Schweizer Haushalt-Panel
Panel suisse de ménages
Swiss Household Panel

RECOMMANDATIONS ET EXEMPLES PRATIQUES CONCERNANT L'APPLICATION DES PONDÉRATIONS

Décembre 2007

Sandra PLAZA

Eric GRAF

Résumé

Pour que les résultats d'analyses sur des données d'enquêtes telles que le Panel Suisse de ménage (PSM) ou de l'enquête sur les Revenus et les Conditions de Vie (SILC) ne soient pas erronés, le chercheur ne doit pas oublier d'utiliser des méthodes adéquates tenant compte des pondérations et de la complexité du plan d'échantillonnage.

L'introduction théorique (explication de notions et présentation de l'univers des enquêtes PSM et SILC) est suivi par la présentation des syntaxes les plus couramment utilisées, dans les logiciels SAS et SPSS, avec d'une part les syntaxes traditionnelles et d'autre part celles prenant en considération le plan de sondage.
Les syntaxes exposées sont celles permettant de réaliser des fréquences et des tableaux croisés, des analyses descriptives (moyenne, écart-type, intervalles de confiance…), des régressions linéaires, des régressions logistiques et des tests paramétriques ou non-paramétriques.

Des exemples d'application de procédures implémentées dans les logiciels mentionnés illustrent les effets des pondérations sur les analyses. Ces exemples montrent l'importance de l'utilisation des poids. On insiste également sur les résultats obtenus avec les procédures dans SAS et SPSS calculant la variance de façon appropriée pour des données d'enquêtes telles que le Panel Suisse de Ménages (PSM) et l'Enquête sur les Revenus et les Conditions de Vie (SILC).

Mots-clé

pondération, calcul de variance, SAS, SPSS, tableaux de fréquences, moyennes, test du Chi-carré, T-test, test de Wilcoxon, régression linéaire, régression logistique, Taylor, PROC SURVEY(…) de SAS et COMPLEX SAMPLE de SPSS

TABLE DES MATIERES

1. Introduction

Lors d'analyses de données d'enquêtes, le chercheur sous-estime souvent l'importance des pondérations et obtient des résultats erronés. L'utilisation des pondérations n'est cependant pas suffisante. Les procédures traditionnelles des différents logiciels statistiques supposent que l'échantillon utilisé est tiré selon un plan aléatoire simple, ce qui n'est souvent pas le cas.
Afin de convaincre les chercheurs sceptiques et d'éviter ces erreurs, ce document a pour objectif de montrer et d'expliquer l'impact des poids dans les analyses les plus courantes (analyses descriptives, tests, régression linéaire, régression logistique) et également de présenter les procédures permettant de tenir compte des spécificités du plan de sondage. Les différents exemples se basent sur les données du Panel Suisse de Ménages (PSM) et/ou de l'Enquête sur les Revenus et les Conditions de Vie (SILC).

1.1. But du document

Les questions de pourquoi, quand et comment utiliser les poids figurent parmi les questions les plus fréquemment posées et sont fondamentales dans les analyses des données d'enquêtes et les estimations y relatives. Les réponses à ces questions peuvent dépendre du contexte et du type d'analyses menées. Il y a parfois des arguments décisifs pour déterminer s'il est approprié ou pas, nécessaire ou pas de pondérer. Dans les cas d'une enquête complexe et de taille limitée comme le PSM (ou comme SILC), nous pensons que dans la plupart des cas la réponse est très claire : *le plus souvent il est tout aussi nécessaire qu'utile de pondérer l'échantillon de données afin de compenser les imperfections dudit échantillon*.
Premièrement, la pondération est construite pour tenir compte des différences dans les probabilités d'inclusion des unités dans l'échantillon considéré. Une telle pondération est essentielle si les différences sont grandes. Les taux importants de non réponse ou d'attrition au cours des vagues d'une enquête panel sont une autre raison importante pour pondérer. La non réponse et l'attrition sont souvent non seulement importantes mais aussi sélectives, et sont par exemple plus marquées pour des ménages dont le revenu se situe dans les extrêmes de la distribution.
D'autre part les poids ont été calés sur des totaux connus de la population suisse (âge, sexe, nationalité, état civil, 7 grandes régions géographiques de suisse), ce qui améliore la représentativité de l'échantillon. Un tel calage réduit les biais de l'échantillon dus à la non réponse, la sous-couverture ou d'autres distorsions, il réduit aussi les variances dans certains cas.

L'utilisation des poids est donc impérative. Si l'on n'applique pas les poids appropriés, les estimations effectuées ne peuvent pas être considérées comme représentatives de la population observée. Le document (Graf, 2007)[1] donne quelques conseils quant au choix de la pondération à utiliser parmi tous les jeux de poids produits.
Dans le présent document nous nous concentrerons sur quelques exemples d'analyses des données du PSM(-SILC) pour illustrer l'utilité mais aussi les effets parfois importants et d'autre fois limités des poids.

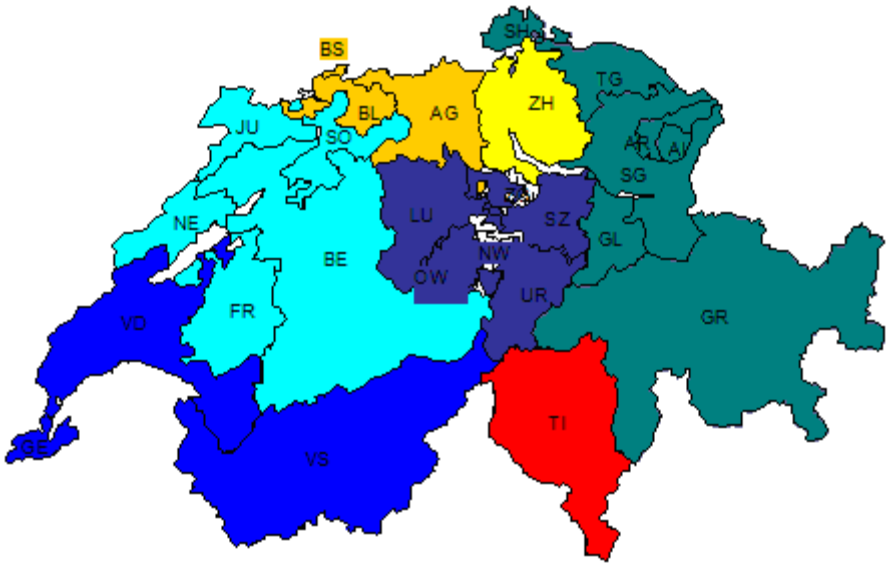
[1] Téléchargeable sur http://www.swisspanel.ch/doc/methodology.php?lang=fr&pid=8

1.2. Univers de l'enquête

Le PSM couvre toutes les personnes résidant dans des ménages privés de Suisse et possédant un raccordement téléphonique (fixe ou mobile) inscrit dans l'annuaire téléphonique. Les personnes vivant dans des homes, dans des institutions, dans des ménages collectifs, en prison ne sont pas couvertes par l'enquête. Les personnes non joignables par téléphone ou possédant un raccordement téléphonique non inscrit ne sont pas couvertes par l'enquête (sous-couverture).
On estime qu'en 2004, lors de la sélection de l'échantillon du PSM_II, 98.5% des ménages privés possédaient un raccordement téléphonique. Par ailleurs le SRH[2] (base de sondage pour les enquêtes auprès des ménages) couvrait quant à lui environ 93% de ces ménages. En 1999, lors de la sélection de l'échantillon du PSM_I, la couverture du SRH était d'environ 95%.

1.3. Echantillon

Figure 1 : les 7 grandes régions géographiques de Suisse (régions NUTS II)



Les échantillons sont stratifiés par grande région, proportionnellement au nombre de numéros de téléphone du SRH par strate, sans sur-échantillonnage régional. Théoriquement un numéro de téléphone égale un ménage, c'est grossièrement vrai.[3] On a tiré selon les proportions suivantes:

[2] Le SRH, registre des enquêtes auprès des ménages de l'OFS, est la base de sondage dans laquelle ont été tirés les échantillons de 1999 et 2004 du PSM.
[3] Il peut arriver qu'on ait des doublons dans le SRH (=deux numéros de téléphone différents menant au même ménage), ou des numéros de téléphone professionnel (2%), de résidences secondaires (<1%).

Tableau 1 : proportions selon lesquelles on a tiré les échantillons des enquêtes PSM et SILC.

STRATES	**PSM_II et SILC (2004, 2005)**	**PSM_I (1999)**
Région lémanique (VD, VS, GE)	18.22%	19.13%
Espace Mittelland (BE, FR, SO, NE, JU)	22.92%	22.40%
Suisse du Nord-Ouest (BS, BL, AG)	13.86%	13.18%
Zurich (ZH)	18.22%	17.30%
Suisse orientale (GL, SH, AR, AI, SG, GR, TG)	13.70%	14.48%
Suisse centrale (LU, UR, SZ, OW, NW, ZG)	8.75%	8.75%
Tessin (TI)	4.33%	4.76%
Total	100%	100%

Les enquêtes SILC et PSM_II ont exactement le même plan d'échantillonnage. Plan qui est identique à celui de 1999 pour le PSM_I à la différence près que les pourcentages définissant les proportions par strates provenaient des recensements respectivement de 2000 (pour SILC, PSM_II) et de 1990 (pour PSM_I).

1.4. Erosion/attrition

Tableau 2 : nombre d'individus/ménages interrogés dans l'enquête PSM. L'attrition relative est donnée par le nombre de cas interrogés l'année n sur le nombre de cas l'année (n-1).

	Niveau individus			*Niveau ménages*			*Niveau individus*			*Niveau ménages*		
Année	PSMI	Attrition relative	Attrition absolue	PSMI	Attrition relative	Attrition absolue	PSMII	Attrition relative	Attrition absolue	PSMII	Attrition relative	Attrition absolue
1999	7799	100%	100%	5074	100%	100%						
2000	7071	91%	91%	4425	87%	87%						
2001	6593	93%	85%	4139	94%	82%						
2002	5691	86%	73%	3581	87%	71%						
2003	5220	92%	67%	3227	90%	64%						
2004	4406	84%	56%	2831	88%	56%	3648	100%	100%	2538	100%	100%
2005	3885	88%	50%	2457	87%	48%	2648	73%	73%	1909	75%	75%
2006	4093	105%	52%	2534	103%	50%	2570	97%	70%	1682	88%	66%

Comme on peut le voir dans le tableau ci-dessus, l'attrition (ou l'érosion) des échantillons au cours des vagues, i.e. la perte des personnes et ménages répondants, est un point très important. A sa 7^{e} vague en 2005, le PSM_I ne contenait plus que la moitié de ses répondants de 1999.

De gros efforts sont faits, d'abord pour limiter la non réponse, ensuite pour améliorer les modèles qui tentent de la corriger dans les pondérations. Malgré cela, l'attrition n'étant pas uniforme, certaines de ses caractéristiques échappent aux modèles. On sait par exemple que pour des variables non ou peu corrélées aux variables utilisées pour modéliser la non réponse et dans les calages de la pondération, l'utilisation ou la non utilisation des poids dans les analyses change peu les résultats. Mais même dans ces cas-là, l'utilisation des poids permet tout de même un calcul de variance plus juste et réaliste (voir section suivante).

1.5. Calcul de variance, tests, intervalles de confiance, …

Les concepts de *variance*, de *coefficient de variation* et d'*intervalle de confiance* fournissent une indication de l'ampleur de la variabilité d'échantillonnage. Cependant, nous rappelons aux utilisateurs que les estimations de ces grandeurs[4] fournies par les procédures classiques de la plupart des logiciels statistiques courants (SAS, SPSS, Stata, MS Excel,…) sont erronées lorsque les données analysées proviennent d'enquêtes complexes comme PSM ou SILC. Ces procédures supposent entre autres qu'on a procédé à un échantillonnage aléatoire simple. C'est donc dire qu'elles ne tiennent pas compte des caractéristiques spéciales du plan de sondage de PSM/SILC comme la stratification, ni des corrections apportées pour la non réponse, ni de la post-stratification (calage sur marge). Même si la plupart de ces procédures permettent le recours à la pondération dans la production d'estimations, elles ne prennent pas correctement en compte toutes les spécificités de nos enquêtes lors du calcul des estimations de la variance, un élément essentiel pour la plupart des tests statistiques. ***En fait, on a tendance à sous-estimer ladite variance et par conséquent les intervalles de confiance liés à des estimations ainsi que les seuils de significativité dans les tests.***

Les procédures PROC SURVEYFREQ, PROC SURVEYMEANS, PROC SURVEYREG PROC SURVEYLOGISTIC, etc. du logiciel SAS permettent de calculer une variance selon la méthode de *linéarisation en série de Taylor* et donnent des résultats très similaires (du moins pour les enquêtes PSM et SILC) à ceux produit par des méthodes de ré-échantillonnage plus lourdes à implémenter comme le *bootstrap* ou le *jackknife*. De ce fait, pour les utilisateurs de SAS, nous recommandons d'utiliser ces procédures qui, assurément, fournissent des résultats bien plus corrects que ceux fournit par les procédures standard (i.e. PROC FREQ, PROC MEANS, PROC REG, PROC LOGISTIC, etc.). Ce document détaille et motive cette recommandation dans la section 2.2 et le chapitre 3.

Le logiciel Stata permet de prendre les mêmes mesures en ajoutant les commandes «`svyset`» et «`svy:`» à la syntaxe conventionnelle. Nous ne détaillerons pas d'exemples d'analyses avec le logiciel Stata dans ce document, mais elles s'appliquent exactement dans le même contexte et pour les mêmes raisons que les PROC SURVEY (…) de SAS. Certains auteurs (Siller, 2005) ont d'ailleurs déjà expérimenté qu'on arrivait ainsi exactement aux même résultats qu'avec SAS.

Tous les exemples d'analyses faits avec les PROC SURVEY (…) de SAS de ce document ont été reproduits avec SPSS qui offre les mêmes possibilités (voir section 2.3 et chapitre 3). Il est donc maintenant à la portée de presque tous les utilisateurs, quel que soit le logiciel qu'ils utilisent, de calculer des variances par une méthode adaptée pour analyser les données d'enquêtes comme PSM et SILC.

1.6. Longitudinal ou transversal

Nous clôturons cette introduction en rappelant que les pondérations transversales se rapportent toujours à l'année en cours, tant pour les ménages que pour les individus, tandis que les pondérations longitudinales (individus) extrapolent toujours à la population résidante en Suisse de 1999 pour le PSM_I, à celle de 2004 pour le PSM_II. La pondération longitudinale combinée PSM_I & PSM_II se rapporte également à la population résidante en Suisse de 2004.

Toutes les pondérations transversales produites extrapolent à la population résidante dans des ménages privés de Suisse en l'année de la vague courante.

[4] Nous pensons aux grandeurs telles que les variances, coefficients de variation, mais aussi les résultats des tests d'hypothèse (par exemple, valeurs de p accompagnant les statistiques t ou statistiques de Pearson).

2. Présentations des procédures

Ce chapitre est consacré à la présentation des différentes procédures utilisées par la suite. On y présentera d'abord les procédures traditionnelles puis les procédures SURVEY de SAS. On consacrera également une section aux différentes procédures du logiciel SPSS, l'objectif étant de montrer qu'il est possible de réaliser exactement les mêmes analyses avec l'un ou l'autre de ces logiciels.

2.1. Les procédures traditionnelles de SAS

On présente ici les commandes disponibles dans SAS pour réaliser des tableaux de fréquences, des analyses descriptives (moyenne, écart-type, minimum, maximum, …), des régressions linéaires, des GLM ainsi que des régressions logistiques. Toutes les commandes présentées ci-dessous font l'hypothèse que l'échantillon a été tiré selon un plan aléatoire simple et elles ignorent la non-réponse. Elles ne fournissent donc pas une bonne estimation de la variabilité de l'échantillon lorsque celui-ci suit un plan de sondage plus complexe.

2.1.1. La procédure pour des fréquences et des tableaux croisés

La procédure PROC FREQ fournit des tableaux de fréquence de une à n entrées et des tableaux de contingence. Lors de tableaux à deux entrées, la procédure calcule des tests et des mesures d'association.

Ci-dessous, nous avons la syntaxe à utiliser pour réaliser cette procédure.

PROC FREQ <options> **data**= *nom de la base de données* ; • Si on ne cite pas la base de données, la procédure va automatiquement prendre la dernière base de données créée. **TABLES** *demande de tableau spécifiques* </options> ; • Permet de spécifier les tableaux de fréquences ou de contingence et demande les tests et mesures d'association.	Commandes requises pour réaliser la procédure
BY *variables* ; • Calcule des tableaux de fréquences pour chaque observation par rapport aux groupes définis par les variables de la commande BY. **EXACT** *options statistiques* </option de calculs> • Demande des tests exacts ou intervalles de confiance pour les statistiques spécifiées. **OUTPUT** <OUT= nom du nouveau jeu de données> options ; • Crée une base de données avec les statistiques calculées par la procédure. **TEST** options ; • Donne les tests asymptotiques pour les mesures d'association spécifiées. **WEIGHT** *variable* </option[5]>; • Identifie la variable qui va pondérer chaque observation.	Commandes supplémentaires
RUN;	

[5] Pour la commande WEIGHT nous avons l'option ZEROS. Cette option inclut les observations qui ont un poids dont la valeur est zéro. Par défaut la procédure ignore les observations qui ont un poids égal à zéro.

2.1.2. La procédure pour des statistiques descriptives (moyenne, écart-type, minimum, maximum…)

Cette procédure sert à calculer des statistiques descriptives telles que la moyenne, l'écart-type, le minimum et le maximum ainsi que des intervalles de confiance de la moyenne. Elle permet aussi de faire des tests sur la moyenne.
Ci-dessous se trouve les commandes à utiliser pour réaliser cette procédure.

PROC MEANS <options> **data**= *nom de la base de données* ; • Si on ne cite pas la base de données, la procédure va automatiquement prendre la dernière base de données créée. **VAR** *variables* ; • Liste les variables pour lesquelles on veut les statistiques.	Commandes requises pour réaliser la procédure
BY *variable*; • Calcule des statistiques séparées pour les observations de chaque groupe. **CLASS** *variables* </options>; • Identifie des variables dont les valeurs définissent des sous-groupes pour l'analyse. **FREQ** *variable*; • Identifie une variable dont les valeurs représentent les fréquences de chaque observation. **ID** *variables*; • Permet d'identifier les observations dans l'output. **OUTPUT <OUT=**nom nouvelle base de données> </options> ; • Crée une nouvelle base de données qui contient les différentes statistiques. **WEIGHT** *variable* ; • identifie la variable dont les valeurs vont pondérer chaque observation.	Commandes supplémentaires
RUN;	

2.1.3. La procédure pour des tests non-paramétriques

Cette procédure permet de réaliser des tests non-paramétriques.
Ci-dessous se trouve les commandes à utiliser pour réaliser cette procédure.

PROC NPAR1WAY <options> **data**= *nom de la base de données* ; • Si on ne cite pas la base de données, la procédure va automatiquement prendre la dernière base de données créée. **CLASS** *variable* ; • Identifie la variable qui permet d'identifier les groupes pour lesquels on va examiner les différences. **VAR** *variables* ; • Nomme les variables qui seront analysées.	Commandes requises pour réaliser la procédure
BY *variable*; • Permet d'obtenir des statistiques séparées pour les observations des groupes définis par la variable. **FREQ** *variable*; • Identifie une variable dont les valeurs représentent les fréquences de chaque observation. **EXACT** *statistic-options </computation-options>*; • Demande des tests exacts pour les statistiques spécifiées. **OUTPUT** **<OUT=**nom nouvelle base de données> </options> ; • Crée une nouvelle base de données qui contient les différentes statistiques.	Commandes supplémentaires
RUN;	

C'est dans les options de la commande "**PROC NPAR1WAY**" qu'on peut définir les analyses qu'on veut faire. Voici les différentes options qui existent:

OPTIONS	TYPE D'ANALYSE
AB	Analyse utilisant les scores d'Ansari-Bradley.
ANOVA	Analyse standard de la variance.
D	Statistique unilatérale de Kolmogorov-Smirnov.
EDF	Statistiques basées sur la distribution empirique de la fonction (test de Kolmogorov-Smirnov, test de Cramer-von Mises et test de Kuiper (si uniquement deux niveaux de classification)).
KLOTZ	Analyse utilisant les scores de Klotz.
MEDIAN	Analyse utilisant les scores de la médiane.
MOOD	Analyse utilisant les scores de Mood.
SAVAGE	Analyse utilisant les scores de Savage.
SCORES=DATA	Analyse utilisant les données de départ comme score.
ST	Analyse utilisant les scores de Siegel-Tukey.
VW	Analyse utilisant les scores de Van der Waerden.
WILCOXON	Analyse utilisant les scores de Wilcoxon.

2.1.4. La procédure pour une régression linéaire

La procédure PROC REG est utilisée pour réaliser des régressions linéaires et permet d'obtenir le meilleur modèle par la méthode des moindres carrés.
Ci-dessous nous avons les commandes disponibles pour réaliser cette procédure.

PROC REG </options> **data**=*nom du jeu de données*; **MODEL** *variables dépendantes = variables explicatives* </options>; • Identifie les variables du jeu de données qui forment le modèle.	Commandes requises pour construire le modèle.
BY *variable* ; • Effectue une régression différente pour les observations de chaque groupe. **FREQ** *variable*; • Identifie une variable numérique dont les valeurs représentent la fréquence de chaque observation. **ID** *variable*; • Avec CLI, CLM, P, R ou INFLUENCE, identifie les observations par les valeurs de cette variable plutôt que de les numéroter. **VAR** *variable*; • Identifie les variables numériques, qui ne font pas parti du modèle, à inclure dans la matrice des produits croisés. **WEIGHT** *variable*; • Identifie une variable numérique dont les valeurs serviront à pondérer les observations.	Commandes facultatives indiquées une fois pour toute la procédure. Elles doivent apparaître avant le premier RUN.
ADD *variable*; • Ajoute des variables indépendantes au modèle de régression. Ces variables doivent apparaître dans les instructions MODEL ou VAR. **DELETE** *variable*; • Enlève des variables indépendantes du modèle de régression. **MTEST** *équation*; </options> • Teste les hypothèses pour les modèles de régression multiple où il y a plusieurs variables dépendantes. Si aucune équation ni option n'est spécifié la commande teste l'hypothèse que les variables sont égales à zéro. **OUTPUT OUT**=*nom du nouveau jeu de données* <statistique>=*nom de la variable ou des variables*; • Nomme un nouveau jeu de données qui contiendra les statistiques spécifiées. **PLOT** • Semblable à la procédure PLOT. **PRINT** <options4><ANOVA><MODELDATA> ; • Imprime les options, le tableau ANOVA et les données pour les variables utilisées dans le modèle.	Commandes qui peuvent apparaître n'importe où après la commande MODEL
RUN;	

2.1.5. La procédure pour des GLM (General Linear Model)

La procédure GLM utilise la méthode des moindres carrés pour avoir des modèles linéaires généraux. Parmi les méthodes statistiques disponibles dans cette procédure, nous avons entre autres les régressions simples et multiples, l'analyse de variances (ANOVA), l'analyse de covariances, l'analyse multivariée de variance (MANOVA) et les corrélations partielles.
Dans le tableau qui suit, nous avons la syntaxe à utiliser pour réaliser cette procédure.

PROC GLM <options> ;	
CLASS *variables*</option> ; • indique les variables de classification.	Cette commande doit apparaître avant la commande MODEL
MODEL *variables dépendantes = effets indépendantes* </options> ; • Seulement un modèle peut être défini.	Commande requise pour réaliser la procédure
CONTRAST *label effet values*<... effect values> </options> ; • construit et test des fonctions linéaires des paramètres. **ESTIMATE** *label effect values*<... effect values> </options> ; • estime des fonctions linéaires des paramètres. **LSMEANS** *effets* </options> ; • calcule les moyennes (marginales) par la méthode des moindres carrés. **TEST** <**H**=effets> **E**=effet </options>; • construit des tests en utilisant la somme des carrées pour les effets et l'erreur spécifiée. **MANOVA** <test-options> </détail-options> ; • accomplit une analyse multivariée de variance. **MEANS** *effets* </options>; • calcule et compare (en option) des moyennes arithmétiques. **OUTPUT** <OUT=nom de la nouvelle base de données> mot-clé = noms <…mot-clé= noms> </options> ; • crée une nouvelle base de données qui contient les diagnostiques pour chaque observation. **RANDOM** *effets* </options> ; • indique les effets certains pour être aléatoire et calcule expected means squares. **REPEATED** *facteur de spécification* </options> ; • accomplit des mesures répétées multivariée et univariée d'analyse de variance.	Commandes supplémentaires. Attention, l'ordre de leurs apparitions présenté ici doit être respecté.
ABSORB *variables* ; • amortit les effets de classification du modèle. **BY** *variables* ; • spécifie les variables qui définissent les sous-groupes pour l'analyse. **FREQ** *variable* ; • Identifie une variable dont les valeurs représentent les fréquences de chaque observation. Si cette commande n'est pas spécifiée, on assigne à chaque observation une fréquence de 1. **ID** *variables* ; • identifie des observations dans l'output. **WEIGHT** *variable*; • spécifie une variable pour les pondérations.	Commandes supplémentaires.
RUN;	

2.1.6. La procédure pour une régression logistique

La régression logistique est utilisée pour étudier la relation entre des variables discrètes (dichotomiques, ordinales ou nominales) et une série de variables explicatives. Cette procédure utilise la méthode du maximum de vraisemblance pour faire correspondre un modèle de régression logistique aux variables discrètes.
Ci-dessous, nous avons la syntaxe pour réaliser cette procédure dans SAS.

PROC LOGISTIC <options> **data**= *nom de la base de données*. • Si on ne cite pas la base de données, la procédure va automatiquement prendre la dernière base de données créée. **MODEL** *variable dépendante=variables indépendantes* </options>; *événement/épreuve = variables indépendantes* </options>; • Seulement un modèle peut être défini mais deux types de modèles peuvent être spécifiés. Le premier est utilisé lorsque la variable dépendante est dichotomique, ordinale ou nominale. L'utilisation de la deuxième forme est limitée au cas de variables dépendantes dichotomiques. Dans ce deuxième cas, on spécifie deux variables qui contiennent les informations sur une expérience ayant deux résultats possibles. Les deux variables sont séparées par un slash. La première identifie le nombre d'événement positif et la deuxième indique le nombre d'épreuves.	Commandes requises pour réaliser la procédure
BY *variables* ; • Permet d'obtenir des analyses séparées sur les observations des groupes définis par les variables de cette commande. **CLASS** *variable* <(v-option)> <*variable*<(v-option)>…> </v-options>; • Nomme les variables de classification qui seront utilisées dans l'analyse. Cette commande doit précéder la commande MODEL. **CONTRAST** *'label' effect values* <…*effect values*> </options>; • Fournit un mécanisme pour obtenir des tests d'hypothèses sur mesure. Le paramètre *label* identifie le contraste dans le résultat, le paramètre *effect* identifie un effet qui apparaît dans la commande MODEL et *values* sont des constantes qui sont des éléments de la matrice L associée à chaque effet. **EXACT** <*label*><*intercept*><*effects*></options>; • Accomplit des tests exacts des paramètres pour les effets spécifiés. Les analyses exactes ne sont pas exécutées quand la commande WEIGHT est spécifiée. **FREQ** variable; • Identifie une variable qui contient les fréquences des événements de chaque observation. Cette procédure examine chaque observation comme si elle apparaît *n* fois et *n* est la valeur de la variable définie dans cette commande. Par défaut une fréquence de 1 est assignée à chaque observation. **OUTPUT** <OUT=*nom de la nouvelle base de données*> </options> ; • Crée une nouvelle base de données qui contient toutes les variables de la base de données de départ **STRATA** *variable*<option> <*variable*<option>...> </options> ; • Nomme les variables qui définissent les strates à utiliser dans la régression logistique stratifiée conditionnelle. **WEIGHT** *variable* </option[6]> ; • Pondère chaque variable de la base de données de départ. Si cette commande n'apparaît pas, chaque observation a un poids égal à 1.	Commandes supplémentaires

[6] L'option **NORMALIZE** peut être ajoutée à la commande WEIGHT. Elle doit être placée après un slash (/). Cette option fait que les poids spécifiés par la variable de la commande soient normalisés.

SCORE <options> ; • Crée une base de données qui contient toute les données avec les probabilités postérieures et les intervalles de confiance prédits. Pour les variables dichotomiques, cette commande peut être utilisée créer une base de données contenant les informations pour la courbe de ROC (voir l'option OUTROC). **TEST** *équation 1* <,..., <*équation k*>> </option> ; • Teste des hypothèses linéaires par rapport aux coefficients de régression. **UNITS** *independent 1 = list 1* <... *independent k = list k*> </option> ; • Permet de spécifier des unités de changement pour les variables explicatives continues pour pouvoir estimer des *odds ratio.* Le terme *independent* correspond au nom de la variable explicative et le terme *list* représente une liste d'unités de changement.	
RUN ;	

2.2. Les procédures SURVEY de SAS

Dans les procédures traditionnelles de SAS, les différentes statistiques sont calculées sous l'hypothèse que l'échantillon a été tiré selon un plan de sondage aléatoire simple et en ayant accès à toute la population. Comme nous l'avons dit précédemment, ce type de procédure ne calcule pas correctement la variance d'un estimateur si l'échantillon a été tiré avec un plan de sondage complexe. C'est pour cela que SAS fournit maintenant des procédures permettant d'analyser des données provenant de plans de sondage complexes.

Les procédures SURVEYFREQ, SURVEYMEANS, SURVEYREG, et SURVEYLOGISTIC utilisent la méthode de linéarisation en série de Taylor pour estimer les erreurs d'échantillonnage d'estimateurs basés sur des plans de sondages complexes. Cette méthode est appropriée pour tous les plans où le premier niveau de l'échantillon est sélectionné avec remise ou quand la fraction d'échantillonnage au premier niveau (avec ou sans remise) est petite, comme c'est souvent le cas en pratique. La méthode de linéarisation en série de Taylor obtient une approximation linéaire de l'estimateur, puis elle effectue un calcul de variance pour cette approximation pour estimer la variance dudit estimateur. Quand on a à faire à des grappes ou à des unités d'échantillonnage primaires (PSUs), les procédures estiment la variance à partir de la variabilité inter-PSU. Quand il s'agit d'un plan stratifié comme pour l'enquête PSM, les procédures mettent en commun les estimations de la variance des strates pour calculer l'estimation globale de la variance.
Pour un plan de sondage à plusieurs niveaux, la méthode d'estimation de la variance dépend seulement du premier niveau. Il suffit donc d'identifier le premier niveau autant bien pour un plan par grappes que pour un plan stratifié. Il n'y a pas besoin d'informations supplémentaires sur d'autres niveaux de l'échantillon. L'annexe I montre les formules utilisées par le logiciel SAS pour calculer les variances des procédures SURVEYFREQ, SURVEYMEANS, SURVEYREG et SURVEYLOGISTIC.
Mis à part la méthode de linéarisation en série de Taylor, il existe d'autres méthodes pour calculer la variance d'estimateurs sur des données d'enquêtes, comme le Jackknife ou le bootstrap. Ces méthodes donnent généralement des résultats similaires. Des vérifications empiriques l'ont montré pour l'enquête PSM. Les procédures SURVEYFREQ, SURVEYMEANS, SURVEYREG, et SURVEYLOGISTIC utilisent actuellement uniquement la méthode de linéarisation en série de Taylor.

2.2.1. Les fréquences et les tableaux croisés avec la procédure SURVEYFREQ

La procédure PROC SURVEYFREQ fournit des tableaux de fréquence de une à n entrées et des tableaux de contingence. Ces tableaux incluent l'estimation des totaux de la population et les proportions avec les écart-types correspondants. Cette procédure calcule l'estimation de la variance sur la base du plan de sondage utilisé dans l'enquête. Pour les données du PSM il s'agit d'un plan stratifié par grandes régions (NUTSII). Elle calcule également des tests et des mesures d'association entre les variables.

Voici les commandes à disposition pour la procédure PROC SURVEYFREQ.

PROC SURVEYFREQ <options> **data**= *nom de la base de données* • Si on ne cite pas la base de données, la procédure va automatiquement prendre la dernière base de données créée. **N**= value | base de données; • Identifie le nombre total des unités d'échantillonnage primaires (PSU) ou une base de données donnant les totaux de chaque strate dans une variable nommée _TOTAL_[7]. **TABLES** *variables* </options> ; • Spécifie les tableaux de fréquences ou de contingence que l'on veut et les statistiques pour ces tableaux. Le terme *request* peut être composé par une variable ou plusieurs variables séparées par des astérisques	Commande requise pour réaliser la procédure.
BY *variables* ; • Donne des analyses séparées pour les observations des groupes définis par les variables citées dans cette commande. **CLUSTER** *variables ;* • Nomme les variables qui identifient les grappes de premier niveau dans un plan de sondage par grappes. **STRATA** *variables* </option> ; • Identifie les variables qui forment les strates dans un plan de sondage stratifié. Si le plan de sondage contient plusieurs niveaux, il ne faut définir que le premier. **WEIGHT** *variable* ; • Nomme la variable qui contient les valeurs des pondérations des observations. Par défaut la procédure donne un poids égal à 1 à chaque observation. Si une observation a un poids négatif ou manquant, elle est exclue de l'analyse.	Commandes supplémentaires.
RUN ;	

[7] Création de la base de données contenant les totaux des strates (syntaxe à utiliser) :

DATA *nom de la base de données qui donne les totaux des strates*;
INPUT NOM DE LA VARIABLE UTILISEE POUR LA STRATIFICATION _TOTAL_;
DATALINES;

Après il faut indiquer dans une première colonne les numéros correspondant aux strates et dans une deuxième les tailles de ces strates.

2.2.2. Les statistiques descriptives avec la procédure SURVEYMEANS

La procédure SURVEYMEANS produit des estimations de la moyenne et de totaux de la population d'enquête. Elle produit également des estimations de la variance, des intervalles de confiance, et d'autres statistiques descriptives. Pour calculer ces estimations, la procédure prend en considération le plan de sondage utilisé pour la sélection de l'échantillon.

PROC SURVEYMEANS <options> **data**= *nom de la base de données* • Si on ne cite pas la base de données, la procédure va automatiquement prendre la dernière base de données créée. **N**= value | base de données; • Identifie le nombre total des unités d'échantillonnage primaires (PSU) ou une base de données donnant les totaux de chaque strate dans une variable nommée _TOTAL_. **VAR** *variables* ; • Liste les variables pour lesquelles on veut les statistiques.	Commandes requises pour réaliser la procédure.
BY *variables* ; • Donne des analyses séparées pour les observations des groupes définis par les variables citées dans cette commande. **CLASS** *variables* ; • Nomme les variables qui seront analysées comme des variables catégorielles. Avec cette commande, la procédure va estimer la proportion pour chaque catégorie à la place d'une moyenne globale. **CLUSTER** *variables* ; • Nomme les variables qui identifient les grappes dans un plan de sondage par grappes. S'il s'agit d'un plan de sondage à plusieurs niveaux, il ne faut identifier que le premier. **DOMAIN** *variables* ; • Nomme les variables qui définissent les domaines pour les analyses de sous-populations. **RATIO** *variables/variables* : • Demande des analyses de rapport entre moyennes ou proportion. Les moyennes des variables qui se trouve avant le slash (/) vont être les valeurs pour le numérateur et les moyennes des variables qui se trouvent après le slash celles du dénominateur. **STRATA** *variables* </option> ; • Nomme les variables qui forment les strates dans un plan de sondage stratifié. Si le plan de sondage contient plusieurs niveaux, il ne faut définir que le premier. **WEIGHT** *variable ;* • Nomme la variable qui contient les poids de l'échantillon. Par défaut la procédure donne un poids égal à 1 à chaque observation. Si une observation a un poids négatif ou manquant, elle est exclue de l'analyse.	Commandes supplémentaires.
RUN ;	

2.2.3. La régression linéaire avec la procédure SURVEYREG

Cette procédure réalise des analyses de régression pour des données provenant d'échantillons à plans de sondage complexes comme des plans de sondages stratifiés, en grappes ou à probabilités inégales. La procédure calcule entre autre les coefficients de régression et leur matrice de variance-covariance. Pour estimer la matrice de variance-covariance des coefficients de régression, la PROC SURVEYREG utilise la méthode de linéarisation en série de Taylor. Elle fournit aussi des tests de significativité pour les effets du modèle et pour n'importe quelles fonctions linéaires estimables des paramètres de celui-ci. En utilisant le modèle de régression, la procédure peut également calculer les valeurs prédites pour les données de l'enquête.

PROC SURVEYREG<options> **data**= *nom de la base de données* • Si on ne cite pas la base de données, la procédure va automatiquement prendre la dernière base de données créée. **N**= value | base de données; • Identifie le nombre total des unités d'échantillonnage primaires (PSU) ou une base de données donnant les totaux de chaque strate dans une variable nommée _TOTAL_. **MODEL** *variable dépendante =variables indépendantes* </options> ; • Spécifie la variable dépendante et les variables ou combinaison de variables indépendantes.	Commandes requises pour réaliser la procédure.
BY *variables*; • Calcule des analyses séparées pour les observations de chaque groupe défini par les variables. **CLASS** *variables*; • Spécifie les variables de classification qui seront utilisées dans le modèle. Cette commande doit apparaître avant la commande MODEL. **CLUSTER** *variables*; • Spécifie des variables qui identifient les grappes dans un plan de sondage par grappes. S'il s'agit d'un plan à plusieurs niveaux, il ne faut identifier que le premier. **CONTRAST** '*label*' effect values <...effects values></options>; **ESTIMATE** '*label*' effect values <...effects values></options>; **STRATA** *variables* </options>; • Spécifie les variables qui forment les strates dans un plan de sondage stratifié. Si le plan de sondage contient plusieurs niveaux, il faut uniquement identifier le premier. **WEIGHT** *variables* ; • Spécifie la variable qui contient les poids de l'échantillon. Si cette commande n'est pas spécifiée, la procédure assigne à toutes les observations un poids égal à 1.	Commandes supplémentaires.
RUN;	

2.2.4. La régression logistique avec la procédure SURVEYLOGISTIC

La régression logistique étudie la relation entre une variable dépendante discrète et une série de variables explicatives. La procédure SURVEYLOGISTIC utilise la méthode du maximum de vraisemblance pour estimer les paramètres du modèle. Cette procédure, contrairement à la traditionnelle, permet de tenir compte de la complexité du plan de sondage utilisé dans les enquêtes telles que le PSM et SILC.

PROC SURVEYLOGISTIC <options **data**= *nom de la base de données* • Si on ne cite pas la base de données, la procédure va automatiquement prendre la dernière base de données créée. **N**= value | base de données; • Identifie le nombre total des unités d'échantillonnage primaires (PSU) ou une base de données donnant les totaux de chaque strate dans une variable nommée _TOTAL_. **MODEL** *variable dépendante= variables explicatives* </options>; *événements/épreuves = variables explicatives* </options>; • Deux types de modèle peuvent être spécifiés. Le premier est utilisé lorsque la variable dépendante est dichotomique, ordinale ou nominale. L'utilisation de la deuxième forme est limitée au cas de variables dépendantes dichotomiques. Dans ce deuxième cas, on spécifie deux variables qui contiennent les informations sur une expérience ayant deux résultats possibles. Les deux variables sont séparées par un slash. La première identifie le nombre d'événement positif et la deuxième indique le nombre d'épreuves.	Commandes requises pour réaliser la procédure.
BY *variables* ; • Donne des analyses séparées pour les observations des groupes définis par les variables citées dans cette commande. **CLASS** *variable* </v-option> < *variable* </v-option>…> <v-options> ; • Nomme les variables de classification qui seront utilisées dans les analyses. Cette commande doit précéder la commande MODEL. **CLUSTER** *variables*> ; • Nomme les variables qui identifient les grappes dans un plan de sondage par grappes. S'il s'agit d'un plan de sondage à plusieurs niveaux, il ne faut identifier que le premier. **CONTRAST** '*label*' effect values <...effects values></options>; • Fournit un mécanisme pour obtenir des tests d'hypothèses adaptés. Cette commande doit apparaître après la commande MODEL : **FREQ** *variable* ; • Identifie une variable dont les valeurs représentent les fréquences de chaque observation. Si cette commande n'est pas spécifiée, on assigne à chaque observation une fréquence de 1. **STRATA** *variables* </options> ; • Nomme les variables qui forment les strates dans un plan de sondage stratifié. Si le plan de sondage contient plusieurs niveaux, il ne faut définir que le premier. **TEST** *équation 1*<…<*équation k*>> </option> ; • Teste les hypothèses linéaires par rapport aux coefficients de régression. Chaque équation spécifie une hypothèse linéaire. **UNITS** *variable explicative 1 = liste 1* <…*variable explicative k = liste k*> </option> ; • Spécifie des unités de changement pour les variables explicatives continues pour pouvoir estimer des *odds ratio.* Le terme *liste* représente une liste d'unités de changement pour la variable explicative indiquée. **WEIGHT** *variable* </option> ; • Nomme la variable qui contient les poids de l'échantillon. Par défaut la procédure donne un poids égal à 1 à chaque observation. Si une observation a un poids négatif ou manquant, elle est exclue de l'analyse.	Commandes supplémentaires.
RUN ;	

2.3. Les procédures analogues dans SPSS

Dans cette section, on va présenter les syntaxes de SPSS correspondant à celles qu'on vient de voir pour le logiciel SAS. On va commencer par la présentation des syntaxes classiques et dans un deuxième temps, on va voir les syntaxes permettant de réaliser des analyses analogues à celles réalisables avec les procédures SURVREY.

2.3.1. Les procédures traditionnelles dans SPSS.

Toutes les commandes présentées ci-dessous font l'hypothèse que l'échantillon a été tiré selon un plan aléatoire simple. Elles ne fournissent donc pas une bonne estimation de la variabilité lorsque ce dernier est plus complexe. Les tableaux ci-dessous montrent les syntaxes pour obtenir des tableaux de fréquences, des analyses descriptives (moyenne, écart-type, minimum, maximum…), des régressions linéaires et des régressions logistiques, sans et avec les poids.

	SANS LES POIDS	AVEC LES POIDS
FREQUENCES	FREQUENCIES VARIABLES=*variables* **.**	WEIGHT BY *poids* **.** FREQUENCIES VARIABLES=*variables* **.** WEIGHT OFF **.**

Cette commande fournit les fréquences et les pourcentages pour chaque modalité de réponse de la variable ou des variables d'intérêt. Elle ne donne cependant pas les écart-types de ces pourcentages. Pour utiliser les pondérations, il faut utiliser la commande WEIGHT BY. Pour éviter que les pondérations s'appliquent à toutes les analyses qui suivront, il est nécessaire de mettre la commande WEIGHT OFF.

	SANS LES POIDS	AVEC LES POIDS
AUTRES ANALYSES DESCRIPTIVES	DESCRIPTIVES VARIABLES=*variables* /STATISTICS=MEAN MIN MAX SEMEAN .	WEIGHT BY *poids* . DESCRIPTIVES VARIABLES=*variables* /STATISTICS=MEAN MIN MAX SEMEAN . WEIGHT OFF.

En utilisant cette syntaxe, on obtient différentes statistiques descriptives. Dans l'exemple ci-dessus, les statistiques demandées sont la moyenne, le minimum, le maximum et l'écart-type de la moyenne. Pour les pondérations, il faut utiliser les mêmes commandes que celles présentées pour les fréquences.

	SANS LES POIDS	AVEC LES POIDS
REGRESSION LINEAIRE	REGRESSION /MISSING LISTWISE /STATISTICS COEFF OUTS R ANOVA /CRITERIA=PIN(.05) POUT(.10) /NOORIGIN /DEPENDENT variable dépendante /METHOD=ENTER *variables explicatives* .	REGRESSION /MISSING LISTWISE /REGWGT=*poids* /STATISTICS COEFF OUTS R ANOVA /CRITERIA=PIN(.05) POUT(.10) /NOORIGIN /DEPENDENT *variable dépendante* /METHOD=ENTER *variables explicatives* .

La syntaxe ci-dessus est utilisée pour réaliser des régressions linéaires. Différentes statistiques sont disponibles. Dans l'exemple, on a choisi d'obtenir les estimations des paramètres, le R^2 ainsi que les analyses de variance (ANOVA). Contrairement aux procédures précédentes, la commande utilisée pour les pondérations est définie dans la syntaxe. Il s'agit de la commande REGWGT.

	SANS LES POIDS	AVEC LES POIDS
REGRESSION LOGISTIQUE	LOGISTIC REGRESSION *variable dépendante* /METHOD = ENTER *variables explicatives* /CRITERIA = PIN(.05) POUT(.10) ITERATE(20) CUT(.5) .	WEIGHT BY *poids* . LOGISTIC REGRESSION *variable dépendante* /METHOD = ENTER *variables explicatives* /CRITERIA = PIN(.05) POUT(.10) ITERATE(20) . WEIGHT OFF .

Cette commande permet de réaliser des modèles de régression logistique. La variable dépendante doit être dichotomique. Concernant les pondérations, il faut utiliser la commande WEIGHT BY. Comme expliqué précédemment, pour que les pondérations ne s'appliquent pas à toutes les analyses qui suivront, il faut mettre la commande WEIGHT OFF à la fin.

2.3.2. Les procédures du module COMPLEX SAMPLE

Avec SPSS, il est possible de réaliser des estimations analogues à celles produites par les PROC SURVEY(…) de SAS grâce aux commandes se trouvant dans le module COMPLEX SAMPLE. Pour ce faire, il faut au préalable créer un fichier contenant les caractéristiques déterminant le plan de sondage utilisé dans l'enquête. Ce fichier est appelé PLAN. Il contient les informations concernant les strates ou les grappes (selon le type de plan de sondage), la variable pour la pondération ainsi que le choix de la méthode d'estimation qui dépend du type de tirage du plan (avec ou sans remise). Dans le cas d'un tirage sans remise, il faut préciser les probabilités d'inclusion ou la taille de la population. La marche à suivre pour construire un fichier PLAN est illustrée en annexe (voir annexe II).

Ci-dessous, nous avons les syntaxes du module COMPLEX SAMPLE pour les fréquences, les statistiques descriptives (les moyennes, écart-type,…), les régressions linéaire et les régressions logistiques. **Elles nous fournissent les mêmes résultats que les PROC SURVEY de SAS**. Toutes ces procédures peuvent également être réalisées à l'aide des menus du programme (voir annexe III), en cliquant.

	SYNTAXE DANS LE MODULE COMPLEX SAMPLE
FREQUENCES	* Complex Samples Frequencies. CSTABULATE /PLAN FILE = *emplacement du fichier PLAN* /TABLES VARIABLES = *variables* /CELLS POPSIZE TABLEPCT /STATISTICS SE CIN(95) /MISSING SCOPE = TABLE CLASSMISSING = EXCLUDE.

Cette procédure donne la taille et le pourcentage pour chaque cellule. Les statistiques demandées sont l'écart-type et l'intervalle de confiance (95%)

	SYNTAXE DANS LE MODULE COMPLEX SAMPLE
AUTRES ANALYSES DESCRIPTIVES	* Complex Samples Descriptives. CSDESCRIPTIVES /PLAN FILE = *emplacement du fichier PLAN* /SUMMARY VARIABLES = *variables* /MEAN /STATISTICS SE COUNT POPSIZE CIN (95) /MISSING SCOPE = ANALYSIS CLASSMISSING = EXCLUDE.

Cette procédure calcule la moyenne. Les statistiques demandées sont l'écart-type, l'intervalle de confiance (95%), la taille de la population et la taille de la population pondérée.

	SYNTAXE DANS LE MODULE COMPLEX SAMPLE
REGRESSION LINEAIRE	* Complex Samples General Linear Model. CSGLM *variable dépendante* WITH *variables explicatives* /PLAN FILE = *emplacement du fichier PLAN* /MODEL *variables explicatives* /INTERCEPT INCLUDE=YES SHOW=YES /STATISTICS PARAMETER SE CINTERVAL TTEST /PRINT SUMMARY VARIABLEINFO SAMPLEINFO /TEST TYPE=F PADJUST=LSD /MISSING CLASSMISSING=EXCLUDE /CRITERIA CILEVEL=95.

Les statistiques demandées sont ici l'estimation des paramètres du modèle, les écart-types, les intervalles de confiance et les T-test.

	SYNTAXE DANS LE MODULE COMPLEX SAMPLE
REGRESSION LOGISTIQUE	* Complex Samples Logistic Regression. CSLOGISTIC *variable à modèliser*(LOW) *variable explicatives* /PLAN FILE = *emplacement du fichier PLAN* /MODEL *variables explicatives* /INTERCEPT INCLUDE=YES SHOW=YES /STATISTICS PARAMETER SE TTEST /TEST TYPE=F PADJUST=LSD /MISSING CLASSMISSING=EXCLUDE /CRITERIA MXITER=100 MXSTEP=5 PCONVERGE=[1e-006 RELATIVE] LCONVERGE=[0] CHKSEP=20 CILEVEL=95 /PRINT SUMMARY VARIABLEINFO SAMPLEINFO.

Les statistiques demandées sont ici l'estimation des paramètres du modèle, les écart-types et les T-test.

3. Exemples d'applications dans SAS et SPSS

Dans ce chapitre, nous allons illustrer les différentes procédures vues au chapitre précédent. L'objectif est de montrer l'impact des pondérations et de montrer les résultats qu'on obtient en utilisant les procédures PROC SURVEY de SAS et COMPEX SAMPLE de SPSS. Pour ces deux types de procédures, on présente, dans la section 3.1, les points importants à prendre en compte pour les appliquer correctement.

3.1. Eléments à prendre en compte pour l'application des procédures SURVEY de SAS et COMPLEX SAMPLE de SPSS

Pour réaliser ces deux types de procédures, différents points doivent être définis au préalable. Pour les utiliser correctement, il est primordial de bien s'informer sur la méthode d'échantillonnage utilisée dans l'enquête. Dans le cas du PSM et de SILC nous avons à faire à des échantillons aléatoires simples stratifiés (voir section 1.3). On a besoin de connaître la variable utilisée pour stratifier ainsi que de la grandeur de chacune des strates. Dans notre cas, on stratifie par les sept grandes régions de Suisse. Selon qu'on travaille sur les données du PSM_I, du PSM_II, de SILC_I ou de SILC_II, la grandeur de ces strates n'est pas toujours exactement la même, du fait que les différents échantillons n'ont pas été tirés tous au même moment[8]. Le tableau ci-dessous indique les tailles de ces dernières.

Tableau 3 : tailles des strates aux moments des différents tirages d'échantillons

STRATES	PSM_I, 1999	PSMII et SILC_I, 2004	SILC_II, 2005
Région lémanique (VD, VS, GE)	714725	648'590	648'176
Espace Mittelland (BE, FR, SO, NE, JU)	837452	784'266	785'951
Suisse du Nord-Ouest (BS, BL, AG)	484667	455'833	456'925
Zurich (ZH)	646469	587'850	599'145
Suisse orientale (GL, SH, AR, AI, SG, GR, TG)	531731	493'606	495'755
Suisse centrale (LU, UR, SZ, OW, NW, ZG)	313548	306'605	308'626
Tessin (TI)	180623	160'123	159'622
TOTAL	**3'709'215**	**3'436'873**	**3'454'200**

Pour utiliser les procédures SURVEY de SAS, on doit premièrement définir dans la commande STRATA la variable identifiant la stratification de notre échantillon. Il faut également créer une base de données dans laquelle on donne la taille de ces strates définies par une variable du même nom que celle renseignant la commande STRATA (voir annexe IV: exemple de création de cette base de données). Dans la syntaxe, on se réfère à cette base de données avec l'option "`N= nom_base_de_données`". Un dernier élément à ne pas oublier est d'indiquer avec la commande WEIGHT la variable de pondération. Il est important de choisir le bon poids. Le choix se fait premièrement selon le type d'analyses désirées (longitudinales ou transversales, section 1.6), deuxièmement, selon qu'on travaille sur les individus ou les ménages et troisièmement selon qu'on utilise les données du PSM_I, du PSM_II, de SILC_I ou SILC_II séparément ou combinées entre elles.

[8] On constate une diminution du nombre de numéros entre 1999 et 2004. Il y a deux raisons principales qui expliquent cette diminution. Premièrement, en 1999, on avait encore les numéros de téléphones non inscrits dans l'annuaire, ce qui n'était plus le cas en 2004. Deuxièmement, les procédures de nettoyage du SRH se sont améliorées pour éliminer les doublons et les adresses ne correspondant pas à des ménages privés.

Pour les procédures COMPLEX SAMPLE de SPSS, il faut avoir, attachée aux données, une variable qui contient pour chaque observation la grandeur de la strate lui correspondant. Cette information est ensuite introduite lors de la création du PLAN (voir paragraphe 2.3.2). N'oublions pas que lors de l'utilisation des procédures COMPLEX SAMPLE, il est essentiel de définir le fichier "plan" car il contient toutes les informations sur le plan de sondage de l'échantillon et sur la variable à utiliser pour les pondérations. Pour résumer ce "PLAN" contient les informations que l'on donne à SAS dans WEIGHT, STRATA et N=.

Attention, quelque soit l'analyse et quelque soit le logiciel utilisé, il ne faut pas oublier que les tailles des strates sont différentes selon qu'on travaille sur le PSM_I, le PSM_II, SILC_I ou SILC_II. Voici un tableau récapitulatif des tailles des strates à utiliser selon l'enquête(s) choisie(s) pour l'analyse:

Tableau 4: tailles de strates à utilisées selon l'enquête

Enquêtes utilisées		**Tailles des strates**	**Nombre de strates**
PSM_I	Longitudinal Transversal	Tailles des strates de 1999	7 strates
PSM_II	Longitudinal Transversal	Tailles des strates de 2004	7 strates
SILC_I	Longitudinal Transversal	Tailles des strates de 2004	7 strates
SILC_II	Longitudinale Transversale	Tailles des strates de 2005	7 strates
PSM_I & PSM_II	Longitudinal Transversal	Tailles des strates de 1999 pour les observations du PSM_I et de 2004 pour celles du PSM_II	14 strates
SILC_I & SILC_II	Longitudinal Transversal	Tailles des strates de 2004 pour les observations de SILC_I et de 2005 pour les observations de SILC_II	14 strates
PSM_I, PSM_II & SILC_I	Longitudinal Transversal	Tailles des strates de 1999 pour les observations du PSM_I et de 2004 pour celles du PSM_II et de SILC_I	14 strates
PSM_I, PSM_II, SILC_I & SILC_II	Longitudinal Transversal	Tailles des strates de 1999 pour les observations du PSM_I, de 2004 pour celles du PSM_II et de SILC_I et de 2005 pour celles de SILC_II	21 strates

3.2. Les tableaux de fréquences

Dans cette section sont étudiées les procédures permettant d'obtenir des tableaux de fréquences. Comme statistique descriptive, les tableaux de fréquences sont un outil important car ils permettent de familiariser le chercheur avec les variables qu'il va utiliser dans ses analyses. Il est d'ailleurs indispensable de commencer toute analyse par une analyse descriptive des variables.
Pour illustrer ces procédures, on a choisi d'étudier dans un premier temps la variable *p05c01*. Il s'agit de l'état subjectif de santé de la personne interrogée et contient les réponses à la question "comment allez-vous en ce moment ?". Les différentes modalités de réponses vont de "1" (très bien) à "5" (très mal). Dans un deuxième temps, on va s'intéresser à la variable *educat05* (niveau de formation le plus élevé).
Les procédures PROC FREQ et PROC SURVEYFREQ fournissent, entre autre, des tableaux de fréquences. L'élément supplémentaire de la procédure SURVEYFREQ est qu'elle nous permet de tenir compte du plan de sondage et de calculer de manière juste des écart-types de pourcentages selon la méthode de linéarisation en séries de Taylor.

Les syntaxes utilisées dans SAS sont les suivantes.

- La procédure FREQ :

`proc freq data= nom_de_la_base_données ;`
`  tables variables;`
`  weight poids;`
 	
`run;`
	- Indique la variable qui permet de pondérer.

- La procédure SURVEYFREQ :

`proc surveyfreq data= nom_de_la_base_données`
`  N= nom_base_de_données;`
 	
`  tables variables;`
`  strata strates;`
 	
`  weight poids;`
 	
`run;`
	- Indique le nom de la base de données contenant les tailles des strates définies par la variable « *strates* » de la commande STRATA
	
	- Pour un plan stratifié, indique la variable qui forme les strates.
	- Indique la variable qui permet de pondérer.

Les analyses sont également réalisées dans SPSS, en particulier les analyses correspondant à la procédure SURVEYFREQ. Pour cela, on utilise la commande FREQUENCIES du module COMPLEX SAMPLE sans oublier de défini le plan correspondant à l'analyse désirée.

L'analyse de la variable *p05c01* (état de santé) se réalise sur 6524 observations qui correspondent aux personnes de 14+ans ayant répondu au questionnaire individuel et dont les poids sont supérieurs à zéro. Les poids utilisés sont les poids transversaux individuels, extrapolant à la taille de la population en 2005 pour le PSM_I et le PSM_II combinés (wp05t1p). Des valeurs manquantes ont été enlevées. Il s'agit de six individus se trouvant dans la modalité "ne sait pas" et de deux individus se trouvant dans la modalité "pas de réponse"
Le tableau 5 présente les fréquences de la variable *p05c01* (état de santé).

Tableau 5. : Fréquences et pourcentages de la variable p05c01 (après exclusion des catégories manquantes)

réponses	n sans poids	n avec poids	SAS					SPSS	
			PROC FREQ			PROC SURVEYFREQ		FREQUENCIES du module COMPLEX SAMPLE	
			Sans poids		Avec poids wp05t1p	STRATA / CLUSTER WEIGHT : wp05t1p			
			%	Ecart type[9]	%	%	Ecart type	%	Ecart type
1 : très bien	1400	1298603	21.46	0.508	21.19	21.19	0.605	21.19	0.605
2 : bien	4158	3891061	63.73	0.595	63.50	63.50	0.713	63.50	0.713
3 : comme ci, comme ça	826	795642.5	12.66	0.412	12.98	12.98	0.511	12.98	0.511
4 : mal	130	126274.9	1.99	0.173	2.06	2.06	0.212	2.06	0.212
5 : très mal	10	16471.6	0.15	0.048	0.27	0.27	0.105	0.27	0.105
TOTAL	6524	6128053	100.00		100.00	100.00		100.0	

Source : PSM_I et PSM_II

[9] L'écart-type est utilisé pour calculer les intervalles de confiance. L'intervalle de confiance à 95% du pourcentage estimé se calcule de la façon suivante:

$$\text{pourcentage estimé} \pm 1.96 * \text{écart - type}$$

Avec l'utilisation des poids, les fréquences des modalités de réponses n'extrapolent plus aux proportions d'individus présents dans l'échantillon mais à la population résidant en Suisse de 14+ ans de 2005. Les pourcentages, quant à eux, varient très peu entre les différentes procédures utilisées. Avec la procédure PROC FREQ, on voit une légère diminution des pourcentages des modalités "1" et "2" ainsi qu'une légère augmentation des autres modalités lors de l'utilisation des poids. La procédure PROC SURVEYFREQ, quant à elle, fournit les mêmes pourcentages que la procédure PROC FREQ avec les poids.
C'est dans le calcul des écart-types correspondants qu'on trouve une différence. Comme on peut le constater, les écart-types fournis par la procédure PROC SURVEYFREQ sont plus grands que ceux de la procédure PROC FREQ sans les poids[10].
Dans la procédure PROC SURVEYFREQ, la commande CLUSTER avec comme variable l'identifiant ménage (idhous05) est utilisée. Pour l'analyse de cette variable on suppose que dans un même ménage l'état de santé peut être similaire entre les différentes personnes. Souvent lorsque quelqu'un a, par exemple, la grippe d'autres personnes autours d'elle vont aussi l'avoir.

Dans le deuxième exemple (analyse de la variable *educat05)*, on travaille sur 3344 observations. Il s'agit d'individus appartenant au PSM_I âgés d'au moins 14 ans, ayant répondu au questionnaire individuel et dont les poids sont supérieurs à zéro. Les poids utilisés ici sont les poids individuels longitudinaux extrapolant à la taille de la population de 1999 (wp05lp1p). On se trouve dans le cas d'une analyse longitudinale. Le tableau 6 montre les résultats obtenus

[10] La procédure PROC FREQ ne fournit pas les écart-types des pourcentages. Pour les obtenir, il faut utiliser la procédure PROC SURVEYFREQ dans laquelle les commandes WEIGHT et STRATA ne sont pas utilisées. Attention, il n'est pas possible d'obtenir des écart-types pour la procédure PROC FREQ avec les poids.

Tableau 6. : Fréquences et pourcentages de la variable educat05

réponses	n sans poids	n avec poids	SAS					SPSS	
			PROC FREQ			PROC SURVEYFREQ		FREQUENCIES du module COMPLEX SAMPLE	
			Sans poids		Avec poids wp05lp1p	STRATA / CLUSTER WEIGHT : wp05lp1p			
			%	Ecart type	%	%	Ecart type	%	Ecart type
0 : école obligatoire inachevée	25	72084	0.75	0.149	1.22	1.22	0.249	1.22	0.249
1 : école obligatoire, formation prof. Élémentaire	304	873959	9.09	0.497	14.81	14.81	0.914	14.81	0.914
2 : stage ménager, 1 année d'école commerciale courte	150	308151	4.49	0.358	5.22	5.22	0.461	5.22	0.461
3 : école de formation générale	32	56649	0.96	0.168	0.96	0.96	0.189	0.96	0.189
4 : apprentissage	1140	2086001	34.09	0.820	35.35	35.35	1.003	35.35	1.003
5 : école professionnelle à plein temps	220	376719	6.58	0.429	6.38	6.38	0.474	6.38	0.474
6 : maturité	403	630853	12.05	0.563	10.69	10.69	0.588	10.69	0.588
7 : formation prof. supérieure	255	345594	7.63	0.459	5.86	5.86	0.399	5.86	0.399
8 : école technique ou professionnelle	122	176902	3.65	0.324	3.00	3.00	0.291	3.00	0.291
9 : école prof. Supérieure	205	289666	6.13	0.415	4.91	4.91	0.370	4.91	0.370
10 : université, haute école	488	685238	14.59	0.611	11.61	11.61	0.610	11.61	0.610
TOTAL	3344	5901816	100.00		100.00	100.00		100.00	

Source : PSM_I

Avec l'utilisation de la variable de pondération wp05lp1p, les fréquences des différentes modalités se réfèrent aux proportions dans la population résidant en Suisse de 14+ ans de 1999.
Dans cet exemple, nous voyons de grandes différences pour les pourcentages obtenus lors de l'introduction des poids dans l'analyse. Ici, l'utilisation ou non des poids a une grande importance car elle change la distribution de la variable. Par ailleurs, il n'y a pas de différences de pourcentages entre la procédure PROC FREQ avec poids et la procédure PROC SURVEYFREQ (avec poids).
On voit que les écart-types de toutes les modalités où le pourcentage a augmenté (modalités 0, 1, 2, 4) ou est resté le même (modalité 3), ont augmenté lors de l'utilisation de la procédure PROC SURVEYFREQ. Pour les autres modalités où le pourcentage a diminué, l'écart-type a fait de même. Cependant, un élément important à relever est que, dans la majorité des cas, l'intervalle de confiance obtenu avec la procédure PROC FREQ (sans les poids) et celui obtenu avec la procédure PROC SURVEYFREQ (avec les poids) ne se recoupent même pas (voir annexe V).
Il est donc ici vital d'utiliser les poids et de calculer correctement les variances.
Pour cet exemple, on a utilisé dans la procédure PROC SURVEYFREQ la commande CLUSTER avec comme variable l'identifiant ménage (idhous05). On suppose, ici, que dans un même ménage on retrouve des personnes avec un niveau de formation similaire.

Pour nos deux exemples, nous avons réalisé les mêmes analyses avec le logiciel SPSS. On arrive exactement aux mêmes résultats que la procédure SURVEYFREQ de SAS, en utilisant la procédure FREQUENCIES du module COMPLEX SAMPLE. Il est donc possible dans SPSS de tenir compte de la complexité de l'échantillon lors de la réalisation d'analyses descriptives telles que les fréquences.

3.2.1. Test du Chi-carré : test d'indépendance entre variables catégorielles

Les procédures PROC FREQ, PROC SURVEYFREQ de SAS et FREQUENCIES du module COMPLEX SAMPLE de SPSS fournissent également la possibilité de réaliser un test d'indépendance entre variables, le test du Chi-carré. Ce test se construit à partir d'un tableau de contingence dans lequel on trouve un facteur R avec r catégories et un facteur C avec c catégories.
Concernant les syntaxes, on prend les mêmes que dans la section 3.2 et on rajoute aux procédures PROC FREQ et PROC SURVEYFREQ l'option `"/chisq"`[11] après la commande `"tables"`.

[11] L'option change selon le test qu'on fait. Dans la procédure PROC SURVEYFREQ, trois types de test du Chi-carré sont possible, le "Rao-Scott Chi-square test" avec l'option `"/chisq"`, le "Rao-Scott modified Chi-square test avec l'option `"/chisq1"` et le "Wald Chi-square test avec l'option `"/wchisq"`. Dans le cas de la procédure PROC FREQ, il n'y a que l'option `"/chisq"` qui correspond au test du Chi-carré de Pearson.

La formule qui suit montre le calcul à réaliser pour un test du Chi-carré de Pearson (test conventionnel) :

$$Q_P = \sum \frac{(n_{rc} - e_{rc})^2}{e_{rc}}$$

Où :

- n_{rc} : Nombre d'observation se trouvant dans la cellule (r, c).
- e_{rc} : Les fréquences théoriques[12] se trouvant dans la cellule (r, c).

Lorsque l'échantillon est de grande taille, cette valeur est distribuée approximativement selon une loi χ^2 avec $(r-1)(c-1)$ degrés de liberté. On rejette l'hypothèse nulle d'indépendance entre les variables si :

$$Q_P > \chi^2_{(\alpha,dl)}.$$

Avec l'utilisation de la procédure PROC SURVEYFREQ, trois tests du Chi-carré sont possibles. Il s'agit des tests du Chi-carré de Rao-Scott, et du test du Chi-carré de Wald.

Le test du Chi-carré de Rao-Scott est calculé de la façon suivante:

$$Q_{RS} = Q_P / D$$

Il s'agit du test conventionnel de Pearson (Q_P) basé sur les totaux estimés, ajusté par une correction (D). Cette correction ne tient compte que du plan de sondage. Elle utilise les informations sur les strates (commande STRATA) et sur les tailles des strates (option N=). Elle tient compte également de poids de sondage qu'elle recalcule elle-même, cette correction n'utilise donc pas les poids finaux fournis dans la commande WEIGHT. Les poids donnés par la commande WEIGHT sont utilisées pour le calcul des fréquences permettant de remplir les cellules du tableau de contingence.
La procédure PROC SURVEYFREQ fournit deux types de correction pour le test du Chi-carré de Rao-Scott. La première utilise les proportions estimées. Il s'agit du test du Chi-carré de Rao-Scott et on l'obtient avec l'option `chisq`. La deuxième forme de correction utilise les proportions sous l'hypothèse nulle. Ce test s'appelle le test du Chi-carré de Rao-Scott modifié et on l'obtient avec l'option `chisq1`. Pour plus d'information sur la correction (D), se référer à l'aide de SAS (recherche: "SURVEYFREQ procedure, chi-square test").

La procédure PROC SURVEYFREQ offre la possibilité de calculer un test du Chi-carré de Wald. Ce test se base sur la différence entre les fréquences pondérées observées et théoriques. Cette statistique teste l'indépendance entre les variables en ligne et en colonne d'un tableau à deux entrées en tenant compte du plan de sondage. Sous l'hypothèse nulle d'indépendance, le test du Chi-carré de Wald suit approximativement une distribution du Chi-carré avec $(r-1)(c-1)$ degrés de liberté pour des échantillons très grands. Pour obtenir ce test, il faut utiliser l'option `wchisq`.

---

[12] $e_{rc} = \frac{n_{r.} n_{.c}}{n}$

Il s'agit du total de la ligne r fois le total de la colonne c divisé par le total des observations. Les fréquences théoriques sont calculées sous l'hypothèse d'indépendance.

Le test du Chi-carré de Wald est calculé de la façon suivante:

$$Q_{Wald} = \hat{Y}'\left(\mathrm{H}\,\hat{\mathrm{V}}(\hat{\mathrm{N}})\,\mathrm{H}'\right)^{-1}\hat{Y}$$

Où:

- $\hat{Y}$: Matrice des différences entre les fréquences pondérées observée et théoriques.
- $\left(\mathrm{H}\,\hat{\mathrm{V}}(\hat{\mathrm{N}})\,\mathrm{H}'\right)$: L'estimation de la variance de $\hat{Y}$.
- $\hat{\mathrm{V}}(\hat{\mathrm{N}})$: Matrice de covariance de l'estimation des totaux pondérés.

Avec la matrice de covariance, cette statistique tient d'avantage compte des poids qu'on livre que les tests du Chi-carré de Rao-Scott.

Pour illustrer ces procédures, on prend les variables *p05n43* (membre d'un parti politique) et *p05a01* (avoir une activité physique), le sexe et le niveau de formation en trois catégories (voir annexe XVI). L'objectif est de tester premièrement l'indépendance entre la variable *p05n43* et les variables du niveau de formation et du sexe, deuxièmement de tester l'indépendance entre la variable *p05a01* et le sexe selon quatre catégories d'âge[13].

Pour l'analyse de la variable *p05n43*, on ne prend en considération que les personnes répondant "membre actif" ou "membre passif", âgées d'au moins 18 ans et dont le poids est supérieur à zéro. L'échantillon est alors constitué de 684 individus.
Pour l'analyse de la variable *p05a01*, on s'intéresse aux personnes de 14+ ans dont les poids sont supérieurs à zéro. Des valeurs manquantes apparaissent. Il s'agit de quatre personnes se trouvant dans la catégorie "ne sait pas". Ces quatre observations sont enlevées pour les analyses et l'échantillon ne contient plus que 6528 observations pour cet exemple. Dans les deux exemples les poids utilisés sont les poids transversaux individuels où la taille de l'échantillon de 2005 reste inchangée pour le PSM_I et le PSM_II réunis (wp05t1s).
Les tableaux 7.a), 7.b) 7.c) montrent les résultats des tests du Chi-carré.

Tableaux 7.a) Membre actif-passif d'un parti politique - niveau de formation, résultats des tests du Chi-carré selon les procédures

		Khi2	p-value
PROC FREQ	Pearson Chi-square	6.5186	0.0384
PROC FREQ WEIGHT (WP05T1S)	Pearson Chi-square	4.5318	0.1037
PROC SURVEYFREQ STRATA WEIGHT: wp05t1s	Rao-Scott Chi-square Test Rao-Scott modified Chi-square Test Wald Chi-square Test	3.3381 3.4946 3.6792	0.1884 0.1742 0.1597

N=684

Source: PSM_I et PSM_II

[13] Les quatre catégories d'âge utilisées:
1ère catégorie: 14-30 ans, 2ème catégorie: 31-45ans, 3ème catégorie: 46-65 ans, 4ème catégorie: 65+ ans.

Tableaux 7.b) Membre actif-passif d'un parti politique – sexe, résultats des tests du Chi-carré selon les procédures

		Khi2	p-value
PROC FREQ	Pearson Chi-square	1.5298	0.2161
PROC FREQ WEIGHT (WP05T1S)	Pearson Chi-square	2.2276	0.1356
PROC SURVEYFREQ STRATA WEIGHT: wp05t1s	Rao-Scott Chi-square Test Rao-Scott modified Chi-square Test Wald Chi-square Test	1.9471 1.9436 2.2124	0.1629 0.1622 0.1374

N=684

Source: PSM_I et PSM_II

Tableaux 7.c) Avoir une activité physique ou pas – sexe selon les catégories d'âge, résultats des tests du Chi-carré selon les procédures.

		14-30 ans		31-45 ans		46-65 ans		65+ ans	
		Khi2	p-value	Khi2	p-value	Khi2	p-value	Khi2	p-value
PROC FREQ	Pearson Chi-square	6.1426	0.0132	0.0694	0.7922	0.4666	0.4946	4.5030	0.0338
PROC FREQ WEIGHT (WP05T1S)	Pearson Chi-square	6.1048	0.0135	0.0329	0.8560	1.1230	0.2893	4.8346	0.0279
PROC SURVEYFREQ STRATA WEIGHT (WP05T1S)	Rao-Scott Chi-square	4.3055	0.0380	0.0208	0.8854	0.7716	0.3797	3.6492	0.0561
	Rao-Scott modified Chi-square	4.2553	0.0391	0.0208	0.8853	0.7743	0.3789	3.6916	0.0547
	Wald Chi-square	3.9919	0.0459	0.0215	0.8835	0.8123	0.3675	3.3668	0.0669
		N=1453		N=1965		N=2190		N=920	

Source: PSM_I et PSM_II

L'hypothèse d'indépendance entre les variables est rejetée si la valeur calculée du test est plus grande que la valeur théorique. La valeur théorique du test dépendant des degrés de liberté et du seuil α qui est fixé pour ces exemples à 0.05 (voir annexe VI).

Dans le tableau 7.a) "membre actif-passif d'un parti – niveau de formation", les résultats obtenus indiquent que l'hypothèse nulle d'indépendance est rejetée lors de l'utilisation de la procédure PROC FREQ sans les poids mais elle n'est pas rejetée avec les autres procédures. Ici, l'utilisation ou non des poids change totalement la conclusion du test.
Dans le tableau 7.b) "membre actif-passif d'un parti – sexe", tous les résultats indiquent que l'hypothèse d'indépendance n'est pas rejetée. Il paraît qu'il n'y a pas de lien entre le sexe et être membre actif ou passif d'un parti. Dans cet exemple, l'utilisation poids et de la correction tenant compte du plan de sondage ne change pas le résultat du test suffisamment pour avoir une modification de la conclusion.
Le tableau 7.c) montre les résultats des tests d'indépendance pour les variables "avoir une activité physique ou pas et le sexe" selon les catégories d'âge. Pour les catégories 31-45 ans et 46-65 ans l'hypothèse d'indépendance n'est jamais rejetée, le risque de première espèce étant toujours plus grand que le seuil α fixée à 0.05 et ceci quelque soit la procédure et le test utilisé. Les résultats obtenus pour la catégorie des 14-30 ans indiquent que l'hypothèse nulle est rejetée. Les valeurs obtenus par les différentes procédures sont toujours supérieurs à la valeur théorique avec un degré de liberté et $\alpha = 0.05$. Le risque de première espèce est lui toujours plus petit que le seuil α. Il faut rester attentif au résultat obtenu par le test de Wald dans la procédure PROC SURVEYFREQ. Il est très proche de la valeur théorique.
Les conclusions pour la catégorie des 65+ ans varient selon les procédures. Avec les procédures PROC FREQ sans ou avec poids, l'hypothèse nulle est rejetée. Avec la procédure PROC SURVEYREG et la prise en compte du plan de sondage, l'hypothèse d'indépendance n'est pas rejetée.

3.3. Autres statistiques descriptives

Les procédures PROC MEANS et PROC SURVEYMEANS permettent de calculer diverses statistiques descriptives. Entre autre, elles fournissent la moyenne et l'écart type ainsi que la taille de l'échantillon, le minimum, le maximum, ainsi que les limites inférieures et supérieures de l'intervalle de confiance du calcul de la moyenne.
Dans les procédures classiques la moyenne d'une variable d'intérêt *Y* est calculée de la façon suivante :

$$\bar{y} = \sum w_j y_j \Big/ \sum w_j \, .$$

Où

- w_j est le poids de l'observation *j*.

Si aucune indication n'est donnée quant à la variable à utiliser pour les pondérations, un poids égal à un est attribué, par défaut, à chacune des observations.

Avec l'utilisation de la procédure PROC SURVEYMEANS, la formule de la moyenne devient :

$$\bar{y} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij} y_{hij} \Big/ w_{...} .$$

Où

- $w_{...}$ représente la somme des poids sur toutes les observations de l'échantillon.
- h : le numéro de la strate avec un total de H strates
- i : le numéro de la grappe dans la strate h avec un total de n_h grappes
- j : le numéro de l'unité dans la grappe i de la strate h avec un total de m_{hi} unités.

Les syntaxes utilisées dans SAS sont les suivantes :

- La procédure MEANS :

`proc means data= nom_base_de_données`
`n vardef=df fw=8 maxdec=2`
`alpha=0.05 mean stderr`
`clm min max ;`
`var variable;`
`class variable_sousgroupes;`
 	
`weight poids;`
 	
`run;`
	- Indique la variable qui permet de faire des analyses sur des sous populations.
	- Indique la variable qui permet de pondérer.

- La procédure SURVEYMEANS :

`proc surveymeans data = nom_base_de_données`
`N=nom_base_de_données`
 	
`NOBS alpha=0.05 MEAN`
`STDERR clm MIN MAX ;`
`domain variable_sousgroupes;`
 	
`cluster grappes;`
 	
`strata strates;`
 	
`var variable;`
 	
`weight poids;`
 	
`run;`
	- Indique le nom de la base de données contenant les tailles des strates définies par la variable « *strates* » de la commande STRATA
	
	- Indique la variable qui permet de faire des analyses sur des sous populations.
	- Indique la variable qui identifie les grappes.
	- Pour un plan stratifié, indique la variable qui forme les strates.
	- Indique la variable qui permet de pondérer.

Dans la commande « **proc means** » et « **proc surveymeans** » on a introduit différentes options. Voici les explications pour chacune d'elles.

N, nobs	Permet d'avoir le nombre d'observation
Vardef[14]	Permet de spécifier le dénominateur pour le calcul de la variance
Maxdec	Permet d'indiquer le nombre maximal de décimales
Alpha	Permet de choisir le seuil de significativité
Mean	Donne une estimation de la moyenne
Stderr	Donne l'écart type de l'estimation de la moyenne
Clm	Donne la limite inférieure et supérieure de l'intervalle de confiance de la moyenne
Min, max	Indique le minimum et le maximum

Nous réalisons également, dans SPSS, les analyses descriptives correspondant à la procédure SURVEYMEANS. Pour cela, nous utilisons la commande DESCRIPTIVES du module COMPLEX SAMPLE, sans oublier de créer, au préalable, le plan.

Pour illustrer ces procédures, nous allons présenter deux exemples. La variable utilisée dans le premier exemple est la variable *i05htyn* (revenu annuel net du ménage). Nous allons travailler, ici, sur les données du PSM_I et du PSM_II réunies et la variable de pondération utilisée est la variable wh05t1s (poids transversaux ménage pour le PSM_I et le PSM_II combinés où la taille de l'échantillon de 2005 reste inchangée). Dans le deuxième exemple, les analyses vont être faites sur la variable *p05p01* (intérêt pour le politique). Ici nous allons travailler uniquement sur les données du PSM_II et la variable de pondération est la variable wp05tp2p (poids transversaux individuels pour le PSM_II extrapolant à la taille de la population de 2005).

Pour l'analyse de la variable *i05htyn*, on prend en compte les ménages dont le poids est supérieur à zéro. On observe des valeurs manquantes (voir annexe VII). Pour l'analyse, elles ne vont pas être prises en compte. L'échantillon sur lequel on travaille est constitué de 3697 observations.
Pour cette variable, les analyses vont se porter sur des sous-populations. Les sous-populations choisies sont les ménages suisses versus les ménages étrangers[15]. On trouve en annexe la syntaxe utilisée pour la création de la variable permettant de différencier les ménages dits "suisses" des ménages dits "étrangers"(voir annexe VIII).

[14]En utilisant les poids transversaux du ménage pour le PSM_I et le PSM_II où la taille de l'échantillon reste inchangée (wh05t1s), on met comme option : vardef=df. Dans ce cas la formule utilisée pour la variance aura n-1 au dénominateur. Par ailleurs, on obtient quasiment le même résultat lorsqu'on travaille avec les poids transversaux ménage pour le PSM_I et le PSM_II extrapolant à la taille de la population de 2005 (wh05t1p) si on choisit l'option : vardef= wdf. Dans ce cas formule utilisée pour la variance aura la somme des poids moins un au dénominateur.

[15] Définition d'un ménage suisse : un ménage est dit « suisse » si toutes les personnes le formant ont au moins une de leur nationalité qui est suisse.
Définition d'un ménage étranger : un ménage est dit « étranger » dès qu'au moins une personne du ménage n'a aucune de leurs nationalités qui est suisse.
Nous obtenons 3411ménages suisses et 286 ménages étrangers.

Les tableaux ci-dessous présentent les résultats obtenus avec la procédure PROC MEANS (sans et avec poids), la procédure PROC SURVEYMEANS et la procédure DESCRIPTIVES du module COMPLEX SAMPLE de SPSS, pour l'ensemble des ménages puis pour les ménages suisses et les ménages étrangers distinctement (tableaux 8 à 10).

Tableau 8. : Tableau récapitulatif des résultats obtenus pour la variable i05htyn pour tous les ménages

Procédure et options utilisées		Mean	stderr	LCL	UCL
SAS	PROC MEANS	96885.394	1142.752	94644.908	99125.879
	PROC MEANS WEIGHT : wh05t1s	91259.119	1104.625	89093.384	93424.854
	PROC SURVEYMEANS STRATA WEIGHT : wh05t1s	91259.119	1167.811	88969.500	93548.738
SPSS	DESCRIPTIVES du module COMPLEX SAMPLE	91259.119	1167.811	88969.500	93548.738

N : 3697
Min : 5000
Max : 1800000

Source : PSM_I et PSM_II

On constate que la moyenne de la variable *i05htyn* change selon la procédure choisie. Elle diminue lorsqu'on utilise les pondérations. Ceci indique que pour l'analyse de cette variable, l'utilisation ou non des poids a de l'importance. En effet, on pourrait arriver à des conclusions différentes étant donné les différences dans l'estimation de la moyenne.
Pour les procédures PROC MEANS avec les poids et PROC SURVEYMEANS, on voit que l'estimation de la moyenne est la même. C'est au niveau des écart-types qu'une différence apparaît. On constate en effet une légère augmentation de ceux-ci avec l'utilisation de la procédure PROC SURVEYMEANS.

Le tableau 9 montre les résultats obtenus avec les procédures PROC MEANS, PROC SURVEYMEANS et DESCRIPTIVES du module COMPLEX SAMPLE de SPSS pour les ménages "suisses" et le tableau 10 montre les résultats obtenus pour les ménages "étrangers".

Tableau 9. : Tableau récapitulatif des résultats obtenus pour la variable i05htyn pour les ménages suisses

Procédure et options utilisées		Mean	stderr	LCL	UCL
SAS	PROC MEANS	97788.625	1130.341	95572.410	100004.840
	PROC MEANS WEIGHT : wh05t1s	92313.227	1085.233	90185.454	94441.001
	PROC SURVEYMEANS STRATA WEIGHT : wh05t1s	92313.227	1136.542	90084.851	94541.603
SPSS	DESCRIPTIVES du module COMPLEX SAMPLE	92313.227	1136.542	90084.851	94541.603

N : 3411
Min : 5000
Max : 1800000

Source : PSM_I et PSM_II

Tableau 10. : Tableau récapitulatif des résultats obtenus pour la variable i05htyn pour les ménages étrangers

Procédure et options utilisées		Mean	stderr	LCL	UCL
SAS	PROC MEANS	86112.937	6012.388	74278.618	97947.256
	PROC MEANS WEIGHT : wh05t1s	85332.370	5041.857	75408.370	95256.371
	PROC SURVEYMEANS STRATA WEIGHT : wh05t1s	85332.370	4286.000	76894.420	93770.320
SPSS	DESCRITIVES du module COMPLEX SAMPLE	85332.370	4286.000	76894.420	93770.320

N : 286
Min : 9600
Max : 162000

Source : PSM_I et PSM_II

Pour le tableau 9, les constatations sont les mêmes que celles présentées pour l'analyse du tableau 8.
Dans l'exemple du tableau 10, une différence apparaît. On observe une diminution des écart-types au lieu d'une augmentation. Ceci s'explique par le fait que les pondérations donnent moins de poids aux valeurs des extrêmes de la distribution qui devient alors moins étalée.

Dans le deuxième exemple (variable *p05p01*: intérêt pour le politique), l'échantillon est restreint aux personnes âgées d'au moins 18 ans (personnes ayant atteint la majorité et pouvant avoir le droit de vote) et ayant un poids supérieur à zéro. La variable contient des valeurs manquantes. Il s'agit d'une personne se trouvant dans la catégorie "pas de réponse" et d'une autre se trouvant dans la catégorie "ne sais pas". Ces valeurs ont été enlevées pour la réalisation des analyses. L'échantillon sur lequel on travaille a 2471 observations. L'analyse porte sur toute la population sans distinction de nationalité et également sur la sous-population des personnes ayant la nationalité suisse. Dans cet exemple, on introduit la commande CLUSTER avec comme variable l'identifiant ménage (idhous05), lors de la réalisation de la procédure PROC SURVEYMEANS. On suppose ici que les personnes d'un même ménage vont avoir des comportements similaires ce qui inclut l'intérêt pour le politique.

Les résultats obtenus se trouvent dans le tableau 11 (pour toute la population sans distinction de nationalité) et le tableau 12 (pour les personnes de nationalité suisse).

Tableau 11. : Tableau récapitulatif des résultats obtenus pour la variable p05p01 pour toute la population âgée d'au moins 18 ans

Procédure et options utilisées		Mean	stderr	LCL	UCL
SAS	PROC MEANS	5.688	0.057	5.577	5.800
	PROC MEANS WEIGHT : wp05tp2s	5.534	0.060	5.417	5.651
	PROC SURVEYMEANS STRATA / CLUSTER WEIGHT : wp05tp2s	5.534	0.077	5.382	5.685
SPSS	DESCRIPTIVES du module COMPLEX SAMPLE	5.534	0.077	5.382	5.685

N : 2471

Source : PSM_II

On constate qu'il n'y a quasiment pas de différence dans les résultats obtenus entre les différentes procédures utilisées. Les moyennes et les écart-tapes changent très peu et les intervalles de confiance calculés avec ou sans les poids se recoupent.

Le tableau suivant, montre les résultats obtenus pour la population suisse uniquement.

Tableau 12. : Tableau récapitulatif des résultats obtenus pour la variable p05p01 pour personnes de nationalité suisse, âgée d'au moins 18 ans

Procédure et options utilisées		Mean	stderr	LCL	UCL
SAS	PROC MEANS	5.841	0.058	5.727	5.955
	PROC MEANS WEIGHT : wp05tp2s	5.878	0.058	5.763	5.993
	PROC SURVEYMEANS STRATA / CLUSTER WEIGHT : wp05tp2s	5.878	0.067	5.746	6.010
SPSS	DESCRIPTIVES du module COMPLEX SAMPLE	5.878	0.067	5.746	6.010

N : 2196

Source : PSM_II

Les résultats obtenus changent également très peu lorsqu'on ne s'intéresse qu'à ce sous groupe de la population. On obtient même des écart-types similaires pour chacune des procédures.

Cet exemple montre que selon les variables choisies, il arrive que l'utilisation des pondérations n'ait que peu d'influence sur les résultats des analyses. Dans notre cas, ceci s'explique par le fait que la variable *p05p01* (intérêt pour le politique) n'est corrélée avec aucunes variables utilisées dans la construction des pondérations.

Par ailleurs, les résultats obtenus dans nos deux exemples avec la commande DESCRIPTIVES du module COMPLEX SAMPLE de SPSS sont exactement les mêmes que ceux obtenus avec la procédure SURVEYMEANS de SAS.

3.3.1. T-Test

Les procédures PROC MEANS, PROC SURVEYMEANS de SAS et DESCRIPTIVES du module COMPLEX SAMPLE de SPSS fournissent également la possibilité de réaliser un test sur la moyenne, le T-test. Il teste l'hypothèse nulle que la moyenne de la population est égale à une valeur μ_0. Par défaut, cette valeur est zéro dans les procédures PROC MEANS et PROC SURVEYMEANS de SAS. Dans la procédure DESCRIPTIVE du module COMPLEX SAMPLE de SPSS, il est possible de déterminer une valeur μ_0 autre que zéro[16]. Cette statistique se calcule de la façon suivante :

$$t(\hat{\bar{y}}) = \frac{\hat{\bar{y}} - \mu_0}{StdErr(\hat{\bar{y}})} \ ,$$

où $StdErr(\hat{\bar{y}})$ est l'erreur standard de la moyenne pour la variable *y* considérée.

[16] Dans SAS, il est également possible de déterminer une valeur μ_0 autre que zéro en utilisant la procédure PROC UNIVARIATE.

Ce test se base sur certaines hypothèses de départ. Ces pré-requis sont :

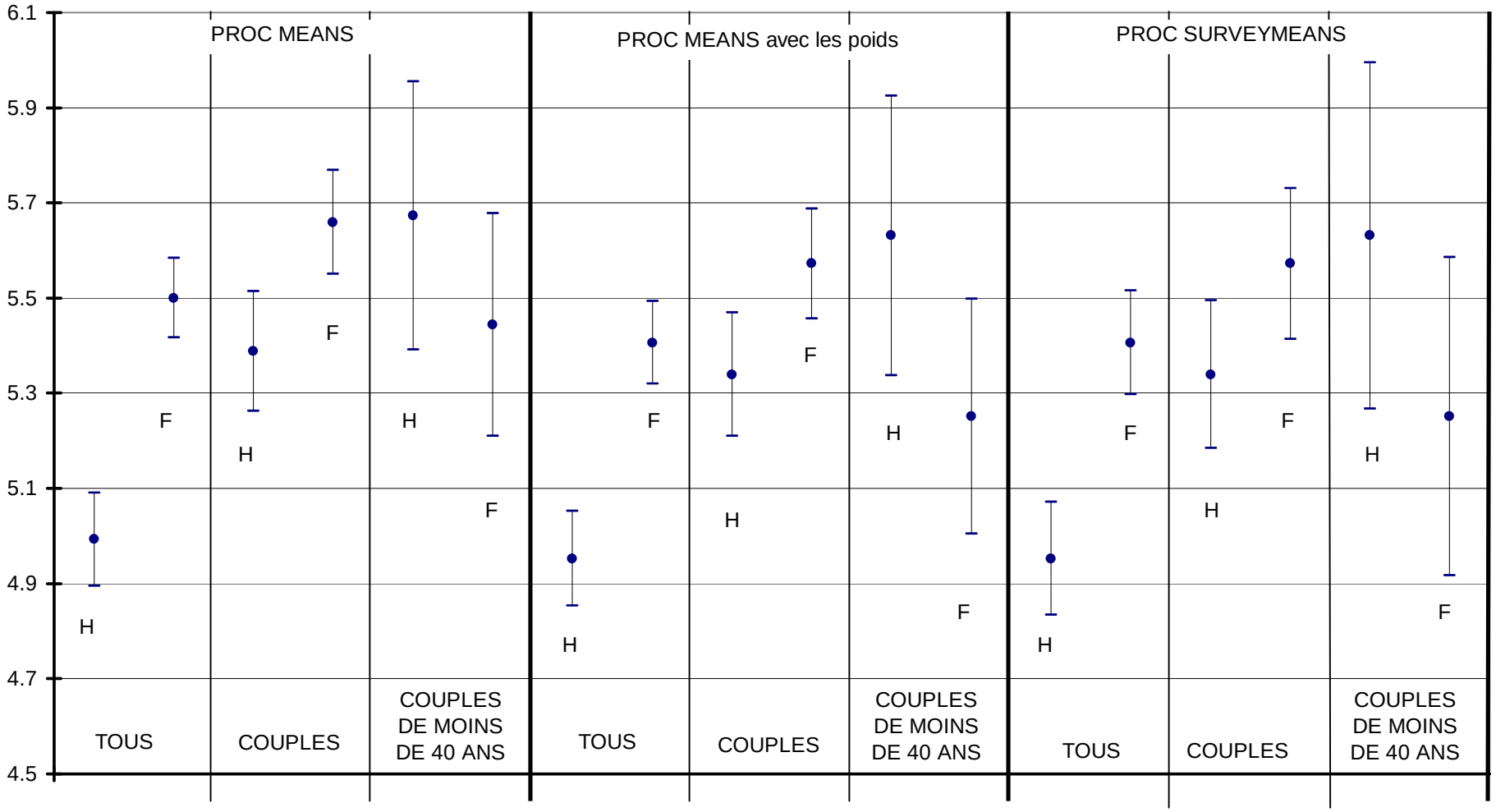
- La normalité : chaque groupe d'observations doit être distribué normalement.
- L'homogénéité des variances à l'intérieur des groupes.

Les syntaxes utilisées sont les mêmes que celles présentées dans la section 3.3, à ceci près que dans les commandes "**proc means**" et "**proc surveymeans**", on rajoute l'option "T". Pour la procédure PROC MEANS, il faut également rajouter l'option "PROBT" si l'on veut obtenir les p-values.

Pour illustrer ces procédures, on utilise la variable *p05p20* (Est-ce que l'on a le sentiment que les femmes sont défavorisées par rapport aux hommes dans certains domaines en Suisse ?). Les modalités de réponses vont de 0 à 10, où 0 signifie "pas du tout défavorisées" et 10 "tout à fait défavorisées". Les personnes considérées sont celles ayant répondu au questionnaire individuel (*status05*=0), de 14+ ans et dont les poids sont supérieurs à zéro. Des valeurs manquantes ont été enlevées. Il s'agit de 7 personnes se trouvant dans la modalité "pas de réponse" et de 53 autres personnes se trouvant dans la modalité "ne sait pas". L'échantillon comporte 6384 observations. La variable utilisée pour les pondérations est la variable wp05t1s (poids transversaux individuels où la taille de l'échantillon de 2005 reste inchangée) et on travaille sur les données du PSM_I et du PSM_II combinés.
Les analyses sont réalisées sur trois groupes de répondants différents. On s'intéresse, dans un premier temps, à tous ceux ayant répondu au questionnaire individuel. Ensuite, on se restreint aux personnes mariées (*cohast05*=1) et dans un troisième temps, on prend en considération uniquement les personnes mariées de moins de 40 ans. Le graphique 1 présente les intervalles de confiance autour de la moyenne pour les différents groupes d'intérêts selon les procédures et le sexe.

Graphique 1: Intervalles de confiance de la variable p05p20

IC variable p05p20 selon les procédures, le sexe et les groupes d'intérêt



Source: PSM_I et PSM_II

On constate que pour le groupe "TOUS" l'intervalle de confiance des hommes ne se recoupe absolument jamais avec celui des femmes et ceci quelque soit la procédure utilisée. On peut dire que les moyennes entre hommes et femmes pour ce groupe d'intérêt ne sont pas les mêmes. Pour le groupe "COUPLES", les intervalles de confiance des hommes et des femmes ne se recoupent pas avec les procédures PROC MEANS avec ou sans les poids. Cependant avec la procédure PROC SURVEYMEANS on observe un léger recoupement de ceux-ci. Les intervalles de confiance pour les hommes et les femmes du dernier groupe "COUPLES DE MOINS DE 40 ANS" se recoupent toujours quelque soit la procédure utilisée.

Sur la base de ces constatations, l'objectif est donc de tester s'il y a une différence ou pas entre la moyenne des hommes et celle des femmes et ceci pour les différents groupes d'intérêts et les différentes procédures. L'hypothèse nulle est donc :

$$H_0 : \bar{x}_{\text{hom}\,mes} - \bar{x}_{femmes} = 0$$

versus

$$H_1 : \bar{x}_{\text{hom}\,mes} - \bar{x}_{femmes} \neq 0$$

Les tableaux 13 à 15 montrent les résultats obtenus pour les différentes populations.

Tableau 13: T-test pour toute la population

		moyenne	T-test	p-value
PROC MEANS	Hommes	5.023	-9.8	<.0001
	Femmes	5.515		
PROC MEANS WEIGHT: wp05t1s	Hommes	4.964	-8.52	<.0001
	Femmes	5.400		
PROC SURVEYMEANS	Hommes	4.964	-7.19	<.0001
	Femmes	5.400		

N=6384

Source : PSM_I et PSM_II

Tableau 14: T-test pour les couples

		moyenne	T-test	p-value
PROC MEANS	Hommes	5.388	-4.25	<.0001
	Femmes	5.659		
PROC MEANS WEIGHT: wp05t1s	Hommes	5.340	-3.53	0.0004
	Femmes	5.573		
PROC SURVEYMEANS	Hommes	5.340	-2.93	0.0034
	Femmes	5.573		

N=3520

Source : PSM_I et PSM_II

Tableau 15: T-test pour les couples de moins de 40 ans

		moyenne	T-test	p-value
PROC MEANS	Hommes	5.673	1.6	0.1106
	Femmes	5.444		
PROC MEANS WEIGHT: wp05t1s	Hommes	5.632	2.55	0.0113
	Femmes	5.251		
PROC SURVEYMEANS	Hommes	5.632	2.04	0.0423
	Femmes	5.251		

N=794

Source : PSM_I et PSM_II

Ces trois tableaux montrent les résultats obtenus. L'hypothèse nulle (hypothèse d'égalité de moyennes) est rejetée lorsque la p-value est plus petite que le seuil α fixé à 0.05 dans ces exemples.
Cependant, on doit se retenir d'interpréter les résultats obtenus car, après vérification, les pré-requis de départ ne sont pas satisfaits. La variable *p05p20* n'a pas une distribution normale. Il est tout de même possible de tester l'hypothèse d'égalité des moyennes en utilisant des tests moins contraignant, comme les tests non-paramétriques (voir section suivante). Pour tester la normalité d'une variable ou de sous-groupes, on utilise le test de Kolmogorov-Smirnov disponible dans la procédure PROC UNIVARIATE[17]. Le test de kolmogorov-Smirnov permet de tester l'hypothèse nulle que les valeurs d'une variable sont un échantillon aléatoire suivant une distribution théorique donnée, dans notre cas une distribution normale.

On s'est intéressé à un deuxième exemple pour illustrer les procédures permettant de faire des T-test. Il s'agit de l'analyse de la variable *i05ptotn* (revenu total personnel annuel net). Pour cette variable, on restreint l'échantillon aux personnes de plus de 20 ans, travaillant à plein temps, n'ayant eu aucune modification d'activité au cours de l'année de l'interview, étant salarié[18] et dont les poids sont supérieurs à zéro. Les poids qu'on utilise sont les poids individuels transversaux où la taille de l'échantillon de 2005 reste inchangée (wp05t1s). Des valeurs manquantes ont été enlevées pour l'analyse (voir annexe IX). L'échantillon pour cet exemple comporte 1634 observations.
L'objectif ici est de tester s'il y a une différence des moyennes selon les niveaux de formation. L'hypothèse nulle est donc:

$$H_0 : \bar{x}_{\text{niveau de formation a}} - \bar{x}_{\text{niveau de formation b}} = 0$$

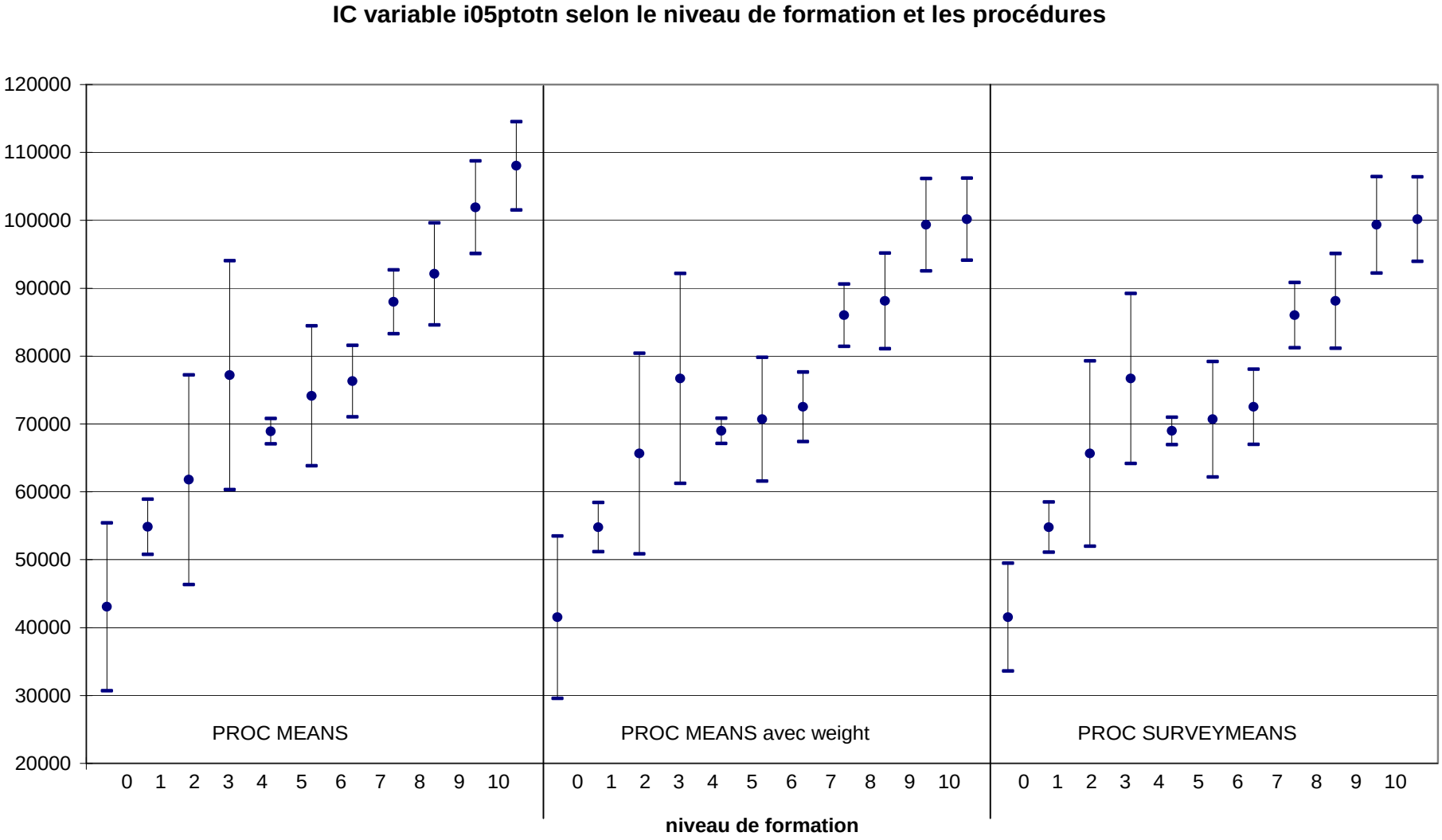
versus

$$H_1 : \bar{x}_{\text{niveau de formation a}} - \bar{x}_{\text{niveau de formation b}} \neq 0$$

Le graphique 2 présente les intervalles de confiance de la variable i05ptotn selon les procédures et le niveau de formation.

[17] Pour pouvoir tester la normalité, on met l'option "`normal`" à la commande "`proc univariate`".
[18] Plein temps: p05w39=2; Sans modification de l'emploi: p05w177=2; Salarié: p05w29a=2

Graphique 2: intervalle de confiance de la variable i05ptotn



Source: PSM_I et PSM_II

Les niveaux de formation 0, 2, 3 ne vont pas être pris en compte pour les tests car ils ont trop peu d'observations (5, 17 et 13 respectivement). Pour cet exemple, les pré-requis ne sont pas satisfaits. On n'a pas de normalité dans chaque groupe d'observation. Il n'y que le groupe correspondant au niveau de formation 6 (maturité) qui apparaît comme étant distribué normalement. Dans ce cas aussi, il faut réaliser les tests avec, par exemple, un test non-paramétrique de Wilcoxon (section suivante).

3.4. Test de Wilcoxon

La procédure PROC NPAR1WAY permet de réaliser des tests non-paramétriques. Ce type de tests est souvent utilisé lorsque certaines hypothèses des tests paramétriques ne sont pas satisfaites. Ici, on s'intéresse au test de Wilcoxon. Ce dernier ne requiert pas de distribution normale des données et il permet de tester si deux fonctions de répartition sur des (sous)échantillons sont égales.

La syntaxe utilisée est:

`proc npar1way wilcoxon`
 	
`data=nom_base_de_données;`
`class variable_de_classification;`
 	
`var variable;`
`run;`
	- Indique le type de test que l'on veut faire.
	
	- Indique la variable qui permet d'identifier les groupes pour lesquels on va examiner les différences.

Pour illustrer cette procédure, on va reprendre les exemples précédents (voir paragraphe 3.3.1.).
Concernant la variable *p05p20*, l'hypothèse nulle reste la même; on cherche à savoir si la moyenne pour les hommes est la même que celle pour les femmes.
Le tableau 16 montre les résultats obtenus pour le test de Wilcoxon avec la procédure PROC NPAR1WAY.

Tableau 16: résultats du test de Wilcoxon pour les différents groupes de population de l'analyse de la variable p05p20

		Scores moyens	Test de wilcoxon	p-value
TOUS	Hommes	3011.840	8598804.000	<.0001
	Femmes	3338.656		
COUPLES	Hommes	1704.680	2756467.500	0.0024
	Femmes	1807.931		
COUPLES DE MOINS DE 40 ANS	Hommes	390.057	127607.500	0.2512
	Femmes	408.998		

Source PSM_I et PSM_II

Les valeurs obtenues pour le test de Wilcoxon indiquent que pour les groupes de population "TOUS" et "COUPLES", l'hypothèse d'égalité des moyennes entre les hommes et les femmes est rejetée. C'est clairement le cas pour le groupe de population "TOUS" où le risque de première espèce[19] est plus petit que 0.01%. Pour le groupe de population "COUPLES" on rejette également l'hypothèse nulle avec un risque inférieur à 0.25%. Pour les "COUPLES DE

[19] Risque de première espèce: probabilité de rejeter l'hypothèse nulle alors qu'elle est vraie.

MOINS 40 ANS", la valeur obtenue par le test de Wilcoxon indiquent que l'hypothèse est acceptée. Le risque de première espèce est ici clairement supérieur à 5%, seuil qu'on a fixé pour les analyses. Il se situe à 25%.
Il est important de souligner que la procédure PROC NPAR1WAY n'offre pas la possibilité d'utiliser les pondérations.
Si on compare les résultats obtenus avec la procédure PROC MEANS sans les poids du paragraphe 3.3.1 (T-test), on voit que les conclusions sont les mêmes. Pour les groupes "TOUS" et "COUPLE", on rejette l'hypothèse et pour le groupe de répondants "COUPLES DE MOINS DE 40 ANS" on ne rejette plus l'hypothèse. Ceci vient également confirmer les constatations réalisées sur les résultats du graphique 1.

Pour la variable i05ptotn, on teste l'hypothèse nulle d'égalité entre les moyennes de deux niveaux de formation donnés. Le choix des différents tests qui vont être réalisés, se fait à l'aide du graphique 2 (paragraphe 3.3.1). Il nous renseigne sur les recoupements ou non des intervalles de confiance. Si les intervalles de confiance entre deux niveaux de formation se recoupent ou sont très proche, on va tester l'égalité des moyennes. Le tableau 17 présente les résultats obtenus avec l'utilisation du test de Wilcoxon.

Tableau 17:résultats du test de Wilcoxon pour la variable i05ptotn selon le niveau de formation

		Niveaux de formation	n	Scores moyens	Test de Wilcoxon	p-value
1	Niveau de formation 1 et niveau de formation 4	1	87	219.328	19081.500	<.0001
		4	620	372.898		
2	Niveau de formation 4 et niveau de formation 5	4	620	342.434	24019.000	0.5295
		5	67	358.493		
3	Niveau de formation 4 et niveau de formation 6	4	620	359.783	53330.500	0.0005
		6	123	433.581		
4	Niveau de formation 5 et niveau de formation 6	5	67	85.388	5721.000	0.0616
		6	123	101.008		
5	Niveau de formation 5 et niveau de formation 7	5	67	90.784	6082.500	<.0001
		7	176	133.884		
6	Niveau de formation 6 et niveau de formation 7	6	123	134.053	16488.500	0.0077
		7	176	161.145		
7	Niveau de formation 6 et niveau de formation 8	6	123	96.211	11171.000	0.0019
		8	91	122.758		
8	Niveau de formation 7 et niveau de formation 8	7	176	130.497	12810.500	0.3030
		8	81	140.774		
9	Niveau de formation 7 et niveau de formation 9	7	176	145.724	29298.500	<.0001
		9	155	189.023		
10	Niveau de formation 8 et niveau de formation 9	8	91	108.445	9868.500	0.0110
		9	155	132.339		
11	Niveau de formation 8 et niveau de formation 10	8	91	155.533	14153.500	0.0018
		10	280	195.902		
12	Niveau de formation 9 et niveau de formation 10	9	155	209.474	32468.500	0.2928
		10	280	222.720		

Source: PSM_I et PSM_II

On rejette l'hypothèse nulle d'égalité des moyennes lorsque la p-value est plus petite que 0.05. C'est le cas des analyses numéro 1, 3, 5, 6, 7, 9, 10 et 11. Pour les autres analyses, les résultats obtenus ne permettent pas de rejeter l'hypothèse d'égalité. Le graphique 2 illustre bien les

résultats obtenus. Les personnes ayant fait un apprentissage ou une école professionnelle à plein temps ont un revenu annuel net moyen similaire, de même pour ceux ayant fait une école professionnelle à plein temps ou la maturité. Mais on ne peut pas dire que les personnes ayant fait une maturité et ceux ayant fait un apprentissage ont un revenu annuel net moyen similaire. Par ailleurs, on peut dire que les personnes ayant suivi une formation professionnelle supérieure ou une école technique ou professionnelle ont un revenu moyen annuel net semblable, de même pour les personnes ayant fait une école professionnelle supérieure, l'université ou une haute école.
Comme la procédure PROC NPAR1WAY n'offre pas la possibilité d'utiliser les poids, il n'est malheureusement pas possible de voir si les différences apportées par les pondérations (voir graphique 2) changent ou confirment les résultats obtenus.

3.5. La régression linéaire

Les procédures PROC REG et PROC SURVEYREG nous permettent de faire des régressions simples ou multiples par la méthode des moindres carrés. Rappelons que, pour que la méthode de régression puisse s'appliquer, certaines hypothèses de départ doivent être prises en compte. Les hypothèses du modèle sont :

- La linéarité du modèle :

  Il doit s'écrire : $y_i = \beta_0 + \sum_{j=1}^{p-1} \beta_j x_{ij} + \varepsilon_i$ où β représente les paramètres à estimer.

- Les erreurs ε_i sont indépendantes et identiquement distribuées selon une loi normale centrée et de variance σ^2 (variance homoscédastique).

Pour vérifier ces hypothèses, il faut contrôler :

- La linéarité de la relation entre y et les x_j.
- La constance de la variance des erreurs
- L'indépendance des erreurs
- La normalité des erreurs

Ces vérifications se font essentiellement par le biais de graphiques. Le graphique entre les résidus et les valeurs prédites nous fournit des informations pour la validation des hypothèses de linéarité et d'homogénéité (variance constante). Le QQ-plot des résidus nous permet de vérifier la normalité de la distribution des erreurs. Si les erreurs sont normalement distribuées, les points sur le graphique doivent être à peu près alignés sur la droite d'équation $e_i = q_i$.

Par ailleurs, la régression par la méthode des moindres carrés n'est pas une méthode robuste. Cela signifie que les résultats peuvent être fortement influencés par certaines observations extrêmes (« outliers »). Cependant les « outliers » contiennent souvent des informations importantes qui peuvent nous amener à découvrir des éléments intéressants. Il ne faut donc pas les éliminer automatiquement sans tenter d'en comprendre la source.
Il faut également faire attention à la corrélation possible entre les variables explicatives car l'existence de telles corrélations entraîne des erreurs dans l'estimation des paramètres lors de l'application de la méthode des moindres carrés. Un tableau de corrélation entre les différentes variables explicatives nous permet de vérifier la présence ou l'absence de colinéarité. Une autre façon de remédier à ce problème est d'utiliser une des méthodes de sélection des variables (forward, backward ou stepwise). Mais il est fortement conseillé de ne pas faire une confiance aveugle en la procédure automatique de sélection des variables pour éliminer toutes les colinéarités potentielles.

Pour illustrer les procédures PROC REG, PROC SURVEYREG de SAS et GENERAL LINEAR MODEL du module COMPLEX SAMPLE de SPSS, nous avons choisi la variable *p05c46* (le poids en kg), comme variable dépendante. Les variables explicatives sont la taille en cm, (*p05c45)*, l'âge, (*age05)* le sexe, (*sex)* ainsi que les variables *maigreur* et *obésité*[20].
La population d'intérêt comprend 6532 individus. Il s'agit de personnes de 14+ ans appartenant au PSM_I et au PSM_II et dont les poids sont supérieurs à zéro. Les poids choisis pour cet exemple sont les poids transversaux individuels permettant d'exploiter le PSM_I et le PSM_II combinés où la taille de l'échantillon de 2005 reste inchangée (wp05t1s).

Avant de se lancer dans la construction du modèle de régression, il est important de réaliser des analyses descriptives des variables entrant en jeu.
Des valeurs manquantes apparaissent pour les variables *p05c46* (poids en kg) et *p05c45* (taille en cm). Dans les deux cas, ce sont des observations se trouvant dans les modalités de réponses "pas de réponse" et "ne sais pas". Le détail se trouve en annexe (annexe X). Ces observations ont été éliminées et nous restreignons la population aux 6464 personnes pour lesquelles nous avons une réponse pour toutes les variables choisies.
Pour l'estimation du modèle de régression linéaire, on s'intéresse à la sous-population des 14-20 ans. Ceci restreint notre échantillon à 730 individus. On trouve, en annexe, une description des différentes variables pour toute la population et pour le sous-groupe (voir annexe XI et annexe XI bis).

Pour réaliser les différents modèles de régression, nous utilisons, dans SAS, les commandes suivantes.

- La procédure REG :

`proc reg data=nom_base_de_données;`
`  model var_dependante=var_explicatives;`
`  by variables_categories;`
 	
`  weight poids;`
 	
`  output out=nom_nouvelle_base_de_données`
`  predicted= y_predict residual= residu;`
 	
`quit; run;`
	- Indique la variable permettant de faire des analyses sur des sous populations.
	- Indique la variable qui permet de pondérer.
	- Indique le nom de la base de données produite qui contiendra les statistiques calculées.

[20] Les variables maigreur et obésité sont construites sur la base de l'IMC (indice de masse corporelle) et prennent la valeur 1 si le critère de sous poids ou respectivement de sur poids est satisfait. Une personne entre dans la catégorie « maigreur » si son IMC est < 18.5 et une personne entre dans la catégorie « obésité » si son IMC >25.

La procédure SURVEYREG :

`proc surveyreg data= nom_base_de_données`
`N= nom_base_de_données;`
 	
`model var_dependante=var_explicatives`
`/anova solution vadjust=none;`
`by variables_categories;`
 	
`strata strates;`
 	
`weight poids;`
 	
`quit; run;`
	- Indique le nom de la base de données contenant les tailles des strates définies par la variable « *strates* » de la commande STRATA
	
	- Indique la variable permettant de faire des analyses sur des sous populations.
	- Pour un plan stratifié, indique la variable qui forme les strates.
	- Indique la variable qui permet de pondérer.

Nous avons également réalisé les analyses de régression linéaire correspondant à la procédure PROC SURVEYREG de SAS avec le logiciel SPSS. Pour cela nous avons utilisé la commande GENERAL LINEAR MODEL du module COMPLEX SAMPLE (pour la syntaxe voir paragraphe 2.3.2). Par ailleurs, SAS utilise un ajustement de la variance (voir annexe I, la procédure SURVEYREG). Pour obtenir les mêmes résultats dans SAS et SPSS il faut omettre cet ajustement et mettre dans la syntaxe de SAS l'option `vadjust=none`.
Le tableau 18 montre les résultats obtenus pour la régression linéaire sur la sous-population des 14-20 ans

Tableau 18 : Récapitulatif des résultats obtenus pour la régression pour les 14-20 ans

Procédure et options utilisées		**Estimations des paramètres**		**Ecart-type**	**p-value**	**R^2**
SAS	PROC REG	Taille Age Sexe Maigreur Obésité	0.738 0.295 -1.258 -10.754 17.468	0.026 0.097 0.450 0.469 0.745	<.0001 0.0026 0.0053 <.0001 <.0001	81.59%
	PROC REG WEIGHT :wp05t1s	Taille Age Sexe Maigreur Obésité	0.736 0.257 -1.125 -11.045 16.410	0.026 0.095 0.452 0.481 0.757	<.0001 0.0067 0.0130 <.0001 <.0001	
	PROC SURVEYREG STRATA WEIGHT :wp05t1s,	Taille Age Sexe Maigreur Obésité	0.736 0.257 -1.125 -11.045 16.410	0.032 0.108 0.513 0.438 0.902	<.0001 0.0173 0.0285 <.0001 <.0001	80.49%
SPSS	GENERAL LINEAR MODEL du module COMPLEX SAMPLE	Taille Age Sexe Maigreur Obésité	0.736 0.257 -1.125 -11.045 16.410	0.032 0.108 0.513 0.438 0.902	<.0001 0.0173 0.0285 <.0001 <.0001	80.49%

N : 730

Source : PSM_I et PSM_II

Dans cet exemple, nous voyons tout d'abord que l'estimation des paramètres change très peu entre les différentes procédures. Nous constatons, par ailleurs, une augmentation des écart-types avec l'introduction des poids dans la procédure PROC REG et l'utilisation de la procédure PROC SUREVEYREG amène une nouvelle augmentation de ces derniers. Ceci a pour effet d'augmenter les p-value.
Dans notre modèle de régression linéaire, malgré l'augmentation de leurs écart-types et de leurs p-value les variables âge et sexe, qui sont les plus sensibles à ces variations, restent significatives. Notons que dans certains cas, le fait d'utiliser un calcul de variance approprié comme c'est le cas avec SURVEYREG pourrait rendre certaines variables non significatives et donc changer le modèle et par là son interprétation.

Par rapport aux résultats obtenus dans SPSS, nous constatons qu'ils coïncident avec ceux obtenus dans la procédure SURVEYREG de SAS.

Cependant, pour être sûr que nos résultats sont valides, il faut vérifier certaines hypothèses de départ du modèle. Entre autre, nous avons vérifié qu'il n'y ait pas de problème de colinéarité entre les variables entrant dans le modèle. Pour cela nous examinons les corrélations entre celles-ci (tableau 19).

Tableau 19 : Tableau de corrélations entre les différentes variables du modèle de régression, pour les 14-20 ans, (sans pondération)

	Poids en kg	**Taille en cm**	**Age**	**Sexe**	**Maigreur**	**Obésité**
Poids en kg	1	0.69	0.29	-0.46	-0.52	0.43
Tailles en cm		1	0.21	**-0.61**	-0.12	-0.00
Age			1	-0.01	-0.27	0.02
Sexe				1	0.05	-0.04
Maigreur					1	-0.12
Obésité						1

Source : PSM_I et PSM_II

Tout d'abord, nous voyons que les variables taille (en cm), maigreur et obésité sont bien corrélées avec la variable dépendante. Il s'agit des variables qui apparaissent comme étant les plus significatives dans le modèle de régression quel que soit la procédure utilisée. Mais on constate aussi que la variable "sexe" est très corrélée avec une autre variable explicative, la variable "taille". On a ici un problème de colinéarité. Ceci a motivé la réalisation de deux analyses de régression linéaire selon le sexe. Le tableau 20 nous montre les corrélations des différentes variables pour les hommes de 14-20 ans

Tableau 20 : Tableau de corrélations entre les différentes variables du modèle de régression, pour les hommes de 14-20 ans, (sans pondération)

	Poids en kg	**Taille en cm**	**Age**	**Maigreur**	**Obésité**
Poids en kg	1	0.65	0.41	-0.60	0.44
Tailles en cm		1	0.33	-0.24	-0.01
Age			1	-0.35	0.06
Maigreur				1	-0.13
Obésité					1

Source : PSM_I et PSM_II

Pour ces nouvelles corrélations, on voit que les différentes variables explicatives du modèle sont bien corrélées avec la variable dépendante et on ne constate plus de problèmes de colinéarité.
A la vue de ces résultats, un nouveau modèle de régression linéaire a été recalculé pour uniquement les hommes de 14-20 ans. Le tableau 21 récapitule les résultats obtenus pour cette sous-population.

Tableau 21 : Récapitulatif des résultats obtenus pour la régression pour les hommes de 14-20 ans

Procédure et options utilisées		**Estimations des paramètres**		**Ecart-type**	**p-value**	**R^2**
SAS	PROC REG	Taille	0.756	0.036	<.0001	79.01%
		Age	0.429	0.161	0.0081	
		Maigreur	-11.895	0.783	<.0001	
		Obésité	17.574	1.077	<.0001	
	PROC REG. WEIGHT :wp05t1s	Taille	0.749	0.036	<.0001	
		Age	0.427	0.158	0.0043	
		Maigreur	-12.136	0.819	<.0001	
		Obésité	16.665	1.081	<.0001	
	PROC SURVEYREG STRATA WEIGHT :wp05t1s	Taille	0.749	0.044	<.0001	78.07%
		Age	0.427	0.173	0.0140	
		Maigreur	-12.136	0.724	<.0001	
		Obésité	16.665	1.328	<.0001	
SPSS	GENERAL LINEAR MODEL du module COMPLEX SAMPLE	Taille	0.749	0.044	<.0001	78.07%
		Age	0.427	0.173	0.0140	
		Maigreur	-12.136	0.724	<.0001	
		Obésité	16.665	1.328	<.0001	

N : 370

Source : PSM_I et PSM_II

Dans ce tableau, on voit que l'estimation des coefficients de régression change très peu avec l'utilisation des différentes procédures. En ce qui concerne les écart-types, on observe, avec l'introduction des poids dans la procédure PROC REG, une légère augmentation de ceux-ci pour toutes les variables à l'exception de l'âge. Avec l'utilisation de la procédure PROC SURVEYREG, on voit de nouvelles variations des écarts-type. On constate également que la p-value de la variable "âge" est sensible aux différentes procédures. Dans le cas de l'utilisation de la procédure SURVEYREG sa valeur se rapproche du seuil de 0.05. Dans cet exemple toutes les variables restent significatives mais n'oublions pas que dans certains cas, l'utilisation d'un calcul approprié de la variance, pourrait en rendre certaines non significatives et changer le modèle et son interprétation.

Dans cet exemple aussi, les résultats obtenus avec la procédure GENERAL LINEAR MODEL du module COMPLEX SAMPLE de SPSS sont exactement les mêmes que ceux obtenus avec la procédure SURVEYREG de SAS.

Nous avons également réalisé la régression pour les femmes de 14-20 ans. Pour ce deuxième groupe, la variable "âge" est non significative pour tous les modèles. Dans le tableau 22, on a les corrélations entre les différentes variables pour la sous-population des femmes de 14-20ans.

Tableau 22 : Tableau de corrélations entre les différentes variables du modèle de régression, pour les femmes de 14-20 ans, (sans pondération).

	Poids en kg	**Taille en cm**	**Age**	**Maigreur**	**Obésité**
Poids en kg	1	0.45	**0.22**	-0.54	0.49
Tailles en cm		1	0.18	0.04	-0.08
Age			1	-0.20	-0.02
Maigreur				1	-0.12
Obésité					1

Source : PSM_I et PSM_II

Ce tableau nous montre que la variable "âge" est effectivement peu corrélée avec la variable dépendante du modèle.
Nous avons donc refait les analyses en enlevant cette dernière. Les conclusions qui en ressortent sont les mêmes que pour les modèles présentés précédemment. Par ailleurs, pour le modèle sans l'âge des femmes de 14-20 ans, toutes les variables restantes sont significatives avec une p-value <.0001.

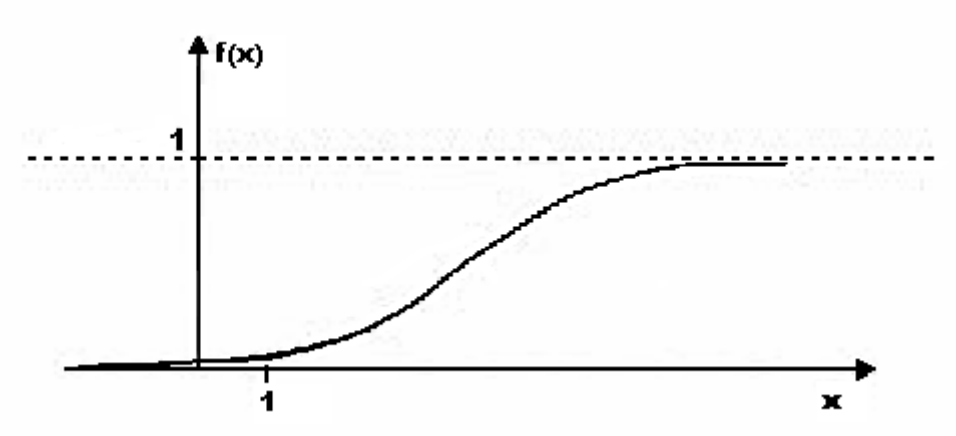
Vérifions une autre hypothèse de départ du modèle. Il s'agit de la normalité des erreurs. Pour cela nous avons réalisé les QQ-plots des résidus standardisés, premièrement pour le modèle avec toute la population de 14-20 ans et deuxièmement pour les hommes de 14-20 ans. Les graphiques se trouvent en annexe (voir annexe XII et XIII). Dans les deux cas l'hypothèse est vérifiée.

3.6. La régression logistique

Les procédures PROC LOGISTIC, PROC SURVEYLOGISTC de SAS et LOGISTIC REGRESSION du module COMPLEX SAMPLE de SPSS permettent de faire des régressions logistiques. Ce type de régression est utilisé lorsque la variable dépendante n'est pas quantitative mais qualitative ou catégorielle. En particulier, la régression logistique permet d'expliquer sous forme de probabilité la relation entre une variable dichotomique et une ou plusieurs variables explicatives. Celles-ci peuvent être dichotomiques ou quantitatives. Dans le cas où les variables explicatives sont quantitatives, le modèle suppose que leurs distributions soient normales. Par ailleurs, la relation entre la probabilité modélisée et les variables explicatives est supposée linéaire (à un terme près).
Ci-dessous, nous avons la formule générale du modèle de la régression logistique :

$$P(Y|X_i) = f(x) = \frac{1}{1+e^{-(\alpha+\sum\beta_i X_i)}} = \frac{e^{(\alpha+\sum\beta_i X_i)}}{1+e^{(\alpha+\sum\beta_i X_i)}}$$

où $(Y|X_i)$ signifie la probabilité de survenue de la catégorie d'intérêt de la variable dépendante Y en fonction des variables explicatives ou facteurs X_i. Une probabilité prenant des valeurs comprises entre 0 et 1, la régression linéaire est inutilisable. La fonction logistique ($f(x)$) a comme représentation une forme sigmoïdale.



L'estimation du modèle de régression logistique se fait généralement par la méthode du maximum de vraisemblance. Il suppose que les observations individuelles sont indépendantes. En comparaison avec la régression linéaire, l'hypothèse de normalité de la distribution n'est pas appropriée. On suppose également les ε_i d'espérance nulle mais leur variance n'est pas constante (elle dépend de x).
La régression logistique se distingue de la régression linéaire par les points suivant :

- La distribution de la variable dépendante n'est pas Normale.
- Le modèle de régression est non linéaire.
- La variance des ε_i est hétéroscédastique.
- L'estimation des paramètres se fait par la méthode du maximum de vraisemblance.

Pour illustrer ces procédures, on a choisi de modéliser les bas revenus.
La variable dépendante représentant les bas revenus est construite à partir de la variable *i05ptotn* (le revenu total personnel annuel net). Les variables explicatives sont le niveau de formation (*educat05)*, le sexe (*sex*), l'âge (*age05*), le nombre d'enfants de 0 à 17 ans dans le ménage (*nbkid05),* le nombre de personnes dans le ménage (*nbpers05)* et l'état civil

(*civsta05*). Pour ces analyses on va prendre les données combinées du PSM_I et du PSM_II ainsi que celles de l'enquête SILC et on ne va s'intéresser qu'aux personnes âgées d'au moins 21 ans, travaillant à plein temps et n'ayant eu aucune modification d'activité au cours de l'année de l'interview[21].
Suite à quelques analyses descriptives, on sait que la variable *i05ptotn* a des valeurs manquantes (voir annexe XIV). L'annexe XV, quant à elle, nous donne plus d'informations sur les autres variables du modèle.

Pour réaliser nos analyses, nous avons enlevé de la variable *i05ptotn* les valeurs manquantes. Il reste 2916 observations valides. Pour la construction de la variable dépendante (bas revenu versus autre), on fixe le seuil des bas revenus à 60% du revenu médian. Dans notre cas, il s'agit des 60% de 72375 francs. Le seuil se situe donc à 43425 francs[22] par année. On obtient 10.36% de personnes se trouvant dans la catégorie des bas revenus.
Par ailleurs, on a recodé chacune des modalités des différentes variables en une variable dichotomiques (pour plus d'information voir annexe XII).

La syntaxe utilisée dans SAS pour réaliser les procédures PROC LOGISTIC et PROC SURVEYLOGISTIC est la suivante :

- La procédure LOGISTIC :

`proc logistic data=nom_base_de_données;`
`  model var_dépendante=var_explicatives`
`  weight poids;`
`run;`
	- Indique la variable qui permet de pondérer.

- La procédure SURVEYLOGISTIC :

`proc surveylogistic data= nom_base_de_données`
`  N= nom_base_de_données;`
 	
`  model var_dépendante=var_explicatives`
`  /vadjust=none ;`
`  strata strates;`
 	
`  weight poids;`
`run;`
	- Indique le nom de la base de données contenant les tailles des strates définies par la variable « *strates* » de la commande STRATA
	
	- Pour un plan stratifié, indique la variable qui forme les strates.
	- Indique la variable qui permet de pondérer.

On a également réalisé les mêmes analyses de régression logistique dans SPSS avec l'analogue à la procédure SURVEYLOGISTIC de SAS. Pour cela on a utilisé la commande LOGISTIC REGRESSION du module COMPLEX SAMPLE après avoir défini le plan comme expliqué au paragraphe 2.3.2. Pour obtenir les mêmes résultats avec SAS et SPSS, il faut dans la syntaxe de SAS mettre l'option `vadjust=none.` On omet ainsi l'ajustement de la variance calculé par SAS (voir annexe I, la procédure SURVEYLOGISTIC).

[21] La variable utilisée pour déterminer si le travail est à plein temps est la variable *p05w39* et celle utilisée pour déterminer s'il y a eu des changements professionnels est la *p05w177*.
[22] Définition d'un bas revenu : est considéré comme bas revenu tout revenu <= 43425 francs

Lors de la réalisation de ces procédures, il faut mettre pour chacune des variables une modalité de référence (voir annexe XVI, les modalités en évidence). On a d'abord fait les analyses en introduisant toutes les variables dans le modèle (voir tableau 23) et ensuite, on n'a retenu que les variables pertinentes (voir tableau 24). Les poids utilisés pour ces analyses sont les poids transversaux individuels, où la taille de l'échantillon de 2005 reste inchangée, pour SILC_I, SILC_II, PSM_I et PSM_II combinés (wp05t2s), supérieur à zéro.
La commande CLUSTER avec comme variable l'identifiant ménage (idhous05) est introduite dans la procédure PROC SURVEYLOGISTIC. Dans le modèle de régression logistique, on utilise différentes variables pour lesquelles l'appartenance à un ménage implique que les personnes de celui-ci aient des caractéristiques similaires.

Tableau 23 : Résultats de la régression logistique avec introduction de toutes les variables dans le modèle

Procédure et options utilisées	**Estimations des paramètres**		**Ecart-type**	**p-value**
PROC LOGISTIC	**Nivform_bas**	**1.524**	**0.220**	**<.0001**
	Nivform_moyen	**0.587**	**0.168**	**0.0005**
	Age21_30ans	**0.514**	**0.232**	**0.0267**
	Age31_40ans	-0.042	0.215	0.8445
	Age41_50ans	-0.074	0.198	0.7081
	Enfant_0	0.127	0.353	0.7202
	Enfant_1	-0.150	0.353	0.6706
	Enfant_2	-0.222	0.315	0.4805
	Celibataire	0.156	0.205	0.4469
	Veuf	0.840	0.411	0.0406
	Separe	**-0.234**	**0.254**	**0.3571**
	divorce	-1.130	0.763	0.1388
	Femme	**1.273**	**0.141**	**<.0001**
	Persmenage_1	**-0.641**	**0.269**	**0.0172**
	Persmenage_2	**-0.527**	**0.232**	**0.0231**
	Persmenage_3	-0.280	0.242	0.2477
PROC LOGISTIC /STEPWISE	**Nivform_bas**	**1.485**	**0.217**	**<.0001**
	Nivform_moyen	**0.581**	**0.167**	**0.0005**
	Age21_31ans	**0.711**	**0.145**	**<.0001**
	Femme	**1.148**	**0.129**	**<.0001**
PROC LOGISTIC WEIGHT : wp05t2s	**Nivform_bas**	**1.157**	**0.201**	**<.0001**
	Nivform_moyen	**0.531**	**0.162**	**0.0010**
	Age21_31ans	**0.640**	**0.212**	**0.0025**
	Age31_40ans	-0.161	0.200	0.4206
	Age41_50ans	0.055	0.188	0.7678
	Enfant_0	-0.198	0.337	0.5570
	Enfant_1	-0.302	0.328	0.3572
	Enfant_2	-0.229	0.288	0.4279
	Celibataire	0.066	0.185	0.7206
	Veuf	-0.857	0.680	0.2075
	Separe	**0.820**	**0.394**	**0.0376**
	divorce	-0.419	0.249	0.0926
	Femme	**1.238**	**0.131**	**<.0001**
	Persmenage_1	-0.423	0.264	0.1099
	Persmenage_2	-0.223	0.221	0.3120
	Persmenage_3	-0.190	0.226	0.4013

Tableau 23 (suite) : Résultats de la régression logistique avec introduction de toutes les variables dans les modèles)

PROC LOGISTIC WEIGHT : wp05t2s /STEPWISE	**Nivform_bas**	**1.162**	**0.199**	**<.0001**
	Nivform_moyen	**0.550**	**0.161**	**0.0006**
	Age21_31ans	**0.699**	**0.136**	**<.0001**
	Divorce	**-0.485**	**0.233**	**0.0370**
	Femme	**1.155**	**0.124**	**<.0001**
PROC SURVEYLOGISTIC STRATA / CLUSTER WEIGHT : wp05t2s	**Nivform_bas**	**1.157**	**0.244**	**<.0001**
	Nivform_moyen	**0.531**	**0.196**	**0.0066**
	Age21_31ans	**0.640**	**0.245**	**0.0089**
	Age31_40ans	0.161	0.246	0.5126
	Age41_50ans	0.055	0.220	0.8010
	Enfant_0	0.198	0.407	0.6272
	Enfant_1	0.302	0.413	0.4656
	Enfant_2	0.229	0.383	0.5501
	Celibataire	0.066	0.228	0.7711
	Veuf	0.857	0.769	0.2651
	Separe	0.820	0.462	0.0760
	divorce	0.419	0.302	0.1656
	Femme	**1.238**	**0.172**	**<.0001**
	Persmenage_1	0.423	0.305	0.1659
	Persmenage_2	0.223	0.262	0.3951
	Persmenage_3	0.190	0.272	0.4844

N : 2916

Source : PSM_I, PSM_II, SILC_I et SILC_II

Ce premier tableau montre tout d'abord qu'on obtient des modèles différents selon la procédure utilisée et le fait d'introduire des poids ou pas. La régression logistique sans les poids fait apparaître 7 variables significatives. En utilisant les poids, il en apparaît cinq et avec la procédure SURVEYLOGISTIC, le modèle ne retient plus que 4 variables significatives. De même, l'utilisation de la méthode stepwise pour la sélection des variables ne donne pas les mêmes résultats si elle est utilisée sans les poids ou avec ceux-ci. On constate que lorsque toutes les variables sont introduites dans le modèle, le nombre de variables apparaissant comme significatives diminue en introduisant les pondérations et ce nombre est encore plus restreint lorsqu'on utilise la procédure SURVEYLOGISTIC. Par contre le résultat du stepwise aboutit à moins de variables sans les poids qu'avec les poids. Dans les résultats obtenus avec SURVEYLOGISTIC, on voit que le fait d'utiliser un calcul approprié de la variance a pour effet de rendre certaines variables non significatives.
Pour pourvoir comparer les coefficients des paramètres entre les différentes procédures, on a refait les analyses jusqu'à l'obtention du modèle qui ne retient que les variables du niveau de formation (bas et moyen), de l'âge (compris entre 21 et 30 ans) et du sexe (femme). Les résultats se trouvent dans le tableau 24.

Tableau 24 : Résultats de la régression logistique avec uniquement les quatre variables retenues dans le modèle

Procédure et options utilisées		Estimations des paramètres		Ecart-type	p-value
SAS	PROC LOGISTIC	Nivform_bas	1.485	0.217	<.0001
		Nivform_moyen	0.581	0.167	0.0005
		Age21_31ans	0.711	0.145	<.0001
		Femme	1.148	0.129	<.0001
	PROC LOGISTIC WEIGHT : wp05t2s	Nivform_bas	1.128	0.198	<.0001
		Nivform_moyen	0.528	0.161	0.0010
		Age21_31ans	0.766	0.132	<.0001
		Femme	1.103	0.121	<.0001
	PROC SURVEYLOGISTIC STRATA / CLUSTER WEIGHT : wp05t2s	Nivform_bas	1.128	0.241	<.0001
		Nivform_moyen	0.528	0.195	0.0067
		Age21_31ans	0.766	0.167	<.0001
		Femme	1.103	0.152	<.0001
SPSS	LOGISTIC REGRESSION du module COMPLEX SAMPLE	Nivform_bas	1.128	0.241	<.0001
		Nivform_moyen	0.528	0.195	0.0067
		Age21_31ans	0.766	0.167	<.0001
		Femme	1.103	0.152	<.0001

N : 2916

Source : PSM_I, PSM_II, SILC_I et SILC_II

Tout d'abord, on constate un léger changement dans l'estimation des paramètres entre la procédure PROC LOGISTIC sans les poids et celle avec les pondérations. La p-value quant à elle augmente légèrement pour la variable *nivform_moyen*. Lors de l'utilisation de la procédure PROC SURVEYLOGISTIC, on observe une augmentation de tous les écart-types, en comparaison avec les résultats de la procédure PROC LOGISTIC (avec poids) ainsi qu'une nouvelle augmentation de la p-value pour la variable *nivform_moyen*.

Avec la procédure LOGISTIC REGRESSION du module COMPLEX SAMPLE de SPSS, on obtient les mêmes résultats que ceux de la procédure SURVEYLOGISTIC de SAS. SPSS offre donc les mêmes possibilités que SAS…

4. Conclusion

Lors d'analyses de données issues d'enquêtes telles que le Panel Suisse de Ménage (PSM) ou l'Enquête sur les Revenus et les Conditions de Vie (SILC), il est nécessaire d'utiliser des procédures adéquates. Les différents exemples présentés montrent l'effet et l'importance de l'utilisation des pondérations mais également l'importance de la prise en compte de la complexité du plan de sondage pour ne pas obtenir des résultats erronés. Par ailleurs, il est possible de réaliser des analyses où le calcul de la variance est adéquat autant bien avec le logiciel SAS qu'avec le logiciel SPSS.

Les exemples présentés illustrent différents types d'analyses couramment utilisées. Il s'agit de tableaux de fréquences, du test d'indépendance (test du Chi-carré), d'analyses descriptives telles que la moyenne, l'écart-type,…, de tests d'égalité de moyennes (T-test, test de Wilcoxon), de régressions linéaire et de régressions logistique.

Pour les tableaux de fréquences, deux analyses ont été réalisées (l'analyse de la variable p05c01 et de la variable educat05). Pour la variable p05c01 (état de santé), les résultats obtenus montrent peu de différences selon qu'on utilise les poids ou pas, et qu'on calcule la variance de manière standard ou par linéarisation en série de Taylor. Cependant pour les résultats de la variable educat05 (niveau de formation) on constate de grandes différences dans les pourcentages qu'on obtient pour les différentes modalités lorsqu'on utilise les poids. De plus, les intervalles de confiance indiquant la précision des pourcentages obtenus changent énormément avec l'introduction des pondérations dans l'analyse.
Les résultats obtenus pour le test du Chi-carré montrent que selon l'analyse faite l'influence sur le résultat des pondérations et de la prise en compte de la correction tenant compte de la complexité du plan de sondage n'est pas la même. Dans certains cas, le résultat est modifié et dans d'autres les conclusions ne changent pas.
L'analyse descriptive (moyenne, écart-type, minimum, maximum,…) de la variable i05htyn (revenu annuel net du ménage) montre que l'utilisation des pondérations influence énormément les résultats. On observe un changement de la moyenne estimée et les intervalles de confiance pour la moyenne estimée avec poids et sans poids ne se recoupent pas. Contrairement à la variable i05ythn, pour l'analyse de la variable p05p01 (intérêt pour le politique), on ne constate quasiment pas de différence dans les résultats obtenus entre les différentes procédures.
Les exemples choisis pour illustrer le T-test n'ont pas fournit des résultats interprétable du fait que le pré-requis de normalité n'est satisfait dans aucun des cas. Cependant, il est possible de réaliser des tests sur les moyennes en utilisant par exemple le test non-paramétrique de Wilcoxon. Malheureusement, la procédure permettant de faire ce type de test ne permet pas de prendre en compte les pondérations et la complexité du plan de sondage ce qui est le cas avec la procédure permettant de faire le T-test. Par ailleurs, en comparant les résultats obtenus avec le T-test et ceux obtenus avec le test de Wilcoxon, on observe que les résultats sont proches malgré les imperfections.
Pour la régression linéaire, l'estimation des paramètres changent peu selon les procédures utilisées. Cependant, on observe une augmentation des écart-types correspondant lors de l'utilisation des poids et lors de l'utilisation de la procédure PROC SURVEYREG (ou GENERAL LINEAR MODEL du module COMPLEX SAMPLE de SPSS). Dans l'exemple présenté, le choix du modèle n'est pas influencé par les pondérations mais l'augmentation des écart-types pourrait rendre non significatives certaines variables dans d'autres exemples.
Pour le modèle de régression logistique, on obtient des modèles très différents selon qu'on utilise les poids ou pas, qu'on calcule la variance de manière standard ou par linéarisation en série de Taylor et qu'on utilise la méthode de sélection de variables STEPWISE ou pas.

L'utilisation des poids ainsi que de procédures permettant de faire recourt à un calcul approprié de la variance ont pour effet de rendre non-significatives certaines variables ; on peut aboutir à d'autres modèles et donc à d'autres conclusions.

Ces résultats relèvent différents points important. On voit que certaines fois les pondérations ont peu d'effet. On observe ceci, lorsque les variables sont peu ou non corrélées avec les variables utilisées pour la construction des pondérations. Cependant, on observe presque toujours des changements à la hausse des écart-types. Ceci nous amène à dire que même si l'influence des pondérations sur l'estimateur est petite il faut tout de même les utiliser car l'estimation de la variance est plus juste. Dans les cas où de grandes différences apparaissent il est bien sûr vital d'utiliser les poids.
Outre l'utilisation primordiale des poids, comme les enquêtes considérées ne sont pas tirées selon un plan aléatoire simple, les procédures traditionnelles implémentées dans les logiciels statistiques ne fournissent pas un calcul adéquat des variances.
C'est pour cela qu'on recommande d'utiliser les procédures SURVEY de SAS. Elles fournissent des résultats bien plus corrects que ceux fournit par les procédures traditionnelles. Comme on a pu l'observer dans les exemples, il y a une augmentation de la variance estimée avec l'utilisation des procédures SURVEY. En effet, ces procédures ne sous-estiment pas, ou moins, ladite variance, comme c'est les cas avec les procédures traditionnelles. Elles permettent de prendre en compte toutes les spécificités de l'enquête.
Les procédures SURVEY donnent des résultats très similaire (du moins pour les enquêtes PSM et SILC) à ceux produit par des méthodes de ré-échantillonnage comme le bootstrap ou le jackknife.
Tous les exemples d'analyse faits avec les procédures SURVEY de SAS ont été reproduits avec le module COMPLEX SAMPLE de SPSS. Ce logiciel offre les mêmes possibilités que SAS. Les logiciels SUDAAN et STATA permettent également de prendre les mêmes mesures comme l'ont montré certains auteurs (Siller, 2005). On a donc maintenant la possibilité, quelque soit le logiciel utilisé, de calculer des variances par une méthode adaptée pour analyser les données d'enquêtes comme le PSM et SILC et les exemples montrent qu'on obtient les mêmes résultats. Notons encore qu'il faut prendre garde à la manière dont SAS et SPSS traitent les valeurs manquantes, car elle diffère quelque peu.

Pour conclure, il est important, dès que cela est possible, d'utiliser les poids quel que soit les analyses à effectuer. En plus, il existe la possibilité de réaliser simplement des estimations de variance plus justes et réalistes lors d'analyses sur des enquêtes comme le PSM et SILC, grâce aux les procédures SURVEY de SAS ou COMPLEX SAMPLE de SPSS.

5. Bibliographie

Borcard, D. (2006). Régression logistique. Département de sciences biologiques, Université de Montréal, Bio-2042.

Cauchon, C. et Latouche, M. (2007). Pondération du Panel suisse des ménages PSM I Vague 7 PSM II Vague 2 PSM I et PSM II combinés ; description détaillée des tâches. Statistiques Canada.

Chambers, R. l. et Skinner, C. J. (2003). Analysis of survey data. Wiley series in survey methodology.

Dodge, Y. et Rousson, V. (2004). Analyse de régression appliquée. Paris, Dunod.

Preux, P.M. et al. (2005). Qu'est-ce qu'une régression logistique ?. *Rev Mal Respir*[23], 2005, 22 : 159-62.

Renaud, A. (2004). Analyses de données d'enquêtes. Quelques méthodes et illustration avec des données de l'OFS, Neuchâtel, Office fédéral de la statistique.

Sites web :

Graf, E. (2007). Remarques sur les pondérations 2004 PSM et SILC.
http://www.swisspanel.ch/file/methods/Weighting_2004_remarks_F.pdf

Graf, E. (2007). Remarques sur les pondérations 2005 PSM et SILC.
http://www.swisspanel.ch/file/methods/Weighting_2005_remarks_F.pdf

FOVEA: fascicule de méthodologie statistique n° 7 ; La régression logistique.
http://www.fovea-group.com/cro/fr/pdf/fms7.pdf

Siller, A. et Tompkins, L. The big four ; analyzing complex sample survey data using SAS, SPSS, STRATA and SUDAAN.
http://www2.sas.com/proceedings/sugi31/172-31.pdf

What's new in data Analysis : sample survey design and analysis.
http://support.sas.com/rnd/app/da/new/dasurvey.html

[23] *Rev Mal Respir :* revue des maladies respiratoires

6. Annexes

Annexe I: formules des variances utilisées par les procédures SURVEY(…)

Les notations

- h : le numéro de la strate avec un total de H strates
- i : le numéro de la grappe dans la strate h avec un total de n_h grappes
- j : le numéro de l'unité dans la grappe i de la strate h avec un total de m_{hi} unités.
- $w_{...}$ représente la somme des poids sur toutes les observations de l'échantillon.
- w_{hij} : poids pour l'observation j de la grappe i, se trouvant dans la strate h.
- f_h : taux du premier niveau d'échantillonnage pour la strate h.
- $\delta_{hij}(r,c)$: variable indicatrice. Si l'observation (hij) se trouve dans la cellule (r,c) alors la variable est égale à 1 sinon elle est égale à zéro.

La procédure SURVEYFREQ :

Notations :

- $\hat{N}_{rc} = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} \delta_{hij}(r,c) W_{hij}$: Estimation du nombre de personnes se trouvant dans la ligne r et la colonne c.
- $\hat{V}_h\left(\hat{N}_{rc}\right) = \dfrac{n_h(1-f_h)}{n_h-1}\sum_{i=1}^{n_h}\left(n_{rc}^{hi} - \overline{n}_{rc}^{h}\right)^2$: Variance par strates.
- $n_{rc}^{hi} = \sum_{j=1}^{m_{hi}} \delta_{hij}(r,c) W_{hij}$: Nombre de personnes pondéré se trouvant dans la ligne r et la colonne c appartenant à la strate h et à la grappe i.

Variance pour la procédure SURVEYFREQ :

$$\hat{V}\left(\hat{N}_{rc}\right) = \sum_{\mathrm{i}=1}^{\mathrm{H}} \hat{\mathrm{V}}_{\mathrm{h}}\left(\hat{\mathrm{N}}_{\mathrm{rc}}\right) = \sum_{i=1}^{H}\left(\frac{n_h(1-f_h)}{n_h-1}\sum_{i=1}^{n_h}\left(n_{rc}^{hi} - \overline{n}_{rc}^{h}\right)^2\right)$$

La procédure SURVEYMEANS :

Notations :

- $\hat{\bar{Y}} = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij}\, y_{hij} \Big/ w_{...}$: Estimation de la moyenne.
- $\hat{V}_h(\hat{\bar{Y}}) = \dfrac{n_h(1-f_h)}{n_h-1}\sum_{i=1}^{n_h}(e_{hi.}-\bar{e}_{h..})^2$: Variance par strates.
- $e_{hi} = \left(\sum w_{hij}\left(y_{hij}-\hat{\bar{y}}\right)\right) \Big/ w_{...}$
- $\bar{e}_h = \left(\sum_{i=1}^{n_h} e_{hi}\right) \Big/ n_h$

Variance pour la procédure SURVEMEANS :

$$\hat{V}\left(\hat{\bar{Y}}\right) = \sum_{i=1}^{H}\left(\hat{V}_h(\hat{\bar{Y}})\right) = \sum_{i=1}^{H}\left(\frac{n_h(1-f_h)}{n_h-1}\sum_{i=1}^{n_h}(e_{hi.}-\bar{e}_{h..})^2\right)$$

La procédure SURVEYREG :

Notations :

- $G = \dfrac{n-1}{n-p}\sum_{i=1}^{H}\dfrac{n_h(1-f_h)}{n_h-1}\sum_{i=1}^{n_h}\left(e_{hi.}-\bar{e}_{h..}\right)'\left(e_{hi.}-\bar{e}_{h..}\right)$
- $e_{hi.} = \sum_{j=1}^{m_{hi}} e_{hij}$

  $e_{hij} = w_{hij} r_{hij} x_{hij}$

  $r = y - X\hat{\beta}$

Variance pour la procédure SURVEYREG :

$$\hat{V} = \left(X'WX\right)^{-} G \left(X'WX\right)^{-}$$

Le facteur (n-1)/(n-p) se trouvant dans le calcul de la matrice G devrait réduire le biais de l'échantillon associé à l'utilisation de l'estimation de la fonction dans le calcul des variances.
Dans le cas d'un échantillon aléatoire simple, ce facteur contribue à la correction des degrés de liberté appliquée au résidu moyen au carré pour les moindres carrés ordinaires à p paramètres.
Par défaut, la procédure va utiliser cet ajustement lors de l'estimation de la variance. C'est équivalent à spécifier l'option VADJUST=DF dans la commande MODEL. Si on ne veut pas utiliser cet ajustement dans l'estimation de la variance, on peut spécifier l'option VADJUST=NONE dans la commande MODEL pour supprimer ce facteur.

La procédure SURVEYLOGISTIC:

Notations:

- $\hat{Q} = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij}\hat{D}_{hij}\left(diag\left(\hat{\pi}_{hij}\right) - \hat{\pi}_{hij}\hat{\pi}'_{hij}\right)^{-1}\hat{D}'_{hij}$
- $\hat{G} = \frac{n-1}{n-p}\sum_{i=1}^{H}\frac{n_h(1-f_h)}{n_h-1}\sum_{i=1}^{n_h}(e_{hi.} - \bar{e}_{h..})(e_{hi.} - \bar{e}_{h..})'$
- $e_{hi.} = \sum w_{hij}\hat{D}_{hij}\left(diag\left(\hat{\pi}_{hij}\right) - \hat{\pi}_{hij}\hat{\pi}'_{hij}\right)^{-1}\left(y_{hij} - \hat{\pi}_{hij}\right)$

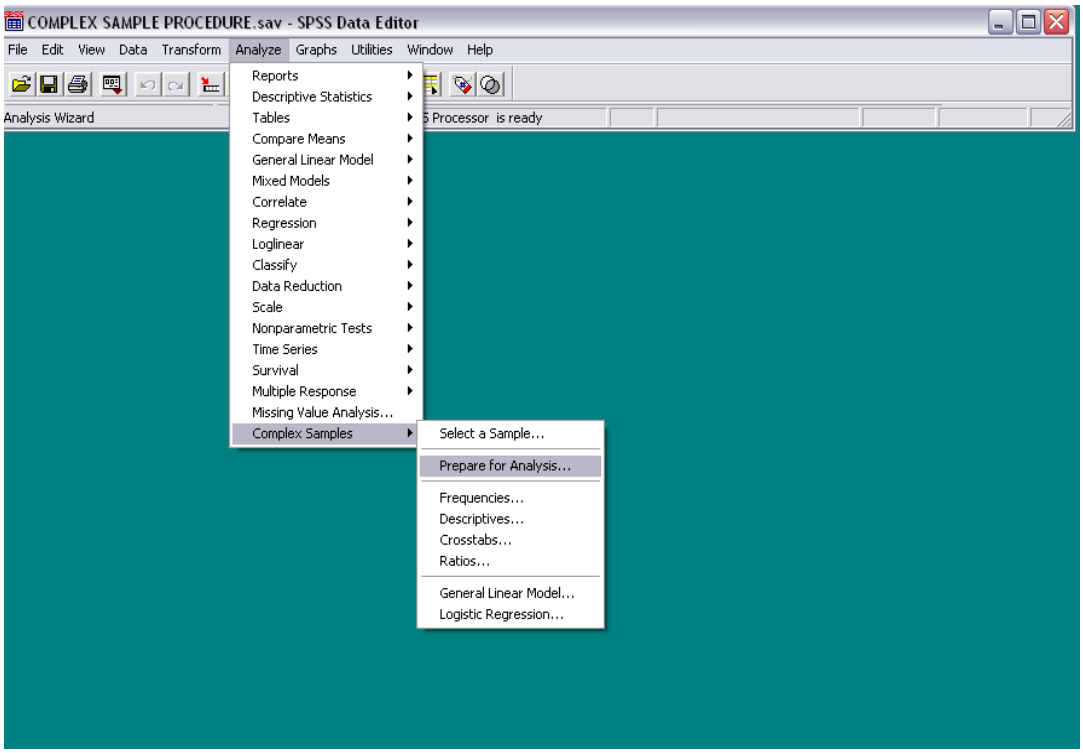
Variance pour la procédure SURVEYLOGISTIC:

$$\boxed{\hat{V} = \hat{Q}^{-1}\hat{G}\hat{Q}^{-1}}$$

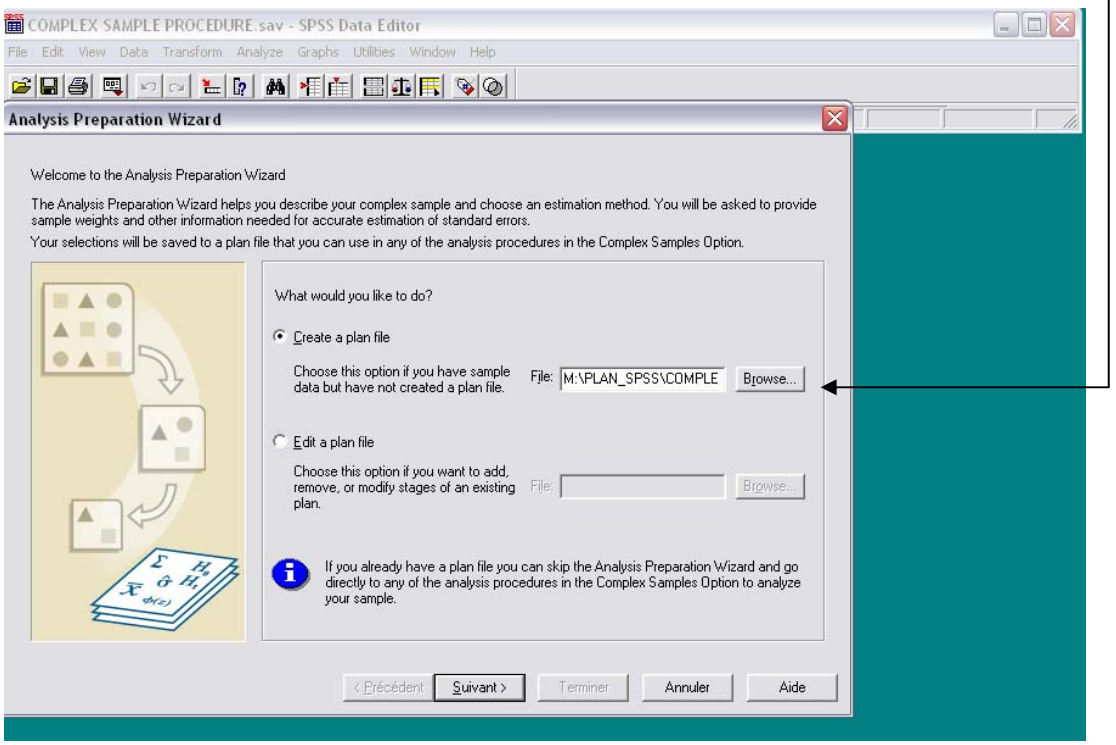
La procédure SURVEYLOGISTIC utilise également le facteur (n-1)/(n-p)dans le calcul de la matrice $\hat{G}$ et ceci pour réduire le biais de l'échantillon associé à l'utilisation de l'estimation de la fonction dans le calcul de la variance.
Dans le cas d'un échantillon aléatoire simple, ce facteur contribue à la correction des degrés de liberté appliquée au résidu moyen au carré pour les moindres carrés ordinaires à p paramètres.
Par défaut la procédure va utiliser cet ajustement lors de l'estimation de la variance. C'est équivalent à spécifier l'option VADJUST=DF dans la commande MODEL. Si on ne veut pas utiliser cet ajustement dans l'estimation de la variance, on peut spécifier l'option VADJUST=NONE dans la commande MODEL pour supprimer ce facteur.

Annexe II: Création du fichier PLAN dans SPSS

Première étape : aller dans le module *Complex Sample* puis dans *Prepare for Analysis* … pour la création du fichier PLAN.

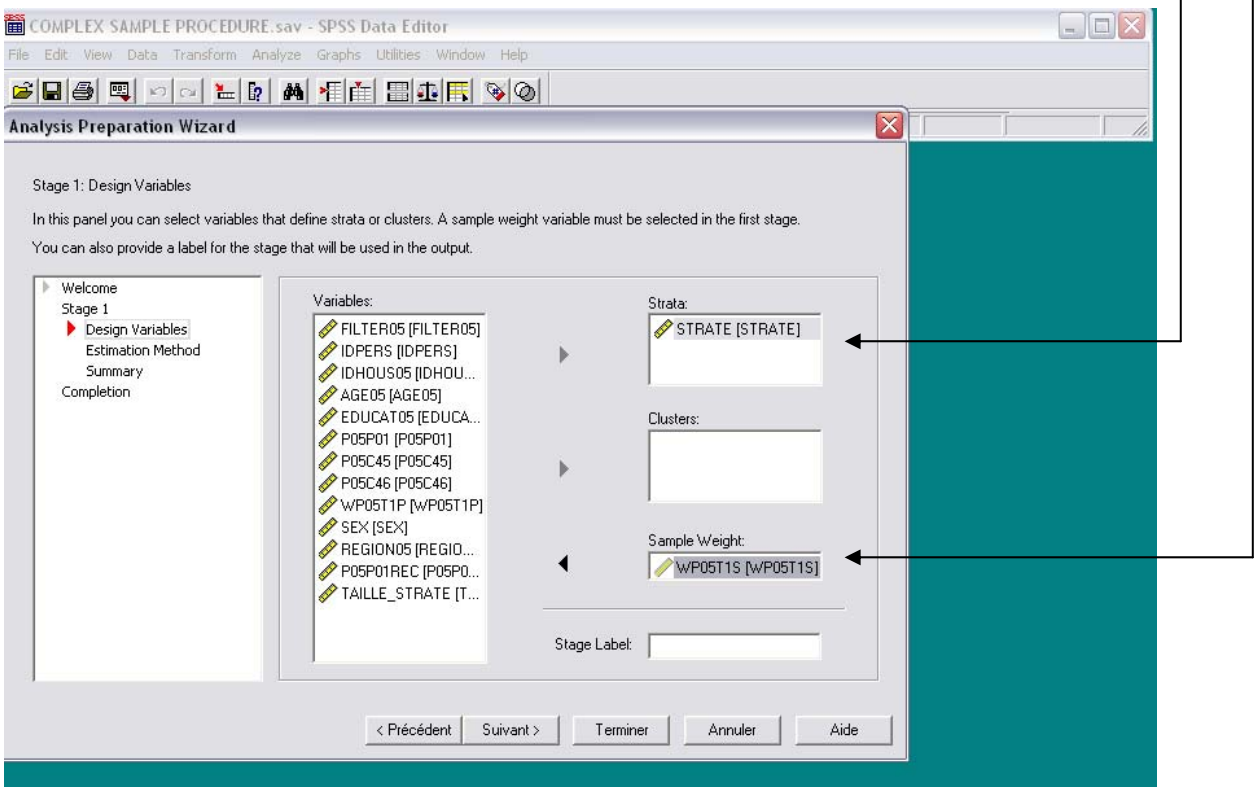


Ici, on a deux options : soit on crée un plan, soit on modifie un plan existant.
Si on crée un plan, il faut déterminer l'emplacement où le fichier sera enregistré.

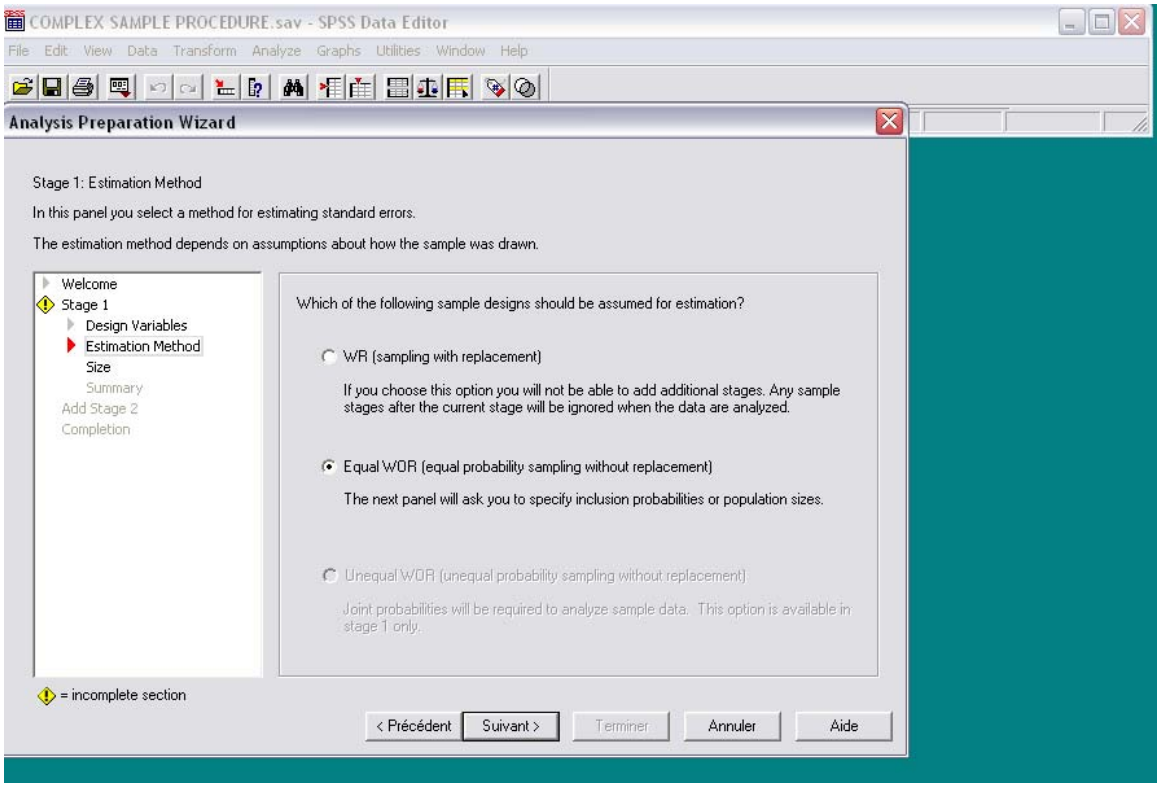


L'étape suivante consiste à :

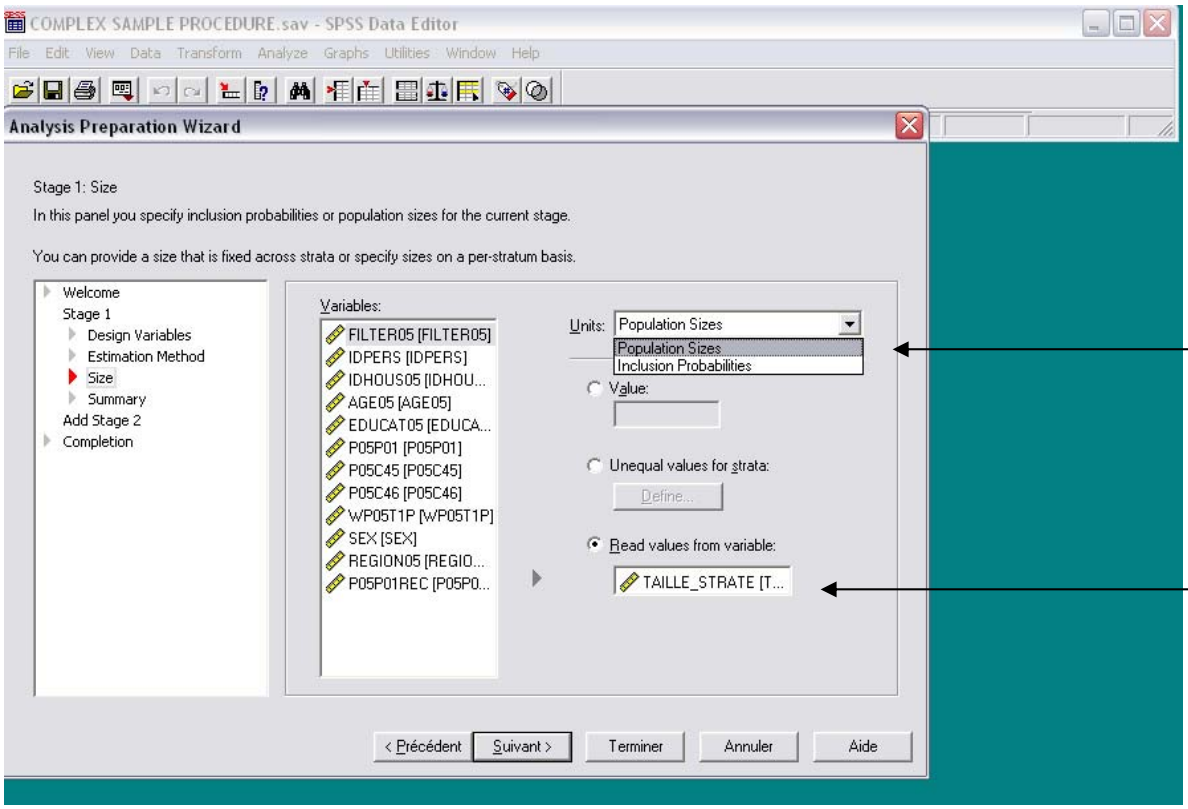
- déterminer la variable qui défini les strates (ou les grappes)
- déterminer la variable qui sera utilisée pour les pondérations (dans les analyses)



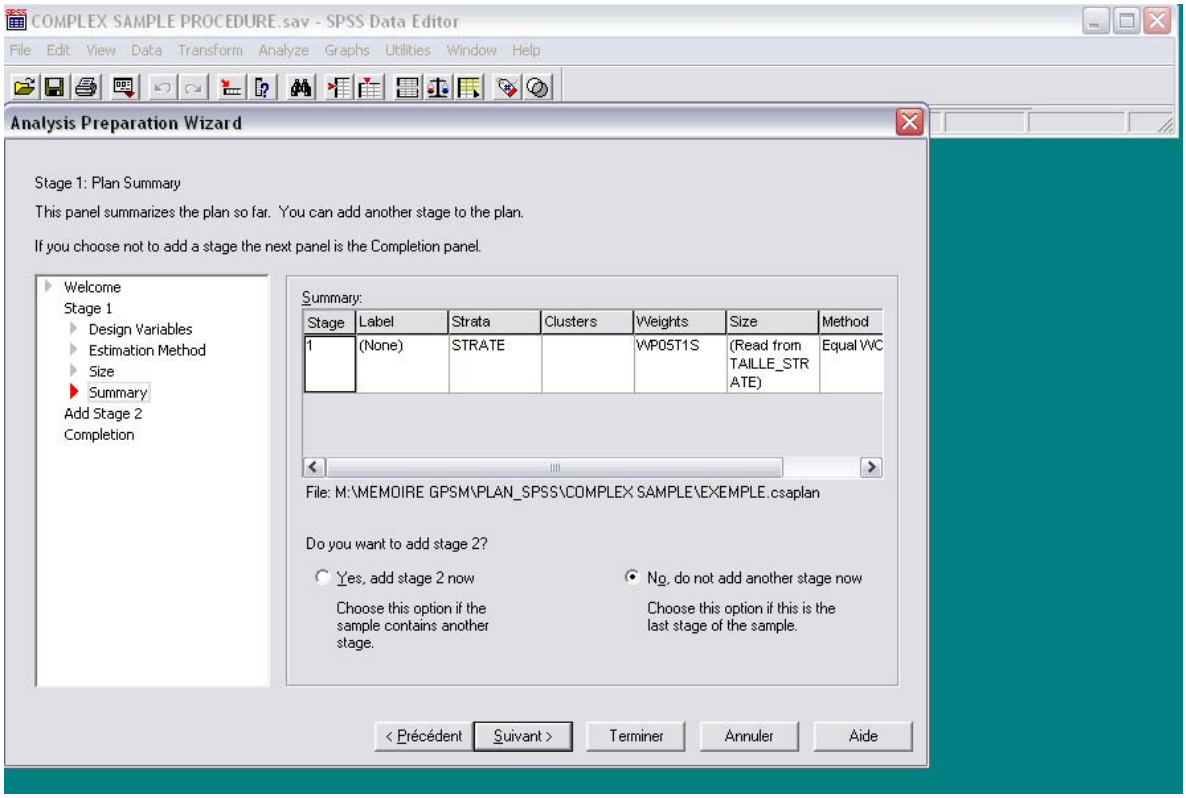
Ensuite, il faut déterminer si le tirage de l'échantillon est avec ou sans remise



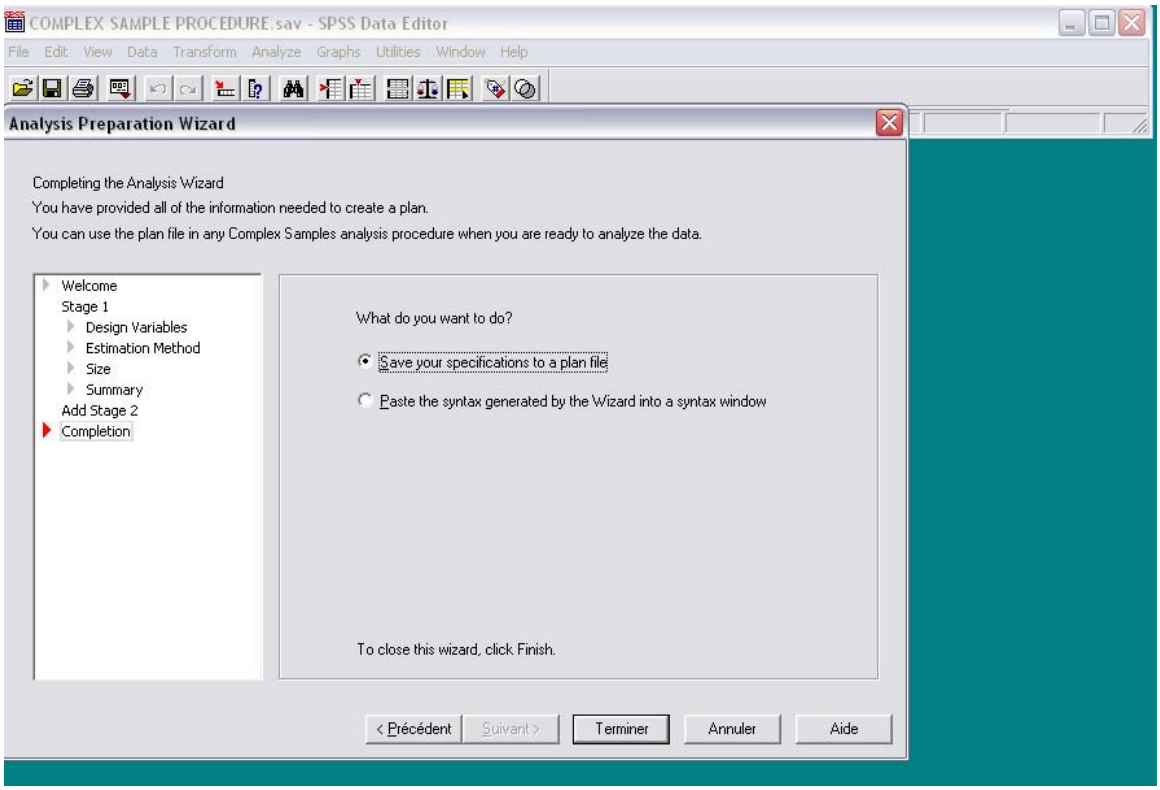
Dans le cas où il s'agit d'un plan sans remise, il faut déterminer les probabilités d'inclusion ou les tailles de la population.
Si on choisit les tailles de la population trois possibilités s'offrent. Dans tous nos exemples nous avons déterminé une variable qui contient les tailles pour chacune des strates.



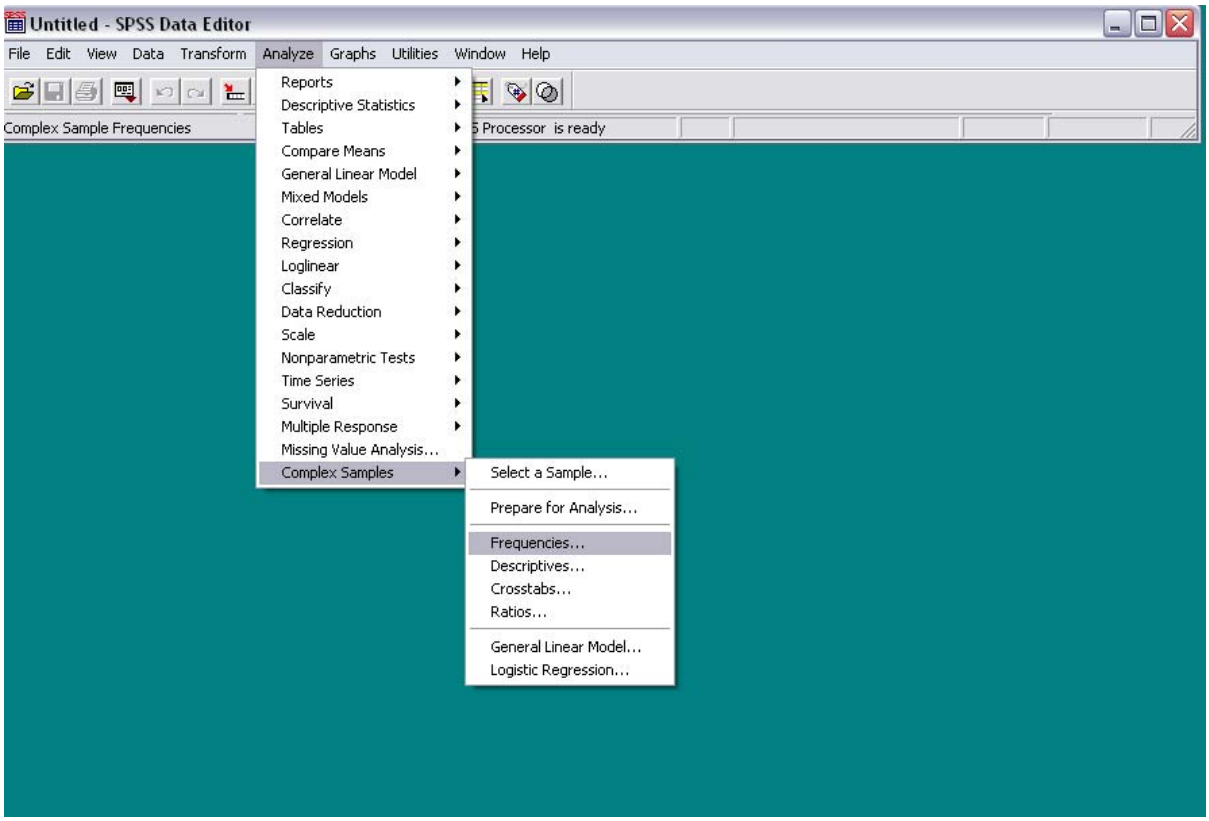
Avant de finir la création du fichier PLAN, nous avons un résumé des différents éléments choisis.



La dernière étape nous permet de choisir entre enregistrer les spécifications dans le fichier déterminé au début ou les coller dans une syntaxe.



Annexe III: Marche à suivre pour accéder aux différentes procédures du module COMPLEX SAMPLE



Annexe VI: syntaxe pour la création de la base de données contenant la variable avec les tailles des strates dans SAS, exemple où on travaille sur les PSM_I et le PSM_II

```
/* création de la base de données contenant la variable avec la taille des strates
              1 à 7: tailles SRH pour PSM_II, 2004
              11 à 77: tailles SRH pour PSM_I, 1999 */
data DataStrates;
   INPUT STRATE _TOTAL_;
   DATALINES;
1 648590
2 784266
3 455833
4 587850
5 493606
6 306605
7 160123
11 714725
22 837452
33 484667
44 646469
55 531731
66 313548
77 180623
;

/* ajustement des formats */
data DataStrates;
      set DataStrates;
      format STRATE 2.;
run;

/* création d'une base de données où au lieu d'avoir la variable Region05 comme variable de
stratification on a la variable créée dans la base de données DataStrates */.
data surveyfreq;
      set indmenmast_freq;
      if filter05 = 1 then do; /* donc PSM_I */
            if Region05=1 then STRATE=11;
            if Region05=2 then STRATE=22;
            if Region05=3 then STRATE=33;
            if Region05=4 then STRATE=44;
            if Region05=5 then STRATE=55;
            if Region05=6 then STRATE=66;
            if Region05=7 then STRATE=77;
      end;
      if filter05 = 0 then STRATE=Region05; /* donc PSM_II */
      format STRATE 2.;
RUN;

/* analyse: réalisation d'un tableau de fréquences*/
proc surveyfreq data=nom_base_de_donnees N=DataStrates;
      tables variable;
      strata STRATE;
      weight poids;
run;
```

Nom de la variable contenant la taille des strates

Nom base de données contenant la taille des strates

Nouvelle variable de stratification

Annexe V: Intervalles de confiance des pourcentages des modalités de réponse de la variable educat05 pour les procédures PROC FREQ sans les poids et PROC SURVEYFREQ

Réponses	Intervalles de confiance pour la procédure PROC FREQ		Intervalles de confiance pour la procédure PROC SURVEYFREQ	
0	0.456	1.040	0.734	1.709
1	8.116	10.066	13.017	16.600
2	3.784	5.188	4.317	6.126
3	0.627	1.287	0.589	1.331
4	32.484	35.698	33.378	37.312
5	5.738	7.420	5.454	7.312
6	10.947	13.155	9.535	11.843
7	6.726	8.526	5.073	6.638
8	3.013	4.284	2.428	3.567
9	5.317	6.944	4.183	5.633
10	13.396	15.791	10.415	12.807

Annexe VI: valeurs théoriques du test du Chi-carré

Degrés de liberté	Valeur du $\chi^2_{(\alpha,dl)}$ pour α=0.05
1	3,84
2	5.99

Annexe VII Fréquences et pourcentages des valeurs manquantes (négatives) de la variable i05htyn, par rapport à toutes les observations (n = 4256).

Réponses	Fréquences	Pourcentages (%)
-8 : autre erreur	166	3.90
-3 : inapplicable	24	0.56
-2 : pas de réponse	186	4.37
-1 : ne sait pas	183	4.30

Nous avons au total 559 observations manquantes, ce qui représente13.13% des ménages de l'échantillon.

Annexe VIII: Syntaxe utilisée pour la création de la variable ménage suisses versus ménages étrangers (men_ch)

Pour construire la nouvelle variable nous avons besoin des variables :

- NAT_1_05
- NAT_2_05
- NAT_3_05
- IDHOUS05

Elles se trouvent dans les données individus.

```
data indiv (keep= idhous05 NAT_1_05 NAT_2_05 NAT_3_05);
      set individu;
      where filter05 in (0 1);
run;
```

```
data indiv2;
      set indiv;
CH1=.;
CH2=.;
CH3=.;
SUISSE=.;
      if NAT_1_05=8100 then CH1=1; else CH1=0;
      if NAT_2_05=8100 then CH2=1; else CH2=0;
      if NAT_3_05=8100 then CH3=1; else CH3=0;
      if CH1=1 or CH2=1 or CH3=1 then SUISSE=1; else SUISSE=0;
run;
```

```
proc sort data=indiv2; by IDHOUS05; run;
```

```
proc univariate data=indiv2 noprint;
      var SUISSE;
      class IDHOUS05 ;
      output out=Somme_par_MEN sum=Nbr_de_Suisses;
run;
```

Cette procédure crée un nouveau fichier qui s'appelle Somme_par_MEN, il contient la variable Nbr_de_Suisses qui donne le nombre de personnes du ménage ayant une nationalité suisse

```
data MEN_CH (keep= idhous05 MEN_CH);
      set Somme_par_MEN;
MEN_CH=.;
      if Nbr_de_Suisses>=1 then MEN_CH=1; else MEN_CH=0;
run;
```

```
proc sort data=MEN_CH; BY IDHOUS05; RUN;
proc sort data=means; by idhous05; run;
```

```
data means_cat;
    merge means MEN_CH;
    by IDHOUS05;
run;
```

Annexe IV: valeurs manquantes de la variable i05ptotn

Réponses	**Fréquences**
-8 : autre erreur	8
-4 : sans revenu personnel	3
-2 : pas de réponse	92
-1 : ne sait pas	16
TOTAL	119
N	1716

Annexe X: Fréquences et pourcentages des valeurs manquantes (négatives) de la variable p05c46 (le poids en kg), par rapport à toutes les observations (n = 6532)

Réponses	**Fréquences**	**Pourcentage (%)**
-2 : pas de réponse	20	0.31
-1 : ne sait pas	31	0.47

Fréquences et pourcentages des valeurs manquantes (négatives) de la variable p05c45 (la taille en cm), par rapport à toutes les observations (n = 6532)

Réponses	**Fréquences**	**Pourcentage (%)**
-2 : pas de réponse	4	0.06
-1 : ne sait pas	14	0.21

Annexe XI: Description des variables après recodage (les valeurs manquantes enlevées, recodage des petites tailles et que les personnes de 14 ans et plus)

Variables	N	Min	Max
Poids en kg	6464	36	172
Taille en cm	6464	118	205
Age	6464	14	94

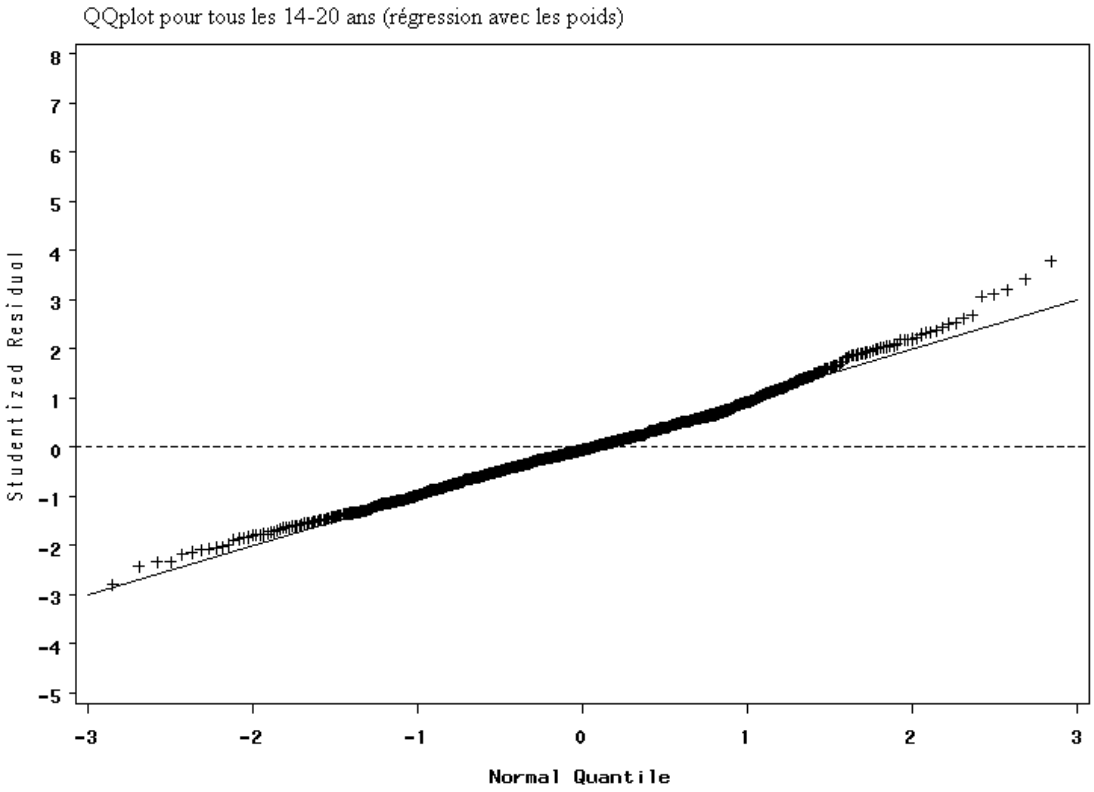
Variables	N	Catégories	n	Pourcentage (%)
Sexe	6464	Homme	2896	44.80
		Femme	3568	55.20
Maigreur	6464	Imc>=18.5	6127	94.79
		Imc<18.5	337	5.21
Obésité	6464	Imc<=25	4311	66.69
		Imc>25	2153	33.31

Annexe XI bis: Description des variables pour la sous-population des 14-20 ans

Variables	N	Min	Max
Poids en kg	730	36	110
Taille en cm	730	150	200
Age	730	14	20

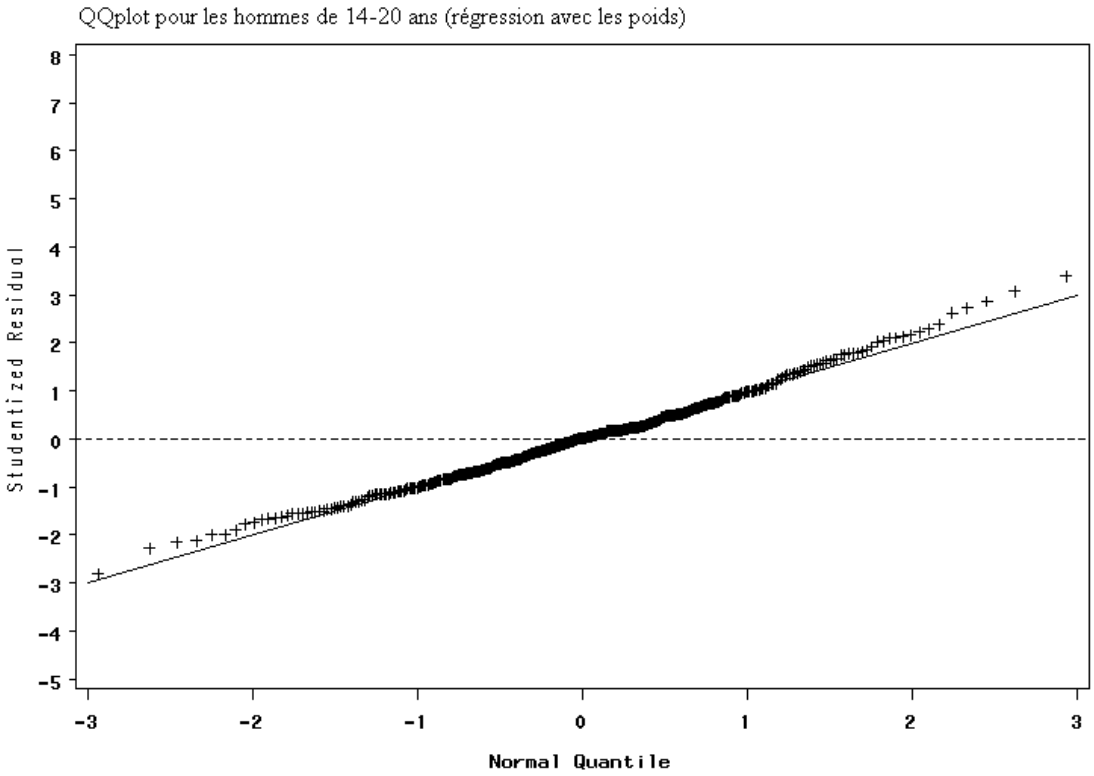
Variables	N	Catégories	n	Pourcentage (%)
Sexe	730	Homme	370	50.68
		Femme	360	49.32
Maigreur	730	Imc>=18.5	590	80.82
		Imc<18.5	140	19.18
Obésité	730	Imc<=25	686	93.97
		Imc>25	44	6.03

Annexe XII: QQ-plot pour tous les 14-20 ans (régression réalisée en tenant compte des poids)



Nous constatons que les points s'ajustent bien à la droite d'équation $e_i = q_i$.

Annexe XIII*:* QQ-plot pour les hommes de14-20 ans (régression réalisée en tenant compte des poids)



Ici aussi les points s'ajustent bien à la droite d'équation $e_i = q_i$.

Annexe XIV: Fréquences des valeurs manquantes (négatives) de la variable i05ptotn pour les personnes de 21 ans et plus travaillant à plein temps et n'ayant eu aucune modification d'activité au cours de l'année de l'interview.

Réponses	**Fréquences**
-8 : autre erreur	43
-4 : sans revenu personnel	20
-2 : pas de réponse	197
-1 : ne sait pas	60
TOTAL	320
N	3236

Annexe XV: Modalités de réponses pour la variable educat05, le niveau de formation le plus élevé, pour les personnes de 21 ans et plus travaillant à plein temps et n'ayant eu aucune modification d'activité au cours de l'année de l'interview.

Réponses	**Explications**	**Fréquences**
0	école obligatoire inachevée	9
1	école obligatoire, formation prof. Élémentaire	189
2	stage ménager, 1 année d'école commerciale courte	39
3	école de formation générale	26
4	apprentissage	1259
5	école prof. à plein temps	118
6	maturité	218
7	formation prof. supérieure	369
8	école technique ou professionnelle	143
9	école prof. Supérieure	325
10	université, haute école	541
	N	3236

L'échantillon étudié comporte 74 % d'hommes et 26% de femme âgés d'au moins 21 ans et travaillant à plein temps.

Annexe XVI: Recodage des différentes modalités des variables appartenant au modèle.

- Niveau de formation :

Au lieu de 11 modalités différentes nous en avons 3.

MODALITES	FREQUENCES	POURCENTAGE
Nivform_bas (contient les modalités 0-3)	228	7.82%
Nivform_moyen (contient les modalités 4-7)	1773	60.80%
Nivform_haut (contient les modalités 8-10)	**915**	**31.38%**
TOTAL	2916	100.00%

- Age :

Nous avons fait 3 catégories de personnes.

MODALITES	FREQUENCES	POURCENTAGE
Age21_30ans	452	15.50%
Age31_40ans	796	27.30%
Age41_50ans	904	31.00%
Age51_81ans	**764**	**26.20%**
TOTAL	2916	100.00%

- Nombre d'enfant de 0 à 17 ans dans le ménage :

Nous avons formé 4 catégories (0, 1, 2, 3etplus enfants).

MODALITES	FREQUENCES	POURCENTAGE
Enfant_0	1836	62.96%
Enfant_1	405	13.89%
Enfant_2	463	15.88%
Enfant_3etplus	**212**	**7.27%**
TOTAL	2916	100.00%

- Etat civil :

Pour cette variable chacune des modalités devient une nouvelle variable.

MODALITES	FREQUENCES	POURCENTAGE
Celibataire	849	29.12%
Veuf	35	1.20%
Separe	57	1.95%
Divorce	270	9.26%
Marie	**1705**	**58.47%**
TOTAL	2916	100.00%

- Sexe :

Chacune des modalités devient une nouvelle variable.

MODALITES	FREQUENCES	POURCENTAGE
Femme	744	25.51%
Homme	**2172**	**74.49%**
TOTAL	2916	100.00%

- Nombre de personnes dans le ménage :

Pour cette variable, nous avons créé 4 catégories différentes (1, 2, 3, 4 et plus personnes dans le ménage).

MODALITES	FREQUENCES	POURCENTAGE
Persmenage_1	563	19.31%
Persmenage_2	924	31.69%
Persmenage_3	445	15.26%
Persmenage_4etplus4	**984**	**33.74%**
TOTAL	2916	100.00%

Chacune de ces nouvelles modalités est une variable dichotomique où 1 signifie que le critère est rempli et où on a 0 sinon.