Traitement des valeurs manquantes et des valeurs aberrantes

Avant de traiter les données, vérifier la qualité des données :

Les données peuvent être :

- manquantes
- aberrantes : la valeur de l'âge est fausse
- en double : le même client est enregistré avec trois identifiants différents

Les doublons donnent plus d'importance aux valeurs répétées.

I. Valeurs manquantes – Généralités

1. Les différents types de valeurs manquantes :

On distingue deux catégories de non-réponse :

- la non-réponse totale, lorsque aucune information n'est recueillie sur une unité échantillonnée
- la non-réponse partielle, lorsque le manque d'information est limité à certaines variables.

Des valeurs manquent parce que elles n'ont pas pu être observées,
elles ont été perdues
elles étaient incohérentes.

3 mécanismes distincts de non-réponse :

a) réponse manquant entièrement au hasard (MCAR, *Missing Completely at random)* :

b) réponse manquant au hasard (MAR, *Missing at random*) :

c) réponse ne manquant pas au hasard (NMAR, *Non missing at random)*.

2. Les différents traitements des données manquantes :

A. ne rien faire
B. utiliser uniquement les enregistrements pour lesquels les données sont complètes
C. utiliser une méthode de repondération
D. imputer une valeur

3. Traitement A : Ne rien faire

Cela oblige à travailler avec un fichier de données incomplet qui ressemble à un morceau de fromage gruyère.

Si les valeurs manquantes sont peu nombreuses, on peut les oublier sans aucun scrupule.

4. Traitement B : utiliser uniquement les enregistrements complets

Si les données sont présentées sous forme de tableau, cela revient à oublier une ligne dès qu'il manque une valeur dans cette ligne : on oublie donc aussi les autres valeurs de cette ligne, qui sont effectivement présentes.

Bien que cette option soit simple et permette d'utiliser un fichier complet, elle présente certains risques. En effet :

- l'échantillon de ceux qui ont répondu à toutes les questions peut être
  - soit trop réduit pour être significatif,
  - soit non représentatif de la population globale.
- elle peut mener à des estimateurs fortement biaisés, à moins que la non-réponse ne dépende d'aucune des variables d'intérêts (cas MCAR).

Cette option ne peut être envisagée que pour une brève analyse descriptive des réponses complètes.

Remarque : On peut agir différemment selon s'il s'agit d'une variable importante ou secondaire.

5. Traitement C : Repondération

Non-réponse totale : Les méthodes de repondération augmentent le poids de sondage appliqué aux répondants pour compenser pour les non-répondants. L'objectif est de produire des estimations approximativement sans biais.

Non-réponse partielle : On peut appliquer des méthodes de repondération mais le principal inconvénient est qu'il faut créer un nouveau poids ajusté pour chaque variable d'intérêt.

Conséquences : Les résultats de diverses analyses peuvent ne pas concorder, c'est pourquoi on utilise peu les méthodes de repondération pour tenir compte de la non-réponse partielle.

II. Valeurs manquantes – différentes méthodes d'imputation :

1. Généralités sur l'imputation :

Définition : L'imputation consiste à produire une « valeur artificielle » pour remplacer la valeur manquante, avec pour objectif de produire des estimations approximativement sans biais.

L'imputation par règle : on applique à une valeur manquante une valeur déterminée suivant une réglementation : Exemple : Calcul montant TTC à partir du montant HT

Les méthodes courantes d'imputation :

- la moyenne
- le ratio
- la régression

} Méthodes déterministes

- le hot-deck aléatoire
- le plus proche voisin

} Méthodes stochastique ou aléatoire

- autre.

Description des méthodes courantes :

- **imputation par la moyenne :** On remplace chacune des valeurs manquantes par la valeur moyenne de l'ensemble de réponses obtenues.
- **imputation par le ratio :** chaque valeur manquante y_i est remplacée par la valeur prévue y_i^* obtenue par régression de y sur x.
- **imputation par régression :** c'est une extension naturelle de l'imputation par la méthode du ratio où l'on se sert de q variables auxiliaires $x_1 \ldots x_q$.
- **imputation par la méthode hot-deck aléatoire :** cela consiste à attribuer la valeur de y fournie par un répondant (donneur), sélectionné au hasard avec remise parmi l'ensemble des répondants, pour remplacer la valeur manquante pour l'unité non-répondante (receveur).
- **imputation par la méthode par le plus proche voisin :** on attribue à l'enregistrement pour lequel la réponse à une question manque la valeur figurant pour cette question dans l'enregistrement obtenu pour le répondant le plus proche, où l'expression « le plus proche » est habituellement définie par une fonction de distance basée sur une ou plusieurs variables auxiliaires.

Méthode d'imputation	Moyenne	Ratio	Régression	Hot-deck aléatoire	Plus proche voisin
Valeur imputée	$y_i^* = \frac{1}{r}\sum_{i \in s_r} y_i = \bar{y}_r$	$y_i^* = \frac{\bar{y}_r}{\bar{x}_r} x_i$	$y_i^* = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \ldots + \hat{\beta}_q x_{qi}$	$y_i^* = y_j$ pour certains $j \in s_r$ tels que $P(y_i^* = y_j) = 1/r$	$y_i^* = y_j$ pour certains $j \in s_r$ tels que $dist(x_i, x_j)$ soit minimal
Variable(s) auxiliaire(s)?	NON	OUI (une)	OUI (une ou plus)	NON	OUI (une ou plus)

Remarques importantes :

- **En général, on procède à un ajustement de poids dans le cas de non-réponse totale et à l'imputation pour une non-réponse partielle.**

- Il faut distinguer les variables principales qui doivent être nécessairement connues de celles qu'il est possible d'imputer.
  Exemple de variables principales pour une entreprise :
  l'activité principale exercée *le nombre de salariés* *le chiffre d'affaires*

- L'imputation permet d'utiliser un poids unique associé à chaque individu, ainsi les résultats de diverses analyses restent cohérents.

- Chaque technique d'imputation conduit à une formule de variance ainsi qu'à une estimation de variance particulière.

- On distingue les groupes de méthodes suivants :

les méthodes déductives : la donnée manquante est déduite des réponses aux autres questions
les méthodes de type "cold-deck" : elles utilisent l'information d'une autre enquête
les méthodes utilisant la prévision par un modèle de régression
les méthodes de type "hot-deck"

- Les procédures d'imputation pour données manquantes sont utilisées depuis plus de 50 ans, surtout par les statisticiens traitant de données d'enquête

Dangers de l'imputation :

1. Même si l'imputation produit un fichier complet de données, l'inférence, en particulier l'estimation ponctuelle, n'est valide que si les hypothèses sous-jacentes sont satisfaites.

2. L'imputation modifie les relations entre les variables.

3. Si les valeurs imputées sont traitées comme des valeurs observées, la variance de l'estimateur risque d'être considérablement sous-estimée, surtout si la proportion de non-réponses est appréciable.

Imputation et hypothèses :

En fait, pour faire une inférence en cas d'imputation, nous n'avons pas d'autre choix que d'émettre des hypothèses quant au mécanisme de réponse et à la variable d'intérêt y.

Il est parfois justifié de supposer que ce mécanisme est du type MCAR
i) la probabilité de répondre à la question y est la même pour toute les individus
ii) les unités répondent à la question y indépendamment l'une de l'autre.

Dans ce cas, l'estimateur imputé est un estimateur sans biais de $\overline{Y}$.

2. Imputation par la moyenne

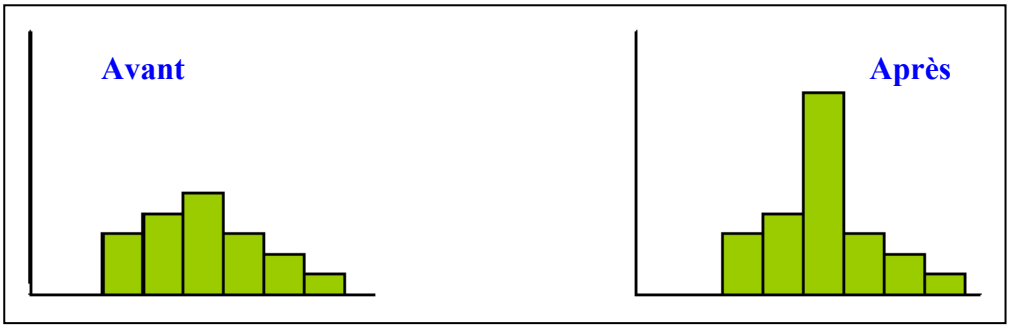
Si les valeurs manquantes sont absentes pour des raisons vraiment aléatoires, on peut sans gros problème les remplacer par la moyenne ou la médiane des variables correspondantes.

Mais souvent, le fait qu'une valeur manque dépend de sa valeur :

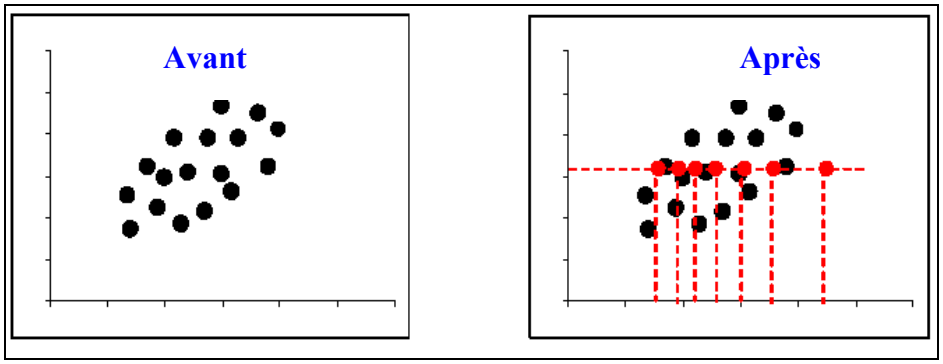
Exemple : On demande le salaire dans un sondage
Les gros revenus hésiteront à répondre : il faut en tenir compte.
La moyenne est alors plus basse qu'elle ne le devrait.

Conséquences :

- Déformation de la distribution marginale de X



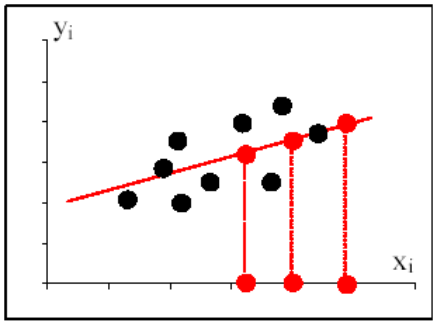
- Déformation des variances et des corrélations avec d'autres variables



3. Cas des variables qualitatives : imputation par le mode

4. Imputation par ratio / régression

Conséquence :



Les corrélations sont augmentées

5. Méthode hot-deck

Rappels concernant cette méthode :

- Son principe = attribuer à une donnée manquante une valeur observée chez un répondant
- Il s'agit donc de trouver parmi les répondants quels sont les donneurs potentiels.

Définition : Une façon simple de procéder est de classer les observations en groupes homogènes appelées cellules d'ajustement ou d'imputation.

On donne à un non-répondant la donnée d'un répondant appartenant à la même cellule d'ajustement.

Construction de ces cellules d'ajustement :

On peut

- utiliser les variables de stratification,
- effectuer des croisements de différentes variables (sexe, classes d'âge, etc.)

Plus on effectue de croisements, meilleures sont les cellules d'ajustement, mais moins nombreux sont les donneurs potentiels.

- modéliser la probabilité de répondre en fonction d'un certain nombre de caractéristiques. Pour chaque observation, on calcule alors la probabilité de répondre. On regroupe ensuite les probabilités obtenues en 5 à 6 classes. Ces classes forment les cellules d'ajustement.

Remarques :

- Si le nombre de valeurs observées par cellules sur la variable à imputer est insuffisant, on procède au regroupement de 2 ou plusieurs cellules au contenu ± similaire.
- Le remplacement de chaque donnée manquante par une valeur observée tirée au hasard
  - Préserve la distribution marginale de la variable
  - Peut fausser les corrélations avec d'autres variables
- C'est une technique particulièrement appropriée pour l'imputation des variables qualitatives. Elle est également intéressante pour des variables discrètes.

6. Imputation par le plus proche voisin :

Plusieurs étapes :

1. Calcul des **distances euclidiennes entre receveurs et donneurs**, pour chaque classe d'ajustement
2. **Recherche de la plus petite distance entre un receveur** (individu ayant des réponses manquantes) **et un donneur**
3. **Attributions des valeurs** des variables spécifiques du donneur au receveur
4. Elimination du receveur utilisé et du donneur

7. Imputation multiple :

Utilité de la méthode : défauts de l'imputation simple

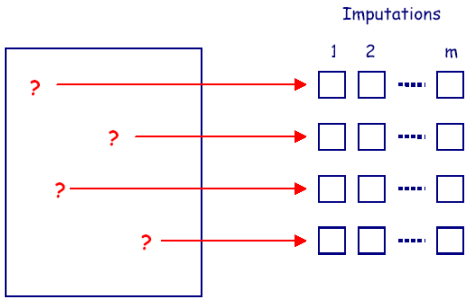
La correction de la non-réponse partielle par une valeur unique présente un défaut majeur. En effet :

- Une unique valeur imputée ne peut pas représenter toute l'incertitude à propos de la valeur à imputer
- Les analyses qui considèrent les valeurs imputées de manière équivalente aux valeurs observées sous-estiment l'incertitude, même si la non-réponse est correctement modélisée et des imputations aléatoires sont générées.
- Ce handicap peut conduire entre autres à des variances nettement sous-estimées.

Description de la méthode « imputation multiple » :

Trois étapes :

1 - Remplacement de chaque valeur manquante par m>1 valeurs simulées



On obtient m bases de données.

2 - Analyse statistique identique de chacune des m bases de données complétées

3 - Combinaison des résultats

Avantage : Cela permet de trouver des estimateurs ponctuels plus efficaces.

Plus le nombre m d'imputations est grand, plus les estimateurs seront précis.

En pratique, on constate qu'on a de bons résultats à partir de 5 imputations.
On constitue pour une variable X à valeurs manquantes, 5 variables $X_1, \ldots, X_5$ *à valeurs complètes.*

8. Conclusion

L'imputation de données manquantes n'est pas une affaire banale.

On constate que les méthodes sont nombreuses et
qu'il n'existe pas de recettes définitives, le statisticien devant agir au cas par cas.

Il est souvent nécessaire d'opérer des va-et-vient entre les données brutes et les données corrigées ou imputées.
En effet il n'est pas toujours possible

- de définir *a priori* les contrôles susceptibles de détecter toutes les incohérences,
- de prévoir à l'avance les méthodes d'imputations les plus pertinentes.

Il faut alors repartir des données brutes pour tester un autre mode de traitement, en veillant à ce qu'il n'interfère pas sur d'autres traitements déjà réalisés.

III. Valeurs aberrantes

Avant d'entreprendre l'imputation des données manquantes, on doit chercher s'il n'y a pas des valeurs aberrantes.

1. Définitions :

Une valeur aberrante est une valeur qui diffère de façon significative de la tendance globale des autres observations quand on observe un ensemble de données ayant des caractéristiques communes.

Soit $x_{(1)}$, $x_{(2)}$,..., $x_{(n)}$ les données ordonnées dans l'ordre croissant. Les valeurs $x_{(1)}$ et $x_{(n)}$ sont respectivement l'observation extrême inférieure et supérieure.

Propriétés :

- Les valeurs extrêmes peuvent être ou ne pas être des valeurs aberrantes.
- Une valeur aberrante est toujours une valeur extrême de l'échantillon.

2. Quelques remarques importantes :

- Les valeurs aberrantes ne sont pas forcément erronées.

  Dans certains cas, la valeur aberrante doit être acceptée comme une indication intéressante.

  Exemple : Prospections minières.

- Il ne faut pas adopter une attitude radicale de rejet,
  ou d'inclusion systématique des valeurs aberrantes.

  - Le rejet systématique peut entraîner la perte d'informations réelles
  - Le rejet des valeurs aberrantes a des conséquences statistiques non négligeables car l'analyse est ensuite faite sur un échantillon censuré qui n'est plus aléatoire.

- En fonction des circonstances, il existe des méthodes, dites robustes, qui prennent en compte toutes les données mais minimisent l'influence des valeurs aberrantes.

- L'apparition de valeurs aberrantes est due à diverses sources de natures différentes, d'où la complexité de l'examen des valeurs aberrantes.

3. Détection des valeurs aberrantes :

a) Contrôle sur le domaine des valeurs :

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 :

- Boxplot
- Histogrammes
- Nuages de points
- diagramme de dispersion des observations classées en fonction de leur rang

c) Tests de cohérence logique

Exemple : On croise des variables comme « Salaire mensuel » et « Loyer mensuel »

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

- On cherche les valeurs aberrantes en dehors de $\left[\bar{x} - 1,5(Q_3 - Q_1) \; ; \; \bar{x} + 1,5(Q_3 - Q_1)\right]$ (Box plot)
- Selon Coulombe et McKay, X_j est une valeur aberrante si $\ln(X_j) > \overline{\ln(X)} + 3\sigma\left(\ln(X)\right)$
- On crée des groupes puis on cherche les valeurs aberrantes en dehors de $\left[\text{Mé}(\text{groupe}) - k\,\sigma(\text{groupe}) \; ; \; \text{Mé} + k\,\sigma(\text{groupe})\right]$ (avec $k > 6$)

e) 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.

f) La méthode des corrélations permet d'analyser les coefficients de corrélations en enlevant une valeur et en évaluant la variation du coefficient entre deux variables marginales. Cette variation permet d'identifier des valeurs aberrantes.

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

Remarques :

- 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
  Montant des salaires bruts payés par employé

- Il est rarement nécessaire de contrôler plus d'une cinquantaine d'individus.

- Toute utilisation de méthodes de détection de valeurs aberrantes par ordinateur doit tenir compte des limites des méthodes fournies par les logiciels.

4. Traitement des valeurs aberrantes :

3 méthodes pour traiter les données aberrantes :

- Les valeurs aberrantes pouvant provenir d'erreurs de saisie, on vérifie si ce n'est pas le cas en retournant au questionnaire papier quand c'est possible et on corrige.
- On les rejette et on applique ensuite une des méthodes d'imputation (moyenne, médiane…) vues pour les valeurs manquantes
- On adopte des méthodes qui diminuent leur impact au cours des analyses statistiques :
  la médiane
  l'écart inter-quartile…

5. Conclusion

Le traitement des valeurs aberrantes est complexe.

Sources :

1. Analyse multidimensionnelle de données incomplètes : utilisation des procédures - Jean-Pierre NAKACHE - Alice GUEGUEN
2. Détecter et corriger la qualité des données avec SAS® Data Quality
3. Détection de valeurs aberrantes lors du traitement des données de la taxe sur les produits et services - Nelson Émond et Guylaine Dubreuil1
4. Deuxième enquête camerounaise auprès des ménages – Présentation des bases de données de l'enquête (Numéro de référence : 2003-9)
5. Enquête nationale auprès des diplômés (END) - http://www.statcan.ca
6. http://www.ipsos.fr/
7. Méthodologie pour le traitement de l'enquête mensuelle sur l'activité auprès des entreprises de travaux publics en Métropôle www.fntp.fr
8. L'imputation des données manquantes, la technique de l'imputation multiple, les conséquences sur l'analyse des données : l'enquête 1999 KOF/ETHZ sur l'innovation - Laurent Donzé
9. Orientations JAMP relatives à l'évaluation des tendances des apports et à la correction des charges - http://www.grappa.univ-lille3.fr/polys/fouille/sortie004.html#toc5
10. Préparation des données : transformations, valeurs manquantes et aberrantes - Vincent Zoonekynd
11. Problèmes de traitement des données dans les enquêtes sur les micro entreprises : l'expérience des enquêtes polonaises SP3 par Bertrand Savoye1
12. Traitement des valeurs aberrantes : concepts actuels et tendances générales Viviane Planchon
13. www.ssc.ca/documents/case_studies/ 2002/missing_dataF.doc