset, y, w, resIC);
********************; /*
 But: estimation de l'écart-type de la médiane pondérée par le biais de
      la méthode de Woodruff, i.e. détermination d'un intervalle de confiance
      dans la fonction de répartition puis retour dans les valeurs de base
      pour déterminer un IC pour la médiane et estimer l'écart-type.

   INput:
      dset: data set
      y   : variable d'intérêt
      w   : poids

   OUTput:
      resIC : data set avec variables
         P_med                        : médiane
         P_low                        : borne inférieure intervalle
         P_up                         : borne supérieure intervalle
         demi1=(P_med-P_low)          : demi intervalle inférieur
         demi2=(P_up-P_med)           : demi intervalle supérieur
         std = max(demi1,demi2)/1.96 : estimation de std(P_med) sans tenir compte du plan
                                       (estimation conservatrice)
 Auteur: Anne Renaud, 2003.
*/

* médiane, avec def variable globale;
proc univariate data=&dset noprint;
 var &y;
 weight &w;
 output out=inter median=med;
run; data _NULL_; set inter; call symput("med", med); run;

* calcul des indicatrices et de la variable;
* num=delta_i*w_i (numérateur ratio);
data indic;
```

```
 set &dset;
 delta=1;
 if &y>&med then delta=0;
 num=&w*delta;
run;

* estimation IC du ratio sum num/sum IBGRS;
proc surveymeans data=indic ratio clm;
 var num &w;
 ratio num/&w;
 ods output ratio=inter;
run;

* sortie des low et up de l'IC en global;
data _NULL_; set inter;
 lowF=LowerCL*100;
 call symput("lowF", lowF);
 upF=UpperCL*100;
 call symput("upF", upF);
run;

* recherche des valeurs de l'IC dans la base;
proc univariate data=&dset noprint;
 var &y;
 weight &w;
 output out=inter pctlpts= &lowF 50 &upF
               pctlpre=P_
               pctlname=low med up;
run;

* résultats;
data &resIC;
 set inter;
 demi1=(P_med-P_low);
 demi2=(P_up-P_med);
 std = max(demi1, demi2)/1.96;
run;

%mend variaMed;
*******************************;
```

A.4 Moyenne et variance de la moyenne pour la LSE98

Le calcul complet de l'estimateur de la moyenne et de son écart-type $\widehat{Std}(5) = \widehat{Std1}_p(\widehat{\overline{y}}_{pond})$ est obtenu avec la procédure SURVEYMEANS et les options "weight" pour tenir compte des poids, "strata" et "cluster" pour tenir compte du plan, "total=" pour tenir compte la correction de la population finie, et "domain" pour tenir compte du fait que les sous-groupes sont des domaines et pas des échantillons indépendants. Nous choisissons de laisser regrouper les strates avec un unique élément et n'utilisons donc pas l'option NOCOLLAPSE.

Input SAS :

```
proc surveymeans data=lse98ex4 mean clm total=TotStraIn;
 var MBLS;
 weight GEWIBGRS;
 strata STRA;
 cluster BURNR_N;
 domain dom;
 ods output domain=inter;
run; proc print data=inter (drop=DomainLabel MBLS) noobs;run;
```

```
[Note: le data set TotStraIn contient 200 observations (strates)
avec les variables STRA et _total_=BURNH (nombre d'entreprises
dans la strate dans le cadre de sondage).]

[Note: variable dom=1 pour les hommes, respectivement dom=2 pour
les femmes, de la région lémanique.]
```

Output SAS :

```
                    The SURVEYMEANS Procedure

                           Data Summary

                Number of Strata                200
                Number of Clusters             7006
                Number of Observations       415787
                Sum of Weights           2068103.36

...

                            MBLS_Lower                    MBLS_Upper
     dom      MBLS_Mean         CLMean    MBLS_StdErr         CLMean
...
       1    6412.449296    6216.251480     100.085024    6608.647112
       2    4887.595161    4768.633148      60.685283    5006.557174
```

A.5 Coefficients de régression et prévisions avec IML pour la LSE98

Prévision de MBLS selon le modèle de la Table 7.4.

```
* coefficient et matrice var-covar de REG et SURVEYREG;
*****************************************************;
proc reg
data=regH3040prev;
 model LNMBLS=EDU1 EDUEXPR EDUTEN EDUEXPRTEN /COVB;
 weight IBGRS;
 ods output ParameterEstimates=regpara CovB=regcovB;
run;

proc surveyreg data=regH3040prev;
 model LNMBLS=EDU1 EDUEXPR EDUTEN EDUEXPRTEN / COVB deff;
 strata STRA;
 cluster BURNR_N;
 weight GEWIBGRS;
 ods output ParameterEstimates=surpara CovB=surcovB;
run;

* Jeu de données pour les prévisions;
************************************;
data inter; /*extrait*/
 input EDU EXPR TEN @@;
 datalines;
8   0   1  ... 16  20  25
   ;
run;

data donpredB;
 set inter;
   interc=1;
   EDU1=EDU/10;
```

```
   EDUEXPR=EDU1*EXPR/10;
   EDUTEN=EDU1*TEN/10;
   EDUEXPRTEN=EDU1*EXPR/10*TEN/10;
run;

* Calcul valeurs prédites et std;
*******************************;
proc iml; /*début IML*/

* lecture données dans IML;
use donpredB;
read all into donpredB;
show contents;
use regpara;
read all into regpara;
use regcovB;
read all into regcovB;
use surpara;
read all into surpara;
use surcovB;
read all into surcovB;

* choix variables;
don=donpredB[,4:8];

* pred et std pred pour reg et surveyreg;
regP=don*regpara[,2];
regStd=sqrt(vecdiag(don*regcovB*t(don)));
surP=don*surpara[,1];
surStd=sqrt(vecdiag(don*surcovB*t(don)));
result=donpredB||regP||regStd||surP||surStd;

* sortie pour continuer dans SAS;
create result from result[colname=
       {EDU  EXPR  TEN interc EDU1  EDUEXPR  EDUTEN  EDUEXPRTEN
        regP regStd surP surStd}];
append from result; show contents;

quit; /* fin IML*/

* Data set pour résultats;
************************;
data predplus;
 set result;
 * mean predicted en MBLS et IC: REG;
 regPex=exp(regP);
 regL=exp(regP-1.96*regStd);
 regH=exp(regP+1.96*regStd);
 * mean predicted en MBLS et IC: SURVEYREG ;
 surPex=exp(surP);
 surL=exp(surP-1.96*surStd);
 surH=exp(surP+1.96*surStd);
run;
```

A.6 Calcul de χ^2_c et χ^2_p pour l'ERC98

Lors de l'utilisation de la procédure PROC FREQ de SAS avec les poids, le calcul de la statistique de Pearson est effectué selon la formule :

$$X^2(N) = \widehat{N} \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(\widehat{p}_{ij} - \widehat{p}_{i+}\widehat{p}_{+j})^2}{\widehat{p}_{i+}\widehat{p}_{+j}} \tag{A.2}$$

avec $\widehat{N} = \sum_{k \in s} w_k$. Les observations sont donc démultipliées selon le poids indiqué. La taille de l'échantillon est prise égale à la taille estimée de la population $\widehat{N}$ et non pas à sa vraie taille n. Il s'agit donc de corriger la valeur $X^2(N)$ comme suit :

$$X_p^2 = \frac{n}{\widehat{N}} X^2(N) \tag{A.3}$$

ou d'appliquer PROC FREQ avec les poids corrigés $w_k^* = w_k(n/\sum_k w_k)$ (voir le code).

Input SAS :

```
* data set avec poids corrigés;
proc univariate data=libkult.evelb98d noprint;
 var XGEWZP;
 output out=inter n=n sum=sum;
run;
data _NULL_;
 set inter;
 call symput('n', n);
 call symput('sum',sum);
run;
data lbcor;
 set libkult.evelb98d;
 poidscor = XGEWZP * &n/&sum;
run;

* test sans poids;
proc freq data=lbcor;
 tables LZUFR2*WOHNTYP2 /chisq;
run;

* test avec poids corrigés;
proc freq data=lbcor;
 tables LZUFR2*WOHNTYP2/chisq;
 weight poidscor;
run;
```

Output SAS :

```
* test sans poids;
               Frequency|
               Percent  |
               Row Pct  |
               Col Pct  |1.Mieter|2.Eigent|  Total
                        |        |ümer    |
               ---------+--------+--------+
                      0 |   2657 |   1526 |   4183
                        |  60.81 |  34.93 |  95.74
                        |  63.52 |  36.48 |
                        |  94.76 |  97.51 |
               ---------+--------+--------+
                      1 |    147 |     39 |    186
                        |   3.36 |   0.89 |   4.26
                        |  79.03 |  20.97 |
                        |   5.24 |   2.49 |
               ---------+--------+--------+
               Total        2804     1565     4369
                           64.18    35.82   100.00


            Statistics for Table of lzufr2 by wohntyp2

       Statistic                     DF       Value      Prob
       ------------------------------------------------------
```

Chi-Square	1	18.6422	<.0001
Likelihood Ratio Chi-Square	1	20.2226	<.0001
Continuity Adj. Chi-Square	1	17.9735	<.0001
Mantel-Haenszel Chi-Square	1	18.6379	<.0001
Phi Coefficient		-0.0653	
Contingency Coefficient		0.0652	
Cramer’s V		-0.0653	

Fisher’s Exact Test

Cell (1,1) Frequency (F)	2657
Left-sided Pr <= F	5.359E-06
Right-sided Pr >= F	1.0000
Table Probability (P)	2.964E-06
Two-sided Pr <= P	1.024E-05

Sample Size = 4369

```
* test avec poids corrigés;
```

Frequency Percent Row Pct Col Pct	1.Mieter	2.Eigentümer	Total
0	2558.3 58.56 61.17 94.62	1623.8 37.17 38.83 97.51	4182.1 95.72
1	145.36 3.33 77.79 5.38	41.496 0.95 22.21 2.49	186.85 4.28
Total	2703.68 61.88	1665.32 38.12	4369 100.00

Statistics for Table of lzufr2 by wohntyp2

Statistic	DF	Value	Prob
Chi-Square	1	20.9441	<.0001
Likelihood Ratio Chi-Square	1	22.5727	<.0001
Continuity Adj. Chi-Square	1	20.2454	<.0001
Mantel-Haenszel Chi-Square	1	20.9393	<.0001
Phi Coefficient		-0.0692	
Contingency Coefficient		0.0691	
Cramer’s V		-0.0692	

```
...
```

Sample Size = 4368.9999831

Annexe B

Trucs dans Wesvar

Le logiciel Wesvar existe sous la forme d'un logiciel complet sur PC et sous la forme d'un module de SPSS.

Wesvar fonctionne par le biais de menus. Les données doivent être transférées sous le format Wesvar. Durant cette étape, on indique le nom des variables identifiant les strates et, si nécessaire, les grappes ainsi que le poids de chaque élément.

Différentes contraintes doivent être prise en compte lors de la préparation des données. Il est par exemple demandé d'avoir un identificateur unique (1,2 ,3 ..,n) dans le jeu de données dans son ensemble et une numérotation des unités dans chaque strate (ex. strate 1 : observations 1, 2, 3, n_1, strate 2 : observations 1, 2, 3 .., n_2).

La méthode de reéchantillonnage est choisie au moment de la construction du jeu de données contenant les poids de "replications". Dans le cas de strates avec plus de 2 éléments, il s'agit de choisir la méthode "JKn". Un traitement spécial est exigé si certaines strates n'ont qu'un seul élément : liste à entrer dans Wesvar ou traitement externe.

La création des poids de reéchantillonnage dans Wesvar demande de grandes ressources en temps et place si le fichier possède beaucoup d'unités du premier niveau (grappes ou éléments). En cas de gros fichier, il est donc préférable de préparer les replicates hors du programme WesVar après avoir effectué des regroupements d'éléments.

Une fois que le fichier est préparé avec les replicates, il peut être utilisé pour différentes analyses tels que les estimations de moyennes, de médianes ou encore des estimations de coefficients dans des modèles de régression.

Bibliography

Ardilly, P. (1994). *Les techniques de sondage*. Editions Technip, Paris.

Bragner, K., Gazareth, P., Modetta, C., Röthlisberger, P., Schmid, B., Schön-Bühlmann, J. and Tillmann, R. (2002). Revenu et bien-être. Niveau de vie et désavantages sociaux en Suisse. OFS.

Brandenberg, A. (1997). Lohnvergleich zwischen dem privaten und öffentlichen Sektor (Bund) in der Schweiz. Rapport non publié.

Brogan, D. J. (1998). *Pitfalls of using standard statistical software packages for sample survey data. In : Encyclopedia of Biostatistics*. John Wiley & Sons, eds. Armitage and Colton. New York. edition.

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

Chambers , R. L. and Skinner, C. J. Eds (2003). *Analysis of Survey Data*. John Wiley & Sons.

Cochran, W. G. (1977). *Sampling Techniques*. John Wiley & Sons.

Cornali, A. (2000). Enquête sur les revenus et la consommation 1998 : Développement des méthodes pour l'analyse de données. Mémoire pour le diplôme postgrade, Université de Neuchâtel.

DHEW (1975). Distribution and properties of variance estimators for complex multistage probability samples. An empirical distribution. Series 2 : Data Evaluation and Methods Research 65, U.S. Departement of Health, Education, and Welfare.

DHHS (1982). A statistical methodology for analysing data from a complex survey : The first national health and nutrition examination survey. Series 2 : data evaluation on methods research 92, U.S. Departement of Health and Human Services.

Fay, R. E. (1985). A jackknife chi-square test for complex samples. *Journal of the American Statistical Association*, 80 :148–157.

Gabler, S., Hoffmeyer-Zlotnik, H.P. and Krebs, D. (Hrsg.) (1994). *Gewichtung in der Umfragepraxis*. Westdeutscherverlag.

Graf, M. (2002). Enquête suisse sur la structure des salaires 2000. plan d'échantillonnage, pondération et méthode d'estimation pour le secteur privé. Rapport de méthodes 338-0010, OFS.

Holt, D., Smith, T.M.F. and Winter P.D. (1980). Regression analysis of data from complex surveys. *J. R. Statist. Soc., A*, 143 :474–487.

Kish, L. (1992). Weighting for Unequal P_i. *Journal of Official Statistics*, 8(2) :183–200.

Kish, L. and Frankel, M. (1974). Inference from complex samples. *Journal of the Royal Statistical Society*, 36 :1–37. (with discussion).

Korn, E. and Graubard, B. (1990). Simultaneous testing of regression coefficients with complex survey data : Use of Bonferroni *t* statistics. *The American Statistician*, 44(4) :270–276.

Lee E. S. , Forthofer, R. N. and Lorimer, R. J. (1989). *Analysing Complex Survey Data*. Sage Publications.

Little, R. (2004). To model or not to model ? Competing modes of inference for finite population sampling. *Journal of the American Statistical Association*, 99(466) :546–556.

Lohr, S. L. (1999). *Sampling : Design and Analysis*. Duxbury Press.

Nathan, G. and Holt, D. (1980). The effect of survey design on regression analysis. *J. R. Statist. Soc. B*, 42 :377–386.

OFS (1999). Enquête sur les revenus et la consommation 1998 (ERC98). Bases. Actualités OFS.

OFS (2000a). Enquête sur les revenus et la consommation 1998 (ERC98). Premiers résultats. Actualités OFS.

OFS (2000b). L'enquête suisse sur la structure des salaires 1998. Résultats commentés et tableaux. Statistique de la Suisse. OFS.

Park, I. and Lee, H. (2001). The design effect : Do we know all about it ? *Proceedings of the Annual Meeting of the American Statistical Association*.

Park, I. and Lee, H. (2002). Nouvel examen des effets de plan de sondage dans le cas de l'échantillonnage avec probabilités inégales. *Le Statisticien d'enquête*, Juillet 2002.

Peters, R. (1997). Enquête sur la structure des salaires en Suisse : Établissement du plan d'échantillonnage. Rapport de méthodes, OFS.

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

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review*, 61(2) :317–337.

Rao, J. and Scott, A. (1981). The analysis of categorical data from complex sample surveys : Chi-squared tests for goodness of fit and independence in two-way tables. *Journal of the American Statistical Association*, 76(374) :221–230.

Rao, J. and Scott, A. (1984). On chi-square tests for multiway contingency tables with cells proportions estimated from survey data. *The Annals of Statistics*, 12 :46–60.

Rao, J. and Scott, A. (1987). On simple adjustments to chi-square tests with sample survey data. *The Annals of Statistics*, 15 :385–397.

Rubinfeld, D. L. (2000). Reference guide on multiple regression. In : Reference manual on scientific evidence, Second edition, Federal Judicial Center, http ://air.fjc.gov/public/fjcweb.nsf/pages/16.

Saporta, G. (1990). *Probabilités, Analyse Des Données et Statistique*. Editions Technip.

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

Skinner, C. J. (1989). *Domain Means, Regression, and Multivariate Analysis. In : Analysis of Complex surveys*, chapter 3, pages 59–88. John Wiley & Sons, Edited by C. J. Skinner, D. Holt and T.M.F. Smith edition.

Skinner , C. J. , Hold, D. and Smith, T.M.F. Eds (1989). *Analysis of Complex Surveys*. John Wiley & Sons.

Skinner, C. J., Holmes, D. J. and Smith, T. M. F. (1986). The effect of sample design on principal component analysis. *Journal of the American Statistical Association*, 81(395) :789–798.

Thomas, D. R. , Singh, A. C. and Roberts, G. R. (1996). Tests of independence on two-way tables under cluster sampling : An evaluation. *International Statistical Review*, 64 :295–311.

Tillé, Y. (2001). *Théorie des sondages. Echantillonnage et estimation en populations finies*. Dunod, Paris.

Westat (2000). *WesVar^{TM} 4.0 User's Guide*. Westat, Rockville, MD.

Wolter, K. M. (1985). *Introduction to variance estimation*. Springer Series in Statistics.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Publikationsprogramm BFS

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

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

Programme des publications de l'OFS

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

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

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

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

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

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

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

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

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

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

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

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

L'analyse de données d'enquêtes par échantillon requiert l'utilisation de méthodes spécifiques qui tiennent compte du plan d'échantillonnage et de la pondération des observations.

Une introduction théorique et des exemples d'application sont présentés dans le cadre de trois types d'analyses courantes: les statistiques descriptives (moyennes et médianes), l'estimation de coefficients dans des modèles de régression linéaire et les tests d'indépendance entre variables catégorielles (chi-carré).

Les exemples basés sur les données de l'Enquête sur les revenus et la consommation de 1998, en particulier le module «Conditions de vie», et celles de l'Enquête suisse sur la structure des salaires de 1998 permettent de voir l'effet de l'utilisation de méthodes adéquates en comparaison avec des méthodes classiques. Ces dernières produisent en effet des résultats biaisés qui peuvent conduire à des conclusions erronées.

N° de commande: 338-0024
Commandes: 032 713 60 60 Fax: 032 713 60 61 E-Mail: order@bfs.admin.ch
Prix: gratuit
ISBN 3-303-00295-9