# Estimation de variance en présence de données imputées : un exemple à partir de l'enquête Panel Européen

Nathalie Caron

**INSTITUT NATIONAL DE LA STATISTIQUE ET DES ÉTUDES ÉCONOMIQUES**
Série des Documents de Travail
Méthodologie Statistique

**N° 9902**

# Estimation de variance en présence de données imputées : un exemple à partir de l'enquête Panel Européen.

**N. Caron**

Novembre 1999

# *Estimation de variance en présence de données imputées : un exemple à partir de l'enquête Panel Européen.*

*Nathalie Caron*
*INSEE, CREST - Rennes, Laboratoire de Statistique d'Enquêtes*
*Campus de Ker Lann - Rue Blaise Pascal - 35170 BRUZ*

**Résumé :**

Dans ce papier, nous évaluons l'impact sur la variance totale de différentes méthodes de correction de la non-réponse partielle par imputation. L'étude est menée à partir de l'enquête Panel Européen, réalisée chaque année par l'Insee depuis 1994, sur la variable revenu d'activité indépendante pour laquelle plus de 10% des valeurs sont imputées. Plusieurs méthodes d'imputation sont envisagées : méthode de hot-deck et deux méthodes économétriques de type revenus simulés. Comme l'enquête Panel Européen est, ainsi que son nom l'indique, une enquête par panel, nous évaluerons aussi avec ce type d'enquête dans une partie préalable le gain obtenu en terme de précision sur des estimations d'évolution par rapport à deux enquêtes indépendantes.

**MOTS CLES :** Correction de la non-réponse par imputation, comparaison de plusieurs méthodes d'imputation, calcul de variance en présence de données imputées.

# Sommaire

# I. Introduction

Les enquêtes par sondage ainsi que les recensements sont confrontés au problème de la non-réponse. On distingue traditionnellement deux grands types de non-réponse : la non-réponse totale lorsqu'on n'observe aucune réponse pour un individu échantillonné (ou que le statisticien juge que le questionnaire est inexploitable) et la non-réponse partielle lorsque l'individu échantillonné a répondu partiellement au questionnaire. Chaque grand type de non-réponse nécessite une technique particulière de correction. La non réponse totale se corrige par des méthodes de repondération (pour plus de détails voir Caron N., 1996). Quant à la non réponse partielle, elle se corrige par des méthodes d'imputation qui consistent à remplacer la donnée absente ou invalide par une donnée « plausible » qui est en général issue ou estimée à partir de la distribution des répondants. Par nature, l'imputation est une opération délicate et n'est réalisée que si la non-réponse n'est pas trop importante dans le fichier. En pratique, le seuil à partir duquel on évite de procéder à des imputations se situe autour de 40%. De plus, lorsque le pourcentage de non-réponse devient important, les méthodes d'imputation sont plus complexes et cherchent en particulier à utiliser le maximum d'informations auxiliaires et à déformer le moins possible la distribution obtenue sur les répondants.

Les méthodes d'imputation sont attractives car lorsque les données manquantes sont imputées, on dispose d'un ensemble de données complet. Cependant, la présence de données imputées n'est pas sans conséquence sur la variance des estimateurs. En effet, le fait d'avoir des données imputées dans le fichier d'exploitation des données augmente en général la variance qu'on aurait obtenue si les données étaient considérées comme réelles c'est-à-dire s'il n'y avait pas eu de non-réponse. De plus, chaque méthode d'imputation de réponses conduit à une formule de variance ainsi qu'une estimation de variance particulière. Ces résultats sont détaillés dans le cas d'un sondage aléatoire simple dans Caron N. (1996). Les problèmes liés à une mauvaise spécification du modèle d'imputation ne seront pas abordés dans ce papier.

L'étude présentée dans ce papier consiste à évaluer l'impact sur la variance totale de la méthode d'imputation utilisée dans le cas d'une enquête dont le plan de sondage est complexe. Elle a ainsi été conduite à partir de l'enquête Panel Européen réalisée par l'Insee en 1994. La variable utilisée est le revenu des indépendants pour lequel plus de 10% des valeurs sont imputées et le paramètre d'intérêt est le revenu moyen des indépendants. Plusieurs méthodes d'imputation sont envisagées : une méthode de type hot-deck et deux méthodes de type économétrique. La première méthode consiste à remplacer la donnée manquante par une valeur observée sur un individu répondant tandis que les méthodes de type économétrique consistent à spécifier un modèle économétrique à partir de la population des répondants et d'imputer pour les réponses manquantes la valeur prédite augmentée d'un résidu non nécessairement observé. Ce dernier peut soit être généré de façon aléatoire à partir d'une distribution normale d'espérance 0 et de variance celle des résidus du modèle obtenus à partir des répondants soit être directement sélectionné parmi l'ensemble des résidus du modèle obtenus sur la population des répondants.
Le principal intérêt des enquêtes par panel est d'obtenir une meilleure précision des évolutions. Comme l'enquête Panel européen est, ainsi que son nom l'indique, une enquête avec réinterrogation, nous évaluerons dans une partie préalable le gain obtenu en terme de précision sur des estimations d'évolution avec ce type d'enquête par rapport à deux enquêtes totalement indépendantes.

## II. Description de la méthodologie de l'enquête

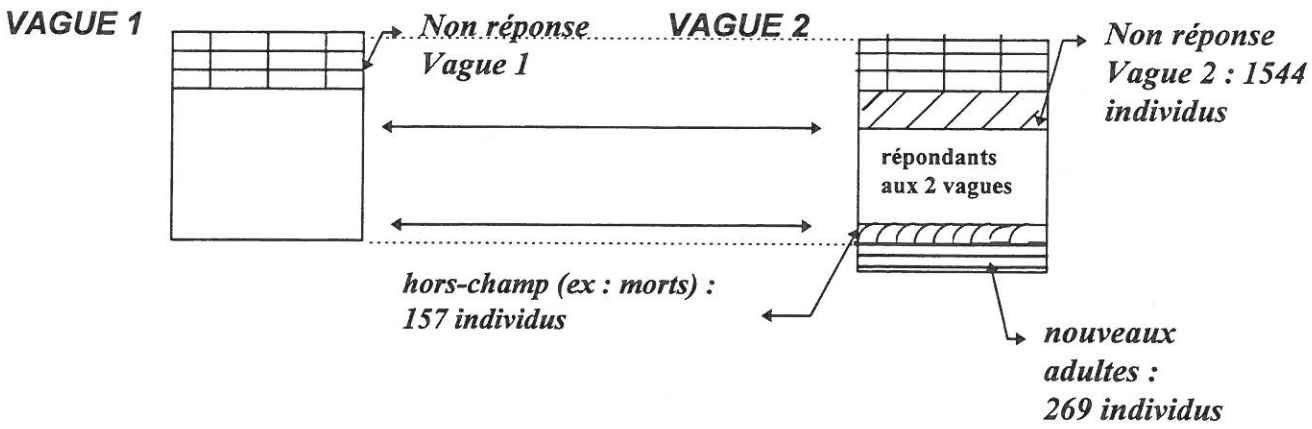
L'objectif principal de l'enquête Panel Européen réalisée par l'Insee depuis 1994 est d'étudier la dynamique d'emploi et de revenus des personnes. Comme pour la majorité des enquêtes ménages de l'Insee, les ménages interrogés pour l'enquête Panel européen ont été extraits de l'Echantillon maître 1990 (E.M. 90), c'est-à-dire qu'ils ont été sélectionnés par un plan de sondage stratifié à plusieurs degrés selon la taille des communes. Parmi les 11080 ménages ainsi choisis, 1461 ont été déclarés hors champ (c'est-à-dire que les logements choisis ne sont pas au moment de l'enquête des résidences principales) et 2275 ont été déclarés non-répondants. L'enquête comporte deux questionnaires, l'un concerne le ménage (conditions de logements,...) l'autre les individus de 17 ans ou plus appelés par la suite individus adultes. Ce sont les individus répondants à la première enquête (appelés par la suite individus Panel) qui constituent ce que l'on appelle le panel et qui seront interrogés plusieurs années de suite. Notons que l'on ne peut parler que d'un panel d'individus et pas d'un panel de ménages qui sont des entités fortement changeantes au cours du temps.
La non-réponse totale a été corrigée par la méthode de correction par groupes homogènes. Cette procédure introduit une phase supplémentaire. L'enquête est donc en deux phases. Elle a été suivie d'une correction des fluctuations d'échantillonnage en réalisant un calage de la structure des données de l'enquête obtenue à partir des logements répondants sur celle obtenue au moment de la réalisation de l'enquête Emploi 1994.

Les variables de calage utilisées sont le nombre d'actifs, la Zeat (qui correspond à un regroupement de régions), le nombre d'individus dans le ménage, le nombre d'hommes par tranche d'âge (en sept postes) et le nombre de femmes par tranche d'âge (en six postes). Au niveau ménage, les trois premières variables sont qualitatives ; les deux dernières concernant des comptages d'individus sont quantitatives. Comme le fichier a « subi » un calage sur marges, la théorie des sondages montre que la variance d'échantillonnage du total estimé d'une variable d'intérêt correspond à celle du total estimé des résidus obtenus à partir de la régression de cette variable sur les variables de calage. Dans cette optique, les variables qualitatives sont mises sous forme numérique et exprimées sous forme d'indicatrices. Pour des variables de type ratio, c'est-à-dire définies comme le rapport de deux totaux estimés (ex: revenu moyen des salariés ou des indépendants), le traitement est un peu plus complexe. En effet, on démontre que $V\left(\frac{\hat{Y}_1}{\hat{Y}_2}\right) = V(\hat{Z})$ où $\hat{Z}$ est le total estimé de la variable linéarisée Z définie pour toute observation k par $\hat{z}_k = \frac{1}{\hat{Y}_2}\left(y_{1k} - \frac{\hat{Y}_1}{\hat{Y}_2} y_{2k}\right)$.

Comme nous l'avons signalé dans l'introduction, le principal intérêt des enquêtes par panel est d'obtenir une meilleure précision des évolutions. Cependant, l'échantillon des répondants est en évolution permanente. En effet, entre deux vagues, les répondants ne sont pas exactement les mêmes. D'une part, les individus panel de la première vague peuvent sortir du champ de l'enquête (différentes causes sont probables : décès ou éventuellement déménagement à l'étranger, entrée dans un établissement collectif, comme un établissement pénitentiaire) et d'autre part ils peuvent ne pas répondre à la seconde vague de l'enquête. Notons que la distinction entre les non-répondants et les hors-champs est primordiale puisque les non-répondants auraient dû répondre à l'enquête tandis que les hors-champ ne sont pas

concernés par l'enquête. De plus, le champ de l'enquête correspond à l'ensemble des personnes de plus de 18 ans. Ainsi, les individus qui atteignent leur majorité entre les deux vagues ne font pas partie du champ de l'enquête lors de la première vague mais entrent dans le champ de la seconde vague. Cependant, si l'on peut négliger dans un premier temps les personnes qui entrent et qui sortent du champ de l'enquête entre deux vagues, il n'en est pas de même pour les non-répondants de la seconde vague. On peut ainsi retenir qu'entre les deux vagues, le nombre d'individus qui sortent du champ de l'enquête est de l'ordre de 150, celui des nouveaux adultes de l'ordre de 250. En revanche, le nombre de non-répondants en vague 2 qui ont répondu en vague 1 est près de 10 fois plus important puisqu'il est d'environ 1500 (voir graphique ci-dessous).

***Schématisation de l'évolution entre les deux vagues***

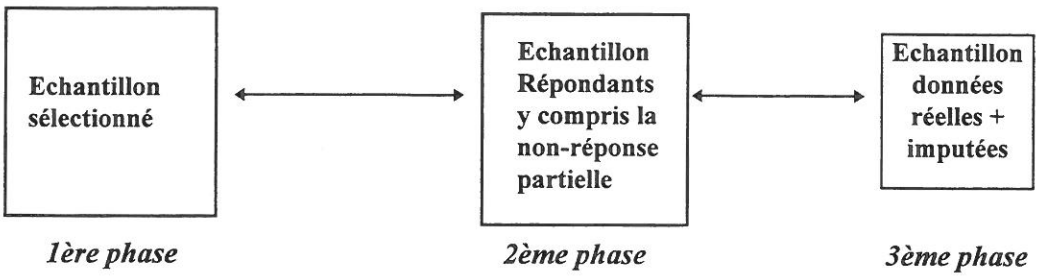



Pour les calculs de précision d'estimateurs d'évolutions dans ce document, la solution adoptée consiste à se restreindre aux répondants sur les deux vagues, c'est-à-dire aux individus pour lesquels on est capable d'évaluer la vraie évolution. Cette solution semble logique puisque pour les autres individus, c'est-à-dire ceux qui ont répondu en première vague mais pas en seconde, on ne possède aucune indication de l'évolution. Ainsi, le nombre d'individus adultes intervenant pour cette partie de l'étude est de 12986.

# III. Modélisation du plan de sondage et estimations de variance

## *III.1. Modélisation de l'enquête*

Pour tous les calculs de précision intervenant dans la suite, l'enquête Panel européen va être considérée comme une enquête en trois phases même si le plan de sondage est un plan de sondage en une seule phase. En effet, la correction de la non-réponse totale, effectuée par la méthode de correction par groupes homogènes, introduit une phase supplémentaire. De plus, la procédure d'imputation peut se modéliser comme une phase supplémentaire.
Détaillons successivement les trois phases.




✶ 1ère phase : la première phase consiste à sélectionner un échantillon $s_1$ dans l'ensemble de la base de sondage. Dans le cadre de l'enquête choisie, le plan de sondage est un plan stratifié à plusieurs degrés selon la taille des communes.

✶ 2ème phase : la seconde phase modélise la correction de la non réponse totale (r répondants parmi l'échantillon $s_1$) effectuée par la méthode des groupes homogènes.

Les probabilités d'inclusion d'ordre 1 des individus relatives à la première phase et à la seconde phase sont respectivement notées $\pi_i = P(\varepsilon_i = 1)$ et $p_i = P(r_i = 1 | s_1)$.

En présence d'une correction de la non-réponse totale, un estimateur sans biais du total $Y = \sum_{i=1}^{N} Y_i$ de la variable $Y$ est : $\hat{Y} = \sum_{i \in r} \frac{1}{p_i} \frac{Y_i}{\pi_i}$. ***Le poids d'échantillonnage des r individus répondants est augmenté.***

En pratique, la distribution de réponse est inconnue. Il faut alors postuler un modèle de comportement des individus face à la non-réponse et ensuite estimer les probabilités de réponse $p_i$.

**La validité de l'inférence, c'est-à-dire l'extrapolation des résultats obtenus sur l'échantillon des répondants à l'ensemble de la population, dépend de la validité du modèle choisi.**

✶ 3ème phase : la troisième phase modélise la méthode d'imputation utilisée.
L'estimateur de Y peut s'écrire sous la forme :

$$\hat{Y}^* = \sum_{i\in r} \frac{1}{p_i}\frac{1}{\pi_i}\left(Y_i\delta_i + (1-\delta_i)Y_i^*\right)$$

où $Y_i^*$ représente la valeur imputée si l'individu i n'a pas répondu à cette question et où $\delta_i = 1$ si l'individu a répondu et 0 si la réponse est imputée

### *III.2. Formule de variance*

Si on s'intéresse à la précision de l'estimateur d'un total obtenu à partir de données imputées et de données réelles, ces données étant issues d'une enquête modélisée comme une enquête en deux phases, nous obtenons[1] :

$$V\left(\hat{Y}^*\right) = V\left(\sum_{i\in r} \frac{1}{p_i}\frac{1}{\pi_i}\left(Y_i\delta_i + (1-\delta_i)Y_i^*\right)\right)$$
$$= V_1\left(E_{2|1}\left(\hat{Y}^*\right)\right) + E_1\left(V_{2|1}\left(\hat{Y}^*\right)\right)$$

avec $E_{2|1}\left(\hat{Y}^*\right) = E_{2|1}\left(E_{3|1,2}\left(\hat{Y}^*\right)\right)$ et $V_{2|1}\left(\hat{Y}^*\right) = V_{2|1}\left(E_{3|1,2}\left(\hat{Y}^*\right)\right) + E_{2|1}\left(V_{3|1,2}\left(\hat{Y}^*\right)\right)$

La variance se décompose en deux : la première partie estime la variance première phase et la seconde estime la variance seconde phase.

***Dans le cas où le mécanisme lié au processus d'imputation (conditionnellement à la sélection de l'échantillon initial et au nombre de répondants au sens de la non-réponse globale) est sans biais, nous obtenons :***

$$E_{3|1,2}\left(\hat{Y}^*\right) = \sum_{i\in r} \frac{Y_i}{p_i\pi_i} = \hat{Y}$$

$$V\left(\hat{Y}^*\right) = V\left(\sum_{i\in r} \frac{1}{p_i}\frac{1}{\pi_i}\left(Y_i\delta_i + (1-\delta_i)Y_i^*\right)\right)$$
$$= V_1\left(E_{2|1}\left(\hat{Y}\right)\right) + E_1\left(V_{2|1}\left(\hat{Y}\right) + E_{2|1}\left(V_{3|1,2}\left(\hat{Y}^*\right)\right)\right)$$

***Les formules de variance et d'estimation de variance se décomposent en deux parties :***

- ***la première regroupant les deux premiers termes correspond à la variance induite par le plan de sondage en deux phases en supposant que les données imputées sont des vraies données (c'est-à-dire qu'il n'y a pas de non-réponse partielle).***
- ***la seconde (= troisième terme) correspond à la variance induite par le mécanisme d'imputation ainsi que sa modélisation.***

[1] les indices 1, 2 et 3 correspondent respectivement au plan de sondage de la première phase, à la modélisation de la correction de la non-réponse globale et à la modélisation de la méthode d'imputation utilisée.

L'hypothèse utilisée dans la recherche de la formule d'estimation de variance concernant l'absence de biais du mécanisme d'imputation (conditionnellement au modèle d'imputation choisi) n'est pas restrictive. En effet, lorsqu'on choisit d'imputer des valeurs, on recherche une méthode d'imputation qui semble a priori sans biais ou en tout cas faiblement biaisée. La « qualité » d'une méthode d'imputation peut s'évaluer par simulation. Plusieurs façons de procéder sont envisageables. La première consiste, par exemple, à sélectionner à partir de la population des répondants un échantillon de non-répondants selon leur probabilité estimée de ne pas répondre à l'enquête, d'imputer une valeur aux non-répondants et de faire différentes comparaisons. Une autre méthode serait de relancer les non-répondants dans le but d'obtenir des réponses pour une fraction d'entre eux et de les comparer à la valeur simulée (Méthode de Hansen et Hurwitz).

La variance liée au plan de sondage en deux phases sera estimée avec le logiciel POULPE (partie IV). Ce logiciel permet d'évaluer la précision de statistiques issues d'enquêtes par sondage complexes, en particulier les enquêtes auprès des ménages réalisées par l'Insee (voir Caron N., 1999, Petit J.-N., 1999 et Caron, N., Deville, J.-C. et Sautory O., 1998). Celle induite par le mécanisme de non-réponse le sera par simulation. Nous comparerons les trois méthodes d'imputation évoquées dans l'introduction.

## IV. Estimation de la variance des deux premières phases

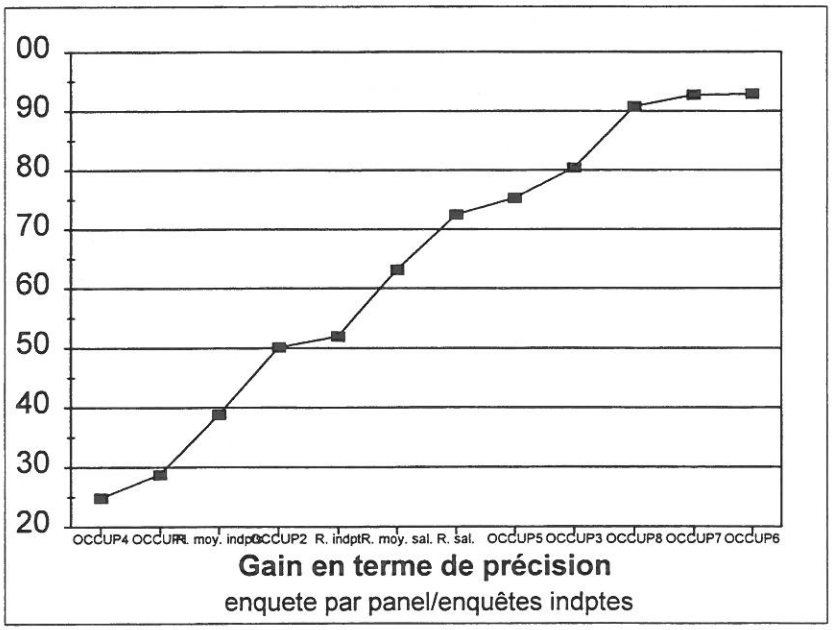
Le logiciel POULPE, écrit en langage macro SAS et développé par l'Unité Méthodes Statistique de l'Insee, va nous permettre d'estimer la variance de la première et de la seconde phases. Dans cette optique, nous considérons que les données imputées sont les vraies données.

Comme nous l'avons déjà précisé dans la partie précédente, la présence de la non-réponse est modélisée comme une phase de sondage supplémentaire. La modélisation choisie est celle d'un tirage stratifié où les probabilités de réponse dans chaque strate sont estimées à partir de l'enquête par le rapport du nombre de répondants à l'enquête et de la taille de l'échantillon.
Comme l'unité d'échantillonnage est le ménage, le principe général consiste à « remonter » sous la forme adéquate les variables du niveau individu au niveau ménage. Dans cette optique, pour chaque variable de type qualitative du niveau individus, on crée autant de variables quantitatives que de modalités de la variable qualitative. On somme ensuite pour chaque ménage les variables quantitatives intervenant au niveau individus ainsi que celles créées pour suppléer les variables qualitatives du niveau individus. Ces "nouvelles" variables qui se situent au niveau ménage apportent donc autant d'informations que celles dont elles sont issues et qui sont présentes au niveau individus.
Dans une enquête par panel, la variance de l'estimation d'une évolution d'un total entre deux dates s'obtient en construisant pour chaque individu une nouvelle variable qui correspond à la différence des valeurs de la variable obtenue à ces différentes dates. En revanche, dans le cas de deux enquêtes indépendantes, cette variance correspond simplement à la somme des variances pour chaque enquête. En principe, comme il existe une corrélation positive entre les variables observées sur le même échantillon à des dates différentes, la variance de l'estimation d'une évolution obtenue à partir d'une enquête par panel est plus faible que celle obtenue à partir de deux enquêtes indépendantes.
Le tableau suivant fournit les résultats pour le nombre de personnes exerçant une profession (occup1), le nombre de chômeurs (occup2), le nombre d'étudiants (occup3), le nombre de militaires du contingent (occup4), le nombre de retraités ou pré-retraités (occup5), le nombre de retirés d'affaires (occup6), le nombre de personnes au foyer (occup7), le nombre d'autres inactifs dont les personnes invalides (occup8), le revenu des indépendants, le revenu moyen des indépendants, le revenu des salariés et le revenu moyen des indépendants.

## *Estimation de la variance des deux premières phases*

| | Estimation 1ère vague | Estimation 2ème vague | Variance vague 1 notée V1 | Variance vague 2 notée V2 | Variance de l'évolution entre les deux vagues noté V12 | Gain[2] en % |
|---|---|---|---|---|---|---|
| **OCCUP1** | 22 282363 | 22 643958 | 377 $10^6$ | 10688 $10^6$ | 7883 $10^6$ | 28.75 |
| **OCCUP2** | 3 355979 | 3 329446 | 8789 $10^6$ | 8750 $10^6$ | 8750 $10^6$ | 50.11 |
| **OCCUP3** | 4 197406 | 4 095288 | 9964 $10^6$ | 11417 $10^6$ | 4181 $10^6$ | 80.44 |
| **OCCUP4** | 215786 | 190997 | 642 $10^6$ | 625 $10^6$ | 952 $10^6$ | 24.85 |
| **OCCUP5** | 245238 | 263990 | 753 $10^6$ | 784 $10^6$ | 379 $10^6$ | 75.33 |
| **OCCUP6** | 10 302686 | 10 621133 | 7530 $10^6$ | 8010 $10^6$ | 1099 $10^6$ | 92.92 |
| **OCCUP7** | 3 533731 | 3 467311 | 7849 $10^6$ | 7886 $10^6$ | 1140 $10^6$ | 92.76 |
| **OCCUP8** | 1 387441 | 1 472683 | 4283 $10^6$ | 4597 $10^6$ | 815 $10^6$ | 90.82 |
| | | | | | | |
| **revenus indpt** | 3.15 $10^{12}$ | 2.95 $10^{12}$ | 1,88 $10^{20}$ | 1,17 $10^{20}$ | 1,47 $10^{20}$ | 51.95 |
| **revenu moyen des indpts** | 128033 | 133816 | 33 $10^6$ | 58 $10^6$ | 56 $10^6$ | 38.86 |
| **revenus salariés** | 2.10 $10^{12}$ | 1.96 $10^{12}$ | 8,41 $10^{21}$ | 4,95 $10^{21}$ | 3,67 $10^{21}$ | 72.51 |
| **revenu moyen des salariés** | 98119 | 98223 | 1,42 $10^6$ | 0,98 $10^6$ | 0,88 $10^6$ | 63.19 |

Nous constatons que le gain en terme de variance lorsque les évolutions sont mesurées à partir d'une enquête par panel plutôt qu'avec deux enquêtes indépendantes varie d'une variable à une autre (voir graphique ci-dessous). Celui-ci est compris entre 28% pour l'estimateur de l'évolution du nombre de personnes exerçant une profession (variable OCCUP1) et 90% pour l'estimateur de l'évolution du nombre de personnes retirées des affaires ou au foyer (variables OCCUP6 et OCCUP7). Par construction de l'indicateur du gain, ce sont les variables qui sont fortement corrélées d'une vague à une autre pour lesquelles le gain est le plus important. Ainsi, par exemple, les retraités d'une année n donnée ont une probabilité très forte d'être encore retraités l'année suivante, ce qui explique que le gain soit maximum.




[2] le gain est mesuré par 1- (V12)/(V1+V2)

## V. Influence de la méthode d'imputation sur la variance

Dans cette partie, nous évaluons la part de la variance totale (pour chacune des deux vagues) qui provient du fait que certaines données ont été imputées ainsi que l'impact de la méthode d'imputation choisie. La variable utilisée est le revenu des indépendants pour lequel plus de 10% des valeurs sont imputées et le paramètre d'intérêt est le revenu moyen des indépendants. Cette variable est recueillie dans le questionnaire revenu (voir annexe) dans lequel on demande l'année n + 1 les revenus en clair de l'année n résultant d'une activité d'indépendant.

### *V.1. Quelques résultats statistiques sur les données imputées*

Dans l'enquête première vague, 884 indépendants ont été interrogés. Parmi eux, 178 n'ont pas répondu à la question sur les revenus (notons que ces individus sont des non-répondants partiels); par conséquent, 178 données ont été imputées[3] ce qui conduit à un taux d'imputation d'environ 20%. Pour l'enquête seconde vague réalisée en 1994, 164 indépendants n'ont pas souhaité répondre à l'ensemble du questionnaire (non-répondants globaux) et sur les 720 indépendants déclarés répondants à l'enquête, 86 n'ont pas répondu à la question sur les revenus. Le taux d'imputation de la seconde vague qui est de l'ordre de 12% est plus faible que celui de la première vague. La méthode d'imputation utilisée est celle des revenus simulés qui est détaillée dans la partie V.2.

***Quelques statistiques sur les revenus des indépendants avant et après imputation***

| | **nb d'obs** | **moyenne simple** | **écart-type** | **minimum** | **maximum** |
|---|---|---|---|---|---|
| Rev. indpts déclarés v1 | 706 | 133472 | 162287 | 700 | 1 728000 |
| Rev. indpts déclarés et imputés v1 | 884 | 131229 | 158414 | 700 | 1 728000 |
| | | | | | |
| Rev. indpts déclarés v2 | 634 | 127360 | 134076 | 700 | 1 000000 |
| Rev. indpts déclarés et imputés v2 | 720 | 135085 | 206257 | 700 | ***3 311814*** |

A partir des données de la première vague, nous constatons que les revenus moyens avant et après imputation sont du même ordre de grandeur. Contrairement aux méthodes d'imputation de type déterministe (voir Caron, 1996), la méthode choisie présente l'avantage de ne pas déformer artificiellement la distribution, c'est-à-dire que les variances empiriques modifiées calculées à partir de l'ensemble des données (les données réelles et les données imputées), sont proches de celles calculées uniquement à partir des répondants. En revanche, les statistiques sur les revenus nous apprennent que le revenu moyen ainsi que son écart-type sont plus importants avec l'ensemble des données complétées qu'avec les répondants. De plus, la

3 la méthode d'imputation utilisée sera précisée ultérieurement

valeur maximale atteinte pour les revenus imputés est trois fois plus importante que celle obtenue sur l'ensemble des répondants.

### *V.2. Description des trois méthodes d'imputation étudiées par la suite.*

Les méthodes d'imputation consistent à remplacer la donnée absente ou invalide par une donnée « plausible » qui est en général issue ou estimée à partir de la distribution des répondants. Par nature, l'imputation est une opération délicate et ne peut être réalisée que si la non-réponse n'est pas trop importante dans le fichier. Les méthodes d'imputation sont attractives car lorsque les données manquantes sont imputées, on dispose d'un ensemble de données complet. Cependant, la présence de données imputées n'est pas sans conséquence sur la variance des estimateurs. En effet, le fait d'avoir des données imputées dans le fichier d'exploitation des données augmente la variance qu'on aurait obtenue si les données étaient considérées comme réelles, c'est-à-dire s'il n'y avait pas eu de non-réponse. Dans cette partie, nous estimons la part de variance due à la correction de la non réponse partielle ainsi que l'influence de la méthode d'imputation choisie sur la précision.
Trois méthodes d'imputation sont comparées par la suite : une méthode hot-deck et deux méthodes économétriques de type résidus simulés. Décrivons successivement ces différentes méthodes.

- ***Les méthodes économétriques***

La méthode utilisée pour imputer les revenus dans l'enquête « panel européen » est celle des résidus simulés. Cette méthode des résidus simulés appliquée à une variable $Y$ (ici le revenu) qui suit approximativement une loi lognormale, consiste à régresser le logarithme de cette variable en fonction d'un ensemble de variables auxiliaires $X$ selon le modèle économétrique $\mathrm{Log}(Y) = X\beta + \varepsilon$ et à remplacer la valeur manquante par $Y^* = \exp\left(X\hat{\beta} + \hat{u}\right)$ où $\hat{u}$ est un résidu obtenu dans la distribution des résidus estimés à partir du modèle sur les répondants. L'ajout du terme contenant le résidu au prédicteur permet d'imputer des valeurs sans biais par rapport au modèle d'imputation. En effet, par construction, en notant $Y$ une variable suivant une loi lognormale. nous avons :

$$\mathrm{Log}(Y) \xrightarrow{\text{loi}} N\left(m, \sigma^2\right)$$

L'espérance de cette variable est $E(Y) = \exp\left(m + \left(\sigma^2 / 2\right)\right)$ ce qui implique que Y n'est pas un estimateur sans biais de $\exp(m)$.
Les résidus intervenant dans les données imputées peuvent être soit générés de façon aléatoire à partir d'une distribution normale d'espérance nulle et de variance celle des résidus obtenus à partir du modèle sur les répondants (méthode 1), soit provenir directement des résidus associés aux répondants (méthode 2). Ces deux variantes seront envisagées par la suite.

En ce qui concerne la **variable « revenu des indépendants »** de la vague 1, les variables explicatives retenues dans le modèle de régression sont :

âge (en 2 postes)
présence d'un aide familial
catégorie socio-professionnelle (en 6 postes)
diplôme (en 6 postes)

durée hebdomadaire du travail (en 3 postes)
patrimoine (en 7 postes)
sexe de la personne

La qualité de la régression peut être mesurée par le coefficient de corrélation linéaire $R^2 = 0.36$, ce qui correspond à une qualité de régression correcte étant donnée la nature des données.

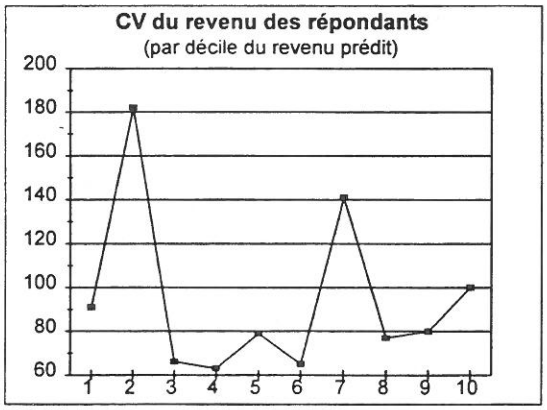
Comme la méthode économétrique fait par construction intervenir dans la valeur imputée un résidu obtenu à partir de la distribution des résidus estimés sur les répondants, on peut se demander quel est l'impact du choix de $\hat{u}$ sur l'imputation, c'est-à-dire quelle est la modification induite par rapport à la valeur du prédicteur. Pour cela, nous avons comparé la valeur imputée $y^* = \exp\left(X\hat{\beta} + \hat{u}\right)$ et la valeur du revenu prédit par le modèle $y_{pred} = \exp\left(X\hat{\beta}\right)$. La différence entre les deux valeurs est comprise entre -442000 et 772000 pour la vague 1 et entre -287338 et 3062084 pour la vague 2.

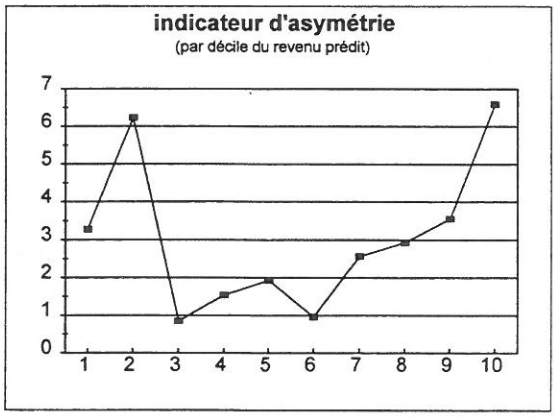
Sur la population des ***répondants de la première vague*** (693 individus), nous obtenons que les résidus estimés sont compris entre -4.07 et 3.32, leur moyenne est nulle et l'écart-type $\sigma$ est 0.8225. Pour les données ***imputées*** (195 individus), les valeurs des résidus utilisés n'ont pas été conservées. Cependant, nous pouvons les déduire en comparant la valeur imputée et la valeur prédite. Ainsi, les résidus utilisés sont compris entre -3.53 et 1.95 et leur moyenne vaut -0.42. Cette moyenne semble anormalement basse puisqu'elle n'appartient pas à l'intervalle de confiance que l'on peut construire à partir de l'ensemble des résidus soit $[-0.11; 0.11]$ qui correspond à $\left[-2\,\sigma/\sqrt{n}; 2\,\sigma/\sqrt{n}\right]$. Ainsi, on peut se demander si la liste de variables explicatives intervenant dans le modèle économétrique utilisé dans cette étude est complète ou si le modèle économétrique est estimable par la méthode des moindres carrés ordinaires.

En ce qui concerne la **variable « revenu des indépendants »** de la vague 2, les variables explicatives retenues dans le modèle de régression sont celles de la vague 1 auxquelles on ajoute deux variables recueillies lors de la vague précédente permettant d'améliorer le modèle de régression. Ces nouvelles variables exprimées dans un système de tranches sont le revenu de l'année précédente (en 7 postes) et le patrimoine total du ménage (en 8 postes). La qualité de la régression est effectivement nettement améliorée puisque le coefficient de corrélation linéaire atteint $R^2 = 0.54$. Ainsi, les variables introduites dans le modèle expliquent environ 54% de la variance. Le modèle ajusté se révèle plus explicatif que celui estimé en vague 1. Cependant, le même problème se pose. En effet, alors que sur la population des répondants (634 répondants) l'écart-type des résidus estimés vaut $\sigma = 0.70$, la moyenne des résidus estimés sur la population des non-répondants (***86*** non répondants) d'une valeur de -0.94 n'appartient pas à l'intervalle de confiance $[-0.15; 0.15]$.

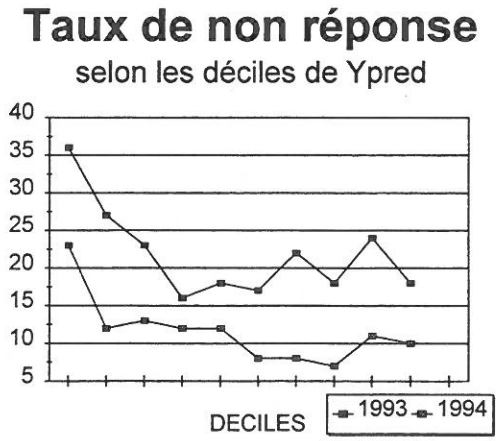
L'étude de la distribution du revenu selon les déciles du revenu prédit nous apprend (voir graphiques ci-dessous) que :

- les distributions sont très concentrées pour 6 déciles. Ce sont les déciles extrêmes qui sont les plus instables, c'est-à-dire qui possèdent un coefficient de variation élevé.
- les distributions sont très asymétriques pour les déciles extrêmes. Pour chacun de ces déciles, la queue de distribution est plus importante. Cependant, la distribution est plutôt asymétrique vers la droite pour les premiers déciles et plutôt vers la gauche pour les derniers déciles.



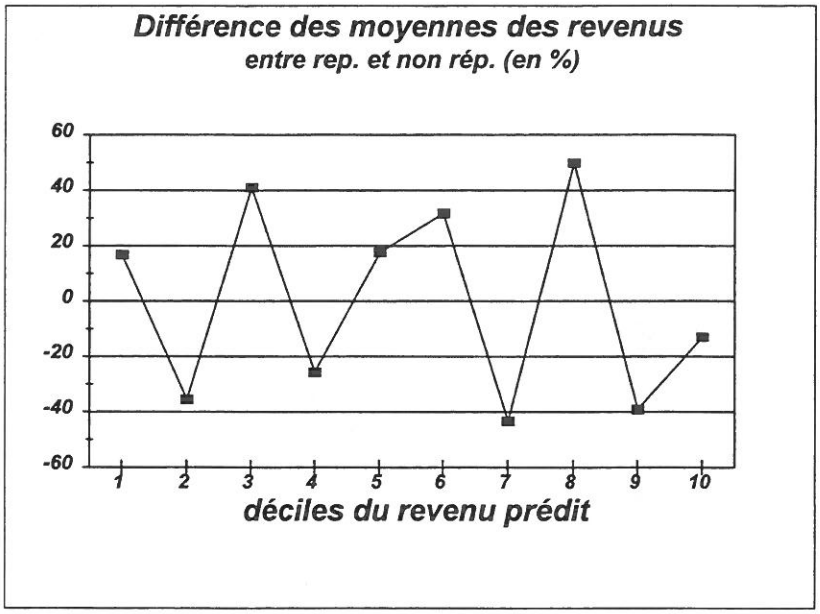

***Etude des taux de non-réponse selon les déciles du revenu prédit :***




- Le taux de non-réponse est très important dans les deux premiers déciles. Il atteint près du double des taux obtenus dans les autres déciles. La différence est plus marquée pour les données de la première vague.

- Dans les autres déciles, le taux de non-réponse est généralement stable et est de l'ordre de 20% pour la première vague et 10% pour la seconde.

Le graphique suivant représente selon les déciles du revenu prédit la différence entre la moyenne des revenus imputés et celle des répondants (en % de la moyenne des revenus des répondants). Ainsi, nous constatons que ce n'est pas dans les déciles du revenu prédit où le taux de non-réponse est le plus important (les deux premiers déciles) que la différence est la plus grande.




- ***Méthode de Hot deck***

L'imputation par la méthode du hot-deck consiste à remplacer la donnée manquante par la valeur observée pour un individu répondant choisi « au hasard ». Les valeurs choisies peuvent par exemple être sélectionnées de manière séquentielle : l'échantillon est classé dans un certain ordre et pour chaque valeur manquante la valeur du répondant qui la précède dans le fichier est imputée. Notons que si la première valeur est manquante, il faudra initialiser le processus par une première valeur a priori (que l'on peut par exemple choisir en considérant le fichier comme « circulaire »). De plus, le critère de classement est important. En effet, si un individu répondant est suivi de nombreux individus non-répondants, la même valeur est imputée plusieurs fois. Le critère de classement doit donc être le moins possible corrélé avec la probabilité de réponse tout en étant fortement lié avec la variable d'intérêt. Dans notre exemple, le critère servant au classement des individus est celui des déciles des revenus prédits.

Comme pour les deux méthodes présentées précédemment, la méthode du hot-deck permet de ne pas déformer la distribution des données complètes (données réelles et données imputées). Ainsi, la variance empirique calculée sur l'ensemble des données est proche de celle calculée à partir des répondants. Notons que cette propriété n'est pas vérifiée par l'ensemble des méthodes d'imputation (comme la méthode d'imputation par la moyenne qui consiste à imputer la moyenne obtenue sur les répondants).

### *V.3. Simulations*

L'objectif de cette partie est, dans un premier temps, d'évaluer par simulation de Monte Carlo la variance induite par le mécanisme d'imputation puis d'obtenir une estimation de la part de cette variance dans la variance globale. Nous comparons trois méthodes d'imputation :

- méthode de Hot deck séquentiel par classes (celles-ci correspondent aux déciles des revenus prédits)
- ***1ère méthode économétrique :***
  La valeur imputée est de la forme $y_i^* = \hat{y}_i + \hat{u}_i$ où $\hat{y}_i$ est la valeur du logarithme du revenu prédit par le modèle et $\hat{u}_i$ est un résidu choisi par sondage aléatoire simple parmi l'ensemble des vrais résidus pris par les répondants.
- ***2ème méthode économétrique*** :
  La valeur imputée est de la même forme que pour la méthode précédente soit $y_i^* = \hat{y}_i + \hat{u}_i$. Cependant, pour cette méthode, le résidu $\hat{u}_i$ est issu de la loi normale $N\left(0, \hat{\sigma}_u^2\right)$ où $\hat{\sigma}_u^2$ la variance empirique des résidus pris par les répondants.

Pour chaque méthode d'imputation, nous avons adopté la démarche suivante :

1. on considère l'ensemble des indépendants
2. on impute pour chaque donnée manquante une valeur
3. on calcule la moyenne des revenus obtenus sur le fichier complété $\bar{y}_\ell$
4. on répète cette démarche 100 fois.
5. on calcule $\bar{y}_{100} = \frac{1}{100}\sum_{\ell=1}^{100} \bar{y}_\ell$ et $V_{100} = \frac{1}{99}\sum_{\ell=1}^{100}\left(\bar{y}_\ell - \bar{y}_{100}\right)^2$

Cette démarche est directement inspirée de la méthode d'imputation multiple proposée par Rubin (1987) qui consiste à remplacer chaque donnée manquante par un vecteur de valeurs possibles et à construire un estimateur qui est fonction de l'ensemble des estimateurs obtenus avec chaque jeu de valeurs imputées. De façon plus précise, en notant $\hat{\theta}_\ell$ et $U_\ell$ les M estimations de données complètes et les variances correspondantes pour un paramètre $\theta$, ces valeurs étant calculées à l'aide des M ensembles de données complétées, l'estimateur de $\theta$ est :

$$\bar{\theta}_M = \frac{\sum \hat{\theta}_\ell}{M}$$

et sa variance est : $V_M = \frac{\sum U_\ell}{M} + \left(1 + \frac{1}{M}\right)\frac{\sum\left(\hat{\theta}_\ell - \bar{\theta}_M\right)^2}{M-1}$.

La variance a deux composantes : la première correspond à la variance intra-imputation, la seconde à la variance inter-imputation.

Lorsque M est suffisamment grand, la variance peut être approximée par la variance inter-imputation soit par :

$$V_M \approx \frac{\sum\left(\hat{\theta}_\ell - \bar{\theta}_M\right)^2}{M-1}$$

Les différentes estimations de variance obtenues selon la méthode d'imputation sont détaillées dans le tableau ci-dessous. Nous constatons que la méthode Hot-deck conduit à une estimation de l'écart-type plus faible que les deux méthodes économétriques. Elle permet de « gagner » plus de 10% sur l'écart-type. Il semble aussi, bien que l'écart soit moins important pour les données de la vague 2, que la méthode économétrique basée sur les résidus obtenus sur l'ensemble des répondants soit plus précise que celle utilisant une distribution normale estimée à partir des résidus des répondants.

***Récapitulation des estimations d'écart-type obtenues par méthodes Monte Carlo***

| | *Hot Deck* | *Econométrie avec vrais résidus* | *Econométrie avec résidus aléatoires* |
|---|---|---|---|
| *Vague 1* | 2396 | 2649 | 3178 |
| *Vague2* | 1630 | 1889 | 1992 |

D'après les formules d'estimation de variance développées dans la partie 3 et les résultats de la partie précédente en ce qui concerne les estimations de variance des deux premières phases, nous en déduisons l'écart-type d'échantillonnage global.

***Ecart-type d'échantillonnage global***

| | *Hot Deck* | *Econométrie avec vrais résidus* | *Econométrie avec résidus aléatoires)* |
|---|---|---|---|
| *Vague 1* | 6290 | 6391 | 6627 |
| *Vague 2* | 8059 | 8147 | 8184 |

Pour la première vague, nous constatons ainsi que près de **15% de la variance** provient du **mécanisme d'imputation,** soit près de **40% de l'écart-type global**. Autrement dit, ne pas prendre en compte le mécanisme d'imputation pour les calculs de précision **revient à sous-estimer** fortement la variance et par conséquent la longueur des intervalles de confiance. Ainsi, l'interprétation des résultats risque d'être faussée.

En ce qui concerne la seconde vague, les résultats sont un peu moins marqués (10% de la variance finale et 32% de l'écart-type final proviennent du mécanisme d'imputation).
En ce qui concerne l'impact de la méthode d'imputation sur l'écart-type, l'effet n'est pas très important. En effet, choisir la méthode d'imputation avec la méthode de Hot-deck ne permet de diminuer l'écart-type global que de 5% par rapport aux autres méthodes économétriques étudiées.

## VI. Conclusion

L'étude présentée a été conduite à partir de l'enquête Panel Européen sur la variable revenu d'une activité indépendante pour laquelle plus de 10% des variables ont été imputées. L'objectif principal était d'évaluer la part de la variance totale due à la méthode d'imputation. Trois méthodes d'imputation sont envisagées : une méthode de type hot-deck (hot-deck séquentiel) et deux méthodes de type économétrique (méthodes des résidus simulés).

Pour la variable étudiée, nous constatons que près de **15% de la variance globale** provient du **mécanisme d'imputation,** soit près de **40% de l'écart-type global**. Autrement dit, ne pas prendre en compte le mécanisme d'imputation pour les calculs de précision risque de fausser l'interprétation des résultats puisqu'on sous-estime fortement la variance. En revanche, l'impact de la méthode d'imputation choisie sur l'écart-type global n'est pas très important. La méthode d'imputation de type Hot-deck conduit à un écart-type global inférieur de 5% par rapport aux deux méthodes économétriques.

Les résultats obtenus sont fragiles puisqu'ils n'ont été établis qu'à partir d'une seule variable. Il serait souhaitable de poursuivre cette étude sur plusieurs variables de cette enquête et sur d'autres enquêtes. L'intérêt est de choisir des variables pour lesquelles le taux de données imputées est important.

## Bibliographie

Caron, N., Deville, J.-C. et Sautory, O. (1998) : "Estimation de précision de données issues d'enquêtes : document méthodologique sur le logiciel Poulpe", document de travail n°9806 de la série Méthodologie Statistique, Insee.

Caron, N. (1996) : "Les principales techniques de correction de la non-réponse", document de travail n°9604 de la série Méthodologie Statistique, Insee.

Caron, N. (1999) : "Le logiciel Poulpe : aspects méthodologiques", Actes des journées de méthodologie statistique, *Insee Méthodes* n°84-85-86.

Rubin, D.B.(1976) : "Inference and missing data", *Biometrika*, 61, pages 581-592

Petit, J.-N. (1999) : "Le logiciel Poulpe : modélisation informatique", Actes des journées de méthodologie statistique, *Insee Méthodes* n°84-85-86.

## Annexe : extrait de la fiche Revenu de l'enquête Panel Européen.

Qui a répondu au questionnaire ?

Prénom :

N° d'ordre ..... NOIRFR

## EN 1996

### 10 - REVENUS DU TRAVAIL

Pour l'activité principale

| | | Montant total annuel (en francs) | Code du revenu dans lequel ce montant est inclus | Nombre de mois couverts par la prestation | Si moins de 12 mois : mois couverts par la prestation (entourer les cases correspondantes) |
|---|---|---|---|---|---|
| • *Revenus d'activité salariée* | | M10RTA | C10RTA | NM10RTA | AN10RTA |
| | | | | | J F M A M J J A S O N D |
| - *Salaires* | 11 | | | | 1 2 3 4 5 6 7 8 9 10 11 12 |
| - *Commissions, pourboires* | 12 | | | | |
| - *Heures supplémentaires* | 13 | | | | |
| - *Mois supplémentaires* | 14 | | | | |
| - *Autres primes* | 15 | | | | |
| - *Congés payés* | 16 | | | | |
| - *Participation sous forme d'actions* | 17 | | | | |
| - *Autres formes de participation, d'intéressement* | 18 | | | | |
| - *Dividendes perçus par les dirigeants salariés de leur entreprise* | 19 | | | | |

• ***Revenus résultant d'une activité d'indépendant***

| | | | | | |
|---|---|---|---|---|---|
| - *Revenus agricoles, BIC, BNC, Rémunération des gérants associés* | 20 | | | | 1 2 3 4 5 6 7 8 9 10 11 12 |

Pour les activités secondaires, épisodiques

| | | | | | |
|---|---|---|---|---|---|
| - *Revenus découlant de ces activités* | 21 | | | | 1 2 3 4 5 6 7 8 9 10 11 12 |

### 30 - CHÔMAGE ET PERTE D'EMPLOI

| | | M30CHA | C30CHA | NM30CHA | AN30CHA |
|---|---|---|---|---|---|
| | | | | | J F M A M J J A S O N D |
| - *Indemnités de licenciement, primes de départ* | 31 | | | | |
| - *Allocations chômage* | 32 | | | | 1 2 3 4 5 6 7 8 9 10 11 12 |
| - *Autres, préciser :* | 33 | | | | 1 2 3 4 5 6 7 8 9 10 11 12 |

## Série des Documents de Travail 'Méthodologie Statistique'

**9601** : 'Une méthode synthétique, robuste et efficace pour réaliser des estimations locales de population'
**G. DECAUDIN, J.-C. LABAT**

**9602** : 'Estimation de la précision d'un solde dans les enquêtes de conjoncture auprès des entreprises'
**N. CARON, P. RAVALET, O. SAUTORY**

**9603** : 'La procédure FREQ de SAS® - Tests d'indépendance et mesures d'association dans un tableau de contingence'
**J. CONFAIS, Y. GRELET, M. LE GUEN**

**9604** : 'Les principales techniques de correction de la non-réponse et les modèles associés'
**N. CARON**

**9605** : 'L'estimation du taux d'évolution des dépenses d'équipement dans l'enquête de conjoncture : analyse et voies d'amélioration'
**P. RAVALET**

**9606** : 'L'économétrie et l'étude des comportements. Présentation et mise en œuvre de modèles de régression qualitatifs. Les modèles univariés à résidus logistiques ou normaux (LOGIT, PROBIT)'
**S. LOLLIVIER, M. MARPSAT, D. VERGER**

**9607** : 'Enquêtes régionales sur les déplacements des ménages : l'expérience de Rhône-Alpes'
**N. CARON, D. LE BLANC**

**9701** : 'Une bonne petite enquête vaut-elle mieux qu'un mauvais recensement ?'
**J.C. DEVILLE**

**9702 :** 'Modèles univariés et modèles de durée sur données individuelles'
**S. LOLLIVIER**

**9703 :** 'Comparaison de deux estimateurs par le ratio stratifiés et application aux enquêtes auprès des entreprises'
**N. CARON, J.C. DEVILLE**

**9704 :** 'La faisabilité d'une enquête auprès des ménages
1. au mois d'août. 2. à un rythme hebdomadaire'
**C. LAGARENNE, C. THIESSET**

**9705 :** 'Méthodologie de l'enquête sur les déplacements dans l'agglomération toulousaine'
**P. GIRARD**

**9801 :** 'Les logiciels de désaisonnalisation TRAMO & SEATS : philosophie, principes et mise en œuvre sous SAS'
**K. ATTAL-TOUBERT, D. LADIRAY**

**9802 :** 'Estimation de variance pour des statistiques complexes : technique des résidus et de linéarisation'
**J.C. DEVILLE**

**9803 :** 'Pour essayer d'en finir avec l'individu Kish'
**J.C. DEVILLE**

**9804 :** 'Une nouvelle (encore une !) méthode de tirage à probabilités inégales'
**J.C. DEVILLE**

**9805 :** 'Variance et estimation de variance en cas d'erreurs de mesure non corrélées ou de l'intrusion d'un individu Kish'
**J.C. DEVILLE**

**9806 :** 'Estimation de précision de données issues d'enquêtes : document méthodologique sur le logiciel POULPE'
**N. CARON, J.C. DEVILLE, O. SAUTORY**

**9807 :** 'Estimation de données régionales à l'aide de techniques d'analyse multidimensionnelle'
**K. ATTAL-TOUBERT, O. SAUTORY**

**9808 :** 'Matrices de mobilité et calcul de la précision associée'
**N. CARON, C. CHAMBAZ**

**9809 :** 'Echantillonnage et stratification : une étude empirique des gains de précision'
**J. LE GUENNEC**

**9810** : 'Le Kish : les problèmes de réalisation du tirage et de son extrapolation'
**C. BERTHIER, N. CARON, B. NÉROS**

**9811** : 'Vocabulaire statistique Français - Chinois - Anglais'
**LIU Xiaoyue, CUI Bin**

**9901** : 'Perte de précision liée au tirage d'un ou plusieurs individus Kish'
**N. CARON**

**9902** : 'Estimation de variance en présence de données imputées : un exemple à partir de l'enquête Panel Européen'
**N. CARON**