Données manquantes

Véronique Tremblay

Valeurs manquantes et valeur absente

Valeur manquante

Il y a une valeur pour cette observation, mais qu'elle n'a pas été capturée.

Valeur absente

Cette variable n'était pas pertinente pour cette observation.

Exemple

Identifiant	Véhicule	Vitesse	Distraction	Ceinture
1	Auto	35	Aucune	Oui
2	Auto	80	Aucune	
3	Vélo	22	Écouteurs	
4	Auto	45		Oui
5	Auto	50	Nourriture	Oui

3 mécanisme de non-réponse

Données manquantes...

1. complètement au hasard (MCAR)

2. au hasard (MAR)

3. pas au hasard (NMAR)

X: une matrice de données de dimension $n \times p$
X_{ij} donne la valeur de la variable j pour l'observation i
θ: un vecteur de paramètres

X_{obs}: les données observée
X_{mis}: les données manquantes
R: est la matrice de réponse de dimension $n \times p$

$$R_{ij} = \begin{cases} 1 \text{ si } X_{ij} \text{ est observé} \\ 0 \text{ si } X_{ij} \text{ est manquant} \end{cases}$$

Mécanisme de non-réponse

On s'intéresse à la distribution de la matrice de réponse:

$$f(R|X, \theta)$$

Mécanisme de non réponse

MCAR (Missing completely at random)

La probabilité de réponse ne dépend pas de X

$$f(R|X,\theta) = f(R|\theta)$$

Exemple MCAR

Un appareil de mesure perd au hasard 30% des valeurs mesurées.

Alors

$$f(R_{ij}|\theta) \sim \text{Bernoulli}(0.7)$$

$\forall i = 1, \dots, n, \quad j = 1, \dots, p$

MAR (Missing At Random)

Lorsque la probabilité de réponse dépend uniquement de la valeur des variables qui ont été observées

$$f(R|X,\theta) = f(R|X_{obs},\theta)$$

Exemple MAR

Identifiant	Véhicule	Vitesse	Distraction	Ceinture de sécurité
1	Auto	35	Aucune	Oui
2	Auto	80	Aucune	
3	Vélo	22	Écouteurs	
4	Auto	45	Aucune	Oui
5	Auto	50		Oui

Exemple MAR

On fait remplir un questionnaire à plusieurs individus, en leur demandant leur sexe et leur salaire.

Si les femmes ont une plus grande probabilité que les hommes d'accepter de fournir leur salaire, alors on aurait un mécanisme MAR.

Mécanisme de non-réponse

NMAR (Not Missing At Random)

La probabilité de réponse dépend de X_{mis}

Exemples NMAR

- La probabilité de réponse dépend de la variable elle-même.
  - Les individus avec des salaires plus élevés ont une plus grande probabilité de refuser de déclarer leur salaire.

- La probabilité de réponse dépend d'une variable non-observée.
  - La probabilité de déclarer son salaire dépend de l'âge de l'individu, mais on n'a pas mesuré cette variable.

Quoi faire?

1. Documenter la non-réponse

2. Identifier le mécanisme de non-réponse
   - Exploration des données
   - Tests pour MCAR
   - Connaissance du domaine

3. Gérer la non-réponse.

Gérer la non-réponse

1. Retirer les variables qui contiennent trop de données manquantes

Perte d'information

2. Retirer les observations avec des données manquantes

Risque de perte de représentativité de l'échantillon si MAR ou NMAR

3. Créer une catégorie «inconnu»
4. Imputer

Résumé

- Distinguer valeur manquante et valeur absente
- 3 Mécanisme de non-réponse
- Gérer la non-réponse