

Données manquantes et imputation

Véronique Tremblay

Remplacer les valeurs manquantes.



Imputer par la moyenne

Remplacer les valeurs manquantes par la moyenne pour chaque variable

☹️ Modifie la distribution

☹️ Altère la structure de dépendance

Exemple

- Échantillon de 60 ($n = 60$) étudiants auquel on a demandé l'âge (X_1) et le niveau d'expérience(X_2), en année.

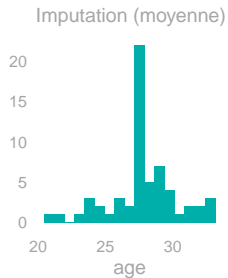
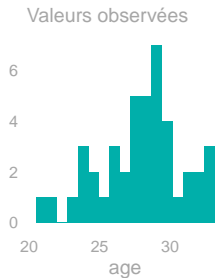
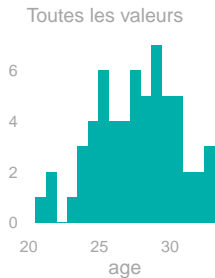
▪

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

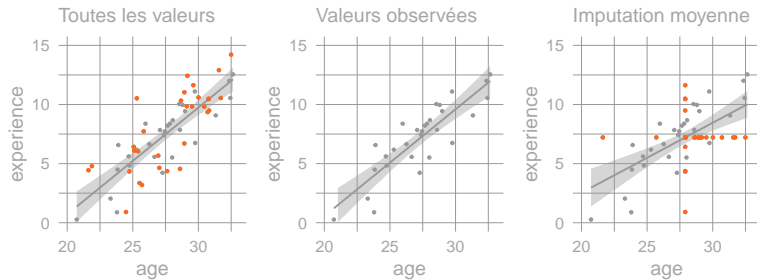
- La *vraie* relation entre les deux variables est

$$X_2 = -20 + X_1$$

Exemple



Exemple



Cas	$\hat{\beta}_0$	$\hat{\beta}_1$
Toutes les données	-17.3	0.9
Valeurs observées	-17.5	0.9
Imputation (moyenne)	-9.3	0.59



Imputer par la distribution

On remplace par une autre valeur pigée au hasard dans les données

😊 Conserve la distribution

😞 Altère la structure de dépendance




Imputer par un modèle

On utilise les autres variables du jeu de données pour prédire les valeurs manquantes

😊 Préserve la distribution et la structure de dépendance

😞 Réduit la variabilité

-  *Utilisez la librairie mice*
- Utilisez R comme variables supplémentaires

- Imputation par la moyenne
- Imputer par la distribution
- Imputer avec un modèle