Calcul des intervalles de confiance pour les EPCV 1996-2004

Le tirage des échantillons est effectué dans l'échantillon-maître issu du recensement de la population et dans la base de sondage des logements neufs achevés entre deux recensements, alimentée par le système de suivi des permis de construire du Ministère de l'Equipement.

Les formules présentées ici ont été établies en supposant que les logements ont été sélectionnés par un plan de sondage aléatoire simple sans remise (SAS) où le taux de sondage est négligeable. Or, dans la réalité ce n'est pas exactement le cas. En effet, les logements sont sélectionnés selon un plan de sondage stratifié (par catégories de communes) à plusieurs degrés, le nombre de degrés étant variable selon les strates. Cependant, en première approximation, pour la grande majorité des indicateurs, on peut considérer que ce plan de sondage complexe est proche d'un sondage aléatoire simple, c'est-à-dire considérer comme négligeable l'"effet de grappe" dû au tirage des logements dans les zones géographiques fixes, dites "unités primaires". On suppose de plus que les non-répondants se comportent comme les répondants.

1. Cas d'un pourcentage ou d'une évolution en point dans la population totale des ménages

a. Calcul pour un pourcentage sur une année

Soient n le nombre de ménages répondants à l'enquête et $\bar{p}$ l'estimateur pondéré de la proportion p des ménages possédant la caractéristique dans la population.

L'intervalle de confiance à 95% s'obtient par la formule :

$$\bar{p} \pm 2\sqrt{\frac{S^2}{n}}$$

où S^2 est la variance statistique modifiée au sein de l'échantillon.

Puisque nous faisons l'hypothèse d'un SAS et nous nous intéressons à l'estimation d'une proportion, nous obtenons :

$$\bar{p} \pm 2\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

b. Calcul pour l'évolution d'une proportion entre deux années consécutives

Soient k le taux de renouvellement de l'échantillon entre les deux années consécutives, ρ le coefficient de corrélation entre les deux estimations sur la partie commune de l'échantillon, S^2 la variance statistique « calculée » sur la proportion et n la taille de l'échantillon sur lequel cette variance est calculée. On suppose S^2 et n constants entre t1 et t2 (sachant que l'on se place sur deux années consécutives, cette hypothèse est réaliste).

Pour un phénomène estimé par une moyenne empirique $\overline{p}_t$, l'estimateur naturel de l'évolution étant noté $\overline{p}_2 - \overline{p}_1$, on démontre que la variance de cette évolution est [1] :

$$V(\overline{p}_2 - \overline{p}_1) = 2(1-\rho(1-k))\frac{S^2}{n}$$

Si l'échantillon est renouvelé par moitié, comme c'est le cas en général dans les EPCV de 1996 à 2004, k=1/2 ; la taille de l'échantillon est supposée constante, et, du fait de l'hypothèse d'un SAS, l'intervalle de confiance à 95% autour de $\overline{p}_2 - \overline{p}_1$ devient :

$$(\overline{p}_2 - \overline{p}_1) \pm 2\sqrt{(2-\rho)\frac{[\overline{p}_2(1-\overline{p}_2)]}{n}}$$

Il arrive qu'il y ait des extensions d'échantillon certaines années (par exemple en octobre 2003). Dans ce cas k est différent de 1/2. De plus n ne peut plus alors être considéré comme constant. Si n_c est l'échantillon commun aux deux années, on a alors la formule suivante :

$$\overline{p}_2 - \overline{p}_1 \pm 2\sqrt{(\frac{1}{n_1}+\frac{1}{n_2}-\frac{\rho}{n_c})*[\overline{p}_2(1-\overline{p}_2)]}$$

c. Calcul pour une évolution d'une proportion entre deux années non consécutives

Dans ce cas, on peut supposer que les deux échantillons sont indépendants. On obtient donc :

$$V(\overline{p}_2 - \overline{p}_1) = \frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}$$

En supposant de plus que le plan de sondage est un plan de sondage aléatoire simple, on obtient comme intervalle de confiance à 95% autour de $\overline{p}_2 - \overline{p}_1$:

$$\overline{p}_2 - \overline{p}_1 \pm 2\sqrt{\frac{[\overline{p}_1(1-\overline{p}_1)]}{n_1} + \frac{[\overline{p}_2(1-\overline{p}_2)]}{n_2}}$$

2. Cas d'un pourcentage ou d'une évolution dans une sous population dans les ménages

Dans ce cas, il faut remplacer n par n_D dans toutes les formules précédentes, n_D étant l'effectif dans l'échantillon de la sous population considérée.

3. Cas d'un pourcentage ou d'une évolution en point parmi les individus

Les individus Kish qui répondent à l'enquête sont tirés aléatoirement au sein des ménages une fois leur composition connue et non dans la population totale (cela correspond à un degré de tirage supplémentaire). Indépendamment du mode de tirage des ménages, il ne s'agit donc pas d'un tirage d'individus par un plan de sondage aléatoire simple au sein d'une base de sondage individu. De plus, il peut y avoir un effet « grappe » sur certaines variables si on sélectionne plusieurs individus par ménage (les individus d'un même ménage pouvant avoir un comportement similaire).

[1] *voir "Estimation dans les enquêtes répétées", Nathalie Caron et Philippe Ravalet, Document de Travail 0005, Insee - Unité Méthodologie Statistique, 2000*

Si on néglige ce possible « effet de grappe », on peut supposer que l'on a réalisé un sondage aléatoire simple d'individus. Dans ce cas, il suffit d'utiliser au niveau « individu » les formules présentées ci-dessus en les adaptant.

En revanche, si l'on cherche à prendre en compte le mode de tirage dans les calculs d'intervalle de confiance, il est alors nécessaire de remonter les variables « individuelles » au niveau de chaque ménage, puis de calculer une variable « synthétique » pour chaque ménage $\hat{u}^k$ et de calculer la variance d'échantillonnage du total de cette variable qui sera une approximation de la variance de la proportion estimée.

De façon plus précise, soit $\overline{p} = \frac{\hat{X}}{\hat{N}_D}$ une proportion estimée dans l'échantillon sur la sous population D à partir des effectifs estimés $\hat{X}$ et $\hat{N}_D$ (par exemple $\hat{N}_D$ est la population estimée de 20-29 ans, $\hat{X}$ le nombre estimé de 20-29 ans ayant subi une agression et $\overline{p}$ le pourcentage de 20-29 ans agressés).

On calcule au niveau de chaque ménage la variable $u^k = \frac{1}{\hat{N}_D}(x_k - \overline{p}\ y_k)$ où x_k est la valeur de la variable X pour le ménage k et y_k est le nombre d'individus dans le ménage appartenant à la sous population. Si N est le nombre de ménages total (pondéré) et n le nombre de ménages enquêtés, nous avons $V(\overline{p}) = V(\hat{U}) = \frac{N^2}{n} S_U^2$. Il suffit donc de calculer la variance S_U^2.

a. Calcul pour un pourcentage sur une année

Nous obtenons l'intervalle de confiance :

$$\overline{p}\ \pm 2\sqrt{\frac{N^2}{n} S_U^2}$$

où N est le nombre de ménages total (pondérés) et n le nombre de ménages enquêtés.

b. Calcul sur deux années consécutives

Si l'échantillon est renouvelé par moitié, k=1/2, $n_1 = n_2 = n$ d'où :

$$\overline{p}_2 - \overline{p}_1 \pm 2\sqrt{(2-\rho)\frac{N^2}{n} S_U^2}$$

En cas d'extension ponctuelle, on a :

$$\overline{p}_2 - \overline{p}_1 \pm 2\sqrt{N^2(\frac{1}{n_1} + \frac{1}{n_2} - \frac{\rho}{n_c})S_U^2}$$

c. Calcul sur deux années <u>non</u> consécutives

Dans ce cas on a :

$$\overline{p}_2 - \overline{p}_1 \pm 2\sqrt{N^2(\frac{S_U^2}{n_1} + \frac{S_U^2}{n_2})}$$

4. Exemples de calculs

Exemples tirés du fichier historique 1996-2003, indicateurs sociaux d'octobre « Participation et contacts sociaux »

% d'individus qui sont allés au cinéma au cours des douze derniers mois pour les années 2002 et 2003

	2002			2003			2002-2003	
Zone d'étude et d'aménagement du territoire (ZEAT)	%	Intervalle de confiance à 95% (SAS)	Effectif brut ayant répondu à la question	%	Intervalle de confiance à 95% (SAS)	Effectif brut ayant répondu à la question	Variation en points	Intervalle de confiance à 95% (SAS)
Sud-Ouest	48,0 %	{ 45,8 ; 50,1 }	652	46,9 %	{ 45,2 ; 48,5 }	1 115	-1,1	{ -3,8 ; 1,6 }
France métropolitaine	52,6%	{ 51,7 ; 53,5 }	5 823	51,6 %	{ 50,9 ; 52,2 }	10 272	-1,1	{ -2,1 ; 0 }

***Lecture** : Dans la ZEAT Sud-Ouest, 46,9% des 1 115 individus interrogés dans l'enquête ont déclaré, pour l'année 2003, être allés au cinéma au cours des douze derniers mois. Le pourcentage « vrai », dans l'ensemble de la population de la ZEAT, a 95% de chances d'être compris en 45,2 % et 48,5 % ; il est donc estimé à + ou – 1,7 points.*
Pour la France métropolitaine, où 10 272 individus ont répondu à l'enquête, le pourcentage « vrai » (estimé à 51,6%) a 95% de chances d'être compris entre 50,9% et 52,2% ; il est donc estimé à + ou – 0,6 point.

Par rapport à l'ensemble du fichier historique, l'année 2003 est une année atypique car l'échantillon sélectionné avait bénéficié d'une importante extension. En comparant avec l'échantillon 2002, nous pouvons constater que plus l'effectif enquêté est important, plus l'intervalle de confiance se resserre. Inversement, chaque utilisateur doit toujours considérer que plus l'effectif est minime, moins les conclusions tirées des résultats obtenus sont fiables.

En réalité, la longueur des intervalles, obtenus par les formules précédentes et proposés dans tous les tableaux, est un peu sous estimée du fait de l'hypothèse d'un sondage aléatoire simple (SAS). Si, à titre d'exemple et en utilisant le logiciel POULPE[2] de l'Insee, on tient compte du fait que le plan de sondage est à plusieurs degrés stratifié par catégories de communes, on obtient des résultats légèrement différents. A titre d'exemple, pour la variable cambriolage, variable des EPCV de janvier, au niveau ménage on obtient exactement le même intervalle. En revanche, pour la variable agression au niveau individu, si pour la France l'intervalle est identique, pour la ZEAT Nord le « vrai » pourcentage est estimé à + ou – 2,1 points, au lieu de + ou – 1,9 points dans le cas d'un SAS.

% en 2004 de ménages cambriolés au cours des deux dernières années

Zone d'étude et d'aménagement du territoire (ZEAT)	%	Intervalle de confiance à 95% (SAS)	Intervalle de confiance à 95% (avec POULPE)	Effectif brut ayant répondu à la question
Nord	2,5 %	{ 1,0 ; 4,0 }	{ 1,0 ; 4,0 }	400
France métropolitaine	2,5 %	{ 2,1 ; 2,9 }	{ 2,1 ; 2,9 }	6 440

***Lecture** : dans la ZEAT Nord comme en France métropolitaine, 2,5% de ménages ont déclaré en 2004 avoir été cambriolés au cours des années 2002-2003. Pour la France, où 6 440 ménages ont répondu à l'enquête, le pourcentage « vrai », dans l'ensemble de la population de la ZEAT, a 95% de chances d'être compris entre 2,1% et 2,9%, il est donc estimé à + ou – 0,4 point et ce, que ce soit en faisant l'hypothèse d'un sondage aléatoire simple ou non (avec POULPE). Dans la ZEAT Nord, où seulement 400 ménages ont répondu, ce même pourcentage « vrai » a 95% de chances d'être compris en 1% et 4%, il est donc estimé à + ou – 1,5 points.*

[2] *voir "Estimations de précision pour l'enquête PCV de janvier 2004 avec le logiciel POULPE et mode d'emploi du logiciel pour des calculs de précision complémentaires (au niveau ménage ou individu)", Sylvie Rousseau, note interne Insee n°070/F410, 26-08-2005*

% en 2004 d'individus agressés au cours des deux dernières années

Zone d'étude et d'aménagement du territoire (ZEAT)	%	Intervalle de confiance à 95% (SAS)	Intervalle de confiance à 95% (avec POULPE)	Effectif brut ayant répondu à la question
Nord	7,8 %	{ 5,9 ; 9,7 }	{ 5,7 ; 9,9 }	833
France métropolitaine	6,7 %	{ 6,2 ; 7,2 }	{ 6,2 ; 7,2 }	11 701

***Lecture** : dans la ZEAT Nord, 7,8 % des 83 individus interrogés dans l'enquête ont déclaré en 2004 avoir été agressés au cours des années 2002-2003. Le pourcentage « vrai », dans l'ensemble de la population de la ZEAT, a 95% de chances d'être compris en 5,9 % et 9,7% ; il est donc estimé à + ou – 1,9 points si l'on fait l'hypothèse d'un SAS. Avec POULPE, ce pourcentage « vrai » est estimé à + ou - 2,1 points.*
Pour la France métropolitaine où 11 701 individus ont répondu à l'enquête, le pourcentage « vrai » (estimé à 6,7%) a 95% de chances d'être compris entre 6,2 % et 7,2%, il est donc estimé à + ou – 0,5 point (avec les mêmes résultats SAS ou POULPE).

La différence peut être plus importante, notamment pour des questions d'opinion comme la préoccupation pour le manque de sécurité dans le quartier, où l'effet de grappe est plus important dans le tirage des individus répondants au niveau des ménages.

% en 2004 d'individus pour qui le manque de sécurité est un problème dans leur quartier

Zone d'étude et d'aménagement du territoire (ZEAT)	%	Intervalle de confiance à 95% (SAS)	Intervalle de confiance à 95% (avec POULPE)	Effectif brut ayant répondu à la question
Nord	17,6 %	{ 14,8 ; 20,4 }	{ 14,0 ; 21,2 }	833
France métropolitaine	13,0 %	{ 12,4 ; 13,6 }	{ 12,1 ; 13,9 }	11 701

***Lecture** : pour la France, les 13% sont estimés à + ou - 0,6 point en faisant l'hypothèse d'un SAS, à + ou - 0,9 point avec POULPE. Pour le Nord, les 17,6% sont estimés à + ou – 2,8 points en faisant l'hypothèse d'un SAS, à + ou – 3,6 points avec POULPE.*