Calcul de variance des enquêtes Elfe : mode d'emploi

Le code permettant de générer les variances suivant le plan de sondage complexe de l'enquête Elfe (variances suivant plan de sondage complet et suivant le plan simplifié) est proposé via le programme « *macro variance.sas* » utilisant le logiciel SAS 9.4 (SAS Institute Inc, 2013). Il s'agit essentiellement de 2 macros :

- macro **%TOTAL** permettant de générer les variances lorsqu'on calcule un total ;
- macro % **RATIO**, permettant de générer les variances lorsqu'on calcule un ratio (ou une prévalence par exemple).

Après avoir exécuté le programme « *macro variance.sas* », l'utilisateur doit générer dans son programme SAS principal, par une étape DATA classique, une table qui doit comprendre a minima les champs suivants (d'autres champs peuvent être présents ici, dont bien évidement les variables sur lesquelles on veut calculer les variances, sans aucune influence sur la suite):

- un champs identifiant les enregistrements (par défaut *ID*...) ;
- un champs identifiant la strate de la maternité (par défaut *M00M1_MATSTRATEC1*) ;
- un champs identifiant la vague de l'enquête (par défaut *M00M1_VAGUE*) ;
- un champs identifiant la maternité (par défaut *M00M1_IDGROUPNAMEALEAC1)* ;
- un champs identifiant le jour de naissance (par défaut *M00M2_JNAISSEALEA*) ;
- un champs identifiant le poids à utiliser (si on veut calculer des variances non pondérées, il suffit de générer une variable valant 1 pour chaque individu et la déclarer ici) ;
- les variables de calage *CS_1* à *CS_13.*

1) Pour estimer un TOTAL

L'utilisateur doit appeler dans son programme la macro **%TOTAL**, par la commande suivante :

```
%TOTAL (table_selection, identifiant, strate, vague, mater, jour, poids, variable, methode) ;
```

Elle nécessite donc de nommer 9 paramètres comme tels :

table_selection=nom de la table où sont les données à traiter
identifiant = champs identifiant les enregistrements (par défaut ID...) ;
strate = champs qui identifie la strate de la maternité (par défaut M00M1_MATSTRATEC1) ;
vague = champs qui identifie la vague de l'enquête (par défaut M00M1_VAGUE) ;
mater = champs identifiant la maternité (par défaut *M00M1_IDGROUPNAMEALEAC1)* ;
jour = champs identifiant le jour de naissance (par défaut *M00M2_JNAISSEALEA*) ;
variable = champs pour lequel on veut calculer le total
methode = pour savoir quelles variables utiliser pour la régression, selon qu'on a effectué la pondération avec les variables de calage CS1 à CS6 ou sur les 13 variables CS1 à CS13.
A priori, si on utilise les pondérations avant les 3 ans ½ de l'enfant, on pose méthode=1. Si on utilise un poids aux 3 ans ½ ou après ou si on utilise un poids généré par la macro *ponderation.sas*, on pose méthode=2.

A titre d'exemple, supposons qu'on veuille estimer le nombre total d'enfants dont la mère souffre de diabète gestationnel (M00X_DIABGEST=1). On va générer la variable DIABETE :

```
data enfant;
set enfant;
if M00X_DIABGEST=1 then DIABETE=1; else DIABETE=0;
run;
```

On lancera alors la commande :

```
%TOTAL (table_selection=enfant, identifiant=IDXX_XX, strate=M00M1_MATSTRATEC1,
vague=M00M1_VAGUE, mater=M00M1_IDGROUPNAMEALEAC1, jour=M00M2_JNAISSEALEA,
poids=M00E_PONDVALC2, variable=DIABETE, methode=1);
```

Cette macro génère une table nommée recapTOTAL_DIABETE (la table est nommée par défaut recapTOTAL suivi du nom de la variable). Cette table comprend les champs suivants :

total	var_calage_effetMAT	var_calage_effetJOUR	variance_NR	variance_TOTALE	variance_PLAN_SIMPLIFIE
55381.874207	1175055.9937	5726768.0034	2416258.0901	9318082.0872	3591314.0838

- Total = estimation du total attendu : ici 55 382 mères souffrent de diabète gestationnel.
- var_calage_effetMAT, var_calage_effetJOUR et variance_NR sont les 3 effets à prendre en compte dans l'estimation de la variance du total :
    - Lorsqu'on prend en compte le plan complet utilisé dans Elfe, on doit sommer ces 3 éléments. C'est ce qui est fait dans « variance_TOTALE ». Ici 9 318 082, soit un écart type estimé de 3052
    - Lorsqu'on prend en compte le plan simplifié préconisé dans Elfe, on doit sommer uniquement les 2 derniers éléments. C'est ce qui est fait dans « variance_PLAN_SIMPLIFIE ». Ici 3 591 314, soit un écart type estimé de 1895

A 95%, en tenant compte de la préconisation faite aux utilisateurs de l'enquête Elfe, on peut donc estimer le nombre total de mères souffrant de diabète gestationnel à [51 667 ;59 096].

Signalons que la macro produit ces mêmes données dans la fenêtre résultats :

resultat estimation TOTAL: données stockées dans la table recapTOTAL_DIABETE								
total	var_calage_effetMAT	var_calage_effetJOUR	variance_NR	variance_TOTALE	variance_PLAN_SIMPLIFIE	etype_PLAN_SIMPLIFIE	borneinf95_PLAN_SIMPLIFIE	bornesup95_PLAN_SIMPLIFIE
55381.87	1175056	5726768	2416258	9318082	3591314	1895.076	51667.52	59096.22

Remarque : on préconise également, dans le document de travail INED associé d'utiliser directement la commande surveyfreq de SAS.

Dans notre exemple,

```
proc surveyfreq data=enfant  ;
table M00X_DIABGEST / nopercent ;
weight M00E_PONDVALC2;run;
```

donne le résultat suivant :

Diabète gestationnel			
M00X_DIABGEST	Fréquence	Fréquence pondérée	Err type de Fréq pond
0-Non	15918	668522	4337
1-Oui	1301	55382	1949
Total	17219	723904	4279
Frequency Missing = 923			

Soit un IC à 95% estimé à [51 562 ;59 202].

Estimons maintenant le total d'enfants pratiquant régulièrement une activité de loisirs à 3 ans ½ (A03R_acextrasc=1).

```
data enfant3A;
set enfant; if A03R_acextrasc=1 then activite=1; else activite=0;
where a03e_pondref >0;
run;
```

```
%TOTAL (table_selection=enfant3A, identifiant=ID25_TS, strate=M00M1_MATSTRATEC1,
vague=M00M1_VAGUE, mater=M00M1_IDGROUPNAMEALEAC1, jour=M00M2_JNAISSEALEA, poids=a03e_pondref,
variable=activite, methode=2);
```

Notons ici méthode=2 et poids=pondération à 3 ans.

Les résultats sont produits dans la table recapTOTAL_ACTIVITE, et sont :

total	var_calage_effetMAT	var_calage_effetJOUR	variance_NR	variance_TOTALE	variance_PLAN_SIMPLIFIE
124453.50142	3292724.9896	8305502.5816	6347557.6093	17945785.181	9640282.5989

On estime donc à 124 454 enfants ceux qui pratiquent régulièrement une activité de loisirs, avec une variance exacte estimée à 17 945 785, et une variance simplifiée estimée à 9 640 282 (soit un écart type plan simplifié = 3104), et donc un intervalle de confiance à 95% pour le plan simplifié de :

borneinf95_PLAN_SIMPLIFIE	bornesup95_PLAN_SIMPLIFIE
118367.93574	130539.0671

Là encore on peut vérifier l'approximation donnée par la procédure surveyfreq avec un écart type estimé à 3076.

```
proc surveyfreq data=enfant3A  ;
table A03R_acextrasc / nopercent ;
weight a03e_pondref;run;
```

Activité régulière de loisir enfant			
A03R_ACEXTRASC	Fréquence	Fréquence pondérée	Err type de Fréq pond
1-OUI	2167	124454	3076
2-NON	9534	639083	5923
Total	11701	763536	5563

2) Pour estimer un RATIO

Les commandes sont identiques aux précédentes, mais on doit donner 2 variables (numérateur et dénominateur). L'utilisateur doit donc appeler dans son programme la macro %**RATIO**, par la commande suivante :

%**RATIO** (table_selection, identifiant, strate, vague, mater, jour, poids, variable1, variable2, methode) ;

Elle nécessite donc de nommer 10 paramètres comme tels :

table_selection=nom de la table où sont les données à traiter
identifiant = champs identifiant les enregistrements (par défaut ID...) ;
strate = champs qui identifie la strate de la maternité (par défaut M00M1_MATSTRATEC1) ;
vague = champs qui identifie la vague de l'enquête (par défaut M00M1_VAGUE) ;
mater = champs identifiant la maternité (par défaut *M00M1_IDGROUPNAMEALEAC1)* ;
jour = champs identifiant le jour de naissance (par défaut *M00M2_JNAISSEALEA*) ;
variable1 = champs pour lequel on veut calculer le numérateur
variable2 = champs pour lequel on veut calculer le dénominateur
methode = pour savoir quelles variables utiliser pour la régression, selon qu'on a effectué la pondération avec les variables de calage CS1 à CS6 ou sur les 13 variables CS1 à CS13.
A priori, si on utilise les pondérations avant les 3 ans ½ de l'enfant, on pose méthode=1. Si on utilise un poids aux 3 ans ½ ou après ou si on utilise un poids généré par la macro ponderation.sas, on pose méthode=2.

Par exemple, si on veut estimer la part des enfants pratiquant une activité de loisirs aux 3 ans ½, on lancera :

```
data enfant3A;
set enfant;
if A03R_acextrasc=1 then activite=1; else activite=0;TOT=1;
where a03e_pondref >0;
run;
```

Puis :

```
%RATIO (table_selection=enfant3A, identifiant=ID25_TS, strate=M00M1_MATSTRATEC1,
vague=M00M1 VAGUE, mater=M00M1 IDGROUPNAMEALEAC1, jour=M00M2_JNAISSEALEA, poids=a03e_pondref,
variable1=activite, variable2=TOT, methode=2);
```

Cette macro génère une table nommée recapRATIO_ACTIVITE (la table est nommée par défaut recapRATIO suivi du nom de la variable au numérateur). Cette table comprend les champs suivants :

numerateur	denominateur	ratio	var_calage_effetMAT	var_calage_effetJOUR	variance_NR	variance_TOTALE	variance_PLAN_SIMPLIFIE
124453.50142	763536.37994	0.1629961646	5.6008327E-6	0.0000146626	0.0000109876	0.000031251	0.0000165884

- Estimation numérateur, dénominateur et du ratio : ici 16.29%
- var_calage_effetMAT, var_calage_effetJOUR et variance_NR sont les 3 effets à prendre en compte dans l'estimation de la variance du ratio :
    - Lorsqu'on prend en compte le plan complet utilisé dans Elfe, on doit sommer ces 3 éléments. C'est ce qui est fait dans « variance_TOTALE ». Ici 0.000031251, soit un écart type estimé de 0.559%
    - Lorsqu'on prend en compte le plan simplifié préconisé dans Elfe, on doit sommer ces 2 derniers éléments. C'est ce qui est fait dans « variance_PLAN_SIMPLIFIE ». Ici 0.0000165884, soit un écart type estimé de 0.407%

Soit un IC à 95% estimé à [15.50% ;17.01%].

Ces calculs sont également fournis dans l'onglet résultats.

RESULTATS estimation ratio: données stockées table recapRATIO_activite

numerateur	denominateur	ratio	var_calage_effetMAT	var_calage_effetJOUR	variance_NR	variance_TOTALE	variance_PLAN_SIMPLIFIE	etype_PLAN_SIMPLIFIE	borneinf95_PLAN_SIMPLIFIE	bornesup95_PLAN_SIMPLIFIE
124453.5	763536.4	0.162996	5.601E-6	0.000015	0.000011	0.000031	0.000017	0.004073	0.155013	0.170979

Remarque : on préconise là aussi dans le document de travail INED associé d'utiliser directement la commande surveyfreq de SAS.

Dans notre exemple,

```
proc surveyfreq data=enfant3A  ;
table  activite / nofreq cl ;
weight a03e_pondref;run;
```

donne le résultat suivant :

Table de activite

activite	Fréquence pondérée	Err type de Fréq pond	Pourcentage	Err type du pourcentage	Intervalle de conf. à95% pour le pourcentage	
0	639083	5923	83.7004	0.4018	82.9127	84.4881
1	124454	3076	16.2996	0.4018	15.5119	17.0873
Total	763536	5563	100.0000			

Soit un IC à 95% estimé à [15.51% ;17.01%].