Données Manquantes
TP 2
Correction de la non-réponse partielle

Exercice 1

Nous nous intéressons à une population de 10 000 individus extraite de l'enquête nationale sur la santé de la population canadienne (CCHS). L'objectif est d'établir les liens entre l'état de santé et les déterminants de la santé.

Nous disposons d'un échantillon de taille $n = 1\ 000$ sélectionné dans U par sondage aléatoire simple. Nous nous intéressons aux variables mesurant le poids physique des individus (POIDS) et la consommation d'alcool (ALCOOL).

Les variables disponibles dans l'échantillon sont décrites dans le tableau 1. Seules les variables POIDS et ALCOOL sont affectées par de la non-réponse partielle. Les autres variables sont observées sur l'ensemble de l'échantillon.

Première partie : variables auxiliaires

1) Créer dans l'échantillon une variable donnant les poids de sondage des individus échantillonnés.
2) Donner les effectifs estimés par province, puis les effectifs estimés par catégorie d'âge.

TABLE 1 – Liste des variables dans l'enquête CCHS

Nom de variable	Valeur	Label
PROVINCE	10	NFLD et LAB.
	11	PEI
	12	NOVA SCOTIA
	13	NEW BRUNSWICK
	24	QUEBEC
	35	ONTARIO
	46	MANITOBA
	47	SASKATCHEWAN
	48	ALBERTA
	59	BRITISH COLUMBIA
	60	YUKON/NWT/NUNA.
AGE	1	12 A 14 ANS
	2	15 A 17 ANS
	3	18 A 29 ANS
	4	30 A 49 ANS
	5	50 A 64 ANS
	6	65 A 74 ANS
	7	75 ANS ET PLUS
SEXE	1	HOMME
	2	FEMME
ALCOOL	1	BUVEUR REGULIER
	2	BUVEUR OCCASIONEL
	3	ANCIEN BUVEUR
	4	JAMAIS BU
POIDS	-	POIDS (KG)

Seconde partie : variable POIDS

3) Créer une indicatrice de réponse pour les variables POIDS et ALCOOL.
4) Analyser le mécanisme de réponse pour la variable POIDS. Le résultat obtenu signifie t-il forcément que le mécanisme de réponse est uniforme ?

Nous notons y_k la variable donnant le POIDS (physique) des individus de la population. Nous nous intéressons tout d'abord à l'estimation du poids moyen

$$\mu_y = \frac{1}{N}\sum_{k\in U} y_k.$$

Nous souhaitons comparer deux méthodes d'imputation :

1. imputation par la moyenne simple,
2. imputation par la moyenne dans des classes.

5) Quel estimateur utiliserait-on pour μ_y si la variable POIDS était renseignée sur tout l'échantillon ?
6) Ecrire le modèle d'imputation associé à la méthode d'imputation par la moyenne simple. Montrer qu'avec cette méthode d'imputation, l'estimateur imputé de la moyenne

$$\hat{\mu}_{yI} = \frac{1}{n}\left(\sum_{k\in S_r} y_k + \sum_{k\in S_m} y_k^*\right)$$

se simplifie sous la forme

$$\begin{aligned} \hat{\mu}_{yI} &= \bar{y}_r, \\ \text{avec } \bar{y}_r &= \frac{1}{n_r}\sum_{k\in S_r} y_k \text{ la moyenne des répondants.} \end{aligned}$$

7) D'après l'étude du mécanisme de réponse, est-ce que cette méthode d'imputation conduit à une estimation non biaisée de μ_y ? Donner l'estimation obtenue avec cette méthode d'imputation.

8) Analyser le lien entre la variable POIDS et les variables auxiliaires de l'échantillon.
9) Nous définissons des classes d'imputation en croisant les variables AGE et

SEXE. Ecrire le modèle d'imputation associé, et mettre en oeuvre .

Nous supposons maintenant que nous souhaitons estimer non seulement la moyenne μ_y, mais également le dernier décile de la variable POIDS.

10) Comparer :
- le dernier décile de la variable y_k estimé à partir de l'échantillon des répondants uniquement,
- le dernier décile de la variable y_k estimé en utilisant la variable imputée par la moyenne simple,
- le dernier décile de la variable y_k estimé en utilisant la variable imputée par la moyenne dans les classes d'imputation.

Commenter les résultats obtenus.

11) Quel mécanisme d'imputation approprié pour le modèle d'imputation vu à la question 9 proposez-vous d'utiliser dans ce cas ? Le mettre en oeuvre, et procéder à l'estimation demandée.
12) Quel est l'inconvénient de cette méthode pour l'estimation du poids moyen ?

Troisième partie : variable ALCOOL2

La variable ALCOOL est regroupée en deux modalités :
- ALCOOL2=1 : Buveur régulier ou occasionnel (ALCOOL=1,2)
- ALCOOL2=0 : Ancien buveur ou non buveur (ALCOOL=3,4)

13) Analyser le mécanisme de réponse pour cette ALCOOL2. Peut-on corriger la non-réponse par une imputation par la moyenne dans des classes d'imputation ?

Nous procédons à des regroupements de variables. La variable PROVINCE est regroupée en 6 modalités selon la région :
- REGION=1 : Atlantique (PROVINCE=10,11,12,13)
- REGION=2 : Québec (PROVINCE=24)
- REGION=3 : Ontario (PROVINCE=35)
- REGION=4 : Prairies (PROVINCE=46,47,48)
- REGION=5 : Colombie-Britannique (PROVINCE=59)
- REGION=6 : Territoires (PROVINCE=60)

La variable AGE est regroupée en 3 modalités :

- AGE2=1 : moins de 18 ans (AGE=1,2)
- AGE2=2 : 18-64 ans (AGE=3,4,5)
- AGE2=3 : 65 ans et plus (AGE=6,7)

14) Procéder aux regroupements demandés, et analyser le lien entre ALCOOL2 et les variables auxiliaires de l'échantillon regroupées (REGION, AGE2 et SEXE).

15) Procéder à une imputation de la variable ALCOOL2 par hot-deck, dans des classes définies en croisant les variables REGION et AGE2.

16) En déduire une estimation du pourcentage de buveurs réguliers, buveurs occasionnels, anciens buveurs et non buveurs.