Cours de traitement des données

Bardin Bahouayila

Institut Africain de la Statistique (IAS)

Eco Stat Consulting (ESC)

TRAITEMENT DES DONNÉES

Rédigé par :

BAHOUAYILA MILONGO Chancel Bardin[1]

[1] E-mail : bardinbahouayila@yahoo.fr / bardin.bahouayila@facebook.com
Tel : 05 075 33 71 / 06 837 81 85

INTRODUCTION

Les valeurs manquantes ou aberrantes sont présentes dans pratiquement toutes les bases de données des applications réelles. Elles peuvent correspondre aux erreurs de saisie ou à la naïveté de l'enquêteur. La mauvaise gestion de ces valeurs peut conduire à l'induction de modèles erronés et à des analyses fallacieuses.

Le traitement des valeurs manquantes et/ou aberrantes est souvent une tâche exigeante, tant du point de vue méthodologique qu'en termes de calcul. L'objectif principal de ce cours est de décrire et de proposer des méthodes de traitement dans chaque cas.

A la fin de ce cours, l'étudiant doit être capable de faire la différence entre une non-réponse (totale ou partielle) et un hors-champs (total ou partiel) ; de reconnaitre une valeur aberrante et d'être apte à traite les données d'une base afin de la rendre exploitable.

De ce fait, nous allons d'abord présenter les différents contrôles qui se font avant la saisie des données. Ensuite, nous exposerons les notions du redressement ; notamment celles de la non-réponse totale, des unités atypiques et de la non-réponse partielle.

CHAPITRE 1 FILTRE OU CONTRÔLE AVANT LA SAISIE

Dans certains cas pratiques, l'étape de la collecte des données se fait au même moment que celle du masque de saisie du questionnaire. Après la collecte des données, il est souhaitable de passer à la saisie des données sur ordinateur à l'aide des logiciels statistiques de saisie des données. Cependant, avant la saisie des données, il est souvent recommandé de revoir les données provenant du terrain afin de s'assurer qu'elles ne sont pas de mauvaise qualité. Parfois, s'il y a lieu, on codifie les questions ouvertes lors de l'enquête.

En dehors de la codification, quelques contrôles peuvent être réalisés pour obtenir des données prêtes à être saisies.

Il existe plusieurs contrôles avant la saisie, notamment : le contrôle uni varié, le contrôle de cohérence interne, le contrôle de vraisemblance et le contrôle agrégé.

LE CONTRÔLE UNIVARIÉ

Comme le nom l'indique, il s'agit d'un contrôle qui consiste à vérifier les variables séparément via des techniques descriptives simples.

- **Pour le cas des variables qualitatives** : vérifier les modalités de chaque variable. Par exemple pour la variable sexe, vérifier qu'elle ne contient que deux modalités (1=Masculin 2=Féminin ; par exemple)
- **Pour le cas des variables quantitatives** : contrôler l'intervalle des modalités de la variable. Par exemple pour le cas d'une enquête sur des personnes âgées de 15 à 24 ans, vérifier que la variable âge n'a que des valeurs comprises entre 15 et 24.

LE CONTRÔLE DE COHÉRENCE INTERNE

Dans ce cas, il existe deux types de contrôle :

- **Le contrôle logique** : il consiste à vérifier une variable en fonction des valeurs d'une autre variable. Exemple, si X est la variable « Avoir un véhicule » et que Y est la variable « Dépense du carburant », la variable Y aura une valeur si l'on répond Oui à la variable X. De ce fait, la variable Y dépend de la variable X.
- **Le contrôle algébrique** : ce contrôle consiste à vérifier l'égalité ou l'inégalité entre deux « groupes » de variables. Exemple, si l'on veut vérifier que le nombre de garçons et de filles d'un ménage est égal au nombre total d'enfants vivant dans ce ménage, on peut considérer que X1= nombre de garçons, X2=Nombre de filles, X=X1+X2 et Y=Nombre total des enfants. Dans ce cas, il faudra vérifier si X=Y.

LE CONTRÔLE DE VRAISSEMBLANCE

Ce contrôle consiste à vérifier si réellement une variable est comprise entre deux autres variables.

Exemple, on peut vouloir vérifier si la VA est réellement entre le CA et la TVA (CA<VA<TVA).

LE CONTRÔLE AGREGÉ

Ce contrôle consiste à vérifier s'il existe des valeurs atypiques au sein d'un **groupe homogène.**

RECAPITULATIF

- Contrôle uni varié
  - **Variable qualitative** : vérifier les modalités de chaque variable : sexe, niveau d'instruction, etc.
  - **Variable quantitative** : contrôler l'intervalle des modalités de la variable : âge, poids, etc.
- Contrôle de cohérence interne
  - Contrôle logique : Contrôler Y à partir de X / X ⟶ Y
  - Contrôle algébrique : Contrôler X et Y / X=Y ou X<Y ou encore X>Y
- Contrôle de vraisemblance : Contrôler Y à partir de X et Z / X<Y<Z
- Contrôle agrégé : contrôle à l'intérieur d'un **groupe homogène** (strates)

CHAPITRE 2 REDRESSEMENT OU CONTRÔLE APRÈS LA SAISIE

Le redressement ou contrôle après la saisie consiste à analyser respectivement les quatre (04) étapes suivantes :

- Faire la différence entre une non-réponse (NR) et un hors-champs(HC) ;
- Traitement de la non-réponse totale (NRT) ;
- Traitement des données atypiques ;
- Traitement de la non-réponse partielle (NRP).

A la fin de ces quatre (04) étapes, on obtient un fichier dit « apuré ».

Nous verrons dans ce chapitre la distinction entre une non-réponse et un hors-champs.

En effet, on est en présence de valeur manquante lorsque, pour au moins une unité de l'échantillon, au moins une question posée (par un enquêteur, dans un questionnaire…) n'a pas reçu de réponse.

Il y a **non-réponse (NR)** quand il y a une valeur manquante dans l'échantillon parce que l'enquêté refuse de coopérer ou que l'enquêteur a oublié de remplir la valeur qu'il fallait. On parle cependant de **hors-champ (HC)** lorsqu'il y a une valeur manquante parce que l'enquêté ne correspond pas à l'unité statistique de l'étude (**hors-champs total : HCT**) ou qu'il ne peut pas répondre à la question posée parce qu'elle ne lui concerne pas (**hors-champs partiel[2] : HCP**).

Pour une unité de l'échantillon, si toutes ou la plupart des variables mesurées sont manquantes, on est en présence de **non-réponse totale (NRT).** Toutefois, s'il n'y a que quelques variables manquantes pour une unité de l'échantillon quelconque, on parle de **non-réponse partielle (NRP).**

De manière générale : $$\boldsymbol{n = n_R + n_{\bar{R}} + n_{HC}}$$

Avec

$\mathrm{n} = \text{taille de l'échantillon}$, $\mathrm{n}_R = \text{effectif des répondants}$, $n_{\bar{R}} = effectif\ de\ non - réponse$

$et\ \mathrm{n}_{\mathrm{HC}} = \text{effectif de hors} - \text{champs}$

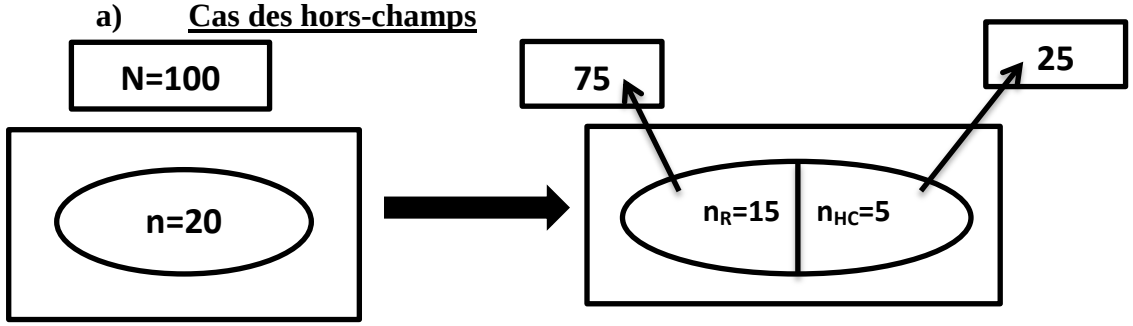
Il est remarqué que le HC n'a aucun effet sur le poids de sondage. Par contre les NR ont des effets sur le poids de sondage.

Exemple :

Si on a les cas suivants :

[2] Cas des sauts dans un questionnaire

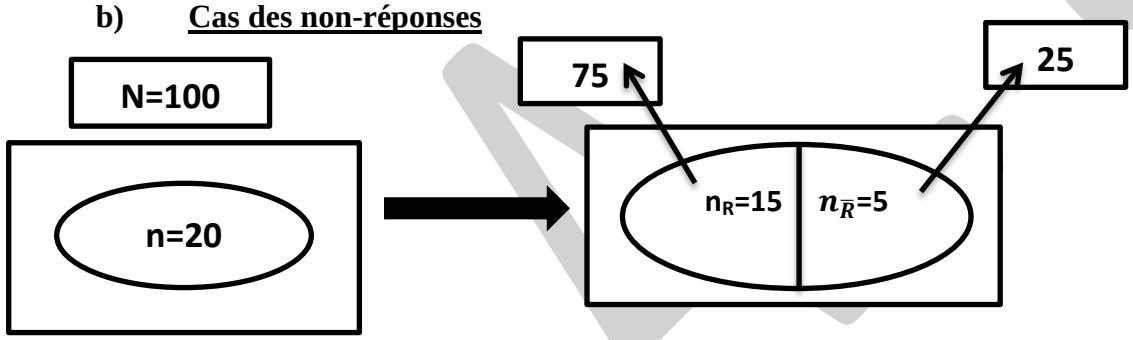
a) Cas des hors-champs



Avant l'enquête, on a choisi d'enquêter un échantillon de 20 individus pour une population totale de 100 personnes. De ce fait, le poids de sondage est $W_1 = \frac{100}{20} = 5$. Ce qui signifie que chaque individu dans l'échantillon représente 5 personnes.

Cependant, après l'enquête on constate 5 HC. Ce qui représente 25 individus de la population totale. Puisque les hors-champs sont considérés comme des individus ne faisant pas partie des unités statistiques voulues, on considère qu'en réalité il n'y a 75 individus qui correspondent à la population totale recherchée. Ainsi, le nouveau poids de sondage est $\boldsymbol{W_2 = \frac{75}{15} = 5 = W_1}$.

b) Cas des non-réponses

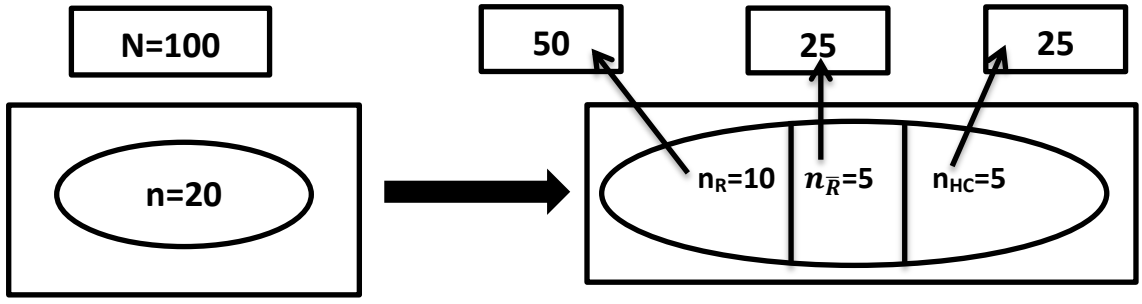


Avant l'enquête, on a choisi d'enquêter un échantillon de 20 individus pour une population totale de 100. De ce fait, le poids de sondage est $\boldsymbol{W_1 = \frac{100}{20} = 5}$.

Cependant, après l'enquête on constate 5 NR. Ce qui représente 25 individus de la population totale. Puisque les NR sont considérés comme des individus faisant également partie des unités statistiques voulues, on considère qu'en réalité la population totale est toujours égale à 100 et que les répondants ne sont qu'au nombre de 15 au lieu de 20. Ainsi, le nouveau poids de sondage est $\boldsymbol{W_2 = \frac{100}{15} = 6,7 \neq W_1}$.

Donc, chaque personne (les 15 qui ont répondu) représente maintenant 6,7 personnes au lieu de 5 seulement.

c) Cas des non-réponses et des hors-champs



CHAPITRE 3 TRAITEMENT DE LA NON-RÉPONSE TOTALE

Nous avons vu précédemment que, d'une part $\boldsymbol{n = n_R + n_{\overline{R}} + n_{HC}}$; et d'autre part que les HC n'ont pas d'impact sur le poids de sondage.

Au regard de toutes ces informations, il en découle que dans le traitement des données, $\boldsymbol{n = n_R + n_{\overline{R}}}$.

a) Cas des variables quantitatives

Dans le cas où nous avons des NRT, il y a deux possibilités de traitement. Soit on supprime l'individu et on **répondère ;** soit on garde l'individu et on cherche à remplir les valeurs manquantes ayant causées des NRT. Dans le second cas, on parle de **l'imputation** (extrapolation en français).

Si on choisit de faire une repondération, le nouveau poids devient $\boldsymbol{w' = \frac{N}{n_R}}$, alors qu'il était avant égal à $\boldsymbol{w = \frac{N}{n}}$. Et $\boldsymbol{w' > w}$.

Au cas où on se penche sur l'imputation, on considère que les individus qui ont des NRT sont des receveurs et on cherche des donneurs qui ont répondu « exactement » comme des receveurs. Il faut noter qu'un donneur ne l'est qu'une seule fois.

Les deux méthodes ont chacune d'elles des avantages et des inconvénients dont nous énumérons dans le tableau ci-dessous.

	INCONVENIENTS	**AVANTAGES**
REPONDERATION	Détérioration de la précision des estimations car le poids de sondage n'est plus le même	Facile à manipuler
IMPUTATION	Erreur de sélection car on n'est pas sûr que le receveur devrait réellement donner la même réponse que son donneur s'il nous répondait en toute sincérité	Stabilité de la précision des estimations car le poids de sondage est toujours le même

Les deux techniques reposent sur l'hypothèse que le profil des répondants est le même que celui des non-répondants. La seule différence qui existe entre les deux, c'est que pour la repondération, la taille de l'échantillon diminue (ce qui fait augmenter le nouveau poids de sondage), alors que pour l'imputation, la taille de l'échantillon reste le même.

b) <u>Cas des variables qualitatives</u>

Dans ce cas, on peut aussi faire une **repondération** ou une **imputation**. Une troisième possibilité peut être ajoutée. C'est celle de la méthode de la **modélisation de la non-réponse**.

La modélisation de la non-réponse consiste à faire un modèle d'économétrie des variables qualitatives, notamment le logit ou le probit binaire. En effet, on a un indvidu $i \in$ à $l'échantillon.$

On construit une variable réponse $Y_i = \begin{cases} 1 \; si \; l'individu \text{ "i" } a \; repondu \\ 0 \; si \; l'individu \text{ "i" } n'a \; pas \; repondu \end{cases}$

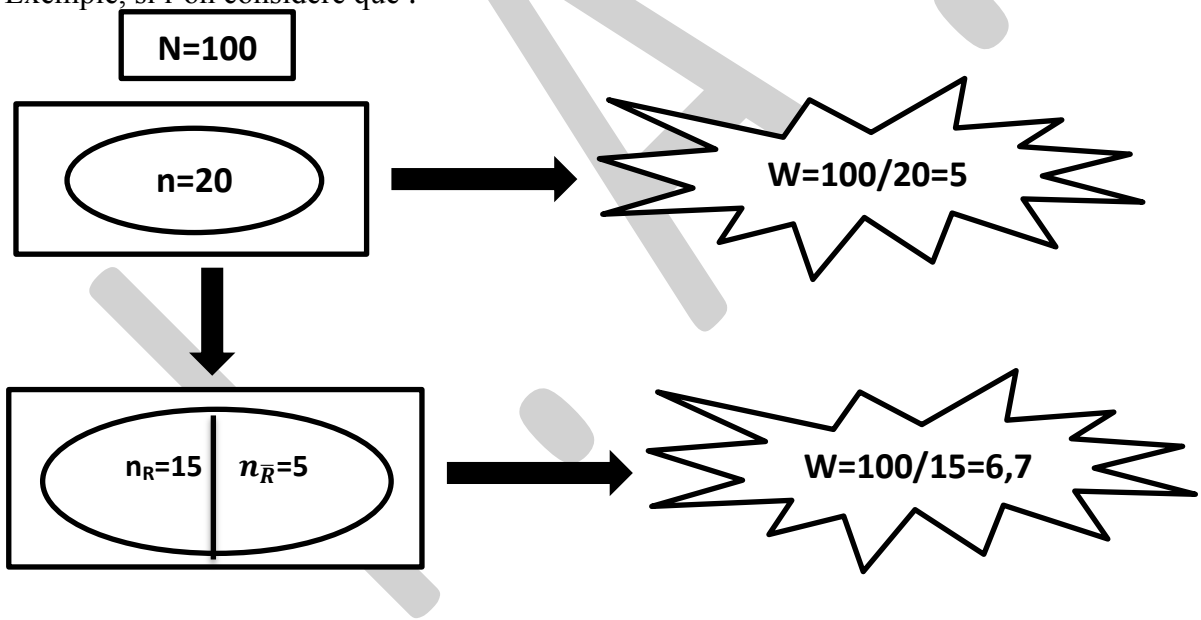
On a X_1, X_2, …, X_k les variables explicatives du modèle. Il faut s'assurer qu'il n'y a pas de variables de la stratification dans la liste des variables exogènes.

On modélise le vecteur X={ X_1, X_2, …, X_k} / $P_i = P(Y_i = 1)$ et $\log\left(\frac{P_i}{1-P_i}\right) = \propto_0 + \propto_1 * X_1 + \cdots + \propto_k * X_k + \varepsilon_i$

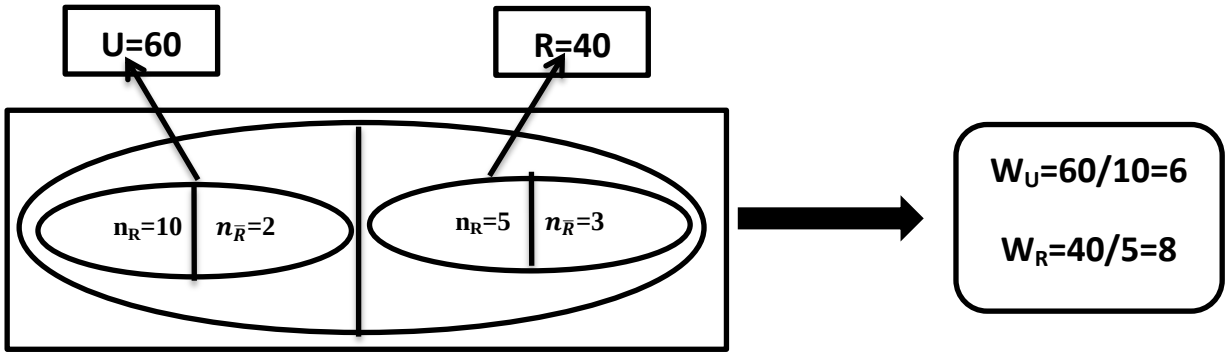
Si $\propto_n \; est \; significativement \; égal \; à \; zéro, alors \; X_n \; n'explique \; pas \; Y.$

Et si c'est le contraire, alors Y est expliqué par X_n.

Supposons par exemple que X_1 explique Y_i et que X_1 est une variable qualitative ayant deux modalités (Rurale et Urbaine). Cela signifie que selon que l'on soit en zone urbaine ou en zone rurale, le taux de la non-réponse n'est pas le même.s
Exemple, si l'on considère que :



On fait deux **<u>poste-stratifications</u>**



CHAPITRE 4 TRAITEMENT DES VALEURS ATYPIQUES

Une valeur aberrante est définie comme une observation ou un sous-ensemble d'observations qui semble(nt) incohérente(s) par rapport aux autres données de l'ensemble.

Il est possible de faire la distinction entre des valeurs aberrantes **unidimensionnelles** (à une variable) et **multidimensionnelles** (à plusieurs variables). En effet, une observation est une valeur aberrante unidimensionnelle si elle est aberrante par rapport à une seule variable. Une observation est une valeur aberrante multidimensionnelle si elle est aberrante par rapport à deux variables ou plus.

Il est peut-être facile, par exemple, de trouver une personne mesurant deux mètres ou une personne pesant 45 kg, mais quelqu'un qui mesure deux mètres et pèse seulement 45 kg est un exemple de valeur aberrante multidimensionnelle.

Il faut également distinguer les valeurs extrêmes et les valeurs aberrantes. En effet, les valeurs extrêmes peuvent être ou ne pas être des valeurs aberrantes. Cependant, une valeur aberrante est toujours une valeur extrême de l'échantillon.

Chaque enquête comprend des valeurs aberrantes pour à peu près chaque variable. Les valeurs aberrantes peuvent provenir de deux manière : l'individu est réellement différent des autres (erreur d'échantillonnage ou erreur due à la méthode d'échantillonnage) ou que la valeur a été mal saisie (une erreur de saisie).

LES CONSÉQUENCES DES DONNÉES ABBERANTES

Statistique descriptive : augmentation de la variance, mauvaise orientation de l'axe principale (analyse factorielle)

Statistique inférentielle : coefficients biaisés, etc.

LES MÉTHODES DE DÉTECTION

a) **Contrôle uni varié** :

➦ **Variable qualitative** : vérifier les modalités de chaque variable : sexe, niveau d'instruction, etc.

➦ **Variable quantitative** : contrôler l'intervalle des modalités de la variable : Exemple : Pour la variable « Total des heures effectuées », une borne maximale (208 heures) est fixée à partir de la convention collective. Les valeurs supérieures à 208 heures sont aberrantes.

b) **Détection graphique** : Pour détecter la présence de valeurs aberrantes On peut utiliser :

- les box plot
- les histogrammes
- les nuages de points

c) **Tests de cohérence logique** : On croise des variables. Exemple : « Salaire mensuel » et « Loyer mensuel »

d) **Détermination de plafonds au-delà desquels il est nécessaire de contrôler les réponses.** *f*

- On cherche les valeurs aberrantes en dehors de $\left[\bar{X} - 1{,}5(Q3 - Q1)\ ;\ \bar{X} + 1{,}5(Q3 - Q1)\right]$
- Selon Coulombe et McKay, X_j est une valeur aberrante si $\ln(X_j) > \overline{\ln(X)} + 3\sigma(\ln(X))$

e) **Les techniques classiques d'analyses multi variées** (analyse discriminante, analyse factorielle des correspondances, analyse en composantes principales) offrent des possibilités d'identification de valeurs anormales.

f) **Normalisation par le test Z** :

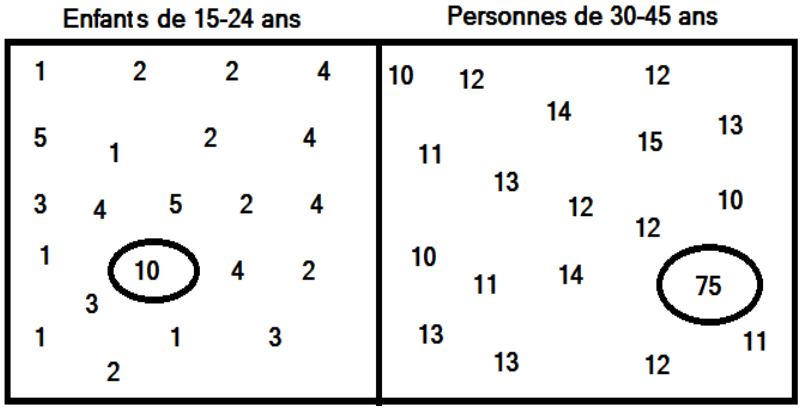
Pour identifier les valeurs aberrantes, on peut utiliser la normalisation par le test Z, souvent une valeur aberrante peut être identifiée parce qu'elle est a une valeur différente des autres en calculant

X*= (x_i-moyenne (x)) / écart type (x). L'inconvénient de cette méthode est que la moyenne et la variance doivent être connues, toutes deux incluses dans la formule de la normalisation par le test Z. De même, ces statistiques sont assez sensibles à la présence de valeurs aberrantes.

Remarque :

- Pour détecter des valeurs aberrantes on peut être amené à calculer de nouvelles variables : Exemples : Total des heures effectuées par employé, total des heures payées par employé ou montant des salaires bruts payés par employé.
- Toute utilisation de méthodes de détection de valeurs aberrantes par ordinateur doit tenir compte des limites des méthodes fournies par les logiciels.
- Une valeur est aberrante si elle engendre un effet de surprise en fonction de ce qu'on attend à partir du modèle. On compare les résultats obtenus à partir du fichier sans la valeur aberrante à ceux obtenus à partir du fichier avec la valeur aberrante.

LA DIFFÉRENCE ENTRE INLIERS ET OUTLIERS



LES MÉTHODES DE TRAITEMENT

Il y a deux (02) possibilités pour traiter les données aberrantes :

- Les valeurs aberrantes pouvant provenir d'erreurs de saisie. Si c'est le cas, on retourne au questionnaire papier quand c'est possible et on corrige. Si on ne retrouve pas le questionnaire, on les supprime et on applique ensuite une des méthodes d'imputation (moyenne, médiane…). Il faut noter que dans la présence d'une mesure aberrante, la médiane des données ne change pas. La médiane est robuste (généralement, il ne varie pas beaucoup) en présence d'un petit nombre de valeurs aberrantes : par contre la moyenne change rapidement.
- Si la valeur a été bien saisie (erreur d'échantillonnage ou due par la méthode d'échantillonnage), on la laisse comme ça et on fait les analyses avec.

CHAPITRE 5 TRAITEMENT DE LA NON-REPONSE PARTIELLE

Nous avons vu précédemment que le traitement des Non-Réponse Totale (NRT) peut se faire par repondération ou par imputation. La chose la plus simple à faire est la repondération car imputer toutes les NRT d'une base de données n'est pas un exercice facile. De manière générale, **on répondère dans le cas des NRT et on impute lorsqu'on a des Non-Réponse Partielle (NRP).**

LES MÉTHODES DE TRAITEMENT

Il existe deux (02) types de méthodes : la méthode déterministe et la méthode aléatoire.

➦ **La méthode déterministe**

Dans ce cas, on y compte plusieurs types d'imputations, parmi lesquelles on peut citer : l'imputation déductive, historique, par la moyenne, du voisin le plus proche, du cold-deck, par ratio et par régression.

☺ *Imputation déductive*

La donnée manquante est déduite des réponses aux autres questions. Ce type d'imputation par règle déterministe est surtout utilisé dans les enquêtes entreprises pour corriger des données intervenant dans des équations comptables.

Exemple 1 : âge <14 ans ⟶ activité professionnelle = inactif.

☺ *Imputation historique*

On impute à l'enquête $t+1$ la valeur de l'enquête précédente multiplié par quelque chose.

Exemple 2 : si le taux de croissance démographique au Congo est stable et est de 2%, et si l'on ne connait pas la population au Congo de 2012 tout en ayant la valeur de 2011 ; alors la valeur estimée de 2012 sera :

$$\boldsymbol{P_{2012} = P_{2011} * 2\%}$$

Cette méthode est beaucoup plus utilisée dans le cas des **données temporelles ou de panel**.

☺ *Imputation par la moyenne totale ou par la moyenne de post-strate*

Les valeurs manquantes de chaque attribut sont remplacées par la moyenne de l'attribut considéré. Il y a deux variantes de l'imputation par la moyenne : l'imputation par la moyenne totale et l'imputation par la moyenne de sous-groupe. Pour l'imputation par la moyenne totale, la valeur absente d'un attribut est remplacée par la moyenne des valeurs de cet attribut de toutes les observations. Pour l'imputation par la moyenne de sous- groupe (classe), la valeur manquante est remplacée par la moyenne du sous-groupe (classe) de l'attribut en question. L'inconvénient de cette méthode est la **sous-estimation** de la variance.

☺ *Imputation par k-plus proche voisin*

La technique de k-plus proche voisin est une technique utilisée pour la substitution des valeurs manquantes, avec la valeur du plus proche voisin dans l'ensemble de données. Pour chaque observation contenant des valeurs manquantes, on recherche ses k plus proches voisines. Dans le cas de **variables continues**, la valeur de remplacement correspond simplement à une **statistique centrale** des valeurs prises par ces k voisins pour la variable en question. Les k plus proches voisins de l'observation considérée, parmi ceux qui appartiennent à **la même classe**, sont alors utilisés pour déterminer la valeur de remplacement.

Exemple 3 :

N°	X1	X2	X3	X4
1	1	10	500	700
2	1	10	300	650
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
30	1	10	300	600
31	2	5	200	550
32	1	10	300	610
33	1	5	500	800
34	1	10	300	
35	2	5	400	500

N°	X1	X2	X3	X4
1	1	10	500	700
2	1	10	300	650
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
30	1	10	300	600
31	2	5	200	550
32	1	10	300	610
33	1	5	500	800
34	1	10	300	**?**
35	2	5	400	500

☺ *Cold-Deck*

On remplace la donnée manquante par une donnée obtenue en dehors de l'enquête.

Exemple 4 :

N°	Pays	Importations	Exportations	PIB
1	Afghanistan	5478,5	5004,1	700
2	Bénin	2145,8	3007,2	650
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
30	Congo	4521,7	3007,3	
31	Gabon	7541,6	2005,9	550
32	Cameroun	6124,2	2001,6	600
33	RCA	3125,4	5003,8	800
34	Tchad	2154,3	3008,3	500
35	Guinée Equatoriale	1248,9	4003,5	500

☺ *Imputation par ratio*

Il existe deux procédures ; on impute Y_k tel que :

a) $\boldsymbol{Y_k = X_k \frac{Y_i}{X_i}}$ si $\frac{Y_i}{X_i}$ est stable quel que soit l'individu i ;

b) $\boldsymbol{Y_k = X_k \frac{\overline{Y}}{\overline{X}}}$ si $\frac{Y_i}{X_i}$ diffère d'un individu à un autre.

Exemple 5 :

1er cas : CA et VA d'une usine de télécommunication entre 1999 et 2006

ANNEES	CA	VA	CA/VA
1999	200 000	100 000	2
2000	40 000	20 000	2
2001	100 000	50 000	2
2002	60 000	30 000	2
2003	450 000		?
2004	150 000	75 000	2
2005	150 000	75 000	2
2006	300 000	150 000	2

2ème cas : Nombre de salariés et VA de quelques usines de télécommunication en 2012

N°	Noms d'entreprise	Salariés	VA	VA/Salariés
1	Alcatel	5 000	100 000	20
2	Nokia	1 000	20 000	20
3	LG	2 000	50 000	25
4	Samsung	1 000	30 000	30
5	Horse	5 000		?
6	Motorola	2 500	75 000	30
7	HTC	3 000	75 000	25
8	Ericson	5 000	150 000	30

☺ *Imputation par régression déterministe*

C'est une approche en deux étapes : d'abord, on estime les rapports entre les attributs, et puis on emploie les coefficients de régression pour estimer la valeur manquante (Frane, 1976). La condition fondamentale de l'utilisation de l'imputation par régression est l'existence d'une corrélation linéaire entre les attributs. La technique suppose également que les valeurs sont manquantes au hasard. Dans le contexte des valeurs

manquantes, deux modèles de régression sont en général employés : **la régression linéaire et la régression logistique**. Cette dernière est beaucoup plus utilisée pour traiter les variables discrètes, alors que la régression linéaire est souvent appliquée sur des variables continues (Little et Rubin, 2002). Pour chacune de ces méthodes, il est possible de tenir compte de l'information de classe en n'utilisant que les observations d'une même classe pour estimer les paramètres de régression. L'inconvénient de cette méthode, c'est les hypothèses qui sont faites sur la distribution des données. Supposer une relation linéaire entre les variables, revient à faire des hypothèses qui sont rarement vérifiées, dans cette situation, le remplacement des valeurs manquantes par des valeurs prédites basées sur un modèle biaisé ne constitue pas un traitement approprié.

Dans la démarche, on considère le modèle $Y = X'\beta + \varepsilon$ et on estime β sur les répondants.

Exemple 6 :

N°	1	2	3	4	5	6	7
X1	379	319	259	199		79	19
X2	4815	4446	4077	3708	3339	2970	2601
N°	**8**	**9**	**10**	**11**	**12**	**13**	**14**
X1		100	110	170	230	290	350
X2	4000	2785	3154	3523	3892	4261	4630

➦ **La méthode stochastique**

☺ *Hot-Deck*

On tire aléatoirement un individu qui a répondu et on impute sa valeur dans la case vide.

☺ *Imputation multiple*

L'imputation multiple est une procédure qui consiste à imputer, pour un non-répondant donné, une valeur provenant de différentes valeurs choisies au hasard avec remise[3] ou sans remise.

Remarque :

- C'est un hot-deck multiple ;
- La méthode de **bootsrap** est l'une des méthodes de cette imputation.

[3] Dans le cas où r < n/2

QUELLES MÉTHODES CHOISIR ?

➦ **Pour corriger de la non-réponse : imputation ou repondération ?**

Comme la repondération consiste à modifier les poids des individus répondants, elle ne peut être envisageable que pour la correction de la non-réponse totale. En effet, il est impossible de travailler avec des poids différents selon la variable utilisée. En revanche, les méthodes d'imputation, plutôt adaptées pour la correction de la non- réponse partielle, peuvent aussi être utilisées pour corriger la non-réponse totale. Mais, au fond, existe-t-il une approche préférable à l'autre ? À l'**Insee**, l'approche adoptée pour la correction de la non-réponse totale est plutôt la repondération pour les enquêtes réalisées auprès des ménages, et plutôt l'imputation pour les enquêtes réalisées auprès des entreprises. Notons qu'en général la méthode de repondération choisie est celle des groupes de réponse homogène et que les variables permettant de définir les groupes sont déterminées par une régression logistique parmi celles disponibles dans la base de sondage (issues du recensement de la population). Il arrive aussi que la correction de la non-réponse soit réalisée par calage. Le choix entre imputation et repondération pour corriger la non-réponse totale est difficile, et il n'existe pas de réponse théorique satisfaisante permettant de nous y aider. Des considérations pratiques peuvent parfois influencer le choix. Par exemple, dans certains pays comme la Suède, l'imputation n'est pas autorisée par la loi pour des données concernant des personnes. Les logiciels développés pour l'imputation sont moins nombreux que ceux pour la repondération. De plus, les développements théoriques concernant l'estimation de variance en présence de données imputées en sont encore au stade de la recherche et conduisent en général à des formules plus complexes que dans le cas de la repondération. Une solution simple consiste à considérer les données imputées comme de vraies données, mais cela conduit à une sous-estimation de la variance et à des intervalles de confiance invalides. Tous ces arguments poussent plutôt à privilégier la repondération.

Il faut noter que les deux types de méthodes ne s'opposent pas systématiquement : il arrive même qu'elles coïncident dans certaines circonstances.

➦ **Pour corriger de la non-réponse partielle : méthode stochastique ou méthode déterministe ?**

Les méthodes déterministes peuvent introduire une forte distorsion dans la répartition des données. De plus, elles peuvent conduire pour des variables quantitatives discrètes à des valeurs non réalistes et ne peuvent pas être appliquées dans le cadre de variables qualitatives. Les méthodes stochastiques présentent quant à elles l'avantage d'imputer une valeur vraisemblable qui soit plausible aussi bien pour des données continues que pour les données discrètes et de pouvoir être utilisées dans le cas de variables qualitatives Cependant, si l'on s'intéresse à des estimations de totaux, les méthodes déterministes sont à privilégier car elles conduisent à des estimateurs plus précis que les méthodes stochastiques. Le choix entre les deux grandes familles de méthodes d'imputation dépend principalement de l'utilisation des données par la suite. Si l'objectif de l'enquête considérée est de ne produire que des totaux, les méthodes déterministes sont recommandées. En revanche, si l'objectif est multiple, il est préférable de choisir des méthodes stochastiques. C'est au prix, souvent important, d'une variance accrue, que l'imputation stochastique conduit à une estimation correcte de la distribution. En effet, l'imputation avec aléa introduit par rapport à l'imputation déterministe un terme de variance supplémentaire dû au mécanisme aléatoire de l'imputation.

Les méthodes d'imputation à utiliser sont, en général, plus complexes que celles présentées ci-dessus ; elles cherchent en particulier à utiliser le maximum d'informations auxiliaires et à conserver certaines relations,

comme par exemple la covariance, avec d'autres variables qui peuvent être aussi partiellement manquantes... la mobilisation d'informations est d'ailleurs indispensable dans les cas où le pourcentage de non-réponse est relativement important (de l'ordre de 30 à 40%). En pratique, il est souvent nécessaire d'imputer plusieurs variables d'intérêt pour un même individu. Modéliser et imputer séparément chacune des variables risque d'introduire des incohérences dans le questionnaire ainsi complété. Par conséquent, dans le but de « préserver » au mieux les relations entre les variables, on choisit plutôt une méthode d'imputation par donneur telle que le hot-deck où le donneur est choisi pour remplacer toutes les variables manquantes d'une unité non répondante. Ce sont donc les mêmes classes d'imputation qui sont retenues pour imputer un ensemble de variables. Il est alors nécessaire de faire des compromis car les variables auxiliaires les plus appropriées pour l'imputation d'une variable donnée ne sont pas forcément celles qui le sont pour les autres variables concernées par l'imputation.

LES TAUX DE RÉPONSE

➦ **Taux de réponse opérationnel**

Ces taux servent à fournir de l'information sur le déroulement du processus de collecte.

Si l'on considère que $\mathbf{n_c}$ **= effectif des réponses complètes** ; $\mathbf{n_{nrp}}$ **= effectif des non-réponses partielles** ;

$\mathbf{n_{nrt}}$ **= effectif des non-réponses totales** et $\mathbf{n_{nc}}$ **= nombre de personnes non contactées,** on a :

☺ *Taux de réussite de l'enquête*

$$T^{(1)} = \frac{n_c + n_{nrp}}{n_c + n_{nrp} + n_{nrt}}$$

☺ *Performance des interviewers*

Ce taux mesure la capacité des enquêtés à compléter leurs rencontres

$$T^{(2)} = \frac{n_c}{n_c + n_{nrp} + n_{nrt}}$$

☺ *Rendement des entrevues :*

$$T^{(3)} = \frac{n_c + n_{nrp}}{n_c + n_{nrp} + n_{nrt} + n_{nc}}$$

☺ *Autre taux*

$$T^{(4)} = \frac{n_c}{n_c + n_{nrp} + n_{nrt} + n_{nc}}$$

Ce taux mesure le rapport entre le nombre d'entrevues complétées et le nombre d'entrevues potentielles.

➦ **Taux de réponse correcteur**

- Donne de l'information sur le processus de collecte ;
- Ramène le sous-échantillon des répondants à l'échantillon initial ;
- Sert à corriger les poids de sondage pour tenir compte de la non-réponse.

$$T = \frac{Nombre\ de\ répondants}{Taille\ de\ l'échantillon}$$

SI VOUS AVEZ BESOIN DE LA CORRECTION DES EXERCICES PROPOSÉS, DE NOS EXERCICES DE TRAVAUX DIRIGÉS, DE NOS SUJETS DE DEVOIR OU D'EXAMEN AVEC SOLUTION, CONTACTEZ NOUS :

95, rue Malanda (Moukondo vers la Tsiémé, Ouenzé)

VOUS POUVEZ AUSSI APPELER OU ECRIRE A L'AUTEUR DE CE DOCUMENT.

E-mail : bardinbahouayila@yahoo.fr / bardinbahouayila@gmail.com

Tel : 05 075 33 71 / 06 837 81 85