RECENSEMENT DE LA POPULATION

LA PRÉCISION DES RÉSULTATS DU RECENSEMENT

La qualité des résultats statistiques d'un recensement dépend de multiples facteurs et en premier lieu de la qualité de la collecte. Elle dépend aussi de la fiabilité des fichiers et répertoires utilisés ainsi que de la qualité des différents traitements mis en œuvre : saisie, contrôles, redressement des anomalies et codification des variables. Enfin, la fraîcheur des données et la comparabilité avec celles d'autres pays sont aussi des critères de qualité. **Cette fiche s'attache plus particulièrement à un des aspects de la qualité : la précision des résultats statistiques issus des sondages.**

Sommaire

1 Le plan de sondage et les échantillons du recensement

Le recensement est basé sur un plan de sondage qui est décrit en détail dans le document « *Comprendre le recensement de la population* »[1] :

- les communes de moins de 10 000 habitants sont enquêtées exhaustivement ;
- dans les communes de 10 000 habitants ou plus, le recensement des ménages est réalisé par sondage sur un échantillon d'environ 40 % des logements sur cinq ans. Les communautés, les habitations mobiles et sans abris sont enquêtés exhaustivement.

Comme lors des recensements généraux traditionnels, le recensement fait l'objet d'une exploitation principale et d'une exploitation complémentaire[2]. L'exploitation principale porte sur l'ensemble des bulletins collectés. L'exploitation complémentaire produit des variables concernant la structure familiale du ménage, l'activité économique et les professions, et porte sur un échantillon pour des raisons de coût de traitement :

Échantillon complémentaire

Population	Communes de moins de 10 000 habitants	Communes de 10 000 habitants et plus
Ménages	20 % (*)	100 % des ménages enquêtés, soit environ 40 % du total des ménages
Individus des communautés	20 % (*)	
Habitations mobiles et terrestres et personnes sans abri	20 % (*)	100 %
Habitations mobiles fluviales (bateliers)	100 %	

(*) Jusqu'à l'enquête annuelle de recensement de 2013, ce taux de sondage était de 25 %.

2 La mesure de la précision des résultats

Pour toutes les variables, qu'elles soient issues de l'exploitation principale ou complémentaire, le sondage pratiqué entraîne une marge d'incertitude sur les résultats. Cette marge est mesurée, pour une variable donnée, par le **coefficient de variation,** noté **CV**. Il renseigne sur l'écart moyen entre la valeur estimée par le recensement et la « vraie » valeur. En termes statistiques il correspond au rapport de l'écart-type à la moyenne. Ce coefficient de variation permet de construire un intervalle de confiance de l'estimation. La vraie valeur est comprise dans 95 % des cas dans la plage de valeurs possibles suivante :

[valeur au recensement x (1 - 2CV) ; valeur au recensement x (1 + 2CV)]

Cette imprécision varie d'une commune à l'autre en fonction du taux de sondage. Elle dépend donc du type de la commune (moins de 10 000 habitants / 10 000 habitants ou plus) et de l'exploitation (principale ou complémentaire) considérés.

Elle varie aussi selon l'effectif de la variable d'intérêt (plus l'effectif obtenu est réduit, plus l'imprécision risque d'être grande car l'estimation de cet effectif repose sur peu d'observations).

1 Insee Méthodes – Hors série – mai 2005

2 Pour plus de détails, se référer à la fiche thématique « *Les exploitations principale et complémentaire* »

3 Les communes de moins de 10 000 habitants

Le recensement est exhaustif dans les communes de moins de 10 000 habitants : les variables issues de l'exploitation principale ne souffrent pas d'imprécision liée au sondage.

En revanche, hormis pour les bateliers, les variables issues de l'exploitation complémentaire présentent une marge d'incertitude liée au taux de sondage T_c qui vaut 25 % lors des enquêtes annuelles de recensement antérieures à 2014, et 20 % à partir de 2014 (cf. tableau §1).

Ainsi, pour un effectif estimé « a » d'une variable issue de l'exploitation complémentaire, l'écart-type, déterminé de façon empirique, est égal à $\sqrt{\frac{a}{T_c}}$ et le coefficient de variation à $\frac{1}{\sqrt{T_c a}}$.

L'intervalle de confiance à 95 % est donné par : $\left[a - 2\sqrt{\frac{a}{T_c}} \; ; \; a + 2\sqrt{\frac{a}{T_c}} \right]$.

Ainsi pour un effectif de 4 900 établi au recensement de la population 2011, basé sur les enquêtes annuelles de 2009 à 2013, l'écart-type vaut 140, le coefficient de variation 3 % et l'intervalle de confiance à 95 % est [4 620 ; 5 180].

4 Les communes de 10 000 habitants ou plus

Dans les communes de 10 000 habitants ou plus, l'échantillon des ménages enquêté représente environ 40 % des logements et des habitants de la commune. En France métropolitaine, le taux de sondage varie d'une commune à l'autre car il dépend de la structure de l'habitat : seules les petites adresses connues sont enquêtées par sondage, les adresses de grande taille et les adresses nouvelles étant enquêtées exhaustivement[3].

Hormis pour les communautés (cf. tableau §1), l'échantillon de l'exploitation complémentaire est le même que pour l'exploitation principale ; la marge d'imprécision est donc la même pour les variables issues des deux exploitations.

Pour les communautés, la précision des variables issues de l'exploitation complémentaire est obtenue par la même formule que dans les communes de moins de 10 000 habitants (cf. §3).

4.1 Résultats sur la population des ménages

La précision de la population des ménages en grandes communes de métropole est calculée en tenant compte du plan de sondage et du calage réalisé à l'Iris[4]. En revanche, elle sous-estime la variance pour plusieurs raisons détaillées dans l'article Brilhault et Caron (2016)[5].

La population des ménages vivant en France métropolitaine est ainsi estimée à 0,05 % près, soit plus ou moins 15 800 personnes. Le tableau suivant, tiré de l'article de Brilhault et Caron (2016), indique la distribution du coefficient de variation associé à la variable population au niveau régional, départemental, et communal.

3 Les petites adresses sont les adresses dont le nombre de logements est inférieur à un seuil propre à chaque commune. Pour plus de détails sur le plan de sondage, voir : Insee Méthodes – Hors série – mai 2005

4 Pour plus de détails, se référer à la fiche thématique « *Les pondérations* »

5 Brilhault et Caron (2016), « *Le passage à une enquête par sondage : quel impact sur la précision du recensement ?* » in Économie et Statistique n°483-484-485, avril 2016.

Distribution des coefficients de variation (en %) de la variable population aux niveaux régional, départemental, et communal

	Région	Département	Commune
75 % Q3	0,20	0,53	1,05
50 % Médiane	0,16	0,35	0,88
25 % Q1	0,12	0,25	0,71

Lecture : pour un quart des régions, le CV de la variable population est inférieur à 0,12 % ; pour la moitié, ce CV est inférieur à 0,16 %, et pour un quart, il est supérieur à 0,20 %.
Champ : communes de 10 000 habitants ou plus de France métropolitaine.
Source : recensement de la population de 2006.

Des éléments plus détaillés sur la précision de la population en grandes communes sont indiqués dans une fiche spécifique[6].

4.2 Résultats sur les autres variables statistiques

Le tableau suivant est extrait du rapport du Cnis sur l'utilisation des données produites par le recensement rénové de la population et leur diffusion de 2005[7]. Il fournit quelques indications sur la précision des résultats dans les communes de 10 000 habitants ou plus pour différentes tranches d'effectifs. Ces coefficients de variation ont été obtenus à partir de tirages dans le recensement exhaustif de 1999 en suivant le plan de sondage envisagé pour le recensement rénové.

Ce tableau a été établi pour le chiffre de l'ensemble de la population d'une **commune** et permet également de juger de la précision des résultats pour des populations ciblées si celles-ci se répartissent de façon homogène sur l'ensemble de la commune. Dans les faits, ce tableau diffère selon les communes puisque la part des adresses enquêtées exhaustivement varie d'une commune à l'autre.

La précision des résultats pour les communes de 10 000 habitants ou plus de France métropolitaine

Tranches d'effectif	Coefficient de variation
50 000 ou plus	< 1,0 %
20 000 – 49 999	1,5 %
10 000 – 19 999	2,0 %
6 000 – 9 999	2,5 %
3 000 – 5 999	3,0 %
2 000 – 2 999	3,5 %
1 000 – 1 999	4,5 %
500 – 999	6,0 %
250 – 499	8,0 %
Moins de 250	> 8,0 %

Lecture : pour une population estimée à 2 700 enfants de moins de 5 ans dans une commune donnée, le coefficient de variation mesurant la précision de cette estimation est de 3,5 %.
Champ : communes de 10 000 habitants ou plus de France métropolitaine.
Source : simulations à partir du recensement exhaustif de 1999.

6 Fiche thématique sur la précision du chiffre de population dans les grandes communes de métropole
7 Cnis (2005), *Utilisation des données produites par le recensement rénové de la population et leur diffusion*, décembre.

Des résultats plus récents sur la précision des variables statistiques ont été calculés au niveau **Iris** à partir du recensement de la population de 2013. L'Iris constitue la plus petite maille géographique diffusée dans une commune. Elle comporte en moyenne 2 000 habitants, taille minimale pour l'analyse infracommunale de variables du recensement. Ce découpage permet notamment d'analyser les disparités au sein d'une commune.

Des coefficients de variation « résumés » sont disponibles pour chaque variable et selon la taille de l'effectif estimé dans l'Iris. Ils sont fournis dans les fichiers de données infra-communales mis à disposition sur le site insee.fr.

5 Calculs de précision : cas pratiques

5.1 Précision des données en niveau

Soit un tableau donnant la répartition par âge de la population d'une commune de plus de 10 000 habitants, obtenue par le recensement. Pour chaque effectif on peut déterminer la précision de l'estimation fournie par le recensement, qui se traduit par une plage de valeurs possibles.

Tranches d'âge	Population au recensement	Coefficient de variation (tableau §4.2)	Plage de valeurs possibles
Moins de 20 ans	4 000	3,0 %	4 000 + ou - 240
De 20 à 39 ans	6 000	2,5 %	6 000 + ou - 300
De 40 à 59 ans	6 000	2,5 %	6 000 + ou - 300
60 ans ou plus	4 000	3,0 %	4 000 + ou - 240
Ensemble	20 000	1,5 %	20 000 + ou - 600

Calcul : pour un effectif donné, la précision mesurée par le coefficient de variation est directement tirée du tableau du § 4.2. La plage de valeurs possible se calcule alors avec la formule :

population estimée + ou – [2 x (population estimée x CV)]

Interprétation : le nombre d'habitants de moins de 20 ans se situe, dans 95 % des cas, entre 3 760 et 4 240.

5.2 Précision des données en structure

Dans la même commune de 20 000 habitants, il s'agit désormais de mesurer la précision de la répartition (en %) de la population par tranche d'âge.

La précision d'un pourcentage dépend à la fois de la précision de son numérateur et de celle de son dénominateur. Le coefficient de variation du pourcentage s'obtient à partir des coefficients de variation du numérateur et du dénominateur se calcule avec la formule suivante :

$$CV_{pourcentage} = \sqrt{(CV_{numérateur})^2 + (CV_{dénominateur})^2}$$

NB : ce calcul ne tient pas compte de la corrélation entre le numérateur et le dénominateur ; si on en tenait compte, la précision serait en réalité meilleure.

Tranches d'âge	Population au recensement	Coefficient de variation (tableau §4.2)	Plage de valeurs possibles
Moins de 20 ans	20 %	3,4 %	(20 + ou - 1,4) %
De 20 à 39 ans	30 %	2,9 %	(30 + ou - 1,7) %
De 40 à 59 ans	30 %	2,9 %	(30 + ou - 1,7) %
60 ans ou plus	20 %	3,4 %	(20 + ou - 1,4) %
Ensemble	100 %		100 %

Calcul : pour les moins de 20 ans, 20 % correspond à un effectif de 4 000 (CV de 3 % d'après le tableau du § 4.2) sur une population de référence de 20 000 (CV de 1,5 % d'après le même tableau). La précision de la part des moins de 20 ans est donc de : $\sqrt{(0,03)^2 + (0,015)^2} = 0,034 = 3,4\ \%$ et la marge d'incertitude de : 0,20 x 0,034 x 2 = 1,4 %.

Interprétation : la proportion de personnes de 20 à 39 ans est donc assurément plus élevée que celle des moins de 20 ans, car la valeur minimale de la première proportion (30 - 1,7 = 28,3 %) est supérieure à la valeur maximale de la seconde (20 + 1,4 = 21,4 %).

5.3 Précision d'un taux

Pour analyser un taux, la démarche est la même que pour des données en structure.

Dans la même commune, pour analyser par exemple la précision d'un taux de chômage de 10 % sur une population de 10 000 actifs, on doit tenir compte de l'imprécision sur le nombre de chômeurs (effectif de 1 000, donc CV de 4,5 % d'après le tableau du § 4.2) et de l'imprécision sur la population des actifs (CV de 2 % d'après le même tableau).

La formule de calcul du CV du taux de chômage de cette commune est donc la suivante :

$$CV_{taux\ de\ chômage} = \sqrt{(CV_{chômeurs})^2 + (CV_{actifs})^2}$$

NB : ce calcul ne tient pas compte de la corrélation entre le numérateur et le dénominateur ; si on en tenait compte, la précision serait en réalité meilleure.

Calcul : la précision du taux de chômage, mesurée par le CV, est donc de $\sqrt{(0,045)^2 + (0,02)^2} = 0,05$, soit 5 %, et la marge d'incertitude est de : 0,10 x 0,05 x 2 = 0,01 soit 1 point sur le taux de chômage.

5.4 Précision des données d'une zone composée de plusieurs communes

Pour analyser une zone constituée de plusieurs communes, il est possible de calculer la marge d'imprécision pour l'ensemble de la zone. Le coefficient de variation pour la zone est donné par la formule suivante :

$$CV_{Zone} = \frac{\sqrt{\sum_c \left(CV_{commune(c)} \times Effectif_{commune(c)}\right)^2}}{\sum_c Effectif_{commune(c)}}$$

Prenons, par exemple, une zone formée de :

- une commune A de moins de 10 000 habitants recensée avant 2014
- une commune B de moins de 10 000 habitants recensée en 2014
- une commune C de 10 000 habitants ou plus.

Dans le cas d'une variable tirée de l'exploitation principale dont les effectifs sont respectivement de 2 000 pour la commune A, de 1 000 pour la commune B et de 5 000 pour la commune C :

- pour les deux communes de moins de 10 000 habitants A et B, il n'y a pas d'imprécision du fait du sondage (voir le § 3) ;
- pour la commune C de 10 000 habitants ou plus, la précision sur un effectif de 5 000 est de 3 % (voir le § 4.2).

Ainsi, le coefficient de variation pour l'ensemble des trois communes est égal à :

$$CV_{Zone} = \frac{CV_{commune\ C} \times 5000}{2000 + 1000 + 5000} = \frac{0{,}03 \times 5000}{8000} = 1{,}9\%$$

La marge d'imprécision associée est de + ou - 3,8 %. L'effectif dans la zone multi-communale est donc compris dans l'intervalle : 8 000 + ou - 304.

Si la variable était tirée de l'exploitation complémentaire, les effectifs des deux communes de moins de 10 000 habitants seraient aussi affectés d'une imprécision (voir le § 3).

Pour les mêmes effectifs, le coefficient de variation deviendrait donc :

$$CV_{Zone} = \frac{\sqrt{(CV_{Commune\ A} \times 2000)^2 + (CV_{Commune\ B} \times 1000)^2 + (CV_{Commune\ C} \times 5000)^2}}{2000 + 1000 + 5000}$$

Avec : $CV_{Commune\ A} = 2 / \sqrt{2000}$ et $CV_{Commune\ B} = \sqrt{5} / \sqrt{1000}$,

$$\text{soit : } CV_{Zone} = \frac{\sqrt{8000 + 5000 + (0{,}03 \times 5000)^2}}{2000 + 1000 + 5000} = \frac{\sqrt{35500}}{8000} = 2{,}4\ \%$$

La marge d'imprécision est alors de + ou - 4,8 %. L'effectif dans la zone multi-communale est donc compris dans l'intervalle : 8 000 + ou - 384 .

6 Application à la comparaison de deux communes

Première approche **:**

Pour comparer deux communes, ou plus généralement, deux zones composées de communes, au regard d'une variable, une première approche consiste à comparer les plages de valeurs possibles pour les effectifs correspondants, selon la méthode exposée plus haut (voir le § 5.2).

Exemple : pour comparer la population des 20-39 ans d'une grande commune A (6 000) à celle des moins de 20 ans d'une grande commune B (7 000), on calcule les intervalles de confiance associés à ces populations à l'aide des CV calculés.

Commune A

Tranches d'âge	Population au recensement	Coefficient de variation (tableau §4.2)	Plage de valeurs possibles
Moins de 20 ans	4 000	3,0 %	4 000 + ou – 240
De 20 à 39 ans	6 000	2,5 %	6 000 + ou – 300
De 40 à 59 ans	6 000	2,5 %	6 000 + ou – 300
60 ans ou plus	4 000	3,0 %	4 000 + ou – 240
Ensemble	20 000	1,5 %	20 000 + ou – 600

Commune B

Tranches d'âge	Population au recensement	Coefficient de variation (tableau §4.2)	Plage de valeurs possibles
Moins de 20 ans	7 000	2,5 %	7 000 + ou – 350
De 20 à 39 ans	13 000	2,0 %	13 000 + ou – 520
De 40 à 59 ans	13 000	2,0 %	13 000 + ou – 520
60 ans ou plus	7 000	2,5 %	7 000 + ou – 350
Ensemble	40 000	1,5 %	40 000 + ou – 1 200

L'effectif maximal de la tranche 20-39 ans de la commune A (6 300) est inférieur à l'effectif minimal de la tranche moins de 20 ans de la commune B (6 650) ; on peut donc considérer que la population des 20-39 ans de la commune A est bien inférieure à celle des moins de 20 ans de la commune B.

Deuxième approche **:**

Pour de telles comparaisons, la démarche la plus rigoureuse consiste à calculer le coefficient de variation de la quantité analysée, en l'occurrence, ici, de la différence entre les deux populations que l'on souhaite comparer (population a de la commune A et population b de la commune B) :

$$CV_{a-b} = \frac{\sqrt{(CV_a \times Eff_a)^2 + (CV_b \times Eff_b)^2}}{Eff_a - Eff_b}$$

On calcule donc le CV associé à la différence entre les deux effectifs d'intérêt 7 000-6 000 = 1 000 ainsi :

$$CV_{20-39\ ans\ Commune\ A - Moins\ 20\ ans\ Commune\ B} = \frac{\sqrt{(0{,}025 \times 6000)^2 + (0{,}025 \times 7000)^2}}{7000 - 6000} = \frac{230}{1000} = 23\%$$

La différence est donc comprise entre (1 000 - 460) et (1 000 + 460). Elle est toujours positive. On peut donc affirmer avec une forte certitude que, malgré les imprécisions liées au sondage, la population des 20-39 ans de la commune A est bien inférieure à celle des moins de 20 ans de la commune B.

La démarche pour la comparaison de deux zones supracommunales est analogue.