Direction des Statistiques d'Entreprises

# E 2010/01

## Comment redresser une enquête thématique ?

Béatrice NEITER et Benoît BUISSON

Document de travail

Institut National de la Statistique et des Études Économiques

Institut National de la Statistique et des Études Économiques

*Série des documents de travail*
*de la Direction des Statistiques d'Entreprises*

**E 2010/01**

# Comment redresser une enquête thématique ?

*Béatrice NEITER et Benoît BUISSON*
*Méthodologues*
*Pôle Ingénierie Statistique des Entreprises*

*Mars 2010*

# Sommaire

# Résumé

L'objectif de ce document est de décrire la démarche pour redresser une enquête - c'est-à-dire corriger la non-réponse et effectuer un calage. A vocation pédagogique, il reprend les différentes étapes de traitement post-collecte d'une enquête, de la phase d'apurement à la livraison d'un fichier définitif de données. Les travaux de mise en place, la correction de la non-réponse, le calage et la livraison du fichier sont ainsi détaillés.

La non-réponse peut être partielle ou totale, selon respectivement que l'unité interrogée ait répondu à l'enquête tout en omettant d'en renseigner quelques parties, ou qu'elle n'ait pas restitué son questionnaire. Les méthodes de correction de la non-réponse décrites dans ce document de travail sont l'imputation (pour la correction de la non-réponse partielle) et la repondération (pour la correction de la non-réponse totale). Elles peuvent être traitées sans ordre prédéfini.

La non-réponse génère un biais (biais d'autant plus fort que le taux de non-réponse est élevé). L'accent est mis sur le fait que l'absence de traitement est préjudiciable à la qualité de l'enquête, et qu'il est donc indispensable de redresser une enquête. En procédant ainsi, on cherche en effet à réduire le biais.

Sans que les notions théoriques soient occultées, il a été apporté un soin particulier dans ce document, à la mise en œuvre pratique des opérations relevant du redressement de la non-réponse.

***Mots clefs :*** *non-réponse, imputation, repondération, typage, calage, groupe de réponse homogène (GRH), bilan qualité, documentation, tests statistiques.*

# Introduction

Quel que soit le type d'enquête et la méthode de sondage utilisée, le statisticien est confronté au problème de la non-réponse. On différencie deux catégories de non-réponse :

- la non-réponse partielle, lorsque l'unité enquêtée répond au questionnaire, en omettant toutefois de renseigner certaines parties dans une proportion plus ou moins importante.
- la non-réponse totale, quand le questionnaire d'une observation échantillonnée n'est pas restitué.

En parallèle de ces deux situations, on notera qu'il existe aussi des individus passés en hors-champ au moment de l'enquête.

Il existe différentes méthodes de correction de la non-réponse (totale ou partielle). Le présent document détaille les étapes permettant de redresser

- d'une part la non-réponse partielle par imputation ;
- d'autre part la non-réponse totale par repondération.

L'ordre de correction n'importe pas : le statisticien peut s'atteler indifféremment à corriger la non-réponse tout d'abord totale, puis partielle. Il s'agit des méthodes utilisées dans le cadre du traitement des enquêtes thématiques entreprises.

Repondérer, c'est compenser la non-réponse totale en augmentant le poids d'échantillonnage des répondants, afin de tenir compte des non-répondants. Imputer, c'est attribuer des valeurs/modalités « plausibles » aux réponses manquantes.
Quelle que soit la méthode retenue, elle a pour objectif de réduire le biais généré par la non-réponse. Par ailleurs, corriger la non-réponse est primordial du fait notamment que les unités non-répondantes présentent des profils différents des unités répondantes.
En outre, les résultats obtenus sont d'autant plus biaisés en présence de non-réponse que le nombre de non-répondants à une question est relativement élevé vis-à-vis du nombre de répondants. La présence de données manquantes influe donc sur la qualité de l'inférence statistique.
Il est donc primordial de corriger la non-réponse, car ne rien faire est nettement plus préjudiciable que d'imputer une estimation. Le choix du modèle d'imputation sera bien sûr déterminant.

Le redressement d'une enquête ne se borne toutefois pas à ces étapes. En effet, ces travaux nécessitent également des opérations de mises en place préalables, souvent sous-estimées mais consommatrices de temps et capitales pour le redressement. De même, après les deux phases décrites ci-dessus, il faut procéder à un calage. Enfin, il y a l'étape de la livraison des données documentées.
Ce document reprend toutes ces étapes, en rappelant les points méthodologiques incontournables et nécessaires à leur bonne compréhension. L'accent est toutefois plus mis sur l'aspect pratique du traitement des données post-collecte. Le lecteur trouvera ainsi au gré des pages, des exemples concrets relatifs aux cas précis d'enquêtes thématiques entreprises, des programmes et sorties SAS commentés, des conseils pratiques, ainsi que des mises en garde envers les pièges à éviter. Ce document a vocation à s'adresser donc au plus grand nombre, concepteurs d'enquête et utilisateurs en premier lieu. Son objectif est de permettre à chacun de mener une enquête à terme, sans pour autant maîtriser toute la théorie statistique.

Le présent document couvre donc les opérations de post-collecte se déroulant entre la phase d'apurement et la livraison d'un fichier redressé.

***Bibliographie***

**Ardilly P.** , 2006, *Les techniques de sondage*, Dunod.

**Brion Ph., Caron N., Piétri-Bessy P.**, « Redresser la non-réponse totale dans les enquêtes auprès des entreprises : les pièges à éviter. Illustration avec l'enquête innovation », Insee.

**Caron N.** , 2005, « La correction de la non-réponse par repondération et par imputation », Insee document de travail M0502, Paris.

**Caron N. et Brilhault G.** , 2004, « Correction de la non-réponse totale : par imputation ou par repondération », Insee document de travail E2004/01

**Caron N.,** 2001, « Présentation des méthodes de calage sur marges et de l'utilisation de la Macro Sas CALMAR », INSEE, Atelier méthodes du 29 décembre 2001.

**Haziza D., 2005**, « Traitement de la non-réponse dans les enquêtes », ENSAI polycopié de la formation continue diplômante des attachés

**Sautory O. et Le Guennec J., 2005**, « La macro Calmar 2 : redressement d'un échantillon par calage sur marges », INSEE

**Tillé Y.** , 2001, *Théorie des sondages*, Dunod.

**Méthodes et pratiques d'enquêtes**, numéro 12-587-XPF, Statistique Canada, notamment le chapitre 10 « traitement »

***Nota :***
*Le site internet du Genes (http://genes.bibli.fr/opac/index.php?lvl=index) met à disposition de l'internaute de nombreux documents en lien avec les thèmes abordés ici, et on y trouvera notamment les travaux de Caron N. cités ci-dessus.*

# I. Étapes préalables au redressement-calage

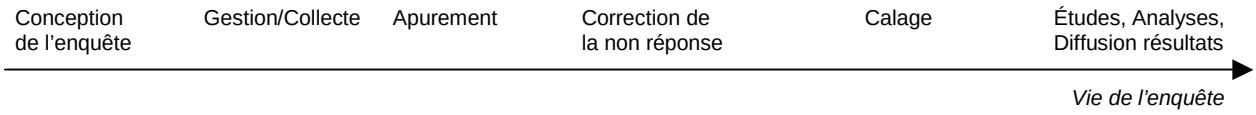
***La phase de redressement-calage est un travail d'envergure, nécessitant une mise en place en amont. Cette étape se place entre la phase de collecte/apurement des données et la phase d'analyse/diffusion des résultats. Il s'agit donc d'une pièce maîtresse indispensable à la constitution d'une base exploitable.***
***Avant de la mettre en œuvre, il faut bien respecter certaines étapes qui vont notamment servir à bien définir les populations suivantes : unités répondantes, unités non-répondantes, unités hors-champ.***

## *1. Le fichier à redresser*

### a. Issu de la phase d'apurement[1]

Dans le déroulement d'une enquête, il s'agit là de l'étape intervenant juste avant celle du redressement de la non réponse. Elle s'effectue après la phase de collecte de l'enquête.

Grossièrement, voilà les différentes étapes entre le moment où une enquête est réfléchie, et celui où ses résultats sont mis à disposition du public :




L'objectif de l'apurement est de permettre la réduction au maximum du nombre d'incohérences dans les questionnaires des unités enquêtées répondantes.

*Exemple issues de l'enquête sur les TIC :*
*Si une entreprise renseigne des montants d'achats réalisés par internet, alors qu'elle a déclaré ne pas posséder d'accès à internet : l'apurement peut forcer la seconde réponse à « oui, l'entreprise possède internet ».*

Pour ce qui est du premier objectif évoqué ci-dessus, une étape de l'apurement consiste par exemple à considérer les réponses dans leur ensemble, afin d'en repérer les valeurs atypiques ou extrêmes. Ceci permet parfois de cerner une incohérence.

*Exemple issu de l'enquête sur les TIC :*
*À l'apurement, il s'est avéré que certaines entreprises répondantes avaient fait une erreur d'unités dans la déclaration de leurs montants (valeurs demandées en k€ et ayant été renseignées en € quelques fois). L'apurement a converti les réponses dans la grandeur monétaire adéquate.*

Pour simplifier : l'apurement ne corrige les incohérences que sur les renseignements fournis par les enquêtés répondants. Cette phase ne corrige pas la non-réponse partielle.
Il est à noter aussi que lors de l'apurement, une réponse très atypique peut être remplacée par de la non réponse partielle.

**Du bon usage d'une indicatrice…**
Il s'agit d'une variable valant 1 si l'unité enquêtée a bénéficié d'un traitement spécifique, et 0 sinon. Ce marquage permet à l'utilisateur de la table de données finale, de « remonter » aux réponses initiales. Elle permet de distinguer ce qui a été fourni par l'unité enquêtée, et ce qui est du ressort des travaux statistiques.
*Par exemple : Une unité peut avoir déclaré ne pas posséder d'ordinateur. Elle renseigne ensuite des montants de ventes effectuées via internet. Si l'étude de cas amène à forcer la première réponse à « oui (possède un ordinateur )», alors on pourra créer une indicatrice selon l'exemple: « corr_ordi = 1 ».*

Chacun des ajustements effectués lors de l'apurement, doit être repérable grâce à une indicatrice.

[1] La phase d'apurement est abordée de façon succincte et très superficielle ici. Il est vivement conseillé au lecteur qui souhaiterait en savoir davantage sur ce sujet, de consulter des documents plus spécifiques à cette étape. A titre d'exemple, il pourra se référer à la note d'apurement de l'enquête TIC 2009 disponible sur le site intranet du pôle ISE.

A la fin de cette phase, le fichier de l'enquête ne contient plus d'incohérences manifestes. Par contre ce même fichier comprend toujours de la non-réponse partielle et de la non-réponse totale.

## b. Les types de variables

Dans une enquête, on distingue différents types de variables :

- **Les variables de lancement**

Elles concernent toutes les données dont on dispose AVANT la collecte. Ces variables ne donnent pas lieu à un redressement.

*Exemples : Coordonnées de l'unité enquêtée, son APE,…*

- **Les variables issues du traitement de l'enquête**

Il s'agit des variables de gestion et d'apurement. Ces variables ne donnent pas lieu à un redressement.

*Exemples : code collecte, date de rappel, indicatrice issue de l'apurement, ….*

- **Les variables quantitatives du questionnaire**

Ces variables issues de l'enquête sont les réponses mesurables demandées aux unités interrogées. Elles seront redressées.

*Exemples : montant des ventes, nombre de points de vente, effectifs, …*

- **Les variables qualitatives et questions filtres du questionnaire**

Ces variables issues de l'enquête sont les réponses non mesurables demandées aux unités interrogées. Elles seront redressées.
Parmi les variables qualitatives, on distingue celles issues de questions filtres, c'est-à-dire dont la réponse influence la suite à donner au questionnaire.

*Exemples :*
*Les réponses de type « oui / non / peut-être ….», « un peu / beaucoup », « jamais / parfois / toujours », génèrent des variables qualitatives.*
*La question A1 de l'enquête TIC « Votre entreprise dispose-t-elle d'un ordinateur ? » est un exemple de filtre. En effet, en réponse, l'unité enquêtée peut cocher :*
*☐ « oui » ou ☐ « non (Passez à la question H1) »*
*Dans le cas d'un « non », elle n'aura pas à répondre aux questions intermédiaires entre A1 et H1. Elle n'y aura pas répondu, mais ce sera dans l'ordre des choses.*

# *2. Prise en compte du champ de l'enquête*

Il est important de **bien définir le champ** de l'enquête au préalable des traitements (avant le début de l'enquête dans l'idéal), cela ayant une incidence sur les travaux à venir.

En effet, **certaines unités enquêtées vont répondre à l'enquête, alors qu'elles sont finalement hors-champ… et vice versa.**

*Reprenons l'exemple de l'enquête TIC ; son champ exclut les entreprises de moins de 10 salariés. La base de sondage ne fera donc intervenir que des unités ayant plus de 10 salariés AU LANCEMENT. Mais, entre la constitution de cette base, et le début de l'enquête, on peut avoir deux cas de figure se situant « à la marge » de ce seuil, pour les entreprises de la population totale:*

- *des entreprises ayant plus de 10 employés peuvent en avoir licencié : elles seront dans le champ au lancement, mais n'en feront plus partie dès la phase de collecte*

- *des entreprises qui avaient un peu moins de 10 salariés peuvent en avoir embauché : hors champ au moment de la constitution de l'échantillon, elles auraient finalement pu être interrogées…*

**Comment peut-on avoir interrogé des unités hors-champ ?**

La base de sondage est constituée à partir de données de lancement (source SIRENE, ou autre). Ces données peuvent avoir déjà varié dans la réalité, sans encore avoir été réactualisées dans ces sources.

Un **ajustement** sera nécessaire lors de la phase d'apurement pour prendre en compte :

- d'une part les unités incluses dans le champ (au lancement) mais qui en sont sorties avant le début de la collecte. Ces unités finalement hors-champ auront tout de même été interrogées.
- d'autre part les unités non détectées dans le champ (lors de la constitution de l'échantillon) alors qu'elles y appartiennent finalement au moment de l'enquête. Ces unités relevant en réalité du champ, n'auront pas été interrogées.

Il ne suffit pas d'éliminer les données fournies par les unités de la première catégorie pour résoudre ce problème de fluctuation d'entreprises dans le champ. En effet, en procédant ainsi, on oublie qu'on ne peut pas « corriger les fluctuations inverses » : autant il est aisé de supprimer les données dont on dispose, autant il est impossible de récupérer les renseignements des unités non interrogées ! Afin de « compenser » ces fluctuations, certains questionnaires de la première catégorie seront malgré tout conservés, pour corriger quelque peu l'absence de questionnaires de la seconde catégorie.

*Exemple tiré de l'enquête TIC 2008 :*
*Comme précisé ci-dessous, le champ de cette enquête n'englobe que les unités de plus de 10 salariés. Que faire lorsqu'une entreprise déclare un effectif inférieur à 10 ? La place-t-on hors champ ou non ? Pour ces cas, les traitements suivants ont été appliqués :*

- *on a décidé de maintenir dans le champ, les entreprises passées naturellement en dessous du seuil (critères : effectif au lancement < 20 et effectif déclaré > 4). Pour ces observations, on a « forcé » l'effectif déclaré à 10.*
- *parmi les unités restantes, on a mis hors-champ les unités pour lesquelles l'effectif au lancement semblait erroné (l'entreprise n'aurait pas dû se trouver dans l'échantillon) ; il s'agit par exemple des agences d'intérimaires, de production audiovisuelle, d'hôtesses, … ces activités conduisent généralement à déclarer plus de personnes qu'il n'y a effectivement de postes dans l'entreprise.*
- *enfin, certains cas particuliers ont été traités individuellement, suite à des recherches dans des sources externes. Cela concernait essentiellement des entreprises ayant déclaré moins de 10 salariés et pour lesquelles l'écart avec l'effectif au lancement était très élevé.*

*(Nota : les traitements décrits dans cette exemple ont en partie été réalisés dès la phase d'apurement.)*

⇨ Il est bon, à ce stade du traitement de l'enquête, d'analyser les critères de définition du champ, afin que les contours de ce dernier soient bien précis.

# *3. La variable « ETAT »*

## a. Comment est-elle définie ?

La variable « ETAT » est construite pour chaque unité de l'échantillon. On distingue 4 situations possibles [2] :

- L'unité enquêtée a restitué son questionnaire, et son appartenance au champ est avérée. Elle est considérée comme **répondante du champ** (modalité « R »).
- L'unité n'a pas restitué son questionnaire, et son appartenance au champ est avérée (confirmation lors de relances téléphoniques par exemple); elle est a priori **non répondante du champ** (modalité « N »).
- L'unité est cessée/morte/inactive, ou on sait finalement qu'elle est **hors champ** (modalité « H »).

[2] Les modalités sont nommées arbitrairement ici, mais seront réutilisées dans la suite du document.

- L'unité n'a pas répondu et on n'a pas davantage d'informations pour la classer dans une des catégories précitées. **Non retour du questionnaire sans information supplémentaire** (modalité « I »).

A ce stade, **la difficulté est donc de bien isoler parmi les non-retours, les unités non-répondantes.**
Dans un premier temps, les unités de l'échantillon sont ventilées grâce au **code-collecte** notamment (renseigné par les gestionnaires). Si l'unité est répondante, on peut aussi utiliser les renseignements fournis (pour confirmer ou non son appartenance au champ).

*Exemple tiré de l'enquête TIC :*
*La codification de la variable ETAT a au préalable été synthétisée dans le tableau ci-après. On utilise notamment les variables de gestion SITU (« état de l'entreprise durant l'enquête ») et G_C (« code collecte ») dont les modalités sont précisées entre parenthèses.*

| ***Condition 1*** | ***Condition 2*** | ***Situation*** | ***ETAT*** |
|---|---|---|---|
| ***SITU='2'*** *(inactive)* | *Tous G_C confondus* | *Entreprises cessées* | ***H*** |
| ***SITU='3'*** *(autres cas-donc hc)* | *Tous G_C confondus* | *Entreprises hors champ* | ***H*** |
| ***SITU='1'*** *(active, par défaut)* | ***G_C='0'****(Demande 1e envoi)* | | ***I*** |
| | ***G_C='1'****(NPAI confirmé)* | | ***I*** |
| | ***G_C='2'****(retour NPAI)* | | ***I*** |
| | ***G_C='3'****(envoi quest. suppl.)* | | ***I*** |
| | ***G_C='4'****(quest. envoyé)* | | ***I*** |
| | ***G_C='4'****(quest. envoyé)* et ***zone doc non vide*** | *Entreprise du champ non cessée car contactée* | ***N*** |
| | *(**G_C='5'****(questionnaire reçu)* ou ***G_C='9'****(questionnaire reçu en saisie externe))* et ***plus de 10 salariés*** | *Entreprises répondantes du champ* | ***R*** |
| | *(**G_C='5'****(questionnaire reçu)* ou ***G_C='9'****(questionnaire reçu en saisie externe))* et ***moins de 10 salariés*** | *Entreprises hors champ* | ***H*** |
| | ***G_C='6'****(non rép. acceptée)* | *Entreprises actives non répondantes* | ***N*** |
| | ***G_C='7'****(envoi de questionnaire suite à NPAI)* | | ***I*** |
| | ***G_C='8'****(exception à l'envoi)* | | ***I*** |

Dans un second temps, on va utiliser des **sources externes** afin de déterminer si la non réponse des unités ventilées dans la dernière catégorie (modalité « I ») s'explique par une volonté délibérée de l'enquêté de ne pas répondre (non répondante du champ, modalité « N ») ou par son incapacité à le faire (si elle n'existe plus ; cas hors champ, modalité « H »).

*Exemple tiré de l'enquête TIC : repérer les entreprises hors champ parmi la catégorie ETAT= « I ».*
*Afin de distinguer, parmi les questionnaires non-retournés, ceux qui concernent des entreprises cessées de ceux qui correspondent à de la non-réponse, ont été considérées comme « cessées », les entreprises*

- *pour lesquelles on n'a aucune information dans l'enquête et*
- n'ayant pas envoyé de déclaration de TVA en 2008 alors qu'elles l'avaient fait en 2007.

*Pour ce faire, la liste des entreprises présentes dans le fichier de lancement mais inactives dans le répertoire SIRENE à la fin de l'enquête a été dressée. Règle de décision adoptée : considérer comme cessée, toute entreprise ne figurant plus dans le répertoire SIRENE.*
*Puis : utilisation du fichier des déclarations de TVA. Cette source présente des problèmes d'identification et une différence de champ avec l'enquête, qui font qu'une entreprise absente du fichier TVA ne doit pas être considérée comme cessée ; par contre, une entreprise présente à une certaine date qui a cessé de l'être, est vraisemblablement sans activité.*

En fin de course, il se peut que la catégorie « I » de la variable ETAT compte encore des entreprises. Deux solutions se présentent alors :

- par commodité, basculer ces entreprises « I » en non-répondantes (ETAT= « N »)
- affecter une probabilité d'être hors-champ à ces entreprises de la catégorie « I ». Autrement dit, répartir leur poids initial sur les unités non-répondantes et hors-champ. Ainsi, les poids des entreprises de la catégorie « R » seraient inchangés, ceux de la catégorie « I » seraient annulés, et impactés sur ceux des catégories « N » et « H » (qui seraient donc modifiés). Cette solution de repondération est certes plus rigoureuse, mais plus coûteuse en temps.

⇨ Lorsque débute le redressement proprement dit, il s'agit de **ne plus avoir d'observation dans la catégorie « I »** (non retour sans information complémentaire)**.**

## b. A quoi sert-elle ?

Cette variable « ETAT » a deux utilités principales :

- Elle permet de calculer le **taux de réponse à l'enquête.** Il s'agit en effet du ratio :

$$\textbf{\textit{Taux de réponse à l'enquête}} = \frac{nb\_de\_"R"}{nb\_de\_"R"+nb\_de\_"N"}$$

*Le taux de réponse ainsi calculé diffère du taux de réponse transmis par le centre de gestion de l'enquête, en lien avec les traitements effectués.*

- La modalité de cette variable permettra de déterminer comment l'unité sera traitée lors de la phase de redressement. A titre d'illustration : seules les unités telles que ETAT = « R » participeront à l'étape de correction de la non-réponse partielle, tandis que l'opération sur la non-réponse totale inclura les unités pour lesquelles ETAT = « N ».

# *4. Passage en non-réponse totale*

**⇨ Lorsqu'un questionnaire réceptionné s'avère être de mauvaise qualité, on peut finalement décider de considérer l'unité dont il est question, comme une non-répondante totale.**

En effet, il arrive que des unités de l'échantillon fassent preuve de mauvaise foi, en retournant un questionnaire mal, peu, voire pas rempli. Il est alors préférable de remettre complètement à blanc ces questionnaires, car ils ne sont pas « bien » exploitables. Si ces unités s'avèrent être dans le champ, elles seront alors considérées comme étant non répondantes totales. En conséquence de cela : au lieu d'avoir la modalité « R » de la variable « ETAT », on leur attribuera la modalité « N ».

# *5. Les unités non substituables*

## a. De quoi s'agit-il ?

**⇨ On qualifie de « non substituable », une unité enquêtée « importante » et/ou très particulière par rapport au(x) thème(s) traité(s).**

*Exemples :*
*Une entreprise peut être jugée « importante » au vu d'un effectif considérable ;*
*Un réseau de franchisés peut être « important » du fait d'un CA très élevé ou d'un nombre de points de vente exceptionnel ;*
*Une entreprise sera traitée « particulièrement » si le thème d'étude est le commerce électronique et qu'elle est connue pour réaliser la quasi-totalitéé de son CA grâce aux services offerts sur son site internet ;*
*Etc.…*

Ces unités non substituables sont désignées
- soit au cas par cas (si l'échantillon n'est pas trop grand)
- soit selon des règles de sélection bien définies (lorsque l'échantillon est conséquent).

A titre indicatif, les non substituables représentent environ 1% (voire 1.5%) des unités figurant dans l'échantillon.

*Exemples :*
*• L'échantillon de l'enquête sur les réseaux de services 2008, comptait un peu plus de 500 réseaux. Parmi eux, 7 ont subjectivement été reconnus « atypiques » de par leur importance en termes de chiffre d'affaires, de montant de ventes, de nombre de points de vente,…*
*• Dans le cas d'une enquête sur les technologies de l'information, pour laquelle on interroge plus de 12000 entreprises, on décide plutôt de mettre en place des règles de sélection automatiques afin de*

*déterminer les non substituables. A titre d'information, les critères de définition de ces unités particulières étaient les suivants (cumuler ces critères) :*

- *être une grande entreprise (variable gen_l= « 01 »)*
- *avoir un chiffre d'affaires au lancement (CAL) supérieur à 2000000 k€ ou que ce CAL représente plus de 5% des CAL de sa strate de diffusion*
- *présenter un effectif au lancement (EFFL) supérieur à 5000 ou que l'unité totalise plus de 5% des EFFL de sa strate de diffusion*
- *avoir un montant des ventes par internet estimé à plus de 100000 k€ au lancement*
- *présenter une estimation du montant des ventes électroniques (i.e. internet et edi) dépassant les 500000 k€ au lancement*
- *avoir un montant des achats électroniques estimé supérieur à 500000 k€ au lancement.*

*Finalement, on obtient 152 unités non substituables dans l'échantillon (c'est-à-dire qu'environ 1.25% des unités de l'échantillon sont non substituables).*

## b. Quelles conséquences pour le redressement ?

### ➢ Poids égal à l'unité

Les unités non substituables ne peuvent représenter qu'elles-mêmes, par conséquent, leur poids doit être égal à l'unité !
A l'inverse, les unités « substituables » représentent « plus qu'elles-mêmes » (elles représentent aussi des unités qui leur sont similaires), et leur poids peut être supérieur à l'unité.

### ➢ Redressement spécifique

Les unités non substituables ne seront pas redressées comme les autres.

En effet, ces unités non substituables non répondantes totales feront l'objet d'un **redressement de la non réponse partielle** par imputation « raisonnée », c'est-à-dire selon des recherches externes ou grâce aux données de l'enquête précédente (dans le cas où celle-ci serait récurrente).

La recherche d'informations au cas par cas nécessite des moyens humains et du temps. Afin de les limiter, il est recommandé de porter une attention toute particulière à la relance des unités non substituables en amont des travaux de redressement.
Par ailleurs, ce « suivi » permettra d'obtenir davantage de réponses directement des unités enquêtées, au lieu d'imputer des résultats, ce qui n'est pas négligeable étant donnée l'importance de ces unités.

Toutefois, lorsque ces procédures ne permettent pas de remplir l'intégralité du questionnaire, celui-ci bascule en non réponse-partielle. Il sera alors soumis aux travaux de redressement communs aux unités « substituables ». Il faut néanmoins limiter ce type de pratique.

Par ailleurs, les renseignements fournis par ce type d'unités ne sont pas utilisés lors de la correction de la non-réponse par imputation aléatoire. En effet, lorsqu'on se sert des unités répondantes pour estimer les renseignements des non-répondantes à une question, on exclura de ce processus, les réponses données par les unités « non substituables ».

Le poids des observations non substituables vaut l'unité. De ce fait, elles sont exclues des corrections par repondération lors du **redressement de la non réponse totale**. Ainsi, leur poids ne sera pas modifié.

***Les données ainsi préparées et prédéfinies, peuvent maintenant être soumises aux travaux d'imputation de la non-réponse proprement dits. Les unités sont désormais ventilées dans trois catégories (variable « ETAT »), et le statisticien a défini les contours des unités particulières (les « non substituables »).***

# II. Redressement de la non réponse partielle

***La correction de la non-réponse partielle consiste à compléter les questions non renseignées par les unités répondantes du champ. Cette opération est un artifice, dans le sens où en aucun cas cela ne remplace une réponse donnée directement par l'unité enquêtée. Toutes les méthodes utilisées ne pourront que s'approcher d'une réponse authentique, mais elles ne remplacent pas la consultation de l'échantillon. Lors de la collecte, il est donc important de mettre tout en œuvre pour obtenir des taux de non-réponse partielle les plus faibles possibles, pour éviter au maximum les « trous » dans les questionnaires retournés (grâce au processus de rappels des unités répondantes, par exemple). Enfin, même si l'imputation d'une réponse ne vaut pas la « vraie » réponse, il reste toujours plus préjudiciable de ne rien faire du tout que d'imputer une réponse approximative.***

⇨ Le redressement de la non réponse partielle peut être indifféremment effectué avant OU après celui de la non réponse totale. L'ordre de passage de ces deux étapes n'importe pas, et leur mise en œuvre se déroule quoi qu'il en soit avant la phase de calage.

## *1. Principe général*

Le redressement de la non réponse partielle concerne les unités répondantes du champ. Cette étape permet de compléter les « quelques » items auxquels l'enquêté n'a pas répondu. Le taux de non réponse partielle est donc calculé *par question*. Il permet de se rendre compte comment chaque interrogation a été perçue par les unités répondantes ; ainsi, si le taux de non-réponse partielle est spécifiquement élevé pour une question, on pourra, en amont d'une éventuelle prochaine vague d'enquête, s'interroger sur la complexité du libellé, ou songer à reformuler ce dernier.
On distingue en général le redressement des variables qualitatives et celui des variables quantitatives. Selon la nature de la variable, le mode de traitement est en effet différent. L'étape de typage est par contre une phase de mise en place commune à ces deux catégories.

## *2. L'étape de typage et les variables de STATUT*

### a. Utilité et mise en place

Pour chaque variable du questionnaire est créée une variable de « statut ». Il est à noter que les variables de statut **ne sont définies que pour les unités du champ répondantes à l'enquête**. On distingue 3 situations possibles[3] :

- l'unité **n'est pas concernée** par la question (statut de modalité « N », comme « non concernée »). Cette question ne sera pas redressée, et restera donc vide.
- l'unité est concernée par la question, et y **a répondu** (statut de modalité « B », comme « réponse brute, bonne réponse,… »). Il n'y a aucune raison de redresser cette question.
- l'unité est concernée par la question, mais n'y a **pas répondu** (statut de modalité « M », pour « réponse manquante »). Cette question va être redressée.

**Être « concernée » ou non par une question...**

Une unité est concernée par une question, lorsqu'on s'attend à ce qu'elle réponde à ladite question. Une unité est souvent « non concernée » après un filtre.
*Soit la formulation : « question 15 - Faites vous partie d'un réseau de franchisés ? (si « non », passez directement à la question 20) », une unité ayant coché « non » en question 15, sera non concernée par les questions 16 à 19 et n'aura pas « obligation » d'y répondre (statuts 16 à 19 : à « N »). Par contre, une autre unité, ayant coché « oui » en 15, et n'ayant pas renseigné les items 16 à 19, aura ses statuts à « M ».*

[3] Les modalités sont nommées arbitrairement ici, mais seront réutilisées dans la suite du document.

De façon arbitraire, dans le présent document, les variables de statuts seront nommées en suffixant le nom de la variable étudiée, par « _S ».
*Ainsi, à la question « Faites-vous partie d'un réseau de franchisés ? », correspondent par exemple :*

- *la variable qualitative du questionnaire « resfran », pouvant prendre ici les modalités « 1 » (« oui »), « 2 » (« non ») ou «  » (absence de réponse).*
- *la variable de statut correspondant, nommée « resfran_S ».*

⇨ En définitive, seules les variables à statut « manquant » (modalité « M ») vont être modifiées suite au redressement de la non réponse partielle. En règle générale, la modalité qui leur sera attribuée sera fonction des réponses des autres unités, à la même question. Ces réponses présentent un statut « bon » (c'est-à-dire de modalité « B »). Dans la table de données finale, les seuls items à blanc seront ceux pour lesquels l'unité n'est pas concernée par la question (statut de modalité « N »).

## b. Calcul des taux de (non) réponse partielle

Cette étape de typage permet donc notamment de cerner les questions à corriger, des non réponses acceptées. Les statuts sont aussi utiles au calcul des taux de réponse (ou non-réponse) partielle, de la façon suivante :

$$\text{Taux de non-réponse à la question} = \frac{nb\_de\_statuts\_"M"}{nb\_de\_statuts\_"M"+nb\_de\_statuts\_"B"}$$

$$\text{Taux de réponse à la question} = \frac{nb\_de\_statuts\_"B"}{nb\_de\_statuts\_"M"+nb\_de\_statuts\_"B"}$$

(Ces deux taux sont les complémentaires l'un de l'autre).

*En reprenant l'exemple du précédent paragraphe, on aura donc :*

$$\textit{Taux de réponse à la question 15} = \frac{nb\_de\_(resfran\_S="M")}{nb\_de\_(resfran\_S="M")+nb\_de\_(resfran\_S="B")}$$

L'étude de ces taux permet parfois de décrire et d'expliquer le comportement de réponse des unités répondantes du champ. On notera par exemple que les questions de la dernière page, connaissent un taux de non-réponse généralement plus élevé que les autres (lassitude des unités enquêtées). Certaines de ces déductions peuvent ensuite être utiles dans la conception ultérieure de questionnaires (pertinence de l'enchainement des questions, choix de la présentation de l'enquête, formulations confuses à revoir,…). En ce qui concerne le redressement de la non-réponse partielle, on sera davantage vigilant sur la méthode à employer et les moyens à déployer pour redresser un item à fort taux de non réponse : en effet, plus le taux de non réponse sera élevé, plus le comportement de réponse qu'on imputera, aura d'impact sur l'exploitation globale de la question.
Toutefois, même s'il est préférable de redresser un item que de laisser vide, il n'en reste pas moins que la réponse imputée est plus ou moins approximative et ne peut remplacer exactement la réponse réelle de l'unité enquêtée.

# *3. Règles déterministes*

On appelle règle déterministe (ou règle logique), toute imputation *« évidente »* permettant de compléter des questions *non* renseignées par l'unité du champ *répondante* au questionnaire.

*Exemple :*
*Dans le cas de l'enquête TIC (technologies de l'information), si l'unité sondée n'a pas répondu à la première interrogation «Possédez-vous un ordinateur ? » alors qu'elle renseigne les autres parties du questionnaire sur l'utilisation qu'elle fait de ses équipements informatiques (internet, intranet, logiciels de bureautique, etc.), on pourra supposer sans trop se tromper qu'il s'agit d'une erreur d'inattention de*

*l'unité sur le premier item, et on remplira celui-ci à « oui, je possède un ordinateur » grâce à une règle déterministe.*
*A noter : Dans la configuration de cet exemple, si l'unité remplit le premier item, mais y déclare « ne pas posséder d'ordinateur », cet item pourra être forcé à « oui je possède un ordinateur »…. lors de la phase d'APUREMENT (en effet, tout comme la non-réponse n'est pas modifiée lors de l'apurement, la réponse n'est quant à elle pas changée lors du redressement).*

Lorsqu'une règle est appliquée, le statut des variables concernées n'est pas modifié. Par contre, une indicatrice sera créée pour chaque règle, afin d'avoir toujours une trace des changements effectués, et de distinguer les réponses « originales » des réponses « imputées de façon logique ». Une documentation rigoureuse est donc indispensable à une bonne utilisation des données (cf. paragraphe « V.3. Documentation »).

*Exemple :*
*En parallèle du cas décrit précédemment, on modifiera aussi la valeur de l'indicatrice adéquate. On peut ainsi créer une variable « règle_ordi » qui vaut 0 par défaut, et 1 lorsque la première question, initialement vide, est complétée à « oui » par la règle déterministe précédemment décrite.*

Les règles logiques peuvent être réfléchies, listées et arrêtées **en amont** du redressement, sur examen approfondi du questionnaire.

**Autre exemple, complet et programmé :**

```
/* REGLE 3: Si l'entreprise n'a pas
déclaré posséder internet (question B1)
et qu'elle n'a renseigné aucune réponse
(valeur absolue ou %) dans la colonne
"site Web" (en F2), alors on met 0 dans
cette colonne. */
DATA ma_table;
SET ma_table;
IF ETAT="R" THEN DO;
regle3=0; /* par défaut*/
IF B1_acces_internet ='2' AND
        MISSING(F2_VENT_INET_VAL) AND
        MISSING(F2_VENT_INET_PCT)
THEN DO;
        regle3=1;
        F2_VENT_INET_VAL = 0 ;
        F2_VENT_INET_PCT = 0 ;
END; END;
RUN;
```

# *4. Redressement des variables qualitatives*

## a. Principe général

De façon schématique : pour corriger la non-réponse à une question donnée, on utilise les renseignements obtenus auprès des unités répondantes d'une catégorie (critères de taille, d'activité, de comportement, etc.), afin de compléter les variables qualitatives non renseignées des unités de la même catégorie non-répondantes à cette question.

Autrement dit, le comportement de réponse des unités du champ ayant complété l'item étudié (statuts à « B ») est analysé et utilisé pour renseigner les « manquants » des autres unités (ayant un statut à « M »). A cet effet, on recherche des variables liées à la variable à redresser : quelles relations statistiques existe-t-il entre les réponses ?

D'autres méthodes peuvent également être proposées, mais elles seront utilisées avec parcimonie.

## b. Recherche de liens

Cette étape permet de déterminer un **comportement de réponse** pour une question, selon les critères de l'unité enquêtée. Grâce au **test du Chi-Deux** et au **V de Cramer**, on peut détecter des liens entre variables. Ces liens seront utiles à l'exécution du macro-programme d'imputation aléatoire de la non-réponse partielle qualitative (présenté dans le prochain paragraphe).

**Détection de lien avec le Khi-deux et Cramer**

Dans un premier temps, il est bon de privilégier les résultats de SAS pour lesquels l'effectif est suffisant par croisement de variables. Dans les sorties, cela revient à écarter les résultats dans lesquels figure la mention « Warning ». On a alors l'assurance de la robustesse des résultats et de la présence d'un lien avec effectif suffisant pour chaque croisement de modalités.
Ensuite, on considère les résultats indiquant un lien au sens du khi-deux (c'est-à-dire présentant une p-value du khi-deux inférieure à 0.05). Le lien sélectionné correspond alors parmi ces observations, au lien le plus fort, repérable grâce à un V de Cramer le plus élevé.
Toutefois, si aucun lien n'est détecté avec le test du Khi-deux, on procède de même, avec des effectifs moins importants (résultats avec la mention « Warning »).

*Prenons pour exemple l' « enquête sur les réseaux dans les services ». Un lien a été détecté entre les deux variables suivantes :*

- *la variable qualitative du questionnaire « redevf » (« Les points de vente indépendants sont-ils soumis à un acquittement de redevances ou cotisations fixes ? »)*
- *la variable de lancement « code_ape » (activité principale exercée - variable discrétisée)*

*C'est-à-dire que les résultats de « redevf » sont différents selon la modalité du code APE du réseau. Ainsi, 3% des réseaux de la tranche d'activité « 3 » (regroupement de plusieurs APE adjacentes), déclarent que leurs points de vente indépendants ne sont pas soumis à un acquittement, tandis qu'ils sont 33% à déclarer cela, lorsqu'ils relèvent de la tranche d'APE « 4 ».*
*Le cas échéant, c'est-à-dire si « redevf » n'était pas liée à l'activité, les réseaux répondraient de façon identique quelle que soit l'APE dont ils sont issus.*

Une variable du questionnaire (soumise au redressement de la non-réponse partielle), peut aussi bien être liée à une variable de lancement, qu'à une autre variable du questionnaire. **Dans cette dernière configuration, l'ordre d'imputation importera** ! En effet, si le lien détecté sur « variable A » fait intervenir « variable B », il faudra *d'abord* imputer « variable B » *puis* redresser « variable A ». Si cette chronologie est insoluble, le statisticien songera, pour une variable, à utiliser un autre lien (bien détecté, mais moins fort que celui choisi initialement), afin de « casser le cercle ».

**Point SAS : Chi-Deux et Cramer**

La commande SAS à utiliser est :

```
PROC FREQ DATA = ma_table;
        WHERE   (variableA_s = 'B' AND variableB_s="B");
        TABLE    variableA * variableB
        / missing chisq deviation expected cellchi2;
RUN;
```

**Quels liens tester ?**
En effet, si on souhaite redresser une variable qualitative, vers quelle(s) autre(s) variable(s) s'oriente-t-on pour détecter un éventuel lien ? Voilà quelques pistes de recherche (liste non exhaustive) :

- On préfèrera dans un premier temps se fier à des variables disponibles pour toutes les unités. Il s'agit essentiellement des données disponibles au lancement et/ou ayant servi à déterminer les strates de tirage, mais on peut aussi penser à des sources externes apportant un renseignement pour chaque unité enquêtée.
- On pourra ensuite aussi se fier à son intuition : quelle variable peut être liée à la variable étudiée ? Quelle autre question peut influencer les réponses à la question considérée ?
- Il est aussi possible, dans le cas d'enquêtes périodiques et pour certaines unités, d'utiliser les réponses de la précédente vague d'interrogation.
- A noter enfin, qu'on peut « créer » des variables potentiellement liées
    - Soit en discrétisant une variable quantitative dont la modalité semble avoir un impact sur la variable qui nous intéresse
      *Exemple :*
      *La variable « nbpvfr » sur le nombre de points de vente peut être discrétisée (ventilée) en 4 tranches. Ces classes de points de vente peuvent ensuite être liés à une variable du questionnaire.*
    - Soit en concaténant deux autres variables (voire plus).
      *Exemple :*
      *La variable « pv_ape », combinaison de la tranche de nombre de points de vente avec la classe de l'activité du réseau, peut être plus fortement liée à une variable du questionnaire que chacune de ces deux variables testées séparément.*

## c. Imputation aléatoire

Le redressement de ce type de variables est grandement facilité par la conception d'un macro-programme par Henri BODET. L'imputation aléatoire se fait en simulant la distribution des réponses observées parmi les unités ayant les mêmes caractéristiques. Ainsi, les liens entre variables sont conservés. Cet outil statistique est d'utilisation aisée ; toutefois, quelques précautions sont à prendre.

Au préalable, il faut veiller à ce que la variable à redresser dispose d'un « statut » (et qu'il soit renseigné pour toutes les unités répondantes du champ).

Par ailleurs, la variable « auxiliaire » choisie (ou variable « liée », autrement dit, celle qui est utilisée pour compléter la variable à redresser) doit respecter quelques formalités :

- elle doit être **qualitative**
- l'utilisateur veillera également à ce que chaque catégorie de la variable auxiliaire présentant des unités à imputer, dispose aussi d'observations imputantes (unités servant de « modèle », de « sources », pour l'opération d'imputation)

- il faut impérativement qu'elle soit **renseignée pour toutes les unités** répondantes du champ.

⇨ Ce dernier point montre bien **l'importance de l'ordre d'imputation des variables à redresser.**

*Exemple :*
*Si la variable « var1 » permet de redresser « var2 », il faudra d'abord s'attarder à compléter les réponses pour « var1 » avant de faire cela pour « var2 » !*

**Que fait le macro-programme ?**

Cet outil d'imputation aléatoire de variables qualitatives se sert des unités imputantes (c'est-à-dire des unités ayant renseigné la variable d'étude, dont le statut est à « B ») pour déterminer la loi de la variable cible conditionnellement à la variable auxiliaire.

Pour chaque modalité de la variable liée, le programme cumule la fréquence d'apparition de chaque modalité de la variable d'étude.

*Exemple :*
*Considérons la modalité « 1 » du code_ape (variable auxiliaire) et la variable d'étude « redevf ».*

| ***Variable cible*** | ***Probabilité d'apparition observée sur les unités imputantes de code_ape = « 1 »*** | ***Paliers de la fonction de répartition*** |
|---|---|---|
| *Modalité « 0 »* | *52.94 %* | *0.5294* |
| *Modalité « 1 »* | *8.82 %* | *0.6176 ( = 0.5294+0.882)* |
| *Modalité « 2 »* | *38.24 %* | *1 ( = 0.5294+0.882+0.3824)* |

Ensuite, pour chaque unité à imputer, on génère un nombre aléatoire suivant une loi uniforme continue sur l'intervalle [0 ;1]. Le programme repère alors le palier inférieur le plus proche de ce nombre, et donne à l'unité, la modalité correspondant à ce palier.

*Dans notre exemple :*
- *si le nombre aléatoire est compris entre 0 et 0.5294, la modalité « 0 » sera imputée à l'unité*
- *si le nombre aléatoire est compris entre 0.5294 et 0.6176, la modalité imputée sera « 1 » à l'unité*
- *si le nombre aléatoire est compris entre 0.6176 et 1, on imputera la modalité « 2 » à l'unité.*

**Comment utiliser la macro ?**

- L'utilisateur se référera à la notice de prise en main de ce programme, conçu par H. BODET, pour mettre en place et « faire tourner » cet outil.
- Après soumission de ce programme, on veillera que les tables « erreur » soient bien vides, signe que le redressement de la variable s'est convenablement déroulé. A cet effet, il est conseillé de modifier le nom des tables SAS en entrée/sortie dans les paramètres de la macro, faute de quoi l'enchaînement de plusieurs appels écraserait à chaque passage les tables créées (et donc non consultables pour vérifications éventuelles).

*Exemples :*

```
%macro_imputation( claspv, vstatut= claspv_s, AUXI = classe_ape,
                        tableIN=a1,
                        tableOUT=a2,
                        rebut = erreurs.rebut1) ;
%macro_imputation( redevf, vstatut= redevf_s, AUXI = pv_ape,
                        tableIN=a2,
                        tableOUT=a3,
                        rebut = erreurs.rebut2) ;
```

## d. Impact du redressement

Afin de s'assurer que le redressement de la non-réponse partielle qualitative est achevé, on pourra vérifier que les variables à statut « M » (renseignement non fourni par l'enquêté) présentent désormais une réponse. Le cas échéant, il faut reprendre l'étape de redressement.

⇨ Dans la table finale de données, on aura donc **toujours simultanément un statut « manquant » et la variable correspondante non vide**. Il s'agira des cas de réponses obtenues par redressement et non directement fournies par l'unité enquêtée. Les seules réponses manquantes acceptées pour les unités du champ, sont celles à statut « N » (unité non concernée par la question).

Une fois cette étape du redressement achevée, l'utilisateur peut **mesurer l'impact de l'opération**. Pour cela, il **compare les résultats avant et après redressement.** On pourra notamment vérifier l'évolution de la répartition des modalités de réponse à une question (exemple : pourcentage de « oui » sur les unités répondantes à l'enquête). Si la différence entre les résultats avant redressement et ceux après redressement est très (trop) nette, un retour aux données s'impose. Le statisticien doit essayer de comprendre cette variation. Soit celle-ci s'explique du fait de la méthode utilisée et des résultats traités, soit un problème de procédure est pointé, auquel cas les données doivent être soumises à un nouvel ajustement.

*Exemple :*
*Impact du redressement de la variable « redevf » de l'enquête « réseaux dans les services ». On mesure les effets de la correction de la non-réponse du point de vue de la population entière (c'est-à-dire qu'on ajoute une pondération[4]).*

*Code SAS et résultats :*

```
TITLE "Réponses sans prendre en compte le redressement - population";
PROC FREQ DATA = reso (WHERE=(etat="1" or nsub=1));
        WHERE redevf_s = "B";
        TABLE redevf;
        WEIGHT poids_l;
RUN;
TITLE;

TITLE "Réponses en tenant compte du redressement - population";
PROC FREQ DATA = reso (WHERE=(etat="1" or nsub=1));
        WHERE redevf_s in ("B" "M");
        TABLE redevf;
        WEIGHT poids_l;
RUN;
TITLE;
```

| | ***Variable cible « Redevf »*** | | |
|---|---|---|---|
| | ***Modalité « 0 »*** | ***Modalité « 1 »*** | ***Modalité « 2 »*** |
| ***Avant redressement*** | *55 %* | *10 %* | *35%* |
| ***Après redressement*** | *54 %* | *11 %* | *35 %* |

*On voit qu'ici, l'imputation aléatoire n'a pas modifié de façon notable la répartition des modalités de la variable « redevf » dans la population étudiée.*

Après cette phase de redressement de la non réponse qualitative, l'utilisateur peut appliquer une seconde fois le **programme d'apurement** ; en effet, la phase d'imputation génère parfois des incohérences de type logique qu'il est bon de détecter rapidement, et de corriger avant de poursuivre.

## e. Variantes

Des **méthodes alternatives** peuvent être utilisées, voire **combinées** entre elles.

Selon l'« importance » à accorder au redressement d'une variable, différents moyens peuvent être mis en œuvre. Le redressement d'une variable dont le **taux de non réponse est très faible**, voire quasi-

[4] *Choix de la pondération :*
*Il faut logiquement indiquer les poids après calage (cf. chapitre ultérieur). Toutefois, à ce niveau, le statisticien ne les a pas encore calculés. Par défaut, il ne dispose donc que de la pondération de lancement. A ce stade, il mesure donc l'impact de la correction de la non-réponse partielle sur la population, avec les poids de lancement. Puis, lorsque ce sera possible, il procédera aux mêmes vérifications, mais avec les poids calés.*

nul, sera nettement moins prioritaire que celui d'une variable à laquelle les unités n'ont majoritairement pas répondu.

Le concepteur de l'enquête peut aussi avoir défini des variables prioritaires, pour lesquelles une attention particulière sera attribuée au moment du redressement.

Il existe d'autres méthodes de correction de la non-réponse partielle ; on citera par exemple la méthode des **unités donneuses** (encore nommée méthode du Hot Deck). Les « donneuses » sont des unités ayant les mêmes caractéristiques que les non-répondantes (selon des critères d'effectif, d'activité, de localisation, etc.). Elles sont donc susceptibles de répondre au plus près de ce que l'unité non-répondante partielle n'a pas confirmé. Si on dispose de plusieurs unités « donneuses », on pourra attribuer à la non-répondante partielle, le mode de toutes les réponses des « donneuses ».

**Questions à choix multiples simultanés :**

Lorsqu'une question propose plusieurs items, le statisticien peut traiter chaque item séparément. Dans ce cas, il veillera parallèlement, à la concordance des réponses entre elles, et sur la suite du questionnaire. Si par malchance, le mécanisme aléatoire de la macro d'imputation n'affecte aucune réponse au final (par exemple, elle n'a donné que des « non » alors qu'on attend au moins un « oui » pour rester logique dans la question), l'utilisateur corrigera « à la main » cette incohérence. Il affectera une réponse au cas par cas, en affectant par exemple un « oui » à l'item le plus fréquemment cité. Il peut songer à d'autres règles d'affectation.
Une autre solution pour corriger un « bloc de questions », est l'utilisation et le redressement d'un vecteur. Cette méthode est plus rigoureuse que la précédente dans le cas d'un nombre peu élevé d'items, mais elle s'avère d'autant plus fastidieuse qu'il s'agit d'un gros bloc. En effet, le principe général, est de considérer le vecteur des n questions du bloc. Par exemple, si le bloc est constitué de 4 questions auxquelles il faut répondre une seule fois « oui » (O) et les 3 autres fois, « non » (N), alors le vecteur de prendra les modalités : « ONNN », « NONN », « NNON », « NNNO ». Le redressement aura donc lieu comme à l'habitude, en prenant comme variable cible le vecteur, et comme modalités les combinaisons précitées.

Dans le cas de peu d'unités réfractaires à une question (par exemple seule une dizaine d'unités de l'échantillon total n'a pas renseigné la question), on songera à redresser **ces cas-là « à la main »,** notamment en recherchant les renseignements manquants dans une **source externe** (autre fichier administratif, sites internet, …). Cette méthode a par exemple été utilisée pour traiter des unités non-substituables de l'enquête « réseaux dans les services », qui étaient au nombre de 7.

**• Cas d'une enquête cyclique**

S'il s'agit d'une **enquête répétitive** (fréquence annuelle,…), on pourra aussi penser à utiliser les résultats de la session passée. Pour cela, on sera vigilant sur les points suivants :

- On repérera bien les **variables homologues** entre sessions d'enquête. Peut-être le nom des variables a-t-il changé, peut-être aussi la question est formulée différemment (dans ce cas, se demander dans quelle mesure on peut utiliser ces résultats).

- Le redressement aléatoire des variables qualitatives **avec variable homologue à la session passée** se déroule en deux phases, décrites ci-après. Ce procédé en deux temps n'est valable que si on suppose que les réponses entre variable cible et variable homologue sont liées.
  Par exemple, dans l'enquête « TIC », on peut considérer que si une entreprise a déclaré posséder un intranet à la précédente session, elle a 90% de « chances » de le posséder encore à la session en cours.

  Schématiquement, le processus est le suivant :

  Tout d'abord, on impute des réponses aux unités non-répondantes en année N (variable à redresser), mais interrogées et répondantes en année N-1 (variable homologue). Pour cela, on s'appuie sur les réponses des unités répondantes non seulement en N, mais aussi en N-1.

  Ensuite, on donne une réponse aux unités non-répondantes en N, non interrogées en N-1 ou interrogées et non répondantes en N-1. Cette fois, on s'appuie sur toutes les unités présentant, en année N, une réponse à statut « B ».

  Mécaniquement, le procédé est mis en œuvre ainsi :

- Un **nouveau statut** (nommons-le « statut1 ») est tout d'abord créé. Si le statut de la variable homologue de la session passée est à « B » et que le statut de la variable à redresser est à « B » ou « M », alors « statut1 » vaut la modalité du statut de la session actuelle.
- On utilise la **macro SAS** pour redresser la variable, en indiquant que le statut à prendre en compte est « statut1 » et que la variable auxiliaire est la variable homologue de la précédente session.

- On utilise ensuite la méthode générale d'imputation aléatoire déjà décrite antérieurement. A cet effet, on crée un **autre nouveau statut** (nommons-le « statut2 ») sur le modèle :
  si le statut de la variable actuelle est à « B », alors « statut2 » est mis à « B » ;
  si le statut de la variable actuelle est à « M » et qu'en plus la variable est vide, alors « statut2 » vaut « M ».
  On note qu'il existe des cas où la variable actuelle, bien que renseignée, a un statut « M ». Ce sont les cas redressés dans la première phase.
- On utilise la **macro SAS** pour redresser la variable. Cette fois, le statut à prendre en compte est « statut2 » ; quant à la variable auxiliaire à utiliser, il s'agit d'une variable repérée comme liée.

*Exemple complet programmé (enquête TIC) :*

```
/* VARIABLE TIC 2008: "G3e_groupement" *** HOMOLOGUE TIC 2007: "A2_groupement_07"*/

/* Création du statut "_S1" */
DATA ma_table1; SET ma_table1;
       IF A2_GROUPEMENT_S_07 = "B" AND G3E_GROUPEMENT_S in ("B" "M")
       THEN G3e_S1 = G3E_GROUPEMENT_S;
RUN;

/* Macro d'imputation aléatoire avec le statut 1 */
%macro_imputation(    G3E_GROUPEMENT,
                      vstatut        = G3e_S1,
                      AUXI           = A2_GROUPEMENT_07,
                      tableIN        = ma_table1,
                      tableOUT       = ma_table2,
                      rebut          = erreurs.rebut1) ;

/* Création du statut "_S2" */
DATA ma_table2; SET ma_table2;
       IF G3E_GROUPEMENT_S = "B" THEN G3e_S2 = "B";
       IF G3E_GROUPEMENT_S = "M" AND MISSING(G3E_GROUPEMENT) THEN G3e_S2 = "M";
RUN;

/* Macro d'imputation aléatoire avec le statut 2 */
%macro_imputation(    G3E_GROUPEMENT,
                      vstatut= G3e_S2,
                      AUXI   = CLASS_SECT,
                      tableIN = ma_table2,
                      tableOUT= ma_table3,
                      rebut  = erreurs.rebut2) ;
```

**• Redressement « par bloc » d'unités ou « par bloc » de variables**

Cette méthode regroupe des questions par « lot », à l'intérieur desquels les unités sont supposées avoir un comportement de réponse homogène par rapport à une question. De fait, on étudie uniquement le comportement de réponse à la question jugée la plus « représentative » du bloc, et on corrige ensuite la non-réponse à toutes les questions de la section, selon le modèle de réponse ainsi préétabli.

Description succincte :

Des « lots » de questions sont constitués, de façon réfléchie (regroupement de questions abordant le même thème par exemple). Dans chaque lot est déterminée une « question pivot ». Il s'agit de LA question la plus pertinente du lot, la plus représentative de cet ensemble. Cette question servira de référence aux autres items, car on suppose que ces derniers sont liés à d'autres variables sur le modèle du pivot. En effet, on étudie les liens de cette variable-clef, et les autres variables sont ensuite aussi redressées selon ce modèle.

*Exemple : enquête « Tic 2009 ».*
*Le module « F » du questionnaire comporte 7 interrogations sur le thème « Utilisation des technologies basées sur l'identification par Radio Fréquence (RFID) ».*
*Après réflexions et recherches, la question F1 (« Votre entreprise utilise-t-elle des instruments basés sur la technologie RFID ? ») peut être désignée comme pivot du lot « F ». Si on détecte un lien entre F1 et une variable auxiliaire « toto » (activité, tranche d'effectifs, autre variable du questionnaire,…), chacune des 7 questions du lot « F » seront redressées en prenant comme variable liée, la variable « toto ».*

## *5. Redressement des variables quantitatives*

Le redressement des variables quantitatives est plus délicat et il est difficile de proposer une méthode « générale », « universelle ».

Au préalable du redressement, on commence par repérer les variables quantitatives et à les classer par « type », par « nature ». En effet, on peut essayer de **regrouper les variables quantitatives** selon leur unité ou leur formulation notamment. Par exemple (liste non exhaustive) :

- valeurs relatives, montants, quantités, …
- données formulées en ratio, en pourcentage,…
- données complémentaires ou compensatoires entre elles
- données bornées (un pourcentage peut être compris entre 0 et 100)
- données de cadrage
- données récurrentes (si enquête cyclique)
- données pour lesquelles il existe des sources administratives ou externes complémentaires
- données quantitatives tributaires d'une réponse qualitative *(exemple tiré de l'enquête « réseaux dans les services » : « Le réseau possède-t-il des marques propres ? » Réponse Oui/Non. « Si oui, à combien estimez-vous la part du CA réalisé avec les marques propres ? »).*

Lors de ces traitements, l'utilisateur ne doit pas perdre de vue **l'objectif d'analyse de la variable** d'intérêt. Il devra s'interroger sur l'angle d'étude qui sera adopté lors de la production des résultats : en effet, s'intéresse-t-on à un de ses **paramètres** de position essentiellement ? Ou optera-t-on plutôt pour une analyse de sa **distribution** ?

*A titre d'illustration : la distribution peut être quelque peu gommée, notamment si le statisticien a d'autorité affecté un paramètre de position telle que la médiane ou (encore plus) la moyenne.*

Quoi qu'il en soit, il est alors intéressant **d'étudier au préalable la distribution** de chaque variable (seulement pour les unités répondantes à la question), afin de visualiser le comportement de réponse, le caractériser et de se rendre compte des ordres de grandeur, des valeurs extrêmes, etc.

**Étude de distribution : vigilance !**
Lors de la recherche d'une variable liée à une variable d'intérêt, grâce à l'étude de la distribution par modalité de la variable d'intérêt, l'utilisateur veillera à ne pas tomber dans les écueils suivants :
- Nombre d'observations suffisant par classe. Ne pas prendre pour un cas général, quelques petits cas particuliers !
- Utilisation prudente de la moyenne ; en effet, ce paramètre est fortement conditionné par les valeurs extrêmes. Bien les détecter au préalable.

Lorsque c'est possible, c'est le moment de mettre en place des **règles déterministes**.
*Exemple tiré de l'enquête « réseaux dans les services »: Si nombre de points de vente déclaré à zéro, alors on impute la valeur nulle au chiffre d'affaires.*

Le **taux de (non) réponse partielle** permet de repérer le degré de priorité du redressement d'une variable : un variable à taux de non-réponse quasi nul préoccupera moins le statisticien qu'une variable grandement ignorée par les unités répondantes, car dans le premier cas, l'imputation aura peu d'impact sur les données prises dans leur ensemble, au contraire du second cas, pour lequel la manière de redresser sera certainement plus visible dans la globalité des réponses.
Ces taux, complémentaires l'un de l'autre, se calculent de la même façon que pour une variable quantitative :

$$\text{Taux de non-réponse à la question} = \frac{nb\_de\_statuts\_"M"}{nb\_de\_statuts\_"M"+nb\_de\_statuts\_"B"}$$

Taux de réponse à la question = $\frac{nb\_de\_statuts\_"B"}{nb\_de\_statuts\_"M"+nb\_de\_statuts\_"B"}$

Dans la suite de ce document, trois méthodes de redressement d'une variable quantitative vont être décrites. Il en existe d'autres que le statisticien pourra mettre en œuvre, selon le « type » de la variable ou sa signification intrinsèque. Ici seront traités :

- la méthode « Discrétisation de la variable quantitative » : la variable quantitative, discrétisée, peut être redressée avec la macro d'imputation aléatoire des variables qualitatives
- la « Méthode générale » avec utilisation de paramètres de position observés sur les réponses fournies par les autres unités enquêtées
- la méthode ayant recours à des sources externes : certaines données sont disponibles sur internet (année de création de l'entreprise par exemple), dans d'autres fichiers administratifs, dans des vagues d'enquête précédente (avec prise en compte d'un taux d'évolution).

Outre ces 3 méthodes que l'on détaillera, il existe d'autres moyens pour redresser une variable quantitative. On les évoquera en dernier lieu de cette partie.

## a. Méthode « Discrétisation de la variable quantitative »

Le statisticien utilisera cette méthode lorsque les réponses données s'avéreront être du même acabit, c'est-à-dire quand le questionnaire demande plutôt des ordres de grandeur sur les réponses quantitatives, ou des nombres correspondant à des bornes de tranches, ou encore quand l'unité utilisée est le pourcentage, etc.…

*Exemple tiré de l'enquête TIC 2008 :*
*Soit la variable « f4_cl_nat_pct » correspondant à la question : « Si votre entreprise a reçu des commandes de biens ou de services via des sites web en 2007, quel pourcentage d'entre elles a été réalisé auprès de clients nationaux ? ».*
*Pour environ 95% des entreprises répondantes du champ à cette question, les déclarations les plus fréquentes sont (par ordre décroissant d'apparition) : 100 %, 95 %, 50 %, 10 %, 80 %, 90 %, 0 %, 70 %, 99 %, 60 %, 40 %, 20 %. On remarque que dans la quasi-totalité de ces réponses, ce sont des nombres « ronds » (dizaines). Les unités enquêtées répondent généralement « approximativement » ; rares sont celles qui détaillent au pourcent près.*

La méthode la plus classique peut être synthétisée dans le tableau suivant[5] :

[5] Notations arbitraires des variables

| Étapes | Remarques et compléments |
|---|---|
| **Règles logiques** | Dans la mesure du possible. |
| Étude de la **distribution** | Utilité : étude des réponses données, leur répartition, les plus fréquentes, les valeurs extrêmes, … |
| **Discrétisation** de la variable à redresser, en se basant sur les réponses renseignées par les autres unités enquêtées.<br><br>*Nommons « toto » la variable quantitative à redresser, et « toto_tr » la variable correspondante discrétisée.* | **Discrétisation d'une variable quantitative**<br>Le but de l'opération est d'obtenir une variable qualitative à partir d'une variable quantitative. Pour ce faire, on découpe la variable quantitative (initiale) en « tranches », en « classes ». Le découpage peut s'effectuer selon les quantiles observés sur la distribution (par exemple). La PROC UNIVARIATE fournit quartiles, déciles, etc. Dans ce cas, toutes les unités situées entre 0 et le 1er quantile, appartiendront à la première tranche, etc.… La tranche des unités sans réponse, aura pour modalité le blanc/point. |
| Création d'un **statut pour la variable discrétisée**.<br><br>*Nommons « toto_S2 » le statut de la variable discrétisée.* | **Comment ce nouveau statut est-il créé ?**<br>Si l'unité est répondante du champ et que la variable discrétisée présente une modalité non vide, alors le statut est à « M ». Si par contre la variable discrétisée propose une modalité vide, le statut est à « B ».<br><br>**En langage SAS :**<br>`DATA ma_table; SET ma_table;`<br>`IF ETAT = "1" AND missing(toto_tr)`<br>`THEN toto_S2 = "M";`<br>`IF ETAT = "1" AND not missing(toto_tr)`<br>`THEN toto_S2 = "B";`<br>`RUN;` |
| Recherche d'une (ou de plusieurs) **variable(s) liée(s)** à cette nouvelle variable discrétisée<br><br>*Cf. aussi paragraphe II.4.b.* | **Comment intuiter qu'il existe un lien** entre deux variables, au vu des données renseignées ?<br>Outre l'étude de liens grâce à l'outil SAS (tests et proc logistic par exemple), on peut se rendre compte de l'existence d'une relation entre deux variables via **l'analyse de la distribution** de la variable à corriger. En effet, si la variable quantitative à redresser présente un paramètre de position avec des valeurs différentes selon les modalités ou les tranches d'une autre variable, on pourra soupçonner avec une grande certitude que ces deux variables sont liées. Le cas échéant : si les valeurs sont quasi-identiques d'une tranche à l'autre, c'est que la tranche (la modalité de la variable « auxiliaire ») n'a pas beaucoup d'enjeu sur la valeur du paramètre. La variable quantitative n'est pas dépendante de l'autre variable.<br>Cette simple observation permet à l'utilisateur d'intuiter certains liens. **Il validera ou infirmera ensuite statistiquement ces relations** entre variables, grâce par exemple au test du khi-deux. |
| Mise en œuvre de la **macro** d'imputation aléatoire de la non-réponse partielle qualitative sur la variable discrétisée. | Cf. aussi paragraphe II.4.c |
| **Passage en quantitatif**.<br>Chaque unité répondante disposant à présent d'une donnée pour la variable discrétisée (qualitative), il s'agit d'obtenir une réponse quantitative. Pour ce faire, l'utilisateur est seul juge de la méthode à adopter : il peut par exemple opter pour imputer le paramètre de position observé sur les unités répondantes de la même classe discrétisée (on applique le plus souvent la médiane), ou, s'il s'agit d'un pourcentage, penser à imputer des valeurs « rondes » (0%, 20%,50%, 90%, etc.). | ***Exemple** (enquête TIC 2008) **:***<br>*Redressement du chiffre d'affaires (variable H1b_CA).*<br>*La discrétisation de la variable H1b_CA a permis d'obtenir la nouvelle variable CA_tr, qui a été redressée grâce à la macro SAS d'imputation aléatoire qualitative. On a ensuite décidé d'affecter à toute unité non-répondante partielle de la classe discrétisée n°1 (CA_tr = « 1 »), la moyenne de la variable H1b_CA observée sur les répondantes de la tranche CA_tr = « 1 ». Le statisticien est seul juge du paramètre de position à imputer (médiane, moyenne, centre, etc.) ; sa décision se fait après analyse de la distribution de H1b_CA sur les répondantes de chaque tranche.* |

***Mise en garde importante :***

*Cette méthode est préférable dans le cas où relativement peu d'unités sont à redresser.*
*En effet, imputer une même valeur à un très grand nombre d'unités (en l'occurrence, d'une même tranche ici), engendre la création d'un « pic », d'une accumulation, dans la distribution de la variable. Ainsi lorsque l'utilisateur accorde une importance à la distribution de la variable, et non uniquement à la moyenne, cette méthode n'est pas à privilégier.*

Une exception toutefois :

Lorsque la variable a redresser, présente de façon « naturelle » un pic sur une ou plusieurs valeurs, cette méthode peut néanmoins être adoptée.
Ainsi, on remarque souvent que pour des réponses à donner en pourcentage, on observe des accumulations spontanées de réponses sur des valeurs telles que 0%, 50%, 95 %, etc.… Il ne sera donc pas choquant d'attribuer de nombreuses valeurs « rondes », et non des valeurs précises telles que 36%, 4%, 98%,…

Au préalable de la correction de non réponse, il est donc **impératif d'effectuer une étude de la distribution des variables** concernées, pour les unités les ayant complétées (donc à statuts de modalités « B »).

Il est proposé ici deux méthodes alternatives à celle décrite ci-dessus :

- Au lieu d'imputer la même valeur à toutes les unités non répondantes d'une même tranche, le statisticien pourra leur attribuer ladite valeur, associée à un résidu aléatoire.

  *Si par exemple, pour une tranche, on obtient, avec la méthode du tableau précédent, la valeur V à imputer, on attribuera pour chaque individu U non répondant de la tranche, la valeur : V +/- E ; où E est un résidu aléatoire, choisie parmi les résidus (écarts à la moyenne) des unités répondantes.*

- La méthode du Hot Deck dans la tranche : il s'agit de trouver, pour chaque unité non-répondante partielle d'une tranche, une unité donneuse, qui soit répondante. Pour choisir l'unité donneuse au sein de la tranche, plusieurs techniques sont possibles parmi lesquelles la méthode du plus proche voisin (selon des caractéristiques à définir) ou encore le Hot Deck séquentiel qui consiste à donner à une unité, la réponse de la « prochaine » unité répondante de la même tranche. Avec ce type de méthode, on ne crée pas de pic de distribution, tout en conservant relativement bien la moyenne dans la tranche.

## b. Méthode « générale »

Cette méthode combine la recherche de variables qualitatives liées, et l'utilisation de paramètres de position et/ou de dispersion.

### ➢ Idée générale

Après étude de la distribution de la variable quantitative à imputer, on s'intéresse à la détection de liens.
Ensuite, on attribue aux unités répondantes du champ n'ayant pas renseigné la variable en question, la valeur d'un paramètre de position (médiane, moyenne, centre, … à choisir pertinemment au regard de la distribution) observé sur les unités répondantes à la question.

*Exemple :*
*Soit la « variable A » liée à la « variable B ». Après étude de la distribution de la variable « A », on décide de retenir la médiane comme paramètre de position. Soient toutes les unités relevant de la modalité « i » de la variable « B » et n'ayant pas rempli la question « A ». A ces unités, on attribue la médiane de « A » observée sur les unités présentant la même modalité « i » de « B ».*

### ➢ Illustration : utilisation d'un ratio

On peut aussi choisir de redresser un RATIO dans lequel intervient la variable d'intérêt, puis seulement ensuite retourner à l'imputation de cette variable d'intérêt, grâce à une règle de trois.

*Exemple (enquête « réseaux dans les services ») :*
*Cette enquête présente les variables suivantes :*
- *le chiffre d'affaires global du réseau (CATTC), ventilé selon différents postes potentiellement générateurs de CA (ex : « Cafran » pour le CA généré par les points de vente franchisés, etc.)*

- *le nombre de points de vente total du réseau sur territoire français (NBPVFR), ventilé aussi selon les mêmes postes que pour le CA (ex : « Pvfran », etc.)*

*Les chiffres d'affaires sont, d'une manière générale, mal renseignés. Cela n'est pas le cas des nombres de points de vente. Afin de redresser la ventilation du CA, on est passé par la création du quotient « CA par PV » pour chaque poste (ex : Ratio_fran = Cafran*100/Pvfran). Ces ratios ont été soumis à la recherche de variables qualitatives liées (zone géo, regroupement APE, tranche de points de ventes en France). La médiane du ratio a ensuite été récupérée, pour chaque modalité de la variable qualitative liée (ex : « MedFran1 » est la médiane de Ratio_fran observée sur les réseaux relevant de la première modalité de la variable liée et pour lesquels ce ratio a été renseigné). Enfin, aux réseaux répondants du champ n'ayant pas de valeur au ratio étudié, on impute une réponse grâce à une règle de trois : ici, Pvfran = MedFran1*Nbpvfr/100*

**Mise en garde** : attention toutefois aux résultats obtenus pris dans leur ensemble ! En effet, on veillera à faire les ajustements nécessaires afin que les valeurs imputées soient concordantes entre elles. En l'occurrence, dans l'exemple, on doit avoir une somme de ratio égale à 100%.

L'**intérêt** de cette manœuvre, est de conserver une certaine homogénéité, une unité d'étude : dans l'exemple, on manipule des « euros par point de vente », ce qui donne un ordre de grandeur par rapport à l'étude brute de montants en « euros ». On tient compte ici du nombre de points de vente nécessaire à chaque réseau pour arriver à un tel CA.

- **Ajout d'un résidu aléatoire :**

Généralement, on impute donc aux réponses manquantes, la médiane (voire parfois la moyenne) observée sur les unités répondantes. Toutefois, on pourrait, pour faire preuve d'encore plus de rigueur statistique, imputer ce paramètre agrémenté d'un résidu.
Pour cela, on calcule, pour chaque unité répondante, la différence entre la valeur qu'elle a déclarée et la valeur de la médiane. A chaque observation répondante i, on a donc un résidu i. Rappelons au passage que la moyenne des résidus est proche de zéro. On choisit ensuite au hasard un de ces résidus (qui peut être positif ou négatif). Le statisticien impute ensuite aux réponses manquantes, la valeur de la médiane, à laquelle on ajoute la valeur du résidu tiré aléatoirement.

## c. Utiliser une (des) source(s) externe(s)

Selon le type de variable, on peut penser à utiliser d'autres sources, voire d'autres données comme par exemple (liste non exhaustive) :

- les variables de lancement
- les variables homologues d'une précédente vague dans le cas d'une enquête répétitive (penser éventuellement à en réactualiser les valeurs)
- une autre source administrative
- des recherches internet

La méthode employée dépend du nombre d'unités à redresser, et du type de variable concernée.

*Exemple : Redressement du chiffre d'affaires de l'enquête « TIC 2008 ».*
*Lorsqu'une entreprise répondante n'a pas renseigné son CA, il a été décidé de le compléter grâce*

- *au CA de lancement,*
- *puis (si ce dernier est absent) grâce aux données disponibles dans la source « FICUS »,*
- *puis par la réponse donnée (voire redressée, si nécessaire) à la précédente session de l'enquête (si les vagues sont assez espacées dans le temps, penser à réactualiser cette valeur au préalable).*

*Exemple (enquête « réseaux dans les services ») : Redressement de l'année de création du réseau.*
*Typiquement, cette variable ne peut pas être redressée aléatoirement, ni selon un quelconque critère ou autre paramètre de position. Les données manquantes ont alors été complétées suite à des recherches effectuées sur internet.*

## d. Rapide tour d'horizon d'autres méthodes

Outre les trois procédures décrites ci-dessus, le statisticien dispose également de la méthode dite « par **Hot-Deck** » ou encore « **méthode avec donneur** ». Elle s'applique à la correction des **variables quantitatives mais aussi qualitatives**.
L'idée est de rechercher le plus proche voisin répondant, d'une unité non-répondante partielle. Les réponses du « proche voisin » seront attribuées à l'observation défaillante.
Cette méthode est peu utilisée. Toutefois, on met davantage en œuvre une de ses « variantes », lors d'enquêtes récurrentes. En effet, on choisit alors comme proche voisin, l'unité partiellement non-répondante elle-même, et on utilise alors les réponses qu'elle a données lors d'une précédente session.

## e. Vérifications et impact du redressement

Une fois qu'il aura obtenu des valeurs imputées, le statisticien **prendra garde à ce que les réponses soient cohérentes dans le questionnaire**, que les résultats ne présentent pas d'incohérences entre eux, que des « règles de bon sens » soient bien vérifiées.

En reprenant les exemples issus de l'enquête « réseaux services » détaillés dans cette section :

- vérifier que la somme des CA ventilés soit bien égale au CA
- dans le même ordre d'idée : vérifier que la somme des ratio soit bien égale à 100%
- vérifier que si le réseau déclare posséder des marques propres, on n'ait pas imputé la valeur nulle à la part du CA réalisé avec ces marques propres

*Pour tester le bon déroulement du redressement de la variable « toto », on pourra créer un **programme SAS** du type :*

```
DATA ma_table; SET ma_table;
        IF not missing(toto) THEN test = 1;
RUN;
PROC FREQ DATA = ma_table;
        TABLE test * toto_s;
RUN;
```

Si après redressement, le questionnaire échappe à certaines de ces évidences, le statisticien devra

- soit reconsidérer son imputation
- soit ajuster « à la main » les cas défaillants, en faisant preuve de bon sens et d'esprit critique (règle de trois, nouvelle règle déterministe, etc.).

Comme pour les variables qualitatives, on vérifiera **après le redressement que toutes les variables de statut « M » sont finalement renseignées**.

**L'impact peut se mesurer** en comparant les paramètres de position/dispersion « avant » et « après » redressement. Pour cela, on ajoutera successivement à la PROC MEANS de SAS (ou PROC SUMMARY, ou PROC UNIVARIATE,…) les instructions suivantes :

- on précisera d'abord

```
WHERE statut_de_la_variable = "B";
```

  ce qui permettra d'obtenir les résultats seulement pour les « véritables » réponses (celles renseignées par l'unité enquêtée, donc la réponse « la plus juste » dont on peut espérer disposer) ;

- puis on indiquera

```
WHERE ETAT = "R";
```

  afin de considérer tous les résultats (les réponses d'origine ET celles imputées).

La confrontation de ces deux jeux de sorties, dévoilera l'impact du redressement. Si les changements sont trop nets, l'utilisateur pourra retourner aux données, essayer de comprendre une telle évolution et l'ajuster si nécessaire.

*Exemple : impact du redressement de la variable « G1b_CA » de l'enquête TIC 2008.*

*Soient les programmes et sorties SAS suivantes :*

```
TITLE "Avant redressement";
PROC MEANS DATA = tic N MEDIAN MEAN STD SUM MIN MAX MAXDEC = 0;
       VAR G1b_CA;
       WHERE G1b_CA_S = "B";
RUN;
TITLE;
```

Avant redressement 15:11 Monday, April 27, 2009

The MEANS Procedure

Analysis Variable : G1b_CA g1b_ca

| N | Median | Mean | Std Dev | Sum | Minimum | Maximum |
|---|---|---|---|---|---|---|
| 9232 | 13327916 | 128383714 | 904922807 | 1.1852384E12 | 14 | 34445121000 |

```
TITLE "Après redressement";
PROC MEANS DATA = tic N MEDIAN MEAN STD SUM MIN MAX MAXDEC = 0;
       VAR G1b_CA;
       WHERE ETAT = "R"; /* équivaut à: G1b_CA_S in ("M" "B") */
RUN;
TITLE;
```

Après redressement 15:11 Monday, April 27, 2009

The MEANS Procedure

Analysis Variable : G1b_CA g1b_ca

| N | Median | Mean | Std Dev | Sum | Minimum | Maximum |
|---|---|---|---|---|---|---|
| 9898 | 12014038 | 128229696 | 950597259 | 1.2692175E12 | 0 | 34445121000 |

*Que nous apprennent ces résultats ?*
*- On voit que 666 ( = 9898 - 9232) réponses ont été imputées pour la variable « G1b_CA »*
*- Le redressement n'a pas notablement modifié la moyenne (toujours de l'ordre de 128 millions d'euros). Idem pour la médiane.*
*- Quant aux valeurs extrêmes constatées, le maximum est le même, par contre le minimum a été abaissé à 0 (diminution de 14 euros ( !)). Après étude des cas à CA nul, on constate que cela concerne une vingtaine d'unités. En revenant au code, ces valeurs s'expliquent par la méthode employée (utilisation de sources administratives externes) :*

```
IF ETAT="R" THEN DO;
       /**** G1B_CA: Si absence de réponse, on prend CAL.
       ***** Si CAL manquant, on prend source FICUS.
       ***** Si CAL manquant, on prend TIC 2007*/
IF G1B_CA_S="M" AND missing(G1B_CA)     THEN DO;
       IF NOT MISSING(CAL)              THEN G1B_CA=CAL*1000;
       ELSE IF NOT MISSING(CATOTAL)     THEN G1B_CA=CATOTAL*1000;
       ELSE IF NOT MISSING (A1_CA_07)   THEN G1B_CA=A1_CA_07*1000;
       ELSE G1B_CA=0;
END;
-  /* ... etc. ... */
```

*Dans cet exemple, force est de constater que le redressement n'a pas bouleversé la grandeur des résultats pris dans leur globalité. Son bilan est donc satisfaisant.*

⇨ Dans la table de données finale, on aura donc **toujours simultanément un statut « manquant » et la variable correspondante non vide**. Il s'agira des réponses obtenues par redressement et non directement fournies par l'unité enquêtée. Pour les unités répondantes du champ, les seuls items à blanc seront ceux des unités non répondantes totales et ceux des unités répondantes partielles non concernées par la question (statut de la variable à « N »). Dans ce dernier cas, on « accepte » la non-réponse ; on ne cherche pas à imputer une donnée (une telle attitude serait un non-sens).

***Au cours des étapes de redressement de la non-réponse partielle (quantitative et qualitative), le statisticien a imputé des réponses aux unités répondantes du champ ayant fait l'impasse sur certaines questions. Ces travaux ont été sans conséquence sur le poids des unités enquêtées. En abordant les phases de redressement de la non-réponse totale et du calage, il s'apprête à modifier le jeu de poids des unités échantillonnées.***

# III. Redressement de la non réponse totale

***Contrairement à la phase de correction de la non-réponse partielle, le redressement de la non-réponse totale par repondération ne cherche pas à compléter explicitement les données manquantes des unités non-répondantes du champ. En effet, l'idée est ici de « transférer » le poids de ces unités, vers les unités répondantes jugées « similaires ».***

## *1. Principe général*

Il s'agit ici de pallier le fait que des unités dans le champ de l'enquête n'aient pas restitué leur questionnaire. Le statisticien dispose essentiellement de deux méthodes pour redresser ces unités non répondantes totales :

- le redressement par imputation
- le redressement par repondération

De façon simplifiée, le **redressement par imputation** attribue à chaque unité non répondante totale, le comportement de réponse des unités qui lui sont similaires (soit au regard par exemple de l'activité, du nombre d'employés, du montant du chiffre d'affaires etc.… s'il s'agit d'entreprises, soit par exemple selon la taille du ménage, sa localisation, ses revenus moyens, etc.… s'il s'agit d'une enquête relative aux ménages). Ainsi, pour chaque catégorie sélectionnée, on aura des unités « donneuses » dont les réponses seront utilisées pour compléter les questionnaires des unités « receveuses ».
Cette méthode ne sera pas détaillée ici.

On s'attardera en effet davantage sur la méthode de **redressement par repondération** appliquée dans le cas des enquêtes thématiques entreprises. Dans ce cas, le poids des unités non-répondantes totales du champ sont « redistribués » aux unités répondantes du champ, de telle sorte qu'au final, les premières auront un poids nul (et pourront même être écartées de la base de données finale), tandis que les secondes auront un poids plus important, du fait qu'elles représenteront alors davantage d'unités.

⇨ Il est à noter que le redressement de la non réponse totale peut être effectué avant OU après celui de la non réponse partielle.

Le redressement des **unités non-répondantes totales du champ** concerne donc les unités telles que ETAT=« N ».
Néanmoins, l'utilisateur peut avoir écarté des unités non-répondantes, afin de leur appliquer un traitement bien spécifique. Ce peut être le cas par exemple, des grandes entreprises non-répondantes totales, ou des unités non substituables non-répondantes totales, que l'on a passées en unités non-répondantes partielles, en récupérant leurs données N-1. Il s'agit en effet d'unités tellement atypiques, qu'on peut préférer leur attribuer leurs réponses passées. Quoi qu'il en soit, ces unités doivent être marquées (création d'une indicatrice) et exclues des étapes de repondération que l'on va décrire ici.

On peut se faire une idée de la **« qualité » de l'enquête** en calculant son taux de réponse global (cf. formules ci-dessous).

$$\text{Taux de réponse à l'enquête} = \frac{nb\_unités\_à\_ETAT\_"R"}{nb\_unités\_à\_ETAT\_"R"+nb\_unités\_à\_ETAT\_"N"}$$

On peut aussi s'intéresser au complémentaire du taux précédent :

$$\text{Taux de non-réponse à l'enquête} = \frac{nb\_unités\_à\_ETAT\_"N"}{nb\_unités\_à\_ETAT\_"R"+nb\_unités\_à\_ETAT\_"N"}$$

## *2. Constitution des GRH*

## a. Qu'est-ce qu'un GRH ?

Les Groupes de Réponses Homogènes (GRH) permettent de ventiler les unités répondantes et non-répondantes totales. Les unités hors-champ sont écartées de cette opération, ainsi que les non-substituables.
Ces regroupements s'opèrent à l'aide de caractéristiques qualitatives disponibles pour *toutes* les observations (généralement des données de lancement), et qui se révèlent avoir une influence significative sur la probabilité de réponse. Ces groupes se définissent donc par rapport à ces caractéristiques.

## b. Comment ces groupes sont-ils définis ?

### ➢ *Méthode générale*

La méthode générale consiste à bâtir un modèle explicatif du comportement de réponse, et à retenir les variables qui apparaissent significatives. Les variables retenues doivent être connues sur les répondants et sur les autres, ce qui exclut le recours aux questions de l'enquête.

> Si on souhaite utiliser des variables *quantitatives* pour expliquer le comportement de réponse, il faudra au préalable discrétiser ces données. On utilisera alors la variable qualitative ainsi obtenue.

### ➢ *Comment ces subdivisions sont-elles obtenues ?*

Il s'agit dans un premier temps, de repérer les variables potentiellement explicatives du comportement de réponse. Le statisticien les choisit parmi les variables - qualitatives ou assimilées - renseignées pour TOUTES les unités répondantes et non-répondantes. Les unités hors-champ n'interviennent pas dans cette partie.
Le but de l'opération, est de déterminer quelles variables ont une incidence sur le comportement de réponse des unités enquêtées. Formellement, il s'agit de modéliser la variable indicatrice de réponse (« l'unité répond à l'enquête » vs « l'unité n'y répond pas ») à partir des variables explicatives sélectionnées.

*A titre d'exemple, détaillons au fur et à mesure, la procédure suivie pour l'enquête « TIC 2009 » :*

*Initialement on a choisi 5 variables qui pouvaient expliquer le comportement de réponse dans le cadre de cette enquête :*

- *Situation année n-1 :*
  - *unité interrogée répondante l'an passé ou*
  - *unité interrogée non répondante ou hors champ l'an passé ou*
  - *unité non interrogée l'an passé.*
- *Localisation : unité localisée - ou non - en région parisienne.*
- *Tranche de l'effectif au lancement (variable quantitative discrétisée).*
- *Secteur d'activité (regroupements de secteurs proches).*
- *Appartenance ou non à un groupe.*

*Ces données sont disponibles pour toute observation de l'échantillon.*

Le **test du khi-deux** peut une fois de plus être utilisé, afin d'effectuer une première sélection parmi ces variables. En effet, ce test permet d'écarter les variables non liées à la variable indicatrice de réponse. L'utilisateur pourra aussi penser à utiliser des croisements de variables (les concaténer, tout en gardant le caractère qualitatif requis pour ces traitements) ou à regrouper certaines modalités. Ces regroupements pourront être motivés par l'argument statistique de similitude des taux de réponse par modalité, ou par le bon sens (modalités de significations connexes). De même, sur les liens détectés, il pourra noter la provenance des plus fortes contributions, ce qui permettra d'apporter des compléments explicatifs et/ou de modifier éventuellement des classes.

*Illustration (exemple commenté issu de « TIC 2009 ») :*

```
/****************** Recherches de liens avec test du khi2 *********************/

DATA    table_tic;
        SET table_tic;
        IF      ETAT in ("R" "N") AND
                nsub ne 1 AND nrp ne 1;       /* Seules les unités répondantes à poids initial
                                                 ** différent de 1, interviennent dans les tests.*/
        indic_rep = "0";                      /* Indicatrice du comportement de réponse */
        IF etat = "R" THEN indic_rep = "1";   /* Vaut "1" si unité répondante, "0" sinon. */
RUN;

/* La table est prête à être soumise aux tests */

PROC FREQ DATA = table_tic;
        TABLE indic_rep;
        TABLE indic_rep * Paris       / CELLCHI2 CHISQ DEVIATION EXPECTED;
        TABLE indic_rep * class_sect  / CELLCHI2 CHISQ DEVIATION EXPECTED;
        TABLE indic_rep * gp          / CELLCHI2 CHISQ DEVIATION EXPECTED;
        TABLE indic_rep * class_effl  / CELLCHI2 CHISQ DEVIATION EXPECTED;
        TABLE indic_rep * tic08       / CELLCHI2 CHISQ DEVIATION EXPECTED;
RUN;
```

A quoi être attentif dans les sorties SAS ?

- Regarder le résultat du test du Khi-deux. Dans le cas détaillé ici, toutes les variables testées présentent un lien.
- Regarder la valeur du V de Cramer, afin de classer les variables, selon l'intensité de leur lien (plus le V de Cramer est élevé en valeur absolue, plus le lien est fort).
- La ligne « Cell Chi-square » indique la contribution de chaque modalité à la statistique du khi-deux. La somme des contributions donne la valeur la statistique du khi-deux (Chi-Square value dans les listing sas).
- Les lignes « Percent », « Row pct » et « Col pct », permettent de détecter les modalités présentant de meilleurs taux de réponse.

*86% des unités testées sont répondantes.*

| indic_rep | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|---|---|---|---|---|
| 0 | 1604 | 13.76 | 1604 | 13.76 |
| 1 | 10050 | 86.24 | 11654 | 100.00 |

Table of indic_rep by Paris

| indic_rep | Paris | | |
|---|---|---|---|
| Frequency<br>Expected<br>Deviation<br>Cell Chi-Square<br>Percent<br>Row Pct<br>Col Pct | 0 | 1 | Total |
| 0 | 1160<br>1267.9<br>-107.9<br>9.1816<br>9.95<br>72.32<br>12.59 | 444<br>336.11<br>107.89<br>34.636<br>3.81<br>27.68<br>18.18 | 1604<br><br><br><br>13.76<br><br> |
| 1 | 8052<br>7944.1<br>107.89<br>1.4654<br>69.09<br>80.12<br>87.41 | 1998<br>2105.9<br>-107.9<br>5.528<br>17.14<br>19.88<br>81.82 | 10050<br><br><br><br>86.24<br><br> |
| Total | 9212<br>79.05 | 2442<br>20.95 | 11654<br>100.00 |

**Avertissement :**
les tests du Khi-Deux et leurs contributions ne sont pas comparables d'un test à l'autre ! Le plus pertinent lorsqu'on souhaite procéder à une comparaison, est de s'intéresser aux résultats de chaque case / chaque modalité, par rapport à la moyenne de la colonne ou de la ligne (chiffres des marges « total »), et d'en conclure un comportement.

***Que dire sur les unités localisées à Paris (colonne Paris = « 1 ») ?***

*On voit que 444 sont non répondantes, alors qu'on en attendant 336 dans ce cas.*

*Leur contribution est de 34.6. Plus la contribution est élevée, plus ce type d'observations explique le lien du khi-deux détecté. A titre illustratif, une contribution peu être importante quand ce qu'on attendait est plus faible que la réalité.*

*En moyenne, 13.76% des unités sont non-répondantes (c'est-à-dire avec indic_rep = « 0 »). Et si on ne considère que les observations parisiennes, on constate que 18.18% d'entre elles n'ont pas répondu à l'enquête ; c'est donc plus élevé que la moyenne nationale.*

Statistics for Table of indic_rep by Paris

| Statistic | DF | Value | Prob |
|---|---|---|---|
| Chi-Square | 1 | 50.8110 | <.0001 |
| Likelihood Ratio Chi-Square | 1 | 47.9675 | <.0001 |
| Continuity Adj. Chi-Square | 1 | 50.3411 | <.0001 |
| Mantel-Haenszel Chi-Square | 1 | 50.8066 | <.0001 |
| Phi Coefficient | | -0.0660 | |
| Contingency Coefficient | | 0.0659 | |
| Cramer's V | | -0.0660 | |

Fisher's Exact Test

| | |
|---|---|
| Cell (1,1) Frequency (F) | 1160 |
| Left-sided Pr <= F | 2.624E-12 |
| Right-sided Pr >= F | 1.0000 |
| Table Probability (P) | 9.379E-13 |
| Two-sided Pr <= P | 4.383E-12 |

Sample Size = 11654

*Ici, toutes les variables soumises au test du khi-deux sont liées à l'indicatrice de comportement de réponse. Le V de Cramer permet de privilégier la variable relative à la « situation en n-1 » (V = 0.23).*
*Toutefois, la modalité des très grosses entreprises au lancement se démarque : 98% des très grandes entreprises sont répondantes (on rappelle toutefois que ce résultat est biaisé ; en effet, pour cette enquête, les grandes unités non répondantes se sont vues attribuer leurs résultats de l'enquête précédente par imputation. Cf. paragraphe 1 de ce chapitre). Il est donc décidé de faire un GRH avec les unités relevant de cette tranche. Les tests suivants ne prendre plus en compte ces entreprises-là.*
*Sans cette catégorie d'unités, la variable d'appartenance ou non à un groupe n'apparaît pas liée au comportement de réponse. Les autres variables, liées, sont classées selon leur V de Cramer. « Situation en n-1 » arrive en tête.*
*On peut confirmer et vérifier ces résultats en utilisant une PROC LOGISTIC.*

La première étape est donc de soumettre les données à un test du Khi-Deux. Ensuite, le statisticien peut mettre en place une **proc logistic**. Cette procédure de SAS permet de modéliser le comportement de réponse en raisonnant « toutes choses égales par ailleurs », ce qui n'est pas le cas d'un test du Khi-Deux. Avec les variables retenues suite au test du Khi-Deux, elle permet donc de détecter les variables qui influencent le comportement de réponse, tout en les classant. Elle fournit également des indications sur les modalités de ces variables qui se distinguent vis à vis du taux de réponse. Les résultats de cette procédure vont donc nous permettre de constituer, par étape, les GRH.
Le lecteur intéressé par cette méthode et la procédure associée, pourra entre autres se référer à la documentation suivante (dans outils, cliquer sur régression logistique).
*En reprenant l'exemple de l'enquête « TIC 2009 », on peut décrire les sorties obtenues :*

```
DATA    table_tic2; /* Nouvelle table de recherche, sans les unités du GRH 1 */
        SET table_tic;
        IF      class_effl ne "5";
RUN;

PROC LOGISTIC DATA = table_tic2 ;
        CLASS paris             (REF = '0')     / PARAM = reference;
        CLASS CLASS_sect        (REF = '07')    / PARAM = reference;
                                /* On préfère ici la classe "07" en référence, car son taux de
                                ** réponse est proche de la moyenne */
        CLASS tic08             (REF = '0')     / PARAM = reference;
        CLASS CLASS_effl        (REF = '2')     / PARAM = reference;
        MODEL indic_rep = paris CLASS_sect tic08 CLASS_effl / SELECTION = forward RSQUARE;
        TITLE 'Modélisation Logistique Forward';
RUN;
```

*Nota : la méthode utilisée est la méthode « forward » ( option « `/ SELECTION = forward` **»**).*
*Il est conseillé de comparer les résultats avec les méthodes « forward », « backward » et « stepwise ».*

Summary of Forward Selection

| Step | Effect Entered | DF | Number In | Score Chi-Square | Pr > ChiSq |
|---|---|---|---|---|---|
| 1 | tic08 | 2 | 1 | 595.0673 | <.0001 |
| 2 | Paris | 1 | 2 | 66.1098 | <.0001 |
| 3 | CLASS_EFFL | 3 | 3 | 29.8787 | <.0001 |
| 4 | CLASS_SECT | 12 | 4 | 30.4148 | 0.0024 |

*La PROC LOGISTIC fait bien intervenir tout d'abord la variable « situation en n-1 » dans le modèle. C'est le même résultat que celui obtenu par l'analyse du Khi-Deux. Ensuite, la variable de localisation parisienne serait la plus influente dans le comportement de réponse. Un test du Khi-Deux ultérieur va permettre de confirmer/infirmer cela, comme nous allons le voir.*

Analysis of Maximum Likelihood Estimates

| Parameter | | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq |
|---|---|---|---|---|---|---|
| Intercept | | 1 | -1.7082 | 0.0804 | 451.3888 | <.0001 |
| Paris | 1 | 1 | 0.5461 | 0.0736 | 55.1093 | <.0001 |
| CLASS_SECT | 03 | 1 | -0.1104 | 0.0869 | 1.6127 | 0.2041 |
| CLASS_SECT | 04 | 1 | -1.3157 | 1.0253 | 1.6467 | 0.1994 |
| CLASS_SECT | 05 | 1 | 0.1373 | 0.2962 | 0.2148 | 0.6430 |
| CLASS_SECT | 06 | 1 | -0.0466 | 0.1067 | 0.1905 | 0.6625 |
| CLASS_SECT | 08 | 1 | 0.1691 | 0.1209 | 1.9540 | 0.1622 |
| CLASS_SECT | 09 | 1 | 0.2983 | 0.1079 | 7.6474 | 0.0057 |
| CLASS_SECT | 10 | 1 | 0.0480 | 0.1277 | 0.1415 | 0.7068 |
| CLASS_SECT | 11 | 1 | -0.4384 | 0.1986 | 4.8742 | 0.0273 |
| CLASS_SECT | 12 | 1 | 0.0348 | 0.2109 | 0.0271 | 0.8691 |
| CLASS_SECT | 13 | 1 | -0.0319 | 0.1227 | 0.0675 | 0.7950 |
| CLASS_SECT | 14 | 1 | 0.1980 | 0.1356 | 2.1321 | 0.1442 |
| CLASS_SECT | 19 | 1 | 0.8004 | 0.5417 | 2.1836 | 0.1395 |
| tic08 | 1 | 1 | 1.4263 | 0.0871 | 268.3741 | <.0001 |
| tic08 | 2 | 1 | -0.6179 | 0.0671 | 84.8618 | <.0001 |
| CLASS_EFFL | 1 | 1 | 0.0947 | 0.0839 | 1.2734 | 0.2591 |
| CLASS_EFFL | 3 | 1 | 0.00293 | 0.0700 | 0.0018 | 0.9666 |
| CLASS_EFFL | 4 | 1 | -0.4204 | 0.1024 | 16.8539 | <.0001 |

*Dans le paragraphe intitulé « Analysis of Maximum Likelihood Estimates », on peut se référer à la dernière colonne, pour détecter les modalités présentant un lien avec l'indicatrice comportementale à modéliser.*
*Noter que le tableau ne liste pas les modalités désignées en référence.*
*Par ailleurs, la colonne « Estimate » indique si les unités de la modalité en question répondent bien ou pas.*

Odds Ratio Estimates

| Effect | | Point Estimate | 95% Wald Confidence Limits | |
|---|---|---|---|---|
| Paris | 1 vs 0 | 1.727 | 1.495 | 1.994 |
| CLASS_SECT | 03 vs 07 | 0.895 | 0.755 | 1.062 |
| CLASS_SECT | 04 vs 07 | 0.268 | 0.036 | 2.001 |
| CLASS_SECT | 05 vs 07 | 1.147 | 0.642 | 2.050 |
| CLASS_SECT | 06 vs 07 | 0.954 | 0.774 | 1.177 |
| CLASS_SECT | 08 vs 07 | 1.184 | 0.934 | 1.501 |
| CLASS_SECT | 09 vs 07 | 1.348 | 1.091 | 1.665 |
| CLASS_SECT | 10 vs 07 | 1.049 | 0.817 | 1.348 |
| CLASS_SECT | 11 vs 07 | 0.645 | 0.437 | 0.952 |
| CLASS_SECT | 12 vs 07 | 1.035 | 0.685 | 1.565 |
| CLASS_SECT | 13 vs 07 | 0.969 | 0.762 | 1.232 |
| CLASS_SECT | 14 vs 07 | 1.219 | 0.934 | 1.590 |
| CLASS_SECT | 19 vs 07 | 2.226 | 0.770 | 6.437 |

*En ce qui concerne la partie intitulée « Odds Ratio Estimates », le lecteur pourra se référer à la documentation citée pour l'interprétation en terme de rapports de probabilités de réponse. On peut juste noter ici que lorsque la valeur 1 n'est pas comprise dans l'intervalle de confiance cela signifie qu'il existe un écart significatif de probabilité de réponse par rapport à la modalité de référence (c'est le cas par exemple pour la localisation à Paris)*

*La variable « Tic08 » est donc la première variable que la procédure a entrée dans le modèle. Il s'agit de la première variable influente. On a donc trois groupes :*

- *sous-population 1 « unité interrogée répondante l'an passé »*
- *sous-population 2 « unité interrogée non répondante ou hors champ l'an passé »*
- *sous-population 3 « unité non interrogée l'an passé ».*

*A partir de là, il est d'usage de considérer les sous-populations de chacune des modalités de cette variable, et de procéder sur ces dernières, à des tests du Khi-Deux avec la deuxième variable détectée par la PROC LOGISTIC (en l'occurrence, la variable « Paris » dans notre exemple). Cela permet de voir si le fait d'être localisé à Paris influence ou non le taux de réponse.*

*Dans notre exemple, les unités de la sous-population 2 ne permettent pas de détecter un lien entre la localisation parisienne et le comportement de réponse. C'est ainsi qu'on obtient un GRH (le numéro 2 dans le tableau récapitulatif ci-après).*
*En s'intéressant à la sous-population 1, on établit un lien entre la variable « Paris » et le comportement de réponse, grâce à un test du Khi-Deux. Ce sous-groupe est donc scindé en deux nouvelles familles :*

- *famille 1 : les unités interrogées répondantes l'an passé, et localisées à Paris.*
- *famille 2 :les unités interrogées répondantes l'an passé, et non localisées à Paris.*

**Que faire si aucun lien n'est détecté ?**

Dans ce cas, on essaye de regrouper certaines modalités de variables potentiellement explicatives (selon des taux de réponse similaires par modalité), et on recommence la procédure de tests…

*A ce stade, on réitère la procédure : sur ces familles, on utilise le Khi-Deux pour détecter d'éventuels liens entre la 3° variable influente selon la PROC LOGISTIC (« class_effl »), et le comportement de réponse.*
*Aucun lien n'est révélé dans la famille 2. Cela permet donc d'obtenir un nouveau GRH (le numéro 3 dans le tableau récapitulatif ci-après).*
*Par contre, un lien est avéré dans la famille 1. On passe donc à la constitution de nouvelles sous-divisions, en s'aidant des modalités de la variable « class_effl ». Et ainsi de suite…*

Parallèlement à cela, l'utilisateur prendra garde à divers caractéristiques des groupes qu'il construit :

- la **taille du groupe** : ne pas envisager de GRH trop petit.
- les **tailles des groupes** : éviter les déséquilibres de tailles entre les groupes.

- le **nombre de groupes** : si les groupes sont trop nombreux, regrouper les GRH proposant des taux de réponse similaires, sous réserve toutefois que leurs caractéristiques aient une signification connexe (un groupe d'unités relevant de l'industrie automobile s'appariera mieux à un groupe d'unités de secteurs industriels que par exemple de la finance).

Finalement nous arrivons aux GRH qui sont décrits dans le tableau suivant :

***Exemple :*** *Tableau récapitulatif des GRH obtenus sur l'enquête annuelle d'entreprises « TIC 2009 » :*

| GRH | *Situation concernant TIC 2008* | *Localisation à Paris (dép 75, 92, 93)* | *Tranche d'effectif au lancement* | *Secteur d'activité* | Taille du GRH | Taux de réponse du GRH (en %) | Coefficient de correction |
|---|---|---|---|---|---|---|---|
| **1** | | | *500 ou plus* | | 1813 | 97.81 | 1.0224 |
| **2** | *Interrogée, hors champ ou NRT* | | *Moins de 500* | | 682 | 47.72 | 2.0954 |
| **3** | *Interrogée répondante* | *Non* | *Moins de 500* | | 3128 | 92.52 | 1.0809 |
| **4** | *Interrogée répondante* | *Oui* | *250 à 499* | | 161 | 93.13 | 7.0737 |
| **5** | *Interrogée répondante* | *Oui* | *Moins de 250* | | 548 | 84.41 | 1.1846 |
| **6** | *Non interrogée* | *Non* | *250 à 499* | | 454 | 90.74 | 1.1020 |
| **7** | *Non interrogée* | *Non* | *Moins de 250* | *Industrie manufacturière, industrie auto, énergie, finances et assurances* | 1173 | 86.86 | 1.1513 |
| **8** | *Non interrogée* | *Non* | *Moins de 250* | *Construction, Commerce, Transport, Immobilier* | 1934 | 83.69 | 1.1948 |
| **9** | *Non interrogée* | *Non* | *Moins de 250* | *Services aux particuliers et aux entreprises* | 858 | 79.32 | 1.2607 |
| **10** | *Non interrogée* | *Oui* | *Moins de 500* | *Industrie manufacturière industrie auto, énergie, finances et assurances* | 175 | 82.84 | 1.2138 |
| **11** | *Non interrogée* | *Oui* | *Moins de 500* | *Construction, Commerce, Transport, Immobilier* | 254 | 69.19 | 1.4452 |
| **12** | *Non interrogée* | *Oui* | *Moins de 500* | *Services aux particuliers et aux entreprises* | 481 | 93.84 | 1.3543 |
| ***Total*** | | | | | ***11 661*** | | |

*Nota : ce tableau ne présente que les unités intervenant dans la repondération, mais l'échantillon total regroupe davantage d'observations, notamment les unités hors champ ou celles à poids initial de 1 qu'on ne souhaite pas repondérer. Il ne faut pas oublier ces unités-là dans la table finale !*
*Note de lecture : l'obtention des deux dernières colonnes est expliquée dans le paragraphe à venir.*

## c. L'étape de repondération

Une fois la proposition de découpage en GRH effectuée, le **taux de réponse** de chaque groupe peut être calculé. Ce taux est en fait le rapport entre le nombre de répondants du groupe, et l'effectif total du groupe (le nombre d'entreprises interrogées actives[6]).
Si, dans un groupe, ce taux s'avère trop faible, le découpage des GRH devra être reconsidéré. Une question importante est celle de la pondération : faut-il calculer un taux de réponse non pondéré par GRH ou pondéré en utilisant les poids de lancement ? Pour l'enquête TIC 09, nous avons privilégié un taux de réponse pondéré afin que par GRH la somme des poids avant correction de la non-réponse et après soit équivalente.

**Calcul des taux de réponse avec SAS**

```
PROC FREQ DATA = table_donnees;
TABLE GRH * indic_rep / outpct OUT= table1;
WEIGHT= poids_l;
RUN;
DATA table2;
        SET     table1;
        KEEP    GRH tx_reponse;
        WHERE   indic  rep = "1";
        tx_reponse = pct_row / 100;
        /* taux en décimales et non en % */
RUN;
```

Le **coefficient de correction** s'obtient ensuite en inversant ce taux de réponse. Ce coefficient est encore appelé « coefficient de dilatation ».

On doit opérer ensuite à une phase de repondération par GRH. Après avoir constitué les groupes de réponses homogènes, l'intérêt de cette étape est de **modifier le jeu de poids.** En fait, il s'agit de modifier judicieusement le poids de sondage des individus répondants. Ceci corrige la non-réponse

[6] cette grandeur peut-être pondérée ou non ; les deux méthodes sont en effet possibles.

totale proprement dite. Pour ce faire, on utilise le taux de réponse calculée précédemment : les « nouveaux » poids sont obtenus en divisant le poids au lancement par cette estimation de la probabilité de réponse, c'est-à-dire en multipliant le poids au lancement par le coefficient de correction.

⇨ **Pour résumer**, si on appelle « poids_nr » le poids tenant compte de la correction de la non réponse totale, alors :

| Valeur de poids_nr | Unités concernées |
|---|---|
| Poids au lancement | Unités pour lesquelles ETAT ne vaut ni « R » ni « N » |
| 1 | Unités non substituables, profilées,… (celles pour lesquelles le poids de lancement est égal à l'unité). |
| (Poids au lancement)/(taux de réponse du grh) c'est-à-dire (Poids au lancement) x (coefficient de correction) | Les autres unités. |

## d. Impacts sur la table de données

Finalement, on peut extraire de la table de données, les unités non répondantes totales (ETAT = « N »), leur poids ayant été redistribué sur les unités répondantes du champ. Si on décide de les conserver dans l'ensemble des données, il faut mette leur poids final à 0, et ne jamais oublier la pondération dans les traitements SAS ultérieurs.

On ne dispose donc plus, dans la table, que de trois catégories d'unités enquêtées :

- les hors-champ (ETAT = « H »), qui ont conservé leur poids de lancement ;
- les répondantes (ETAT = « R »), dont le poids a été modifié ;
- les unités ne représentant qu'elles-mêmes (poids initial et final égaux à l'unité).

L'utilisateur pourra s'intéresser à la distribution des poids, avant et après correction de la non réponse totale. Il remarquera que les poids ont généralement augmenté, sauf dans la situation très particulière où le GRH aurait un taux de réponse maximal (100%). Il portera son attention sur les poids aux valeurs les plus extrêmes.

**Unités basculées de non-répondantes totales à non-répondantes partielles :**

Certaines unités très particulières, non répondantes totales, peuvent gagner à être considérées en unités non-répondantes partielles.

*A titre d'exemple, les entreprises échantillonnées dans TIC 2009 telles que :*

- *non répondantes en 2009*
- *figurant dans l'échantillon 2008*
- *d'effectifs au lancement élevé*
- *n'étant pas « non substituables »*

*ont été basculées en non-réponse partielle, et on donc été traitées comme des unités imparfaitement répondantes.*
*Pour cela, leurs données déclarées en 2008 ont été récupérées ; puis les réponses encore manquantes ont été obtenues par une imputation ordinaire.*

L'intérêt de cette démarche est de privilégier la déclaration à l'enquête précédente pour des grandes unités qui ne représentent qu'elles-mêmes (poids de 1) plutôt que chercher à les repondérer. Cela est possible du fait du caractère annuel de l'enquête.

***L'étape de correction de la non-réponse totale a modifié le jeu de poids. Toutefois il sera nécessaire de réaliser un calage afin de respecter la structure de la population selon certains critères, structure qui n'est plus forcément respectée à la fin des opérations de correction de non-réponse.***

# IV. Calage

***L'étape qu'on s'apprête à aborder ici, permet de rétablir un jeu de poids en adéquation avec la population étudiée.***

## *1. Intérêt du calage*

Les méthodes de calage d'un l'échantillon consistent à changer les poids - poids suite à correction de la non-réponse totale dans le cas présent - pour que les estimations de totaux de variables quantitatives soient égales aux vrais totaux connus par ailleurs sur la population. Nous pouvons également caler afin que les estimations d'effectifs de modalités de variables catégorielles soient égales aux vrais effectifs connus sur la population.

Par exemple, nous décrivons ci-après le calage « naturel » le plus simple que l'on souhaite effectuer, afin de retrouver le nombre d'entreprises présent dans la base de sondage. Au démarrage de l'enquête, la somme des poids de lancement, par strate de tirage de l'échantillon, est théoriquement égale à la population totale (de la base de sondage) que l'on souhaite étudier.
Un calage s'impose lorsque la somme des poids utilisés, ne permet plus d'obtenir l'évaluation de la population d'origine.
En effet, le jeu de poids de l'échantillon a pu évoluer depuis le début des opérations. Le calage permet de remettre l'échantillon en adéquation avec la structure de la population.
Ainsi, le calage va créer un nouveau - et dernier - système de poids (que l'on appellera « poids_cal »), de telle sorte que leur somme soit à nouveau égale à la population étudiée, dans son ensemble et par groupe à définir (par strate de calage).
Il est à noter que le calage peut être effectué, même en présence de non-réponse totale.

## *2. Préparation de la table*

L'étape de calage décrite précédemment peut se résumer à une post-stratification, en utilisant un coefficient qui soit le ratio suivant :

$$\frac{Nombre\_d\_entreprises\_pondéréde\_la\_post\_strate}{Ensemble\_des\_poids\_de\_la\_psot\_strateaulancement}$$

L'utilisateur peut également mettre en œuvre la procédure de calage via l'utilisation de la macro « CALMAR » (CALage sur MARges). Cet outil est l'objet des paragraphes suivants.

Quoi qu'il en soit, en amont de ces opérations, il faut donc définir la ou les variables de calage ainsi que les sous-populations sur lesquelles on souhaite caler. Ce choix est de la responsabilité de la maîtrise d'ouvrage de l'enquête.

*Pour l'enquête TIC 2009 par exemple, la maîtrise d'ouvrage a décidé de caler sur les strates de diffusion au sens d'Eurostat uniquement sur la variable « nombre d'entreprises ».*

On veillera par ailleurs à écarter des opérations à venir, les unités qui ne représentent qu'elles-mêmes et doivent donc avoir un poids final de un (non substituables, profilées,…).
Il est par contre important que les hors-champ participent au calage. En effet, les unités déclarées hors champ de la population de référence suite à l'enquête, sont sensées représenter des entreprises de la base de sondage initiale, qui, si elles avaient fait partie de l'échantillon, auraient aussi été déclarées hors-champ par la suite.
L'intérêt de du calage sur marges est qu'après cette phase, on ait, sur chaque groupe (post-strate) choisi, une somme des poids issus du calage (sur les unités répondantes et hors-champ), identique à la somme des poids initiaux (sur l'ensemble des unités échantillonnées), c'est-à-dire identique au vrai effectif de la population étudiée.

## 3. Outil : la macro CALMAR

Le statisticien est invité ici à utiliser la macro « CALMAR » disponible sur insee.fr. Il trouvera également de la documentation sur ce programme, ainsi que son mode d'emploi.

*A titre d'illustration, voici les étapes du calage de l'enquête TIC 2009 :*

```
/*** Table de base pour calculer les "marges" ;
Calage sur les strates de diffusion Eurostat . ***/

data tablecal;
      SET tic;
      keep siren l1_nomen poids_l effl strate_l nent strate_c;
      IF SUBSTR(strate_l,1,6) = "D10A12"     THEN strate_c = "01";
      IF SUBSTR(strate_l,1,6) = "D13A15"     THEN strate_c = "02";
      /*** ETC… jusqu'à strate_c = « 29 » (Code omis par gain de place) ***/

      if nsub = 1 or nrp = 1 then delete;
      /* Population sur laquelle on cale: écarter les non substituables (dont
      profilées) et les ** unités non répondantes totales passées en non
      répondantes partielles, mais conserver les ETAT = "H" (unités hors champ). */
      nent=1;
run;

/* Avant de poursuivre, vérifier à ce niveau qu'il y a bien assez d'unités par
strate de calage. ** Si trop peu d'observations (ex: 3): regrouper des strates
adjacentes.
** De même, si au final des poids anormalement élevés sont obtenus sur les plus
petites strates, ** faire des regroupements. */

/*** Les marges : élaboration de la table ***/

proc means data=tablecal sum;
      var   nent;
      weight      poids_l;
      class strate_c;
      title "calage sur nombre d'entreprises";
      output out=essai sum=snent;
run;

proc means data=tablecal sum;
      var poids_l;
run;

data essai1;
      set essai (where=(_type_=1));
      nent= round(snent);
      keep strate_c nent;
run;

proc sort data=essai1; by strate_c; run;
proc transpose data=essai1 out=essai2 name=var prefix=mar label=strate; run;

data essai2;
      set essai2;
      n=29 /* nombre de strate_c créées */;
      var="strate_c";
      keep var n mar1-mar29;
run;

/*** Table pour le calage ***/

data verif;
      SET tic;
      keep siren strate_c poids_nr poids_l grh;
      IF SUBSTR(strate_l,1,6) = "D10A12"     THEN strate_c = "01";
      IF SUBSTR(strate_l,1,6) = "D13A15"     THEN strate_c = "02";
      /*** ETC… code identique à précédemment, omis par gain de place. ***/
```

```
      if nsub = 1 or nrp = 1 then delete;
run;

proc sort data=verif; by strate_c; run;

/*** Charger à blanc le code de la MACRO CALMAR (disponible sur www.insee.Fr) */
/*** Appel de la macro calmar : ***/

options mprint;
%calmar(            data         = verif,
                    M            = 2,
                    poids        = poids_nr,
                    poidsfin     = poids_cal,
                    datapoi      = sortie,
                    editpoi      = oui,
                    ident        = siren,
                    datamar      = essai2)

/*** Table finale ***/
proc sort data=sortie;    by siren; run;
proc sort data=verif;           by siren; run;
data final;
      merge verif sortie;
      by siren;
      if poids_cal < 1 then poids_cal = 1;
      keep siren poids_cal strate_c;
run;

/*** Fusion avec table complète ***/
proc sort data=final; by siren; run;
proc sort data=tic; by siren; run;
data tic_calee;
      merge tic(in=a) final;
      by siren;
      if a;
      if missing(poids_cal) then poids_cal=poids_nr;
run;
```

La sortie SAS et son commentaire figurent page suivante:

```
**********************
***     BILAN      ***
**********************

Date : 24 AOUT 2009                Heure : 13:29

Table en entrée : VERIF

Nombre d'observations dans la table en entrée  : 12016
Nombre d'observations éliminées                : 1611
Nombre d'observations conservées               : 10405

Variable de pondération : POIDS_NR

Nombre de variables catégorielles : 1
Liste des variables catégorielles et de leurs nombres de modalités :
  STRATE_C (29  )
Taille de l'échantillon (pondéré) : 166977
Taille de la population           : 166979

Méthode utilisée : raking ratio
Le calage a été réalisé en 5 itérations
Les poids ont été stockés dans la variable POIDS_CAL de la table SORTIE
```

***Que comprendre sur cette sortie SAS ?***

*Il s'agit là du bilan fourni par la macro CALMAR, appliquée à l'enquête TIC 2009 déjà traitée auparavant.*

*Les 1611 observations éliminées englobent uniquement les unités non répondantes, écartées de la phase de calage (au préalable du calage, le poids_nr de ces entreprises - poids issu du traitement des GRH - a été forcé à zéro). Au lieu de forcer leur poids à zéro, ces unités auraient pu être supprimées de la table avant l'opération de calage (il est d'ailleurs préférable de le faire).*
*Dans le cas où une unité ne présenterait pas de modalité pour une variable de calage, cette unité serait placée sur cette ligne (d'où l'intérêt de devoir supprimer au préalable les 1611 observations précitées ; cela permet de mieux détecter d'éventuelles anomalies).*
*On voit bien, par ailleurs, que la ligne « observations en entrée » est la somme des deux lignes suivantes.*

*La taille de l'échantillon pondéré (166977) est la somme de la variable de calage « poids_cal », sur les 29 strates de calage de la table de marges. Il s'agit de la somme des effectifs de toutes les marges que l'utilisateur a introduites au début de CALMAR, autrement dit de l'effectif de la population qui résulte de ces marges.*

*Après cette étape, voilà les résultats que l'on obtient sur la table entière :*
*Somme des poids de lancement = 167 350. Il s'agit de la somme des poids en entrée de CALMAR.*
*Somme des poids calés = 167 352.*
*L'écart de 2 unités est acceptable, et certainement le fait d'arrondis. Ces deux sommes sont proches, le calage s'est bien effectué. Dans le cas d'une différence importante, l'utilisateur cherchera à expliquer ceci, en tentant de trouver l'origine de l'écart*
*On peut juger du bon déroulement du calage, en s'intéressant au critère de convergence : le calage s'est effectué en 5 itérations ici. Si convergence il y a, c'est que le calage s'est bien déroulé.*

*Par ailleurs, considérons la différence entre la somme des poids_cal du bilan (166979) et celle des poids calés sur la table finale (167352). L'écart s'élève ici à 373, et correspond aux poids des entreprises exclues du calage car leur poids de lancement est conservé : les non substituables (nsub = 1), et les non répondantes totales basculées en non répondantes partielles (nrp = 1).*

En complément de ce bilan édité par SAS, l'utilisateur pourra se référer à une autre sortie, celle du rapports des poids par strate de calage. Il est à noter que cette vérification n'a pas pour but de déterminer si l'étape de calage s'est bien déroulée, mais plutôt de détecter des changements de poids importants, qu'il faut s'expliquer.

A titre d'exemple, commentons la sortie SAS suivante :

calage sur nombre d'entreprises
13:22 Thursday, September 2
Méthode : raking ratio
Rapports de poids (pondérations finales / pondérations initiales)
pour chaque combinaison de valeurs des variables

| Obs | strate_c | Effectif combinaison | Rapport de poids |
|---|---|---|---|
| 1 | 01 | 737 | 0.98363 |
| 2 | 02 | 168 | 1.03350 |
| 3 | 03 | 270 | 1.02620 |
| 4 | 04 | 402 | 0.96710 |
| 5 | 05 | 519 | 0.99660 |
| 6 | 06 | 150 | 1.02998 |
| 7 | 07 | 1134 | 0.97973 |
| 8 | 08 | 723 | 1.02645 |
| 9 | 09 | 301 | 0.94029 |
| 10 | 10 | 241 | 0.98532 |
| 11 | 11 | 737 | 0.98732 |
| 12 | 12 | 498 | 0.98320 |
| 13 | 13 | 902 | 0.97157 |
| 14 | 14 | 795 | 1.00688 |
| 15 | 15 | 335 | 0.99253 |
| 16 | 16 | 416 | 1.07187 |
| 17 | 17 | 84 | 0.98546 |
| 18 | 18 | 204 | 1.06316 |
| 19 | 19 | 31 | 1.15487 |
| 20 | 20 | 15 | 1.76814 |
| 21 | 21 | 56 | 1.13327 |
| 22 | 22 | 156 | 0.95250 |
| 23 | 23 | 188 | 0.92974 |
| 24 | 24 | 14 | 1.08183 |
| 25 | 25 | 478 | 1.03776 |
| 26 | 26 | 119 | 1.00866 |
| 27 | 27 | 309 | 0.95181 |
| 28 | 28 | 159 | 1.02607 |
| 29 | 29 | 264 | 0.97817 |

***Que comprendre sur cette sortie SAS ?***

*Il faut repérer les rapports de poids éloignés de l'unité. En effet, dans le cas d'un rapport de poids trop faible, cela signifie que les unités de la strate concernée, ont un poids final (calé) bien inférieur à leur poids initial.*

*Ici, on souligne que la strate strate_c = « 20 », présente un rapport de poids élevé (1.76). Les poids calés des unités de ladite strate sont donc beaucoup plus importants que les poids initiaux. Il revient à l'utilisateur d'expliquer pourquoi, et de détecter une éventuelle anomalie.*

*L'étude des unités de la strate 20 montre qu'il s'agit d'une petite strate : 21 observations, réparties entre 15 répondantes du champ, et 6 non répondantes du champ. Le rapport des poids élevé ici, s'explique par la petite taille de cette strate, et la proportion - importante - de non répondantes, par rapport aux répondantes. Ce phénomène est à imputer au fait que les strates de correction de la non réponse (les GRH), ne sont pas identiques aux strates de calage.*
*Il aurait été ici plus judicieux, de créer une strate de calage plus « grosse » que la numéro « 20 » dont nous disposons.*

La macro SAS de calage minimise les écarts entre poids initiaux et poids calés, mais fournit parfois des poids calés inférieurs à l'unité, ce qui a certes, un sens statistique, mais pas beaucoup de sens « commun » (comment une entreprise ne peut-elle même pas se représenter elle-même entièrement ?). Ces résultats, peu réalistes, doivent alors être rectifiés « à la main », comme cela est expliqué dans le paragraphe ci-dessous.

## 4. Ajustements

Les procédures automatiques de calage autorisent la création de poids calés inférieurs à l'unité ; une telle situation est difficilement explicable aux utilisateurs (comment un individu, une entreprise, peuvent-ils représenter moins qu'eux-mêmes ?). Cela donne lieu à des ajustements, afin d'obtenir des poids au moins égaux à l'unité d'une part, et à rétablir l'équation des poids initiaux/calés d'autre part.

➢ Si l'enquête concerne plusieurs milliers d'unités, l'utilisateur pourra uniquement et simplement bloquer à l'unité, les poids calés inférieurs à 1. En lien avec le nombre d'entreprises concernées, cela ne change quasiment rien pour l'utilisateur. A cette échelle, de tels ajustements auront des répercussions négligeables sur l'ensemble de la table.

➢ Si l'enquête concerne beaucoup moins d'unités, l'utilisateur, après avoir rehaussé à l'unité les poids calés inférieurs à 1, devra entreprendre une étape supplémentaire d'ajustement. En effet, suite à cette première modification, il pourra s'avérer, pour certaines strates de calage, que la somme des poids initiaux de lancement, soit alors différente de la somme des poids calés.

⇨ Ainsi, l'ultime variable de poids issue de la phase de calage est obtenue. C'est cette variable qu'il faudra utiliser dans tous les traitements de diffusion ultérieurs.

***A ce stade, le redressement est terminé. Le calage a permis d'obtenir un nouveau jeu de poids, cohérent. Si l'imputation de la non réponse partielle n'a pas encore été effectuée, il est temps de la mettre en place maintenant. Sinon, l'utilisateur peut entamer les travaux annexes à la phase de redressement, qui permettront de finaliser, documenter, et transmettre le fichier de l'enquête à la maitrise d'ouvrage.***

# V. Vers le fichier définitif

***Tous les travaux de redressement (correction de la non-réponse et calage) ont été mis en œuvre un à un. Il s'agit donc à présent de vérifier qu'ils n'ont pas créé d'incohérence dans la table de données ; en effet, pris dans leur ensemble, ils ne doivent pas, malgré tous les soins du statisticien, avoir abouti à des réponses contradictoires.***
***Une fois cette ultime étape d'ajustements achevée, la table de données est destinée à être exploitée dans le cadre des études. A cet effet, il faut mettre à disposition de tout utilisateur, une documentation complète et détaillée, des différents traitements effectués. Cela permettra d'éviter toute exploitation non adaptée de la table.***
***Enfin, le bilan qualité clôt les travaux relatifs à l'enquête. Il reprend toutes les étapes successives de la collecte au redressement, et en fait une étude objective quant aux moyens et méthodes employés.***

## *1. Derniers ajustements*

Afin de mettre une table à disposition, il est recommandé d'effectuer certains tests de cohérences sur les données ; en effet, il se peut que les travaux relatifs au redressement aient généré des informations contradictoires. C'est possible également, que certaines étapes se soient mal déroulées et que, du fait de l'ampleur de la table, le statisticien n'ait pas perçu cela.

Différentes pistes de vérifications sont proposées ici :

- Les **règles d'apurement** doivent être à nouveau soumises aux données ; cela permettra de repérer si des données imputées séparément, sont à l'origine d'incohérences lorsqu'on considère ces données dans leur globalité.

- Un test simple consiste à **croiser chaque variable du questionnaire avec sa variable de statut**. Cela permet de vérifier que le redressement de la non-réponse partielle s'est bien déroulé. Ainsi, les statuts de modalité « M » doivent finalement tous correspondre à une modalité de la variable ; en effet, désormais, seuls les statuts de modalité « N » coïncident avec une variable vide, « à blanc ». Dans le cas des « N », on pourra également porter un intérêt particulier à la gestion des filtres. Enfin, les statuts à modalité « B » doivent toujours se rapporter à des réponses de la variable.

- Il est aussi intéressant de mesurer **l'impact du redressement**. A titre d'explication, nous utiliserons des exemples.

*Impact du redressement de la non-réponse partielle :*

Il s'agit de savoir dans quelle mesure les réponses imputées, influencent les résultats. Pour ce faire, on compare le résultat d'une variable, en utilisant d'abord uniquement les observations pour lesquelles on a une réponse brute (statuts à « B »), puis en considérant les observations à réponse brute ET redressée (statuts à « B » et « M »).

```
/* Impact redressement variable qualitative */

TITLE "Question G1, réponses brutes";
PROC FREQ DATA = table_tic;
    TABLE G1_VENTES;
    WHERE G1_VENTES_S = "B";
    WEIGHT poids_cal;
RUN;
TITLE;
```

Question G1, réponses brutes
11:01 Tuesday, October

The FREQ Procedure

| G1_VENTES | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|---|---|---|---|---|
| 1 | 20009.67 | 13.05 | 20009.67 | 13.05 |
| 2 | 133335.7 | 86.95 | 153345.4 | 100.00 |

```
TITLE "Question G1, impact du redressement";
PROC FREQ DATA = table_tic;
    TABLE G1_VENTES;
    WHERE G1_VENTES_S in ("B" "M");
    WEIGHT poids_cal;
RUN;
TITLE;
```

Question G1, impact du redressement
11:01 Tuesday, October

The FREQ Procedure

| G1_VENTES | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|---|---|---|---|---|
| 1 | 20580.53 | 13.17 | 20580.53 | 13.17 |
| 2 | 135655.6 | 86.83 | 156236.1 | 100.00 |

*⇨ On voit dans cet exemple, que si on considère uniquement les déclarations des unités interrogées, on trouve 13.05 % des entreprises effectuant des ventes électroniques. Cette répartition est quasi identique (13.17 %) lorsqu'on ne distingue pas l'origine des réponses (brutes et imputées).*

*💣 Des résultats éloignés, doivent alerter le statisticien, et l'inciter à expliquer la différence observée. Un décalage n'est pas forcément signe d'une erreur dans le déroulement des étapes de correction de la non-réponse, dans la mesure où il peut être expliqué. Il doit être justifié.*
*Inversement, des résultats voisins ne sont pas le gage d'un parfait déroulement des opérations.*

```
/* Impact redressement variable quantitative */

TITLE "Question A2 en valeur, réponses brutes";
PROC MEANS DATA = table_tic MEAN MEDIAN SUM N MIN MAX;
    VAR A2_EMP_ORDI_VAL;
    WHERE A2_EMP_ORDI_VAL_S = "B";
    WEIGHT poids_cal;
RUN;
TITLE;
```

Question A2 en valeur, réponses brutes 15
11:01 Tuesday, October 20, 2009

The MEANS Procedure

Analysis Variable : A2_EMP_ORDI_VAL

| Mean | Median | Sum | N | Minimum | Maximum |
|---|---|---|---|---|---|
| 26.3016042 | 7.0000000 | 3633596.35 | 7738 | 1.0000000 | 70000.00 |

```
TITLE "Question A2 en valeur, impact du redressement";
PROC MEANS DATA = table_tic MEAN MEDIAN SUM N MIN MAX;
    VAR A2_EMP_ORDI_VAL;
    WHERE A2_EMP_ORDI_VAL_S in ("B" "M");
    WEIGHT poids_cal;
RUN;
TITLE;
```

```
Question A2 en valeur, impact du redressement                    16
                                          11:01 Tuesday, October 20, 2009

                    The MEANS Procedure

              Analysis Variable : A2_EMP_ORDI_VAL

      Mean          Median             Sum         N         Minimum         Maximum
-------------------------------------------------------------------------------------
27.6580009       7.0000000       3924937.02      8014       1.0000000       70000.00
-------------------------------------------------------------------------------------
```

⇨ *Là encore on remarque que le redressement n'a pas modifié de façon exagérée les résultats. Selon qu'on prenne ou non en compte les effets de l'imputation, on a en moyenne respectivement presque 28 et un peu plus de 26 salariés par entreprise qui utilisent un ordinateur. La médiane, paramètre de position beaucoup moins sensible que la moyenne, ne varie pas. L'essentiel est aussi de s'assurer que le redressement n'a pas généré une valeur extrême, un point aberrant (auquel cas, il faudrait soit l'ajuster, soit l'expliquer). Ici, les bornes n'ont pas changé.*

*Impact du redressement de la non-réponse totale :*

L'effet du redressement de la non-réponse totale, s'observe en modifiant la pondération. On effectue les calculs d'abord avec le poids au lancement, puis - toutes autres choses égales par ailleurs -, avec le poids obtenu suite au travail sur les GRH.

```
/* Impact du redressement de la non réponse totale */

TITLE "Question A2, pondération de lancement";
PROC MEANS DATA = table_tic MEAN MEDIAN;
 VAR A2_EMP_ORDI_VAL;
 WEIGHT poids_l;
RUN;
TITLE;
```

```
Question A2, pondération de lancement
                                          11:01 Tues

                    The MEANS Procedure

              Analysis Variable : A2_EMP_ORDI_VAL

            Mean          Median
-----------------------------------
      41.8522167       8.0000000
-----------------------------------
```

```
TITLE "Question A2, pondération issue de la correction de la non
réponse totale";
PROC MEANS DATA = table_tic MEAN MEDIAN;
 VAR A2_EMP_ORDI_VAL;
 WEIGHT poids_nr;
RUN;
TITLE;
```

```
Question A2, pondération issue de la correction de la non réponse totale
                                          11:01 Tuesday, October 20, 2

                    The MEANS Procedure

              Analysis Variable : A2_EMP_ORDI_VAL

            Mean          Median
-----------------------------------
      38.6068932       8.0000000
-----------------------------------
```

⇨ *Ici, l'impact est perceptible, mais il reste minime. De presque 42 salariés par entreprise utilisant un ordinateur, on passe à près de 39.*

*Modification des poids de calage :*

```
/* Modification des poids de calage */

TITLE "Question A2, pondération issue de la correction de la non
réponse totale";
```

```
PROC MEANS DATA = table_tic MEAN MEDIAN;
 VAR A2_EMP_ORDI_VAL;
 WEIGHT poids_nr;
RUN; TITLE;
```

```
Question A2, pondération issue de la correction de la non réponse totale
                                                    11:01 Tuesday, October 20, 2

                        The MEANS Procedure

                Analysis Variable : A2_EMP_ORDI_VAL

                      Mean           Median
                ------------------------------
                38.6068932        8.0000000
```

```
TITLE "Question A2, pondération suite au calage";
PROC MEANS DATA = table_tic MEAN MEDIAN;
 VAR A2_EMP_ORDI_VAL;
 WEIGHT poids_cal;
RUN; TITLE;
```

```
                Question A2, pondération suite au calage
                                                    11:01 Tuesda

                        The MEANS Procedure

                Analysis Variable : A2_EMP_ORDI_VAL

                      Mean           Median
                ------------------------------
                38.5040529        8.0000000
```

⇨ *Contrairement à la comparaison précédente, les résultats doivent ici peu fluctuer selon qu'on utilise l'une ou l'autre des deux pondérations. Dans l'exemple ci-dessus, les résultats sont similaires.*

Dans tous les travaux qui concernent l'impact du redressement, il est important de pouvoir expliquer les différences de résultats entre la situation avant la correction de la non-réponse et la situation après cette étape. Ces différences s'expliquent en lien avec le profil des non-répondants et le modèle de correction utilisé. Pour l'utilisateur, ces explications sont importantes.

- Les **contributions** pour quelques variables cibles peuvent aussi être calculées. Il s'agit de repérer les unités participant de façon importante à l'obtention des résultats d'une variable quantitative. L'importance de l'entreprise (directement ou via son poids) peut justifier une forte contribution ; ses réponses méritent peut-être toutefois d'être réajustées. Si l'entreprise est de très petite taille et avec un poids faible, il s'agit par contre certainement d'une erreur d'imputation ou de saisie, qu'il faut retrouver et corriger. Dans l'analyse des contributions, il faut bien distinguer les réponses « brutes » des réponses « corrigées » (suite à la correction de la non-réponse). Les fortes contributions de réponses brutes peuvent et doivent se détecter pendant la phase de collecte. Après la correction de la non-réponse, il faut surtout porter une attention aux données « corrigées » (statut à « M »). Toutefois une dernière vérification sur les réponses « brutes » peut s'avérer importante du fait des changements de poids après la correction de la non-réponse totale et le calage.

*Exemple : dans l'enquête TIC, les contributions sont calculées sur les variables relatives au commerce électronique. Ici, on traite la variable G2_vent_web_val :*

```
TITLE "Calcul de l agrégat";
PROC MEANS DATA = table_tic SUM;
     VAR G2_VENT_WEB_VAL;
     WEIGHT poids_cal;
RUN;
TITLE; /* La somme s'élève à 82842847528 */

DATA table_tic;
     SET table_tic;
     IF not missing(G2_VENT_WEB_VAL)
     THEN ctr_web = (G2_VENT_WEB_VAL * 100 *
poids_cal)/82842847528;
RUN;
```

```
/* Vérifier par ailleurs que la somme des contributions
vaut 100% */
```

Observations à contributions les plus élevées
11:01 Tuesday, Oct

| ctr_web | ETAT | G2_VENT_WEB_VAL | G2_VENT_WEB_VAL_S | poids_cal |
|---|---|---|---|---|
| 2.56765 | R | 250431292 | N | 8.494 |
| 2.51307 | N | 2081900000 | | 1.000 |
| 2.11022 | R | 1748166926 | N | 1.000 |
| 2.07392 | R | 76583982 | N | 22.434 |
| 1.99668 | R | 1654110000 | N | 1.000 |
| 1.78231 | R | 1476520000 | N | 1.000 |
| 1.63598 | R | 1355293040 | N | 1.000 |
| 1.61700 | R | 51023125 | N | 26.254 |
| 1.46615 | R | 1214604214 | N | 1.000 |
| 1.23599 | R | 1023931400 | M | 1.000 |
| 1.10007 | R | 911328000 | B | 1.000 |
| 0.87610 | R | 725786111 | M | 1.000 |
| 0.85729 | R | 338136900 | N | 2.100 |
| 0.83105 | R | 4731168 | N | 145.517 |
| 0.78809 | R | 634188600 | M | 1.029 |
| 0.73730 | R | 87000000 | N | 7.021 |
| 0.72891 | R | 603846100 | N | 1.000 |
| 0.70459 | R | 100561990 | N | 5.804 |
| 0.68239 | R | 565310000 | N | 1.000 |
| 0.67598 | R | 560000000 | N | 1.000 |

⇨ *L'attention du statisticien se portera sur les unités à contributions importantes (ctr_web > 1%). Un retour aux données lui permettra de comprendre ce phénomène. Si le statut de la variable est à « B », il s'agit d'une réponse de l'unité enquêtée ; ceci ne peut être dû directement au redressement, mais la « véracité » de la déclaration peut quant à elle être remise en cause. Dans le cas d'une réponse brute, le statisticien pourra notifier les contributions supérieures à 3%. Dans le cas d'une réponse redressée, il pourra revenir sur le résultat en modifiant la tranche d'imputation par exemple. Attention toutefois : des données hors-normes ne sont pas nécessairement à corriger…*

## 2. Livraison et tests

Le statisticien ayant procédé au redressement de la non-réponse, livrera tout d'abord à la maîtrise d'ouvrage, une table *provisoire* de données.
Ces résultats seront alors manipulés, testés, vérifiés, par une division extérieure (généralement la maitrise d'ouvrage de l'enquête), ce qui permettra de détecter d'éventuelles anomalies, de s'interroger sur certains cas de figure, de vérifier divers points.

*Par exemple, pour l'enquête TIC, la maîtrise d'ouvrage a travaillé à l'obtention de résultats demandés par Eurostat, ce qui lui a permis de formuler quelques remarques. Notamment, elle s'est rendue compte qu'une règle de l'apurement n'était pas toujours vérifiée pour la question demandant de cocher « oui/non » face à différents types de connexion. En effet, certaines unités ne proposaient que des « non » (il faut au moins un « oui »). Lorsque tous les items ont un statut « M », la correction mise en place fut la suivante :*

- *Établir pour chaque item, la répartition des modalités « oui/non », uniquement sur les unités ayant un statut « B » à l'item. Cela permet de repérer l'item le plus fréquemment à « oui ».*
- *Pour les unités ne présentant que des « non » à statuts « M », on force alors l'item détecté ci-dessus, à « oui » (et on laisse son statut à « M »).*
- *Si une unité avec tous les statuts à « B » avait dû être corrigée, il aurait fallu créer une indicatrice pour repérer ce changement, et aussi forcer le statut à « M ».*

Toutes ces remarques et questions seront alors transmises à l'unité en charge du traitement de la non-réponse, qui procèdera à des modifications et corrections, permettant ainsi d'obtenir une table *définitive* de données. Pour la plupart des enquêtes thématiques, la table devient définitive suite au retour des remarques d'Eurostat après examen des indicateurs diffusés au niveau européen.
Les deux tables dont il est ici question, sont à livrer avec un « mode d'emploi », autrement dit leur documentation.

## 3. Documentation

La documentation a pour dessein de permettre à n'importe quel usager de la table, de comprendre comment elle a été produite, et de savoir comment manipuler ces données.

Au niveau du redressement, on distinguera deux documentations :

- la première, succincte, sera fournie avec la table provisoire. C'est le « minimum de survie » pour utiliser la table. On y présentera les variables les plus importantes et indispensables à l'exploitation des données (variables de statuts, variable ETAT, quelle pondération utiliser, …), et on veillera aussi à y insérer quelques exemples de programmes, avec leur interprétation statistique.
  *Exemple de la mini-doc de l'enquête TIC disponible sur l'intranet du pôle ISE.*

- la seconde, complète, sera livrée avec la table définitive. Elle aborde tous les points nécessaires à la bonne compréhension de l'enquête : champ de l'enquête, informations sur l'échantillonnage, rappel du questionnaire, liste de toutes les variables initiales et créées suite au redressement, accompagnées de leur définition, méthode de redressement utilisée, cas spécifiques, comptages et répartitions, travaux annexes,… A titre d'exemple, le lecteur est invité à consulter la documentation de l'enquête « TIC » sur l'intranet[7] du pôle ISE. Il est à noter que cette documentation, peut être fournie ultérieurement à la livraison de la table définitive.

Parallèlement à ces notices, on dispose bien évidemment toujours d'autres documents complémentaires, comme par exemple le bilan de collecte/apurement, établi par le service en charge de la gestion de l'enquête.

## *4. Bilan qualité*

Le bilan qualité constitue un document de grand intérêt lors de la livraison d'une enquête. Il a pour vocation d'évaluer la qualité de ladite enquête - notamment si elle est périodique -, et non des méthodes statistiques à proprement parler. Par ailleurs, cette étape est préconisée par le code des « bonnes pratiques », et souhaitable pour des instances comme les CNIS ou encore EUROSTAT.

Une fiche qualité abordera principalement 8 thèmes :

- fiche descriptive de l'enquête (n°visa, périodicité, année, …)
- dates clefs (prévues et effectives)
- préparation de l'enquête (existence d'un comité d'utilisateurs, …)
- présentation de l'échantillon enquêté (plan de sondage, sous-populations de diffusion, …)
- indicateurs liés à la production (nombre de retours, de questionnaires utilisables, d'unités hors champ parmi les retours, taux de non réponse partielle, agrégats et variables cibles, …)
- indicateurs de précision (coefficients de variation pour les estimations de chaque variable cible, …)
- diffusion (archivage, publications, documentation, …)
- moyens relatifs à l'opération (équipes mobilisées, moyens informatiques, …)

Afin de se renseigner davantage sur le bilan qualité, le lecteur curieux peut consulter le diaporama associé à la présentation méthodologique du 19 juin 2008 organisée par le pôle ISE à Nantes. Ce document est disponible sur l'intranet du PISE, à l'adresse : http://10.44.230.102/Laviedupole/PresentationMethodo/2008seance31906bilanqualiteenqueteNouveau_Dossier/Presentationmethodobilanqualite.ppt

A titre d'exemple, le lecteur pourra trouver la fiche qualité relative à l'enquête sur les « déchets non dangereux des établissements commerciaux en 2006 » sur l'intranet du pôle ISE.

[7] *http://10.44.230.102/Travauxnationaux/TIC2008/Documentation.zip*

# Conclusion et pistes de réflexion

Le présent document a donc présenté toutes les étapes de redressement d'une enquête, à mettre en œuvre dès la fin de la phase d'apurement. Ces traitements permettent au final de livrer un fichier définitif de données, soigneusement documenté et commenté.

Le lecteur a pu constater, à travers notamment de nombreux conseils et exemples pratiques, combien un redressement mêlait automatismes et traitements au cas par cas. Pour cette raison, il s'agit d'être toujours vigilant quant aux résultats obtenus, et de faire preuve de pertinence lors des interprétations et observations statistiques.

Que la méthode utilisée soit l'imputation ou la repondération, le comportement des unités répondantes est primordial. En effet, il permet de déterminer des modèles statistiques et de décrire au mieux la distribution inconnue des unités non-répondantes (partielles ou totales). Même si, tout au long des opérations, le statisticien doit être précautionneux de vérifier les résultats intermédiaires obtenus, et d'alimenter ses données de toute information auxiliaire externe disponible, il ne faut pas pour autant oublier que la qualité et la précision d'un redressement se jouent en partie dès la phase de collecte. Il est alors intéressant de mettre en place très en amont, des outils de contrôles et de relances des unités défaillantes. De même il faut porter un soin tout particulier à la conception du questionnaire, étape dont l'enjeu est souvent à tort minimisé par rapport à d'autres étapes.

Les lecteurs confrontés à des redressement d'enquêtes entreprises peuvent contacter en cas de besoin le pôle ISE, notamment pour obtenir des macros SAS facilitant le traitement de la non réponse.