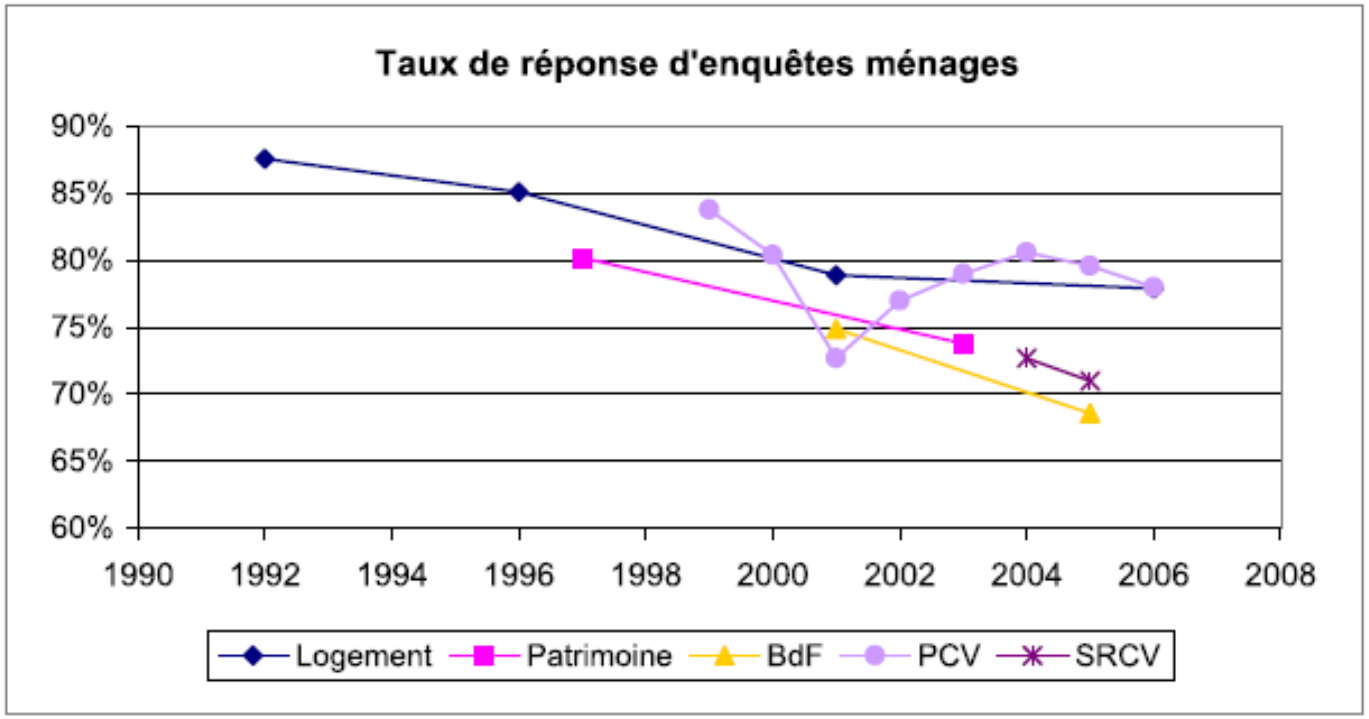
Non-réponse et données manquantes

Sylvie Rousseau & Gilbert Saporta

décembre 2011

EXEMPLES DE TAUX DE RÉPONSE À CERTAINES ENQUÊTES



Pour les enquêtes auprès des entreprises , le taux de non-réponse est de l'ordre de 10 % à 15 % pour les enquêtes obligatoires

DEUX GRANDS TYPES DE NON-RÉPONSE

- **La non-réponse totale** :
  - L'individu n'a pas du tout répondu à l'enquête
  - On peut tout de même avoir de l'information sur le non-répondant
    - les éléments de la base de sondage
    - des informations collectées de visu sur le terrain
      - *Exemple : le type de logement (individuel ou collectif)*
- **La non-réponse partielle** :
  - l'individu refuse de répondre à certaines questions

LES CAUSES DE LA NON-RÉPONSE TOTALE

- Impossibilité de contacter l'individu
  - *Exemple : déménagement, absence, digicode, …*
- Refus de l'individu de répondre à l'ensemble de l'enquête
- Incapacité de l'individu de répondre
  - *Exemple : problème de langue*
- Perte du questionnaire ou impossibilité de l'exploiter
  - *Exemple : difficulté à lire une écriture (localités au recensement), incohérences des réponses*
- Abandon de l'individu au tout début du questionnaire
  - *Exemple : temps d'interview jugé trop long, questions jugées indiscrètes*

LES CAUSES DE LA NON-RÉPONSE PARTIELLE

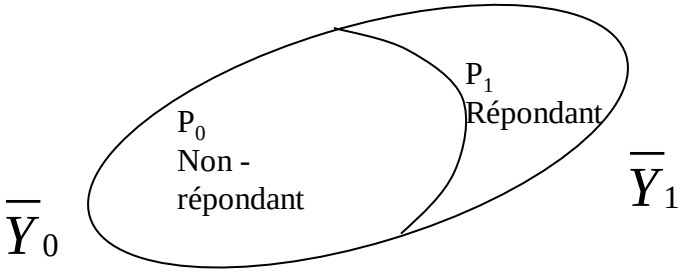
- Refus de répondre à certaines questions jugées indiscrètes
  - *Exemple : enquête sur les revenus, sur les comportements de violence, sur la consommation de produits illicites, sur les marges commerciales, …*
- Non-compréhension de la question posée par l'enquêté
- Non-compréhension de la réponse par l'enquêteur
- Abandon de l'individu en cours d'enquête
  - *Par lassitude par exemple (enquêtes à vague, panels, carnets de dépense)*

LES CONSÉQUENCES DE LA NON-RÉPONSE

- Introduction d'un biais
  - *Exemples*
    - *Enquête sur les compétences à l'écrit et à l'oral où les personnes peu diplômées répondent moins bien*
    - *Enquête sur les revenus avec des ménages d'une personne plus difficilement joignables*
- Perte de précision
- Diminution de la taille de l'échantillon

Biais de non -réponse

Deux strates



$$\overline{Y} = \frac{N_0}{N}\overline{Y}_0 + \frac{N - N_0}{N}\overline{Y}_1 = \frac{N_0}{N}\overline{Y}_0 + \overline{Y}_1 - \frac{N_0}{N}\overline{Y}_1$$

En l'absence d'hypothèse sur le mécanisme des données manquantes, seul $\overline{Y}_1$ peut être estimé

Biais : $\overline{Y} - \overline{Y}_1 = \frac{N_0}{N}\left(\overline{Y}_0 - \overline{Y}_1\right)$

2 PRINCIPES VIS-À-VIS DE LA NON-RÉPONSE

- Minimiser la non-réponse, l'idéal étant de ne pas en avoir
  - Faciliter la collecte
  - Soigner le questionnaire
  - Anticiper avec un plan de sondage adapté

- Après collecte, étudier le comportement de réponse observé

COMMENT LIMITER LA NON-RÉPONSE LORS DE LA COLLECTE ?

- Prévenir l'enquêté
  - *Exemple : envoyer une lettre-avis*
- Établir un rapport de confiance avec l'enquêté
- Informer sur l'aspect confidentiel
  - *Le secret statistique est défini dans la loi n° 51-711 du 7 juin 1951 modifiée sur l'obligation, la coordination et le secret en matière de statistiques*
- Offrir des conditions agréables pour le passage du questionnaire
- Donner des incitatifs, voire rémunérer l'enquêté

COMMENT LIMITER LA NON-RÉPONSE LORS DE LA CONCEPTION DU QUESTIONNAIRE ?

- Poser des questions courtes
- Éviter les questions trop complexes, les tournures négatives
- Eviter les abréviations
- Disposer d'une bonne traduction dans le cas où le questionnaire doit être passé en plusieurs langues
  - *Exemple : le recensement de la population*
- Éviter les questionnaires trop longs (pas plus d'une heure)

Questions sensibles ou indiscrètes: la méthode des questions aléatoires (Warmer, 1965)

<u>Première technique:</u>

On tire au sort dans une urne avec θ boules blanches et 1- θ boules noires la question

Si blanc: question A: « Avez-vous fraudé le fisc? »

Si noire: question : « Je n'ai pas fraudé »

On veut estimer P_A.

On recueille Π = Proba de Oui =

$$\theta P_A + (1-\theta)(1-P_A)$$

% de « Oui »

$$\widehat{P_A} = \frac{\widehat{\Pi} - (1-\theta)}{2\theta - 1} \qquad V\left(\widehat{P_A}\right) = \frac{1}{(2\theta-1)^2} V\left(\widehat{\Pi}\right) \simeq \frac{P_A(1-P_A)}{n} + \frac{1}{n}\frac{\theta(1-\theta)}{(2\theta-1)^2}$$

Inconvénient: $\overline{A}$ aussi indiscrète que A!

Deuxième technique:

Si blanche, question A sensible

Si noire, question B banale

$$\Pi = P_A\theta + P_B(1-\theta) \qquad \widehat{P_A} = \frac{\widehat{\Pi} - (1-\theta)P_B}{\theta}$$

$$V\left(\widehat{P_A}\right) \simeq \frac{\Pi(1-\Pi)}{n\theta^2} + \frac{P_B(1-P_B)(1-\theta)^2}{n\theta^2}$$

P_B peut être connu à l'avance ou estimé par une autre enquête.

Exemple:

A: combien de fois avez-vous avorté?

B: nombre idéal d'enfants?

Exemple: Brown 320 officiers

Consommation de drogue: 2 enquêtes, une anonyme, l'autre à question aléatoire

Drogue	Q. Anonyme	Q. aléatoire
Marijuana	5% (1.2)	9% (4.1)
Hallucinogène	1.6% (0.7)	11.6% (4.1)
Amphétamine	1.9% (0.7)	8% (3.3)
Barbiturique	0.6% (0.7)	7.9% (3.9)
Narcotique	0.3% (0.3)	4% (3.9)

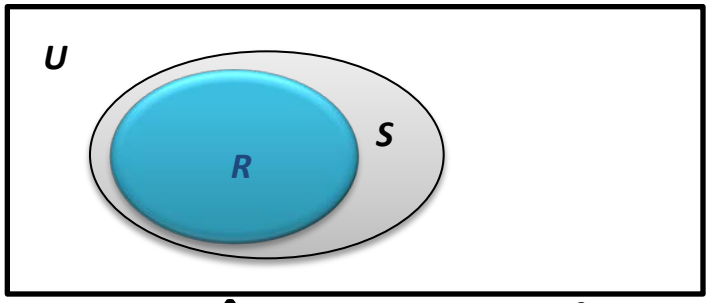
LE TRAITEMENT DE LA NON-RÉPONSE

- **Non-réponse totale : par repondération**
  - On augmente le poids des répondants de sorte à faire les parler aussi pour les non-répondants

- **Non-réponse partielle : par imputation**
  - On renseigne la valeur manquante par une donnée plausible

LES MÉTHODES DE REPONDÉRATION
LE PRINCIPE

- Augmenter le poids des individus répondants
- La non-réponse peut être modélisée comme la deuxième phase d'un sondage :
  - 1ère phase : le choix des enquêtés
  - 2ème phase : le choix des non-répondants



- Il faut donc connaître ou estimer les probabilités de réponse des individus

LES MÉTHODES DE REPONDÉRATION
LA TECHNIQUE

- Pour estimer un total : $Y = \sum_{k \in U} Y_k$
- En l'absence de non-réponse, on utilise l'estimateur d'Horvitz-Thompson : $\hat{Y} = \sum_{k \in S} \frac{Y_k}{\pi_k}$
- En présence de non-réponse, un estimateur sans biais est donné par : $\hat{Y}_r = \sum_{k \in R} \frac{Y_k}{\pi_k p_k}$

  où p_k est la probabilité de réponse de l'individu *k*
- Le poids des individus répondants est augmenté
- Mais les p_k sont inconnues
  - il faut les estimer par un « modèle de réponse »

MÉCANISME DE RÉPONSE GLOBALEMENT UNIFORME

- Chaque individu a la même probabilité de répondre
- Estimée par le taux de réponse empirique : $\hat{p} = \frac{r}{n} \quad \forall \mathrm{k}$

  avec r = nombre de répondants, n = nombre d'enquêtés
- Dans le cas d'un sondage aléatoire simple, cela revient à estimer la moyenne d'une variable par la moyenne calculée sur les répondants
- Epilogue :
  - Ne rien faire revient à postuler un mécanisme de réponse globalement uniforme
  - Cela sous-entend que les non-répondants ont le même comportement que les répondants au regard de la variable d'intérêt
  - Cette hypothèse est très forte
  - *Exemple d'une enquête sur la santé, l'état de santé influe sur le comportement de réponse*

MÉCANISME DE RÉPONSE HOMOGÈNE À L'INTÉRIEUR DE SOUS-POPULATIONS

- On suppose un taux de réponse constant au sein de sous-populations supposées homogènes en termes de comportement de non-réponse : $U = U_1 \cup \ldots \cup U_h \ldots \cup U_H$

$$p_k = p_h \quad \forall \mathrm{k} \in U_h$$

- Cela implique de disposer d'informations sur les non-répondants
- Pour constituer ces sous-populations, on peut utiliser :
  - des méthodes d'analyse des données
  - des méthodes économétriques.
- La probabilité de réponse au sein de la sous-population *h* est estimée par : $\hat{p}_h = \frac{r_h}{n_h} \quad \forall k \in U_h$

avec r_h = nombre de répondants dans U_h

et n_h = nombre d'unités de l'échantillon dans U_h

CALAGE ET NON-RÉPONSE

- Si on a réalisé une correction de la non-réponse totale (par repondération des répondants), ce sont ces nouveaux poids qu'il faut utiliser en entrée de Calmar

- Si on procède directement au calage sur l'échantillon des répondants sans traitement préalable de la non-réponse totale, on montre que ceci permet à la fois de corriger la non-réponse totale et d'améliorer la précision des estimations, ***sous la condition*** que les variables explicatives de la non-réponse soient incluses dans les variables de calage

Données manquantes

- Les mécanismes (Rubin,1976)
  - MCAR (Missing Completely at Random)
    - P(Y manquant) indépendant de Y et du reste
    - Hypothèse forte mais réaliste si volontaire
  - MAR (Missing at random)
    - P(Y manquant/Y,X)=P(Ymanquant/X)
    - Non testable
  - MCAR et MAR: données manquantes ignorables
  - Cas non ignorable: nécessité de modéliser le mécanisme pour obtenir des estimations sans biais
- Ignorer ou estimer les données manquantes?

Supprimer les DM?

- « listwise »
  - Perte d'information
  - Marche pour MCAR et en régression pour les X si MAR selon Y
- « Pairwise »
  - Utilisable pour modéle linéaire, ACP
    - Matrices non positives, statistiques de tests biaisées

Estimer les DM: l'imputation

- Compléter la non-réponse par une valeur plausible.
  - Méthodes implicites
  - modèles
- Attention, en traitant les données imputées comme de vraies réponses :
  - distributions faussées
  - estimateurs de variance incorrects

PRINCIPALES MÉTHODES D'IMPUTATION

- **Méthode déductive :**
  - On déduit la donnée manquante à partir des données renseignées
    - *Exemple : le revenu total obtenu par sommation de ses composantes*

- **Cold-deck :**
  - On remplace la donnée manquante par une donnée obtenue en dehors de l'enquête
    - *Exemple : le nombre de salariés lu dans le répertoire Sirène*

PRINCIPALES MÉTHODES D'IMPUTATION

- **Prédiction par la moyenne :**
  - On impute la moyenne observée sur les répondants
    - La distribution des valeurs est fortement modifiée
      - *Exemple : imputation du revenu moyen*

- **Prédiction par la moyenne par classe :**
  - On partitionne l'échantillon en classes (sous-populations)
  - On impute la moyenne observée sur les répondants de la même classe
      - *Exemple : imputation du revenu moyen par catégorie socio-professionnelle*

PRINCIPALES MÉTHODES D'IMPUTATION

- **Hot-deck ou Hot-deck par classe :**
  - On impute la valeur observée pour un répondant choisi au hasard, par un sondage aléatoire simple avec ou sans remise
    - *Exemple : imputation du chiffres d'affaires d'entreprises de même taille et de même activité*

- **Hot-deck séquentiel :**
  - On impute la valeur du répondant précédent
    - Importance du choix de la variable de tri

- **Hot-deck métrique :**
  - On impute la valeur observée de l'individu le plus proche au sens d'une distance donnée, mesurée sur les caractéristiques disponibles de tous les individus
    - *Exemple : imputation des dépenses pour un produit donné par un ménage de même catégorie socio-professionnelle, de la même commune*

PRINCIPALES MÉTHODES D'IMPUTATION

- **Prédiction par le ratio :**
  - On impute $Y_k = X_k \frac{\bar{Y}_r}{\hat{\bar{X}}_r}$
- **Prédiction par un modèle de régression :**
  - On considère le modèle $Y = X'\beta + \varepsilon$
  - On estimé β sur les répondants

Estimation basée sur des modèles

- Une donnée manquante sur une variable Y est modélisée à partir des variables X selon un modèle de régression
  - **régression simple** en prenant la variable la plus corrélée.
  - **régression multiple**
  - **modèle linéaire général** si X est nominale et la variable à expliquer est quantitative.
  - **Analyse discriminante,** ou **régression logistique** si Y nominal
- Remarque: cas particulier de l'estimation par la moyenne

Algorithme EM (espérance, maximisation)

- étape E: espérance conditionnelle de chaque donnée manquante sachant les données observées, d'où estimation des paramètres.
- étape M calcule les estimateurs du maximum de vraisemblance des paramètres, avec les lois conditionnelles des données manquantes.

convergence vers la valeur la plus probable de chaque donnée manquante pour l'estimation obtenue des paramètres

Maximisation de la cohérence interne, ou de l'homogénéité

- Présentation hollandaise de l'ACM de **G=(G1|G2|…|Gm)** comme la minimisation d 'une fonction de perte:

$$\sigma(X, Y) = \frac{1}{m} \sum_{j=1}^{m} (X - G_j' Y_j)' (X - G_j' Y_j)$$

$$X = \frac{1}{m} \sum_{j=1}^{m} G_j Y_j$$

- Les données manquantes sont complétées pour avoir σ minimal: ACM avec valeurs propres maximales.

MCA with missing data

Unit	Income	Age	Car
1	***x***	young	am
2	medium	medium	am
3	***y***	old	jap
4	low	young	jap
5	medium	young	am
6	high	old	am
7	low	young	jap
8	high	medium	am
9	high	***z***	am
10	low	young	am

Results of the 27 MCA

x	*y*	*z*	λ_1	*x*	*y*	*z*	λ_1	*x*	*y*	*z*	λ_1
l	l	j	.70104	m	l	y	.63594	h	l	y	.61671
l	l	m	.77590	m	l	m	.72943	h	l	m	.66458
l	l	o	.76956	m	l	o	.72636	h	l	o	.65907
l	m	j	.78043	m	m	y	.70106	h	m	y	.70106
l	m	m	.84394	m	m	m	.77839	h	m	m	.74342
l	m	o	.84394	m	m	o	.84394	h	m	o	.74342
l	h	j	.78321	m	h	y	.73319	h	h	y	.68827
l	h	m	.84907	m	h	m	.80643	h	h	m	.74193
l	**h**	**o**	***.84964**	m	h	o	.80949	h	h	o	.74198

- Solution unidimensionnelle peu réaliste:

  $\max (l_1 + l_2 + \ldots + l_k)$

- Recherche exhaustive impossible. Algorithmes itératifs.

deux inconvénients majeurs pour les imputations par modèle:

- risque d'incohérence: si plusieurs données manquantes sont estimées une par une et non conjointement, sans prendre en compte les corrélations
- variabilité sous-estimée: deux unités ayant les mêmes valeurs de **X** auront la même estimation pour la valeur manquante de Y

IMPUTATION MULTIPLE (Rubin)

- imputer chaque donnée par m>2 valeurs obtenues par tirage dans un ou plusieurs modèles d'estimation. Puis analyse des données sur chacun des m jeux de données complétés
- simulation de la distribution a posteriori des données manquantes , variances correctes.
- Mais: complexité des calculs, temps de calcul et volume considérable.

Fusions de fichiers

- Combiner des données provenant de sources différentes:
  - enquêtes, sources administratives, fichiers clients, données socio-économiques agrégées, etc.
- Chaque base peut être constituée d'unités statistiques différentes ou d'agrégation de ces unités à différents niveaux.

- Fusion de fichiers. Cas élémentaire:
- deux fichiers: F1 *p+q* variables mesurées sur n_0 unités, F2 sous-ensemble de *p* variables pour n_1 unités. Souvent n_0 est faible par rapport à n_1 .

X_0	Y_0
X_1	?

- Un cas plus complexe

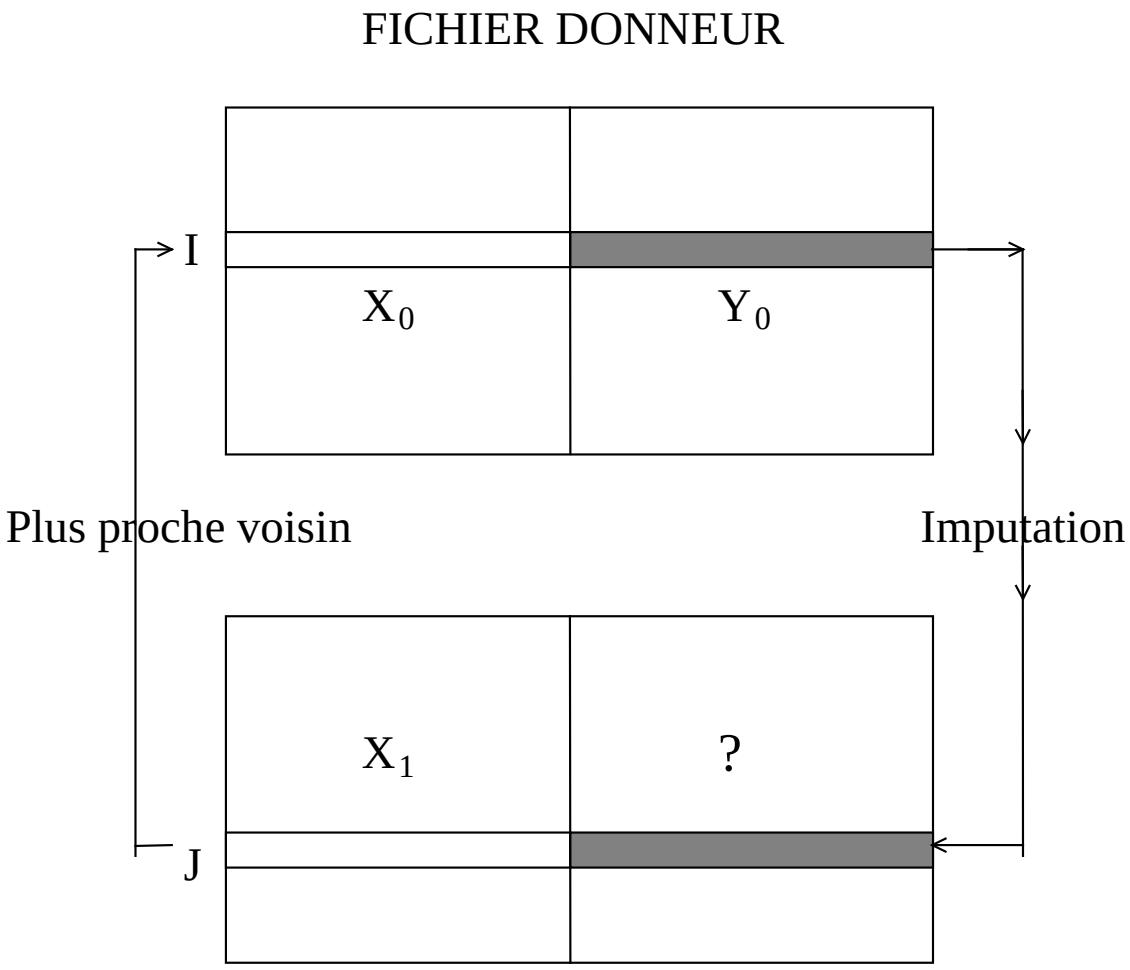
X_0		Y_0
X_1	Z_1	

Modèles et méthodes pour la fusion de données

- Appliquer industriellement une technique de traitement de données manquantes.
- Deux approches:
  - Méthodes d'imputation: compléter la non-réponse par une valeur plausible.
  - Repondération :affecter aux répondants des pondérations pour compenser les non-réponses
- Conditions à vérifier préalablement:
  - la taille de la population du fichier donneur est suffisamment importante par rapport au fichier receveur
  - les variables communes et les variables spécifiques possèdent des liaisons relativement fortes entre elles.

Les méthodes implicites:

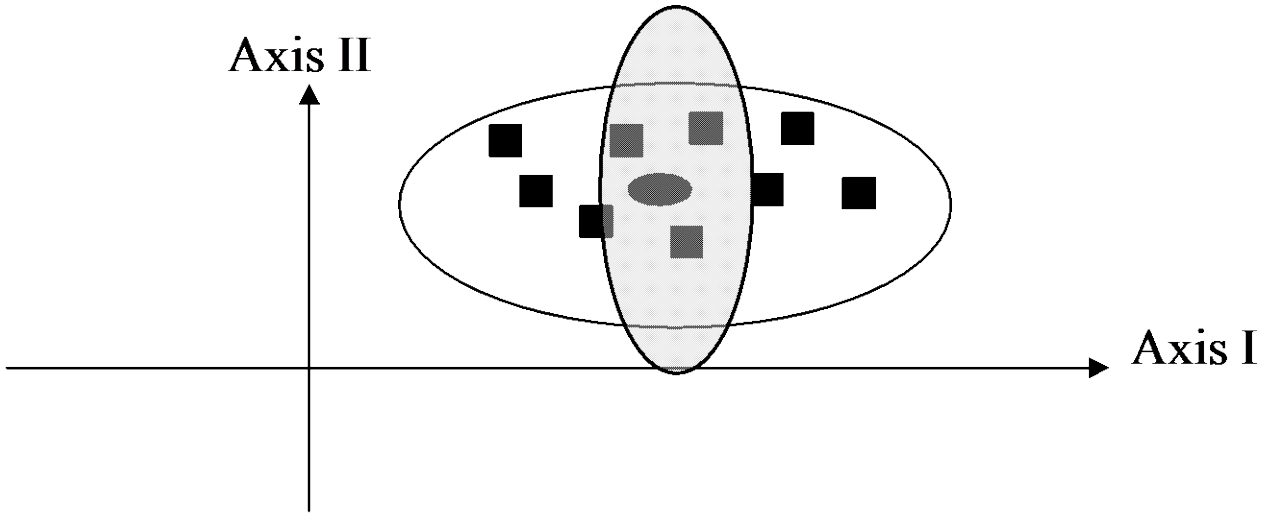
- fusion par appariements intra-cellulaires,
- imputation par Hot-Deck,
- méthode des plus proches voisins etc.…
- donner simultanément aux variables du fichier receveur toute l'information et les renseignements détenus par les variables du fichier donneur.


FICHIER DONNEUR
I
X_0
Y_0
Plus proche voisin
Imputation
X_1
?
J
FICHIER RECEVEUR

La fusion sur référentiel factoriel

- Fréquemment utilisée en France. Son principe (Santini 1984) repose sur :
  - ✍- les variables critiques : servent à déterminer pour l'individu du fichier receveur ses donneurs éligibles.
  - ✍- les variables de rapprochement : une partie des variables communes, par un calcul de distance, permettant de choisir pour chaque receveur le donneur éligible le plus proche

- Réferentiel factoriel: ACM sur l'ensemble des variables critiques ou communes
- Détermination d'un voisinage du receveur
- Choix final parmi les donneurs éligibles selon les variables de rapprochement (sexe, age, …)
- Pénalisation pour éviter de prendre trop souvent les mêmes donneurs (voir fusion par mariage)



Fusion par mariages

- éviter qu'un même donneur transmette son information à plusieurs receveurs (mariages multiples)
- si un donneur est déjà marié à n receveurs, d est pénalisée par :

$$d' = 1 - (1 - d)^n$$

- G. Santini a imaginé 6 types différents de relations de voisinage par “ mariage ”: A receveur, B donneur.
  - le mariage par “ coup de foudre ” (voisins réciproques) : si A est le plus proche voisin de B et si B est le plus proche voisin de A et n'a jamais été marié, alors A et B sont immédiatement mariés.
  - le mariage avec “ l'ami d'enfance ” : si B est le plus proche voisin de A, mais B est déjà marié à A' , alors A sera marié à B' qui est le plus proche voisin de A après B.
  - le mariage par “ adultère ” : variante du cas précédent quand d(B ’,A) est plus grand que la distance pénalisée entre A et B (puisque B est déjà marié a A'). On marie alors A et B.

- Fusion avec collage du vecteur entier du donneur
  - moins bon pour la reconstitution de données individuelles,mais garde la structure de corrélation et évite les incohérences
- Régression variable par variable.
  - C’est l’inverse
- Dans tous les cas il est nécessaire d'avoir:
  - Un nombre suffisant de variables communes
  - Des corrélations élevées entre variables communes et variables à imputer.
  - Une structure commune entre fichier donneur et fichier receveur: distributions comparables des variables communes ou critiques, sinon résultats biaisés. Redressements souvent nécessaires.

Validation

- procédures empiriques où on estime des données connues mais cachées que l'on compare ensuite aux vraies valeurs:validation croisées, bootstrap …
- Indicateurs:
  - reconstitutions de données individuelles
  - prévisions au niveau de groupes
  - reconstitutions de marges, de croisements

Un exemple:

- Données SPAD 992 interviews, divisées aléatoirement en deux fichiers : 800 donneur 192 receveur.
- 4 variables communes:

  Q1 - classe d'age(5 catégories),
  Q2 - taille d'agglomération (5 catégories),
  Q3 - heure de coucher (7 catégories),
  Q4 - age de fin d'études (5 catégories) .
- 3 variables d ’opinion Y à imputer:

  Q5 - La famille est le seul endroit où on se sent bien ? (oui, non)
  Q6 - Plus haut diplôme obtenu (7 catégories),
  Q7 - Taux d'écoute TV (4 catégories).

Table 3 performances individuelles

Méthode	***Classifications correctes***
Aléatoire	*49%*
Homogénéité max.	*54%*
FRF	*47%*

Table 4 performances marginales

Q5	***Vraies marges***	***Homogénéité max***	***FRF***
1	*136*	*136*	*125*
2	*56*	*56*	*67*
Q6			
1	36	6	49
2	70	114	65
3	35	16	27
4	29	23	33
5	4	33	1
6	18	33	15
7	0	0	2
Q7			
1	100	118	100
2	36	18	43
3	37	29	31
4	19	27	18

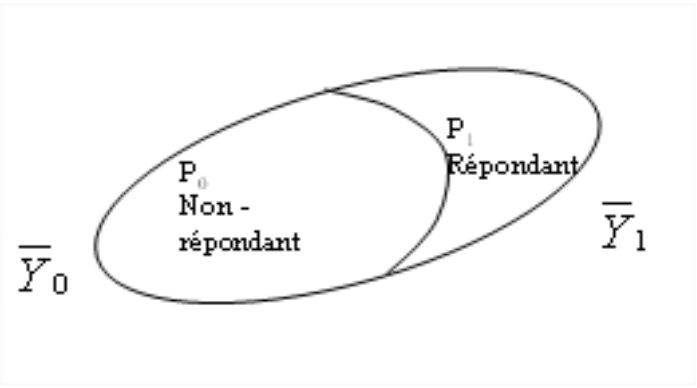
Le score de « propensity»

- Origine: essais cliniques avec affectation non-aléatoire entre traitement et témoin (contrôle)
  - Z=1 traité, Z=0 sinon. p covariables X= (x^1,x^2, ..,x^p)
  - propensity score e(x)=P(Z=1/X)
- Résumé unidimensionnel: permet de stratifier, de chercher des jumeaux (appariement), de repondérer en cas de données manquantes
- Estimé habituellement par une régression logistique

Application: données manquantes

- Si mécanisme ignorable:

$$\hat{\bar{Y}} = \frac{1}{N}\sum_{i=1}^{N}\frac{z_i y_i}{e(\mathbf{x}_i)}$$



cf Horwitz-Thomson

$$\hat{\bar{Y}}_0 \simeq \frac{1}{n_0}\sum_{i=1}^{N}\frac{z_i y_i (1-e(\mathbf{x}_i))}{e(\mathbf{x}_i)}$$

Application en fusion

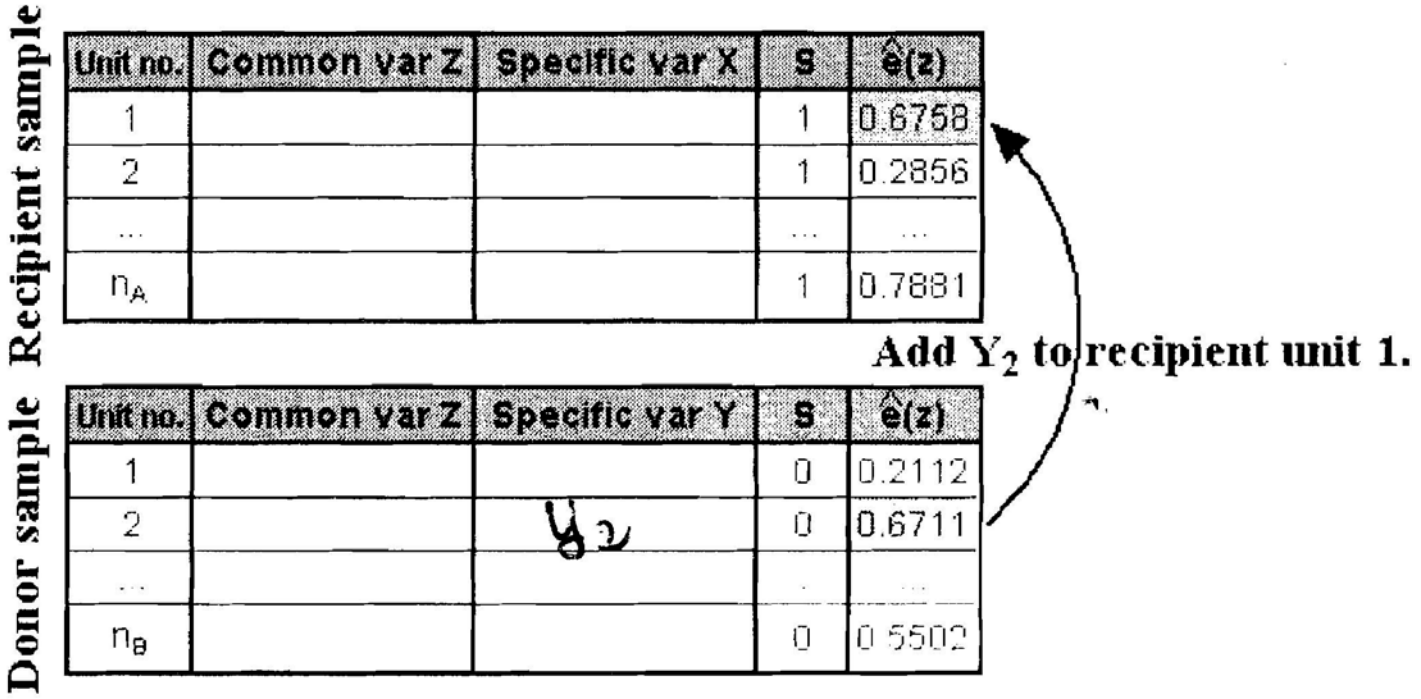


Figure 2.3. Principle of propensity score matching

S.Rässler, 2002

Propriétés (1)

- **Equilibrage:**

  Pour un score donné e(X), on tire des échantillons aléatoires simples parmi Z=1 et Z=0.

  Alors les lois de X dans chaque groupe sont les mêmes:

  $P(X / Z=1,e(X))= P(X / Z=0,e(X))$

- Avantage: facile de fabriquer des échantillons appariés même si X est de grande dimension
  - Si appariement exact impossible : ppv ou strates

Propriétés (2)

- **Consistence** : estimation sans biais de l'effet τ d'un traitement Y :
  - $\tau = E(Y_t)-E(Y_c)$
  - si l'effet de l'affectation traitement-contrôle est ignorable conditionnellement à X (donc à e(X)) et si $0<P(Z=1/X)<1$ (**Y_t et Y_c sont indépendants de Z conditionnellement à X)**
  - alors τ est estimé sans biais par la moyenne des différence entre observations appariées selon e(X)

Propriétés (3)

- **Etudes d'observation non randomisées**
  - résout le problème de l'inférence causale
  - réduit les biais « ouverts »: ex: comparer la mortalité des fumeurs et non-fumeurs alors que les fumeurs sont en moyenne plus vieux que les non-fumeurs
  - à comparer avec la post-stratification

Rééquilibrage d'enquête

- Exemple
  - une enquête de référence aléatoire,
  - une enquête web

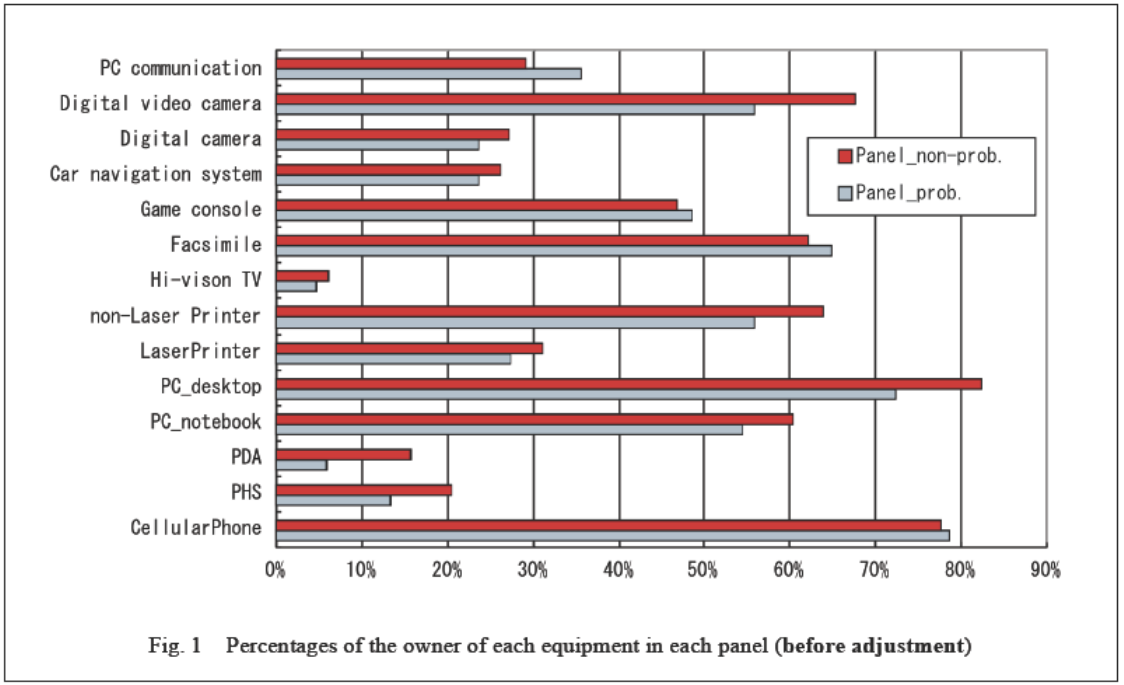


Fig. 1 Percentages of the owner of each equipment in each panel **(before adjustment)**

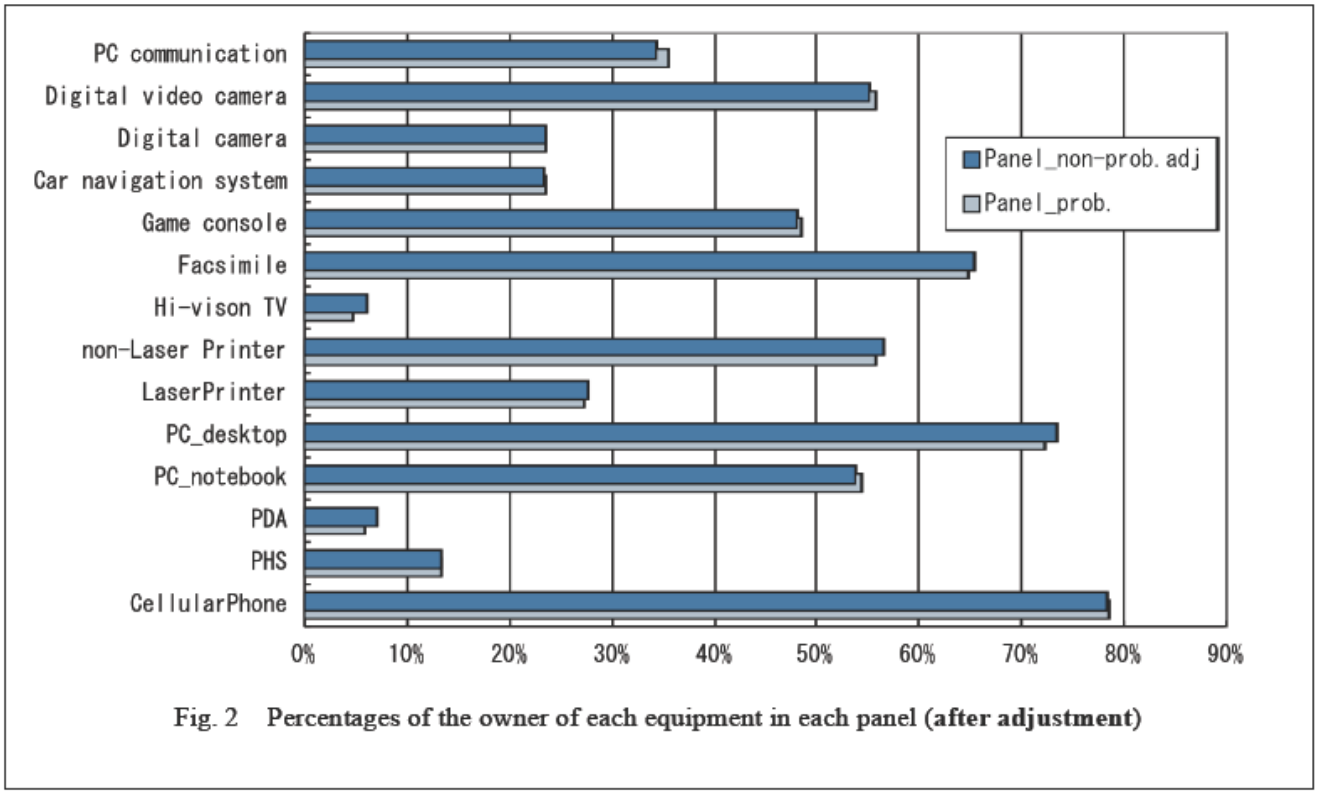


Fig. 2 Percentages of the owner of each equipment in each panel (**after adjustment**)

Yoshimura,

- Plus simple que la post-stratification sur plusieurs variables (calage sur marges)
- Mais hypothèses fortes

Conclusions

- Techniques:
  - L’estimation de données manquantes massives: stimulant pour les statisticiens.
  - besoin réel de fournir à l'utilisateur final une base unique sans “ trou ”.

- Prudence quand on utilise des “ données ” qui sont en réalité des estimations et non des valeurs observées: ne jamais utiliser à un niveau individuel, mais uniquement agrégé.
- Conséquence perverse: un moindre effort de collecte, puisque l'on peut reconstituer des données…
- Nécessité de valider

- Déontologiques (confidentialité et protection de la vie privée) :

  - des données qui n'ont pas été recueillies mais estimées, peuvent être ajoutées dans des fichiers à l'insu des individus concernés. Quid de La loi “ Informatique et Liberté ” ?
  - paradoxe alors que les INS développent des techniques pour assurer la confidentialité

Références

- Allison P. (2002) *Missing data*, Sage Publications
- Caron, N.(2006) : *Les principales techniques de correction de la non-réponse,et les modèles associés*, Série des Documents de Travail « Méthodologie Statistique » de l'Insee, n°9604
- Co V. (1997) *Méthodes statistiques et informatiques pour le traitement des données manquantes.* Doctorat, CNAM. Paris.
- Fischer N. (2004) *Fusion Statistique de Fichiers de Données*. Doctorat, CNAM, Paris.
- Rässler S. (2002), *Statistical matching*, Springer
- Rosenbaum P.R., Rubin D. (1983) The central role of propensity scores in observational studies for causal effects, *Biometrika* 70, 41-55
- Saporta G. (2002) Data fusion and data grafting . *Computational Statistics and Data Analysis*, 38(4),465-473