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.

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 » (**Insee Méthodes - Hors série - mai 2005**) :

- les communes de moins de 10 000 habitants sont enquêtées exhaustivement ;
- dans les communes de 10 000 habitants ou plus, le recensement est réalisé par sondage sur un échantillon d'environ 40 % des logements sur cinq ans.

Comme lors des recensements précédents, le recensement fait l'objet d'une **exploitation principale** sur l'ensemble des bulletins collectés et d'une **exploitation complémentaire** sur un sous-échantillon. Les variables concernant la structure familiale du ménage, l'activité économique et les professions, sont, pour des raisons de coût de traitement, élaborées sur un échantillon dit « complémentaire » composé de la façon suivante :

- pour les communes de 10 000 habitants ou plus : l'ensemble des résidences principales recensées et leurs habitants, soit environ 40 % ;
- pour les communes de moins de 10 000 habitants : 25 % des résidences principales recensées et leurs habitants ;
- pour toutes les communes : 25 % des personnes recensées dans les communautés et les habitations mobiles et 25 % des personnes sans abri.
- l'ensemble des mariniers.

Dans les deux exploitations, principale et complémentaire, le sondage entraîne une marge d'incertitude sur les résultats. Cette fiche donne les informations permettant d'apprécier l'ampleur de cette marge d'incertitude et les conseils d'utilisation des données qui en découlent.

2. Le cas des communes de 10 000 habitants ou plus

2.1. La précision des données

Le sondage entraîne une marge d'incertitude sur les résultats statistiques, mesurée, pour une variable donnée, par le **coefficient de variation** noté : CV. Il renseigne sur l'écart relatif possible entre la valeur donnée par le RP 2006 et la vraie valeur. En termes statistiques il correspond au rapport de l'écart type à la moyenne. La vraie valeur sera comprise dans 95 % des cas dans la plage de valeurs possibles suivante :

[valeur du RP2006 x (1 - 2CV) ; valeur du RP2006 x (1 + 2CV)].

La précision des données du RP2006 a été estimée par la simulation d'un grand nombre de tirages dans le fichier des données du recensement de 1999.

Le tableau suivant montre la précision indicative obtenue pour différents effectifs d'une variable du RP2006 en commune de 10 000 habitants ou plus. Il est valable pour le principal comme pour le complémentaire puisque, pour ces communes, l'échantillon est quasiment le même.

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

Tranches d'effectif	Précision (CV)
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 %

Source : Rapport du CNIS « Utilisation des données produites par le recensement rénové de la population et leur diffusion » - Décembre 2005

Ce tableau a été établi pour l'ensemble de la population d'une commune. La précision des autres variables dépend de la répartition de la population correspondante sur le territoire considéré. Si la répartition est homogène, le tableau donne une bonne indication de la précision ; si elle est peu homogène, la précision est moins bonne.
En toute rigueur, ce tableau peut différer un peu selon les communes. En effet, les grandes adresses et les adresses nouvelles étant enquêtées exhaustivement, leur part dans l'ensemble des adresses est susceptible d'influer sur la précision.

Quelle utilisation concrète peut-on faire de cet indicateur de précision ? Que veut dire un coefficient de variation de 10 % en général ? Quel résultat peut-on énoncer sans risque d'erreur ? Cela dépend du contexte d'utilisation, et de la question que l'on se pose.

2.2. Précision des données relatives à une commune

2.2.1. Des données en niveau

Soit un tableau donnant la répartition par âge de la population d'une commune de 20 000 habitants. Pour chaque effectif on peut déterminer la précision qui se traduit par une plage de valeurs possibles.

Tranche d'âges	Population au RP2006	Précision (CV)	Plage de valeurs possibles
De 0 à 19 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 § 2.1. La plage de valeurs possible se calcule alors avec la formule :

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

Interprétation : Par exemple, le nombre d'habitants de 0 à 19 ans se situe, dans 95 % des cas, entre 3 760 et 4 240.

2.2.2. Précision d'un taux

Autre cas de figure, si on analyse 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) soit, d'après le tableau du § 2.1, 4.5 % et de l'imprécision sur la population des actifs soit 2 %. Le numérateur est compris entre 910 et 1 090 et le dénominateur entre 9 600 et 10 400. En première approximation on peut en déduire une fourchette du taux de chômage : [(910 / 10 400) ; (1 090 / 9 600)] soit : [8,8 % ; 11,4 %].

Un calcul plus précis est le suivant : $CV_{tauxchômagee} = \sqrt{\left(CV_{chômage}\right)^2 + \left(CV_{actifs}\right)^2}$

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

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.

2.2.3. Précision des données en structure

Lorsqu'on veut analyser des données en structure, la démarche est la même que pour le taux de chômage, pour chacune des cases analysées.

Tranche d'âges	Population au RP2006	Précision	Plage de valeurs possibles
De 0 à 19 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) %
De 60 ou plus	20 %	3,4 %	(20 + ou - 1,4) %
Ensemble	100 %		100 %

Calcul : Pour la case des 0–19 ans, 20 % correspond à un effectif de 4 000 sur une population de référence de 20 000. La précision de la part des 0–19 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 : Dans ce cas, on peut conclure par exemple que la proportion de personnes de 40 à 59 ans est assurément plus élevée que celle de 60 ans et plus, 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 %).

2.3. Précision des données infracommunales

La plus petite maille géographique diffusée dans une commune est l'Iris. Elle comporte en moyenne 2 000 habitants, taille minimale pour l'analyse infracommunale de variables du RP. Ce découpage permet notamment d'analyser les disparités au sein d'une commune. Si on analyse un tableau

donnant la répartition de la population par tranches d'âges d'un Iris, on obtient, toujours à partir du tableau du § 2.1, les précisions suivantes :

Tranche d'âges	Population au RP2006	Précision	Plage de valeurs possibles
De 0 à 19 ans	426	8,0 %	426 + ou - 68
De 20 à 39 ans	618	6,0 %	618 + ou - 74
De 40 à 59 ans	469	8,0 %	469 + ou - 75
60 ans ou plus	454	8,0 %	454 + ou - 73
Ensemble	1 967	4,5 %	1 967 + ou - 177

Calcul : Principe similaire à celui des données communales. Seuls les effectifs changent.

Interprétation : On voit que la précision pemet d'effectuer une analyse des disparités dans la commune, en comparant plusieurs Iris. Cependant, un tableau à 10 cases semble correspondre au niveau de détail maximum pour utiliser un Iris dans une analyse infracommunale.

NB : Par ailleurs, à ce niveau, le lien entre la précision et l'effectif est moins net car les effets de grappe, induits par le tirage d'adresses et non de logements ou d'individus, y sont plus sensibles.

2.4. Comparaison avec la précision du recensement de 1999 - Exploitation complémentaire

Pour les communes de 10 000 habitants ou plus, les précisions des résultats de l'exploitation complémentaire au RP2006 et au RP1999 (cf. formule au paragraphe suivant) sont très proches.

Effectifs	Précision (CV) RP2006	Précision (CV) RP1999
Plus de 50 000	< 1,0 %	< 0,9 %
50 000	1,0 %	0,9 %
20 000	1,5 %	1,4 %
10 000	2,0 %	2,0 %
6 000	2,5 %	2,6 %
3 000	3,0 %	3,7 %
2 000	3,5 %	4,5 %
1 000	4,5 %	6,3 %
500	6,0 %	8,9 %
250	8,0 %	13,0 %
Moins de 250	> 8,0 %	> 13,0 %

3. Le cas des communes de moins de 10 000 habitants

Pour les variables du complémentaire, la marge d'incertitude est la même que lors des recensements antérieurs. De façon empirique, pour un effectif « a », l'écart-type est égal à $\sqrt{4a}$ et le coefficient de variation à $2/\sqrt{a}$. La marge d'incertitude est de + ou - $4\sqrt{a}$. Ainsi pour un effectif de 10 000, l'écart-type vaut 200, la précision est de 2 % et l'intervalle de confiance à 95 % de + ou - 400.

C'est pourquoi, les résultats tirés de l'exploitation complémentaire ne sont pas affichés pour les zones de moins de 2 000 habitants. En revanche, les bases téléchargeables contiennent les résultats pour toutes les communes mais à seule fin de permettre des agrégations sur des zones d'au moins 2 000 habitants. Les informations pourront être utilisées avec un niveau de détail d'autant plus grand qu'elles concernent une zone plus peuplée.

4. Le cas d'une zone composée de plusieurs communes

Lorsqu'on analyse une zone constituée de plusieurs communes, il est possible de calculer la marge d'imprécision pour l'ensemble de la zone.

Prenons, par exemple, une zone formée de deux communes A, B de moins de 10 000 habitants et d'une commune C de 10 000 habitants ou plus.

Considérons une variable tirée de **l'exploitation principale** dont les effectifs sont respectivement : 2 000, 1 000 et 5 000.

Pour les communes de moins de 10 000 habitants, il n'y a pas d'imprécision du fait du sondage.

Pour la commune de 10 000 habitants ou plus, la précision sur un effectif de 5 000 est de 3 %.

Le coefficient de variation de la variable pour l'ensemble des trois communes est donné par la formule suivante :

$$\frac{CV_{CommuneC} \text{ x } 5000}{2000+1000+5000} = \frac{0{,}03 \text{ x } 5000}{8000} = 1{,}9\ \%$$

La marge d'imprécision est de + ou - 3,8 %. La valeur est comprise dans l'intervalle : 8 000 + ou - 304.

Si la variable était tirée de **l'exploitation complémentaire**, les effectifs des communes de moins de 10 000 habitants seraient aussi affectés d'une imprécision. Pour les mêmes effectifs, le coefficient de variation serait :

$$\frac{\sqrt{(CV_{CommuneA} \times 2000)^2 + (CV_{CommuneB} \times 1000)^2 + (CV_{CommuneC} \times 5000)^2}}{2000+1000+5000}$$

Avec : $CV_{CommuneA} = 2/\sqrt{2000}$ et $CV_{CommuneB} = 2/\sqrt{1000}$, soit :

$$CV_{Zone} = \frac{\sqrt{(8000)+(4000)+(0{,}03 \times 5000)^2}}{2000+1000+5000} = \frac{\sqrt{34500}}{8000} = 2{,}3\ \%$$

La marge d'imprécision est de + ou - 4,6 %. La valeur est comprise dans l'intervalle : 8 000 + ou - 368.

5. La comparaison de deux communes

Lorsqu'on veut comparer deux communes, ou 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.

Exemple : soit à comparer l'effectif des 20-39 ans de la commune A (6 000) à celui des moins de 20 ans de la commune B (7 000), communes qui présentent les structures par âges suivantes.

Commune A :

Tranche d'âges	Population au RP2006	Précision	Plage de valeurs possibles
De 0 à 19 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
De 60 ou plus	4 000	3,0 %	4 000 + ou – 240
Ensemble	20 000	1,5 %	20 000 + ou – 600

Commune B :

Tranche d'âges	Population au RP2006	Précision	Plage de valeurs possibles
De 0 à 19 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
De 60 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 bien inférieur à l'effectif minimal de la tranche 0-19 ans de la commune B (6 650).

Cependant, pour de telles comparaisons, la démarche la plus rigoureuse consiste à calculer le coefficient de variation de la quantité analysée, en l'occurrence la différence entre les deux valeurs de la variable.

$$CV_{(CommuneA-CommuneB)} = \frac{\sqrt{(CV_{CommuneA} \times 6000)^2 + (CV_{CommuneB} \times 7000)^2}}{7000 - 6000} = \frac{230}{1000} = 23\,\%$$

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, l'effectif des moins de 20 ans de la commune B est bien supérieur à l'effectif des 20-39 ans de la commune A.

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