Une (rapide) introduction aux données manquantes

David Hajage david.hajage@aphp.fr

Sorbonne Université

2026-02-06

SANTÉ SORBONNE UNIVERSITÉ

SANTÉ SORBONNE UNIVERSITÉ

Sommaire

SANTÉ SORBONNE UNIVERSITÉ

1

Pourquoi existe-t-il des données manquantes ?

Non réponse globale (non réponse au niveau de l'individu)

Dans une étude/enquête, certains individus peuvent être injoignables ou refuser de participer.

Non réponse à certains items

Certains individus peuvent ne pas connaître la réponse à des questions spécifiques ou refuser d'y répondre.

Erreurs de collecte ou de traitement des données

Un enquêteur peut, par exemple, oublier de poser une question à un individus, ou de collecter ceraines données.

Données manquantes prévues par le plan d'étude

Certaines questions peuvent être posées uniquement à un sous-échantillon aléatoire d'individus.

Certaines valeurs peuvent être *censurées*

En analyse de survie (ou de délai jusqu'à un événement), l'événement d'intérêt (par exemple le décès) peut ne pas survenir avant la fin de la période de suivi d'un individu.

Données manquantes *vs.* données conditionnellement non définies

Les données manquantes doivent être distinguées des données qui sont conditionnellement non définies.

Par exemple, un répondant sans enfants ne peut pas indiquer l'âge de ses enfants.

Il n'existe pas de solution miracle

Toutes les données manquantes n'ont pas a être prises en compte

- Données manquantes prévues par le plan d'étude
- "Données manquantes" conditionnellement définies

! Important

Pour les autres, il est généralement impossible de procéder de manière rigoureuse sans faire au moins des hypothèses partiellement invérifiables sur le mécanisme générant les données manquantes.

Notations

Rubin a introduit plusieurs distinctions fondamentales concernant les données manquantes.

Soit la matrice X $(n \times p)$ représentant les données complètes pour un échantillon de n observations sur p variables.

- Certaines entrées de X, notées X_{mis}, sont manquantes.
- Les entrées observées de X sont notées X_{obs}.

SANTÉ SORBONNE UNIVERSITÉ

2

Types de données manquantes

2.1 Les mécanismes MCAR, MAR, MNAR
2.2 Mécanismes ignorables et non ignorables
2.3 Données utilisées dans la suite

SANTÉ SORBONNE UNIVERSITÉ

2

Types de données manquantes

2.1 Les mécanismes MCAR, MAR, MNAR

2.2 Mécanismes ignorables et non ignorables

2.3 Données utilisées dans la suite

Données manquantes complètement au hasard (MCAR)

Les données sont dites *manquantes complètement au hasard* (*Missing Completely At Random*, MCAR) si les données observées peuvent être considérées comme un échantillon aléatoire simple des données complètes.

La probabilité qu'une valeur soit manquante est indépendante de cette valeur elle-même et de toute autre valeur (observée ou manquante) du jeu de données.

Données manquantes au hasard (MAR)

Les données sont *manquantes au hasard* (*Missing At Random*, MAR) lorsque le mécanisme de non-réponse dépend des données observées mais, conditionnellement à celles-ci, pas des données manquantes elles-mêmes.

? Exemple

- Certains individus refusent de déclarer leur revenu et diffèrent systématiquement du reste de l'échantillon.
- Si la décision de ne pas déclarer le revenu est indépendante du revenu lui-même, conditionnellement aux informations *observées* (niveau d'éducation, profession, etc.), alors les données sont MAR.

Données manquantes au hasard (MAR) (cont.)

La condition MCAR est un cas particulier plus restrictif de MAR. Exemple :

? Exemple

- **MCAR** : la probabilité que le revenu soit manquant est indépendante du niveau de revenu lui-même, et est la même pour tous les individus.
- **MAR** : la probabilité que le revenu soit manquant est indépendante du niveau de revenu lui-même, et est la même pour tous les individus *ayant un même niveau d'éducation*.

Données manquantes non au hasard (MNAR)

Les données sont *manquantes non au hasard* (*Missing Not At Random*, MNAR) lorsque la probabilité d'être manquante dépend de la valeur manquante elle-même, même après prise en compte des données observées.

? Exemple

Si, à caractéristiques observées égales, les individus à revenu élevé sont plus susceptibles de ne pas déclarer leur revenu, alors les données manquantes sur le revenu sont MNAR.

SANTÉ SORBONNE UNIVERSITÉ

2

Types de données manquantes

2.1 Les mécanismes MCAR, MAR, MNAR

2.2 Mécanismes ignorables et non ignorables

2.3 Données utilisées dans la suite

Mécanismes ignorables et non ignorables

- Si les données sont MCAR ou MAR, il n'est pas nécessaire de modéliser explicitement le mécanisme de données manquantes pour obtenir des analyses valides.
  - Le mécanisme est alors dit *ignorable*.
  - *Cela ne veut pas dire qu'on ne fait rien* pour améliorer la précision ou diminuer les biais (en particulier en cas de mécanisme MAR).

Peut-on déterminer si les données manquantes sont ignorables ?

En dehors de situations particulières (par exemple des données manquantes prévues par le plan d'étude), il est en général impossible de savoir si les données sont MCAR, MAR ou MNAR.

- Montrer que la non-réponse dépend de données observées exclut l'hypothèse MCAR.
- En revanche, ne pas détecter de lien avec les données observées ne prouve pas que les données sont MCAR.
  - Les non-répondants peuvent différer des répondants selon des caractéristiques non observées.
-

Peut-on déterminer si les données manquantes sont ignorables ?

En dehors de situations particulières (par exemple des données manquantes prévues par le plan d'étude), il est en général impossible de savoir si les données sont MCAR, MAR ou MNAR.

- Montrer que la non-réponse dépend de données observées exclut l'hypothèse MCAR.
- En revanche, ne pas détecter de lien avec les données observées ne prouve pas que les données sont MCAR.
  - Les non-répondants peuvent différer des répondants selon des caractéristiques non observées.
- Voyons maintenant quelques méthodes de prise en compte des données manquantes.

SANTÉ SORBONNE UNIVERSITÉ

2

Types de données manquantes

2.1 Les mécanismes MCAR, MAR, MNAR
2.2 Mécanismes ignorables et non ignorables
2.3 Données utilisées dans la suite

Simulations des données

On simule 250 observations issues d'une loi normale bivariée avec :

- Moyennes : $\mu_1 = 10$, $\mu_2 = 20$
- Variances : $\sigma_1^2 = 9$, $\sigma_2^2 = 16$
- Covariance : $\sigma_{12} = 8$

La corrélation entre X_1 et X_2 est : $\rho_{12} = \frac{8}{\sqrt{9 \times 16}} = \frac{2}{3}$

SIMULATIONS DES DONNÉES (CONT.)

Pentes de régression :

- Régression de X_1 sur X_2 : $\beta_{12} = 8/16 = 0.5$
- Régression de X_2 sur X_1 : $\beta_{21} = 8/9 \approx 0.889$

```r
set.seed(20260204)
n <- 250
mu <- c(10, 20)
sigma <- matrix(c(9, 8, 8, 16), 2, 2)
df <- as.data.frame(MASS::mvrnorm(n = n, mu = mu, Sigma = sigma))
names(df) <- c("X1", "X2")
```

La variable X_1 est complètement observée, tandis que X_2 contient des données manquantes, générées selon 3 mécanismes.

SIMULATIONS DES DONNÉES MANQUANTES

- **MCAR** : 100 observations de X_2 sont rendues manquantes au hasard.

```
df$X2_MCAR <- df$X2
df$X2_MCAR[sample(1:nrow(df), 100)] <- NA
```

SIMULATIONS DES DONNÉES MANQUANTES (CONT.)

- **MAR** : la probabilité que X_2 soit manquante dépend de X_1 observée.

```
pmiss <- plogis(-0.5 - (2/3)*(df$X1 - 10)) # proba de NA ↓ quand X1 ↑
df$X2_MAR <- df$X2
df$X2_MAR[as.logical(rbinom(n = nrow(df), 1, pmiss))] <- NA
sum(is.na(df$X2_MAR)) # nombre de NA
```

```
[1] 103
```

Comme X_1 et X_2 sont corrélés positivement, les petites valeur de X_2 sont plus souvent manquantes.

SIMULATIONS DES DONNÉES MANQUANTES (CONT.)

- **MNAR** : la probabilité que X_2 soit manquante dépend de la valeur (potentiellement non observée) de X_2.

```
pmiss <- plogis(-0.5 - 0.5*(df$X2 - 20)) # proba de NA ↓ quand X2 ↑
df$X2_MNAR <- df$X2
df$X2_MNAR[as.logical(rbinom(n = nrow(df), 1, pmiss))] <- NA
table(is.na(df$X2_MNAR)) # nombre de NA
```

```

FALSE  TRUE
  145   105
```

Les petites valeur de X_2 sont également plus souvent manquantes.

Questions clés pour chaque approche

1. Les estimateurs sont-ils sans biais ?
2. Les inférences statistiques sont-elles valides (IC, valeurs p) ?
3. Les données observées sont-elles utilisées efficacement ?

SANTÉ SORBONNE UNIVERSITÉ

3

Méthodes basiques

3.1 Analyses sur cas complets
3.2 Imputation par la moyenne
3.3 Best-worst case analysis

SANTÉ SORBONNE UNIVERSITÉ

3

Méthodes basiques

3.1 Analyses sur cas complets

3.2 Imputation par la moyenne

3.3 Best-worst case analysis

Complete-cases analysis

Les observations avec au moins une valeur manquante sont exclues.

Avantages :

- Simple à mettre en œuvre.
- Estimations sans biais si les données sont MCAR.
- (Estimations sans biais des coefficients de régression si les données manquantes ne dépendent pas de la variable à expliquer.)

COMPLETE-CASES ANALYSIS (CONT.)

Inconvénients :

- Peu efficace : peu mener à l'exclusion d'un nombre important d'observation.

> **?** Exemple
>
> Même si chaque variable ne comporte qu'un faible pourcentage de données manquantes, celles-ci peuvent concerner des observations différentes d'une variable à l'autre. L'exclusion de toutes les observations présentant au moins une valeur manquante peut réduire fortement la taille de l'échantillon analysé.

- Peut conduire à des biais importants si les données manquantes sont MAR ou MNAR.

SANTÉ SORBONNE UNIVERSITÉ

3

Méthodes basiques

3.1 Analyses sur cas complets

3.2 Imputation par la moyenne

3.3 Best-worst case analysis

Moyenne globale

balb

Moyenne prédite

blss

SANTÉ SORBONNE UNIVERSITÉ

3

Méthodes basiques

3.1 Analyses sur cas complets
3.2 Imputation par la moyenne
3.3 Best-worst case analysis

SANTÉ
SORBONNE
UNIVERSITÉ

4

Pondération

Article

seaman-white-2011-review-of-inverse-probability-weighting-for-dealing-with-missing-data.pdf

SANTÉ
SORBONNE
UNIVERSITÉ

5

Imputation multiple

Article

Erler