Transcription de l'épisode 078 – Les résultats (08) – DEFI6M

Analyse des données quantitatives : Données et valeurs manquantes (méthodes d'imputation 1 et 2 sur 5)

Dans une base de données, il arrive que des données et des valeurs soient manquantes : elles ne sont pas renseignées pour tous les cas. Ce qui rend les choses un peu plus complexes, c'est qu'il y a plusieurs manières de considérer et de traiter ces données manquantes selon les cas, on parle d'imputation des données. La plus simple et la moins contraignante serait de supprimer les lignes qui contiennent une valeur manquante. Mais attention, on risque vite d'éliminer beaucoup d'individus et de se retrouver avec des données qui ne sont plus représentatives.

Jetons un coup d'œil à quelques cas standards.

Parfois, le fait que la donnée soit manquante est une information en soi. Si la variable n'est pas renseignée pour certains individus parce qu'ils ne sont pas concernés alors on ne cherchera pas à imputer statistiquement une valeur. Par exemple la variable « date de décès » dans une étude épidémiologique ne sera pas renseignée pour les patients n'étant pas décédés. Dans ce cas, vous n'aurez pas besoin de compléter les valeurs manquantes, ça n'aurait tout simplement pas de sens.

Abordons à présent les 2 premières méthodes d'imputation de données manquantes.

La 1ère solution serait de supprimer les individus qui comportent des données manquantes.

C'est tentant de se débarrasser tout simplement des individus ayant des valeurs manquantes. Certains outils le font automatiquement lorsqu'on exécute un algorithme, attention donc à ne pas se retrouver avec un tout petit échantillon. Si ce sont des valeurs qui ne sont pas trop importantes, par exemple des données sociodémographiques, par-ci par-là, ok. S'il manque les réponses à tout un ensemble d'échelles de mesure ou à un trop grand nombre des items à l'échelle, il paraît possible de supprimer les cas. Mais n'utilisez cette méthode que si vous avez vraiment peu de données manquantes sinon vous risquez de biaiser les données.

La seconde solution serait de remplacer les données manquantes par une valeur fixe.

La méthode la plus simple consiste à remplacer toutes les valeurs manquantes d'une variable par une valeur fixe. Pour choisir cette valeur, on analyse la variable pour les individus ayant

des valeurs renseignées, il peut s'agir de : la moyenne, la médiane, la valeur la plus fréquente, valeur fixe, …

En tant que valeur fixe, vous pouvez aussi attribuer une valeur, par exemple la valeur 9 sous le logiciel SPSS et indiquer que la valeur 9 représente une valeur manquante au logiciel.

Nous aborderons la prochaine fois 3 autres méthodes un peu plus complexes pour traite les données manquantes.

Voilà, c'est terminé pour aujourd'hui et je vous remercie de votre écoute ; je vous invite à visiter mon blog methodorecherche.com et on se retrouve lundi pour un nouvel épisode du podcast de Methodo Recherche. Bon WE et à lundi !

Références :

Enders, C. K., & Guilford Press. (2010). Applied missing data analysis. NewYork; London: The Guilford Press.

Abonnez-vous au Podcast suivant votre préférence d'écoute. Vous trouverez toutes les possibilités et les explications à l'URL :

https://methodorecherche.com/subscribe-to-podcast/

En complément, vous êtes libre de vous abonner à ma newsletter et recevoir gratuitement le bonus "6 clés essentielles pour réussir brillamment votre mémoire de recherche (ou votre thèse)".

http://bit.ly/2RsYpll

A très bientôt,

Christophe