# Analyse de Survie
# Professeur Abdellatif El Afia

# Données manquantes

# Problématique

- Il est quasiment impossible dans une étude réelle d'avoir des données totalement complètes.
- Traiter ces données de façon à ne pas biaiser les résultats..

# Caractéristiques

Données manquantes Intentionnelles ; prévues par l'enquêteur, comme des réponses du type oui ou non ? si oui une autre sous question.

Données manquantes Non intentionnelles ; hors de contrôle de l'enquêteur, un patient oublie ou refuse de répondre à une question.

# Notations

- $Y$ une matrice $n.p$ contenant les valeurs de $p$ variables pour tous les $n$ individues de l'échantillon.
- $R$ l'indicatrice des réponses, une patrice $n \times p$ de $0$ et $1$ telle que :

$$r_{ij} = 1 \text{ si } y_{ij} \text{ est observée, et } r_{ij} = 0 \text{ sinon.}$$

- Les données observées sont désignées par $Y_{obs}$ et celles manquantes par $Y_{mis}$.
- $Y = (Y_{obs}, Y_{mis})$ contient toutes les valeurs de données. Cependant les valeurs de $Y_{mis}$ nous sont inconues elle sont indiquées par $R$.
- Si $Y = Y_{obs}$ c'est à dire que l'échantillonest complètement observé
- Si aucune données n'a été obtenue pour l'individu $i$, la $i^{ème}$ ligne de $Y$ contiendra uniquement l'identifiant de $i$ et, éventuellement, des données administratives (cas de non − éponse totale).

# Mécanismes des données manquantes

- Mécanisme de non-réponse.
- Le processus qui régit les probabilités d'absence des données.
- $P(R/Y_{obs}, Y_{mis}, \psi)$ :Le modèle de données manquantes dont les paramètres sont contenus dans $\psi$ décrit la relation dépendance de la distrubution de R qui dépond de $Y = (Y_{obs}, Y_{mis})$.

# MCAR

- Les données sont dites on MCAR (Missing Completly At Random) si :

$$P(R/Y_{obs}, Y_{mis}, \psi) = P(R = 0/\psi).$$

- MCAR est souvent irréaliste en pratique.

# MAR

- Les donnes sont dites MAR (Missing At Random) si :

$$P(R/Y_{obs}, Y_{mis}, \psi) = P(R = 0/Y_{obs}, \psi)$$

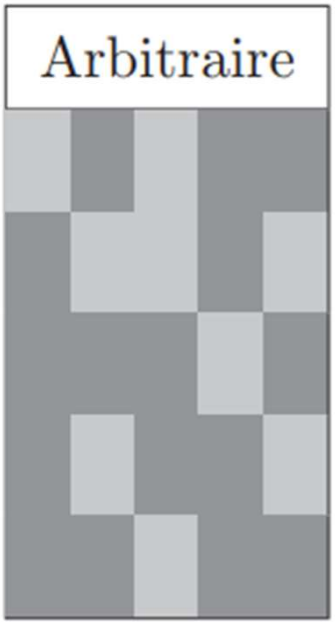
- MAR est plus général et plus réaliste que MCAR

# MNAR

Les données sont dites MNAR (Missing Not At Random) si $P(R/Y_{obs}, Y_{mis}, \psi)$ ne se simplifie pas.

# Configuration des données manquantes





## Flux entrant

Le coefficient de flux entrant $I_j$ est défini par :

$$I_j = \frac{\sum_j^p \sum_k^p \sum_i^n (1 - r_{ij}) r_{ik}}{\sum_k^p \sum_i^n r_{ik}}$$

$I_j$ est le nombre de variables paires $(Y_j, Y_k)$ avec $Y_j$ manquant et $Y_k$ observé, devisé par le nombre totale de cellules observées.

# Flux sortant

Le coefficient de flux sortant $O_j$ est défini par :

$$O_j = \frac{\sum_j^p \sum_k^p \sum_i^n (1 - r_{ik}) r_{ij}}{\sum_k^p \sum_i^n (1 - r_{ij})}$$

$O_j$ est le nombre de variables paires $(Y_j, Y_k)$ avec $Y_j$ observé et $Y_k$ manquante, devisé par le nombre totale de cellules incomplètes.

# Méthodes de gestion de données manquantes

- Analyse des cas complets (Complete case analysis)
- Etude des cas disponibles (Pairwise deletion)
- Procédure de modélisation de la distribution des données manquantes
- Imputation par la moyenne
- Imputation par régression
- LOCF et BOCF
- Méthode de l'indicatrice
- L'imputation multiple

## Analyse des cas complets (Complete case analysis)

- Eliminer les lignes (individus) ayant des données manquantes dans les variables à analyser
- Peut fournir des estimations de la moyenne, des coefficients de régression et de corrélation biaisés.
- Pas toujours mauvaise

# Etude des cas disponibles (Pairwise deletion)

- Ne considérer que les cas où les variables d'intérêt sont complètes.
- La moyenne d'une variable $X_i$ est calculée sur tous les cas ayant des données observées sur $X_i$, et la corrélation et covariance de deux variables $X_i$ et $X_j$ sont calculer sur tous les cas où les deux variables sont observées à la fois.

# Procédure de modélisation de la distribution des données manquantes

- Pas besoin d'imputer les données incomplètes ou les éliminer de l'analyse
- les paramètres de la distribution du modèle sont estimés par maximum de vraisemblance (maximal likelyhood)
- Faire des inférences et estimer les valeurs des données manquantes.
- Evaluées en estimant leur erreur standard.

# Imputation par la moyenne

- Remplacer les valeurs manquantes de chaque variable par sa moyenne, ou bien le mode pour les données catégoriques.
- Donne une variance sous-estimée et dérange les relations entre les variables.
- Donne des estimations biaisées même pour la moyenne quand les données ne sont as MCAR.
- Utilisée uniquement pour les cas où seules quelques valeurs sont manquantes.

## Imputation par régression

- Production d'un modèle à partir des données observées en incorporant les autres variables.
- Avec la condition que les variables soient corrélées linéairement.
- Sous MCAR, cette méthode fournit des estimations non biaisées de la moyenne. Si les paramètres influençant les données manquantes font partie du modèle, les poids de la régression ne sont pas biaisés sous la condition MAR.

# LOCF et BOCF

- Remplacer les valeurs manquantes par la dernière valeur observée (Last Observation Carried Forward).

- Remplacer les valeurs manquantes par la valeur de référence (Baseline Observation Carried Forward).

- Estimations biaisées même sous MCAR

# Méthode de l'indicatrice

- Remplacer chaque valeur manquante par 0.
- Peut produire des résultats biaisés même sous la condition MCAR.
- Les conditions sous lesquelles cette méthode est valide sont difficiles à obtenir en pratique.
- Ne permet pas d'avoir des données manquantes dans les événements d'intérêt.

# L'imputation Multiple

- Créer un nombre $m$ de jeu de données ($m > 1$).
- réalise l'étude statistique souhaitée sur chacun des jeux de données.
- Combiner les estimations de toutes ces études en une seule avec une erreur standard calculée.

# L'imputation Multiple

- Etape 1 : l'imputation

La méthode crée $m$ jeu de données complets en remplaçant chaque valeur manquante par une valeur plausible tirée spécifiquement d'une distribution modélisée à partir des données observées. Les $m$ jeux de données sont identique pour les valeurs observées, mais différentes pour celle à imputer.

## L'imputation Multiple

- Etape 2 : l'analyse

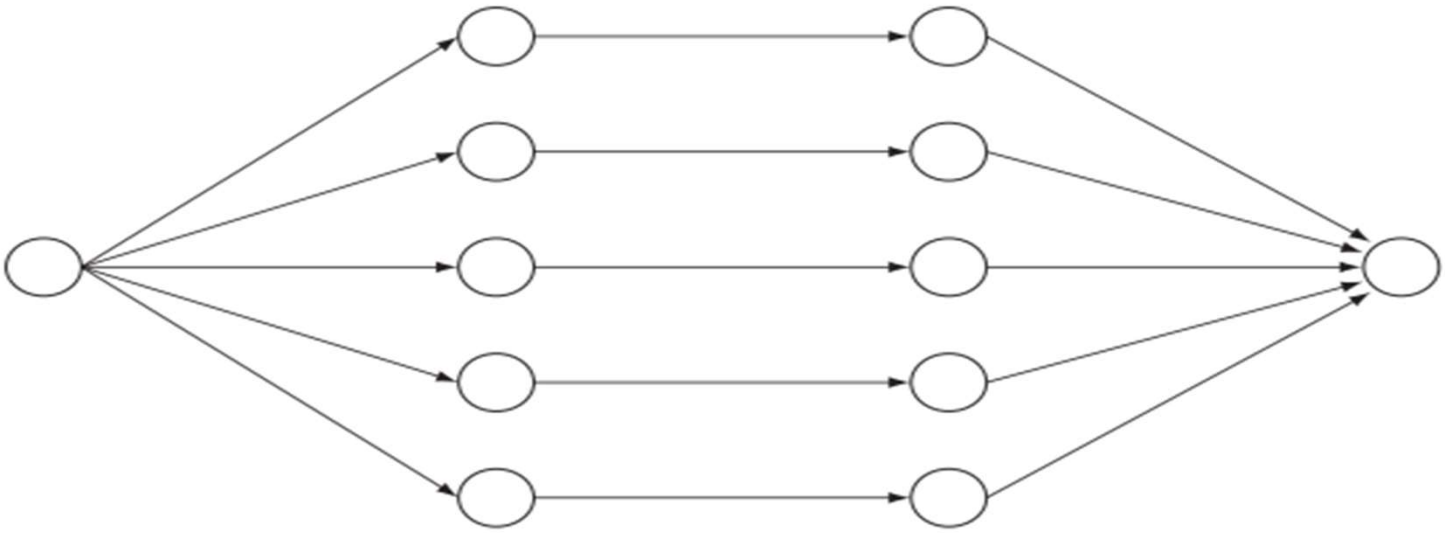
Ensuite une analyse statistique est utilisée sur chaque jeu de données maintenant tous imputés, pour appliquer les méthodes voulues comme si les données étaient complètes et en estimer les paramètres d'intérêt. Ce qui donne $m$ résultats différents. Ces différences coulent de l'incertitude sur les valeurs imputées.

# L'imputation Multiple

- Etape 3 : combinaison des estimations

Les $m$ parametres estimés sont ensuite combinés en une seule dont on estimera la variance. Sous les bonnes conditions, les valeurs estimées ne sont pas biaisées et ont des propriétés statistiques correctes.

# L'imputation Multiple



Données incomplètes | Données imputées | Résultats de l'analyse | Résultats combinés

# L'imputation Multiple Bayésienne

L'imputation multiple bayésienne est une méthodes bayésienne qui permet d'imputer une variable **continue** sous le modèle linéaire normal.

$$\dot{y} = \dot{\beta}_0 + X_{mis}\dot{\beta}_1 + \dot{\epsilon}$$

Le choix idéal pour imputer les variables distribuées normalement.

# L'imputation Multiple

- Permet d'imputé une variable quantitative discrète ou continue.
- Pour chaque valeur manquante prendre un petit nombre de candidats (3 à 10) qui ont des valeurs proches pour les variables observées.
- Remplacer la variable manquante par celle d'un candidat du lot tiré aléatoirement.

# L’imputation Multiple, Variables Qualitatives.

la régression logistique : pour des variable binaires incomplètes.

$$P(y_i = 1|X_i, \beta) = \frac{\exp(X_i\beta)}{1 + \exp(X_i\beta)}$$

Le modèle Logit multinomial pour les variable nominale de k catégories

$$P(y_i = k|X_i, \beta) = \frac{\exp(X_i\beta_k)}{\sum_{k=1}^{K} \exp(X_i\beta_k)}$$

# TP 4 :

- Votre Data contient elle des données manquantes? Sinon, en faire une copie en supprimant quelque données.
- Utiliser une méthode d'imputation multiple pour compléter la copie de Data.
- Reprendre le modèle Cox-PH sur la Data imputée.
- Comparer les résultats