Les données manquantes

Solutions à quelques situations

Hans Ivers, Ph.D.

hans.ivers@psy.ulaval.ca

Formation ACSPUL

27 novembre 2019

FAS-1113

UNIVERSITÉ LAVAL

Thèmes de la formation

1) Impacts des données manquantes (DM)
2) Patrons de DM
3) Approches dans la gestion des DM
4) Situation 1 : questionnaire avec DM
5) Situation 2 : ANOVA/régression avec DM
6) Situation 3 : Modèles latents
7) Situation 4 : Analyses longitudinales
8) Situation 5 : Analyses multiniveaux
9) Références bibliographiques

Impacts des DM

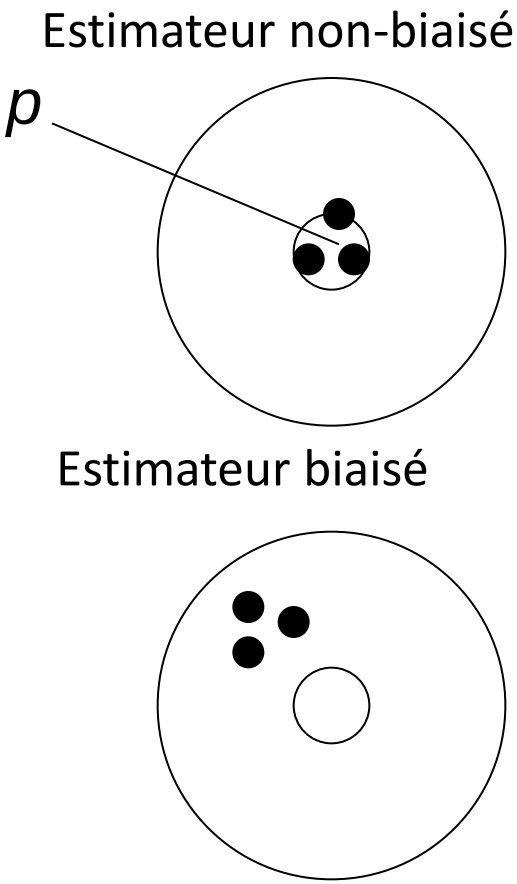
Conséquence #1 : Affecte les paramètres estimés (produit des estimés « biaisés »)

Biais = E(*e*) - *p*

Le biais corresponds à la différence entre la moyenne des estimés (*e*) (si on utilise une infinité d'échantillons) et le paramètre (*p*) dans la population.

Exemples : moyenne, pourcentage, coefficient de régression, ...

Question : Est-ce toujours vrai?



Impacts des DM

Conséquence #2 : Affecte la puissance statistique des tests

Explication : réduction du nombre d'observations = perte de puissance statistique

Conséquence #3 : Affecte la généralisation des résultats

Est-ce vrai dans tous les cas?

Est-ce qu'il y a un pourcentage « acceptable » de DM?

IMPORTANT : Le ***patron*** des données manquantes est bien plus important que la ***quantité*** de données manquantes

Patron de DM

MCAR (missing completely at random) :

<u>définition</u> : la probabilité d'avoir une DM est parfaitement imprévisible.

Elle n'est pas reliée aux données observées ou manquantes.

<u>exemple</u> : une partie des questionnaires complétés a été perdue par la poste

<u>conséquence</u> : l'analyse sur les sujets disponibles est valide, même si moins puissante

Patron de DM

MAR (missing at random) :

<u>définition</u> : la probabilité d'avoir une DM est reliée aux variables déjà observées dans la base de données

Elle n'est pas reliée à la valeur de la donnée manquante.

<u>exemple</u> : les sujets de sexe masculin répondent moins fréquemment à une question sur l'humeur dépressive

<u>conséquence</u> : si on tient compte des variables reliées à la probabilité des DM dans le modèle statistique, l'analyse sur les sujets disponibles est valide.

Patron de DM

MNAR (missing not at random) :

<u>définition</u> : la probabilité d'avoir une DM est reliée aux variables NON observées dans la base de données

Elle est reliée à la valeur de la donnée manquante.

<u>exemple</u> : les sujets avec un revenu élevé ont tendance à ne pas répondre à la question sur le revenu

<u>conséquence</u> : on doit tenter d'identifier des variables qui permettront indirectement de mesurer la probabilité d'avoir une donnée manquante puis faire les analyses selon le patron MAR.

Sinon, analyses de sensibilité (+++ complexe).

Approches pour gérer les DM

Continuum d'approches :

1. Conserver uniquement les observations complètes
   - Retrait des observations incomplètes (approche par défaut dans les logiciels)
2. Faire pour le mieux avec les données disponibles *
   - Utiliser des modèles robustes aux DM
   - Pondération les observations disponibles
3. Compléter les observations incomplètes
   - Imputation des DM

Approches pour gérer les DM

<u>Retrait des observations avec DM</u>

- ✓ retrait total des observations incomplètes (*listwise deletion*)

  par défaut dans SPSS, SAS, etc.
- ✓ retrait des observations par analyse (*pairwise deletion*)

  disponible dans tous les logiciels

Approches pour gérer les DM

Utilisation de modèles robustes aux DM

- ✓ Les modèles utilisant la méthode des moindres carrés (régression linéaire, ANOVA) ne sont pas robustes aux DM (*listwise deletion*)

- ✓ Les modèles mixtes (effets aléatoires) sont robustes aux DM pour des comparaisons de groupe ou des régressions

- ✓ Pour les modèles avec structures latentes, les méthodes d'estimation EM ou FIML sont robustes aux DM.

Approches pour gérer les DM

Pondération des observations

- Approche centrale en théorie des sondages
- Donner un poids aux **données disponibles** correspondant à l'inverse de la probabilité de réponse

Strate	**1**	**2**	**3**
Variable	1, X, X	2, 2, 2	3, X, 3
Probabilité de réponse	1/3	3/3 = 1	2/3
Poids	3	1	3/2

Approches pour gérer les DM

<u>Pondération des observations</u>

Dans chaque calcul (descriptif ou inférentiel), la valeur de chaque observation est multipliée par son poids.

$$\bar{X}_{brut} = \frac{\sum vd}{n} = \frac{1+(2+2+2)+(3+3)}{6} = 2.17$$

$$\bar{X}_{pond} = \frac{\sum (w \times vd)}{\sum w} = \frac{3/1 \times 1 + 3/3 \times (2+2+2) + 3/2 \times (3+3)}{9} = 2$$

Approches pour gérer les DM

Pondération des observations

***Exemple** : Une population contient 40% d'hommes et 60% de femmes. Vous avez invité 80 H et 120 F pour votre étude, mais seuls 20 H et 50 F ont répondu à l'instrument.*

Devez-vous pondérer?

Si oui, quel poids devra être utilisé pour les H et F de votre échantillon?

Approches pour gérer les DM

<u>Imputation SIMPLE des données manquantes</u>

✓imputation par moyenne de la variable

réduit la variance/covariance des données

✓projection de la dernière donnée disponible (*last observation carried forward*) (longitudinal)

direction du biais inconnu (À ÉVITER!!)

Approches pour gérer les DM

<u>Imputation selon un modèle statistique de DM</u>

Méthodes selon maximum de vraisemblance

- expectation-maximization (EM) – 1 imputation par DM
- imputation multiple – *k* imputations par DM ($k < 10$)

Étape 1 Calculer la matrice de variance-covariance (VC) sur les données disponibles	Étape 2 Prédire les données manquantes à l'aide d'une régression multiple et imputer ces valeurs dans le jeu de donnés	Étape 3 Recalculer la matrice de VC et la comparer avec celle de départ. Si différences, refaire l'étape 2 jusqu'à ce que les deux matrices soient similaires

Situation #1 : Questionnaire avec DM

Problème

Un sujet n'a pas complété l'ensemble de son questionnaire (q1 à q5, score total = somme des questions)

Solutions

- Imputation par la moyenne de la question

  Plusieurs problèmes!

- Imputation par la moyenne du sujet

  Attention de bien le faire…

  Exemples SPSS : somme manuelle, fonction somme, fonction somme pondérée

Situation #1 : Questionnaire avec DM

Problème

Un sujet n'a pas complété l'ensemble de son questionnaire (q1 à q5, score total = somme des questions)

Solutions

- Imputation par maximum de vraisemblance (EM)

  Exemple SPSS

  Quoi mettre dans le modèle d'imputation?

  Imputation simple ou multiple?

Situation #2 : ANOVA/régression avec DM

Problème

Un sujet n'a pas de valeurs sur la variable :

- ***indépendante*** (le groupe pour une ANOVA*, ou un score binaire/continu pour une régression linéaire/logistique)
- ***dépendante*** (continue pour ANOVA/régression linéaire, ou binaire pour régression logistique)

* à plan simple ou factoriel seulement

Solutions

Imputation par maximum de vraisemblance, séparément pour VD et VI

Situation #3 : Modèles latents avec DM

Problème

Vous désirez faire des modèles avec des structures latentes (analyse factorielle, acheminatoire, équations structurelles, classes/profils latents, courbe de croissance, etc.) et les données multivariées sont incomplètes.

Solutions

Pour chacune de ces analyses, plusieurs méthodes d'estimation sont disponibles (moindres carrés pondérées, maximum de vraisemblance, full-information ML, EM, etc.). La solution est de choisir une méthode robuste aux DM.

Situation #3 : Modèles latents avec DM

Analyse	Méthode robuste
Analyse factorielle confirmatoire	FIML
Path analysis/équations structurelles	FIML
Classes/profils latents	EM

Ces méthodes sont disponibles dans SPSS AMOS, SAS CALIS ou Mplus (le choix le plus flexible)

Situation #4 : Analyses longitudinales avec DM

Problème

Un sujet a abandonné l'étude ou cessé de répondre à la VD durant l'étude.

Solutions

- Éviter la projection de la dernière donnée disponible (LOCF)!!!
- Imputation par maximum de vraisemblance est un choix intéressant, mais présente plusieurs limites (sensibilité au choix du modèle d'imputation, augmentation artificielle du nombre de degrés de liberté, etc.)

Situation #4 : Analyses longitudinales avec DM

<u>Solutions</u>

- Modèle linéaire robuste aux DM : le modèle mixte

Rappel : limites de l'ANOVA à mesures répétées

- Les participants avec 1+ observation manquante sont retirés (*listwise deletion*)
- Les covariables temporelles ne peuvent pas être incluses (seulement les covariables qui ne varient pas)
- Si la dépendance temporelle ne respecte pas la *symétrie composée*, les corrections réduisent la puissance (H-F, G-G, Box)

Situation #4 : Analyses longitudinales avec DM

Le modèle mixte calcule la statistique F à l'aide d'une méthode par <u>maximum de vraisemblance</u>

- Ne demande pas la présence de données complètes.
- Rend la technique moins sensible aux données extrêmes et aux petits échantillons
- Les conclusions sont généralisables à la population (écart-types ne sont plus disponibles et sont remplacés par des erreurs standards, comme ANCOVA)
- On observe des résultats identiques à l'ANOVA à mesures répétées sur des données complètes
- ATTENTION : postulat que les DM sont MCAR/MAR

Situation #4 : Analyses longitudinales avec DM

Exemple : 2 des 6 patients avec données manquantes

Patient	T1	T2	T2-T1
1	*20*	*12*	*-8*
2	*26*	*24*	*-2*
3	*16*	*17*	*+1*
4	*29*	*21*	*-8*
5	*22*	*XX*	*XX*
6	*XX*	*17*	*XX*
M (listwise)	*22.75*	*18.50*	*-4.25 (n = 4)*
M (dispo)	*22.60*	*18.20*	*-4.40 (n = 5)*

Situation #4 : Analyses longitudinales avec DM

*Produit des moyennes **ajustées***

Estimated Marginal Means

temps

Estimates[a]

temps	Mean	Std. Error	df	95% Confidence Interval	
				Lower Bound	Upper Bound
1	22.466	2.016	7.374	17.747	27.185
2	18.145	2.016	7.374	13.426	22.864

a. Dependent Variable: vdm.

Ces moyennes utilisent l'ensemble des données disponibles et sont ajustées pour le patron des données manquantes (i.e., utilise l'information disponible d'un sujet pour estimer la performance attendue sur le temps manquant de ce sujet).

Situation #4 : Analyses longitudinales avec DM

Les degrés de liberté (ddl) sont ajustés pour le *N effectif* :

- Approximation des ddl du dénominateur par la méthode de Satterthwaite est réalisée par défaut dans SPSS (option dans SAS ou R).
- Cela peut donner des degrés de liberté avec des <u>décimales</u> : 4.27 ddl (3 ddl selon sujets complets et 5 ddl si pas de données manquantes)

Fixed Effects

Type III Tests of Fixed Effects[a]

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	5.390	134.886	.000
temps	1	4.269	4.628	.094

a. Dependent Variable: vdm.

Situation #4 : Analyses longitudinales avec DM

Estimé du changement T1-T2

- Selon la moyenne des scores de changement calculables (n = 4) :

  18.50 - 22.75 = -4.25

- Selon la différence entre les deux moyennes calculées sur l'ensemble des données disponibles (n = 5) :

  18.20 – 22.60 = -4.40

- Selon les moyennes ajustées du modèle mixte :

  18.15 – 22.47 = -4.32

Situation #4 : Analyses longitudinales avec DM

Quelle est la valeur des données manquantes?

- ATTENTION : il faut se rappeler que le **modèle mixte ne fait pas d'imputation** (il ne remplace pas la donnée manquante par une donnée « réaliste »)
- Les moyennes de chaque temps (et non de chaque sujet) sont plutôt ajustées selon un principe similaire à l'ANCOVA :

  Si le participant ayant une DM au T1 a une performance inférieure à la moyenne de son groupe au T2, sa performance "attendue" au T1 sera également inférieure à celle de la moyenne de son groupe au T1 => par conséquence, la moyenne estimée du groupe au T1 sera *ajustée* à la baisse

Situation #5 : Analyses multiniveaux

Problème

On fait des analyses sur des couples mais, pour certains couples, seul(e) un(e) des deux partenaires répond à la VD.

Solutions

- Faire une analyse avec les données disponibles. Deux problèmes :

  - on assume que les DM sont MCAR

  - on ne tient pas compte des cas où les deux partenaires ont répondu (leurs données sont corrélées)

- Imputation multiple est un choix possible mais difficile (composition du modèle d'imputation?)

Situation #5 : Analyses multiniveaux

Solutions

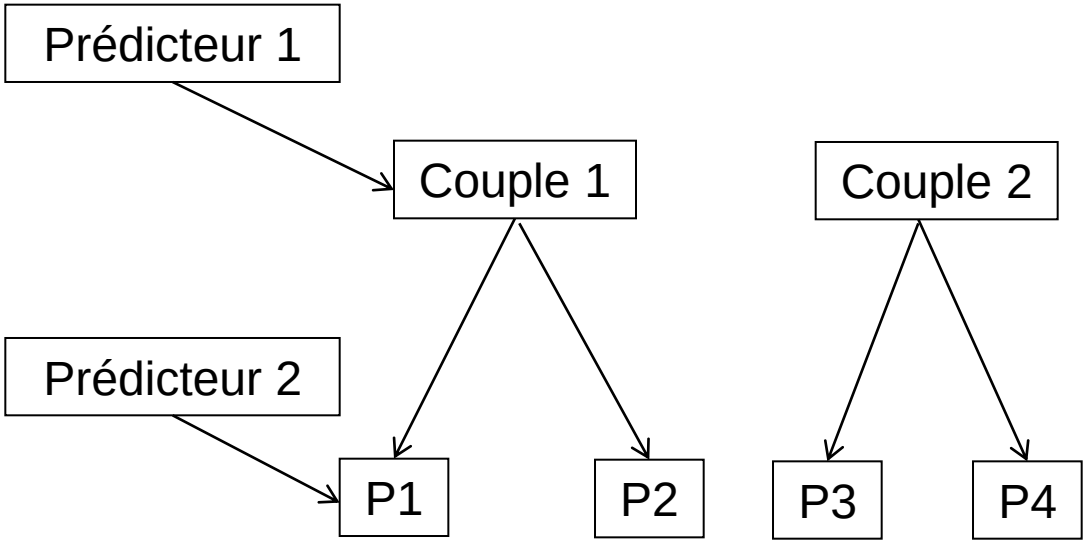
- Utilisation de la capacité du modèle mixte à utiliser l'information disponible pour produire des estimés qui tiennent compte du patron de DM.

Le modèle multiniveau, qui est un type de modèle mixte, permet de tenir compte d'une structure en grappes (*clusters*) des données, qui viole le postulat d'indépendance des observations.

Situation #5 : Analyses multiniveaux

Structure multiniveaux

- Caractéristiques des couples et des partenaires peuvent influencer la VD
- Tient compte de la non-indépendance, i.e., corrélation des réponses dans une grappe (ICC)



Situation #5 : Analyses multiniveaux

Modèle avec un *intercept* aléatoire (la moyenne de la VD varie selon le couple) et un prédicteur de niveau 1 (sexe)

$$(1)\ \mathrm{vd}_{c,p} = \mathrm{int}_c + Bsexe_{c,p} + e_{c,p}$$

$$(2)\ \mathrm{int}_c = \mathrm{int} + \mathrm{var(int)}_c$$

$$c = 1,\ldots,6$$

$$p = 1,2$$

$$\mathrm{var(int)}_c \approx \mathrm{N}(0,\tau^2)$$

$$e_{c,p} \approx \mathrm{N}(0,\sigma^2)$$

Bibliographie

- Allison, P. D. (2002). *Missing data (Vol. 136). Newbury Park, CA: Sage Publications.*
- Graham, J.W. (2009). Missing Data Analysis : Making it work in the real world. *Annual Review In Psychology, 60,* 549-576.
- Molenberghs, G., & Kenward, M. G. (2007). *Missing Data in Clinical Studies. England: Wiley & Sons.*
- Roth, P. L. (1994). Missing data: A conceptual review for applied psychologists. *Personnel Psychology, 47(3), 537-560.*

UNIVERSITÉ LAVAL