Devoir : Méthodes avancées de modélisation

-

Partie 1

Question 1

Sous l'hypothèse que les poids de sondage décrivent directement le nombre de sujets représentés dans la population cible, estimons pour chacune des cohortes la prévalence de la maladie dans sa population-cible.

Les prévalences pour chaque cohorte sont issues de l'estimateur d'Horvitz – Thompson. Il s'agit de calculer une moyenne pondérée par les poids fournis, pour chaque cohorte. Après séparation du tableau de données en cinq sous-tableaux, chacun correspondant à une cohorte, il est possible d'appliquer la formule pour obtenir une estimation ponctuelle.

La variable informant sur la maladie (« statut ») est codée de façon binaire. Elle prend pour valeur 0 en l'absence de maladie et pour valeur 1 en sa présence. L'estimation de la prévalence peut se rapporter au calcul de la moyenne de la variable considérée, ce qui équivaut à l'estimation du pourcentage.

Pour chaque cohorte, on calcule :

$$\hat{x} = \frac{\sum_{i=1}^{n} w_i \cdot x_i}{\sum_{i=1}^{n} w_i}$$

où **x** représente la variable « statut », et **wi** les poids de sondage associés à chaque sujet.

La somme s'effectue pour tous les sujets d'une cohorte, soit :

- Pour tout i de [1 ; 2934] pour la cohorte B
- Pour tout i de [1 ; 112] pour la cohorte C
- Pour tout i de [1 ; 119] pour la cohorte D
- Pour tout i de [1 ; 497] pour la cohorte E

La cohorte A comporte 2686 sujets mais il existe une valeur manquante du poids de sondage pour 69 d'entre eux, soit 2,6% de l'échantillon. Ces sujets ont donc été exclus de l'estimation, ce qui pourrait conduire à un biais. La somme s'effectue donc pour tout i de [1 ; 2617] pour la cohorte A avec exclusion des sujets à valeur manquante.

Les quatre autres cohortes ne présentent pas de données manquantes pour ce calcul.

Ces estimations sont effectuées en utilisant le package « survey » de R. Pour chaque design crée par cohorte, il est possible de calculer la moyenne et son erreur standard (fonction « svymean »). Afin d'obtenir une estimation par intervalle des prévalences, l'approximation par la loi normale pourrait être invalidée en raison de certaines très faibles valeurs de proportions estimées. Il est possible d'utiliser une fonction implémentant le calcul des intervalles de confiance directement par méthode exacte et estimant les pourcentages (« svyciprop »), et permet au passage de vérifier que les calculs par l'estimateur de la moyenne sont cohérents.

Les résultats sont présentés dans le tableau ci-dessous :

Cohorte / Estimation	A	B	C	D	E
Prévalence ponctuelle	0.080	0.106	0.020	0.008	0.084
Intervalle de confiance à 95%	[0.069 ; 0.090]	[0.094 ; 0.120]	[0.006 ; 0.060]	[0.001 ; 0.06]	[0.063 ; 0.110]

Les résultats montrent que la prévalence estimée semble variable d'une cohorte à l'autre (il faudrait tester la significativité de ces comparaisons). Nous observons que :

La prévalence est estimée à 10.6% ([9.4% ; 12.0%]) dans la cohorte B, représentant des femmes et des hommes de 20 à 90 ans.

Elle est estimée à 8.4% ([6.3% ; 11.0%]) dans la cohorte E, représentant des femmes et des hommes de 25 à 70 ans.

Elle est estimée à 8.0% ([6.9% ; 9.0%]) dans la cohorte A, représentant des femmes et des hommes de 20 à 80 ans.

Elle est estimée à 2.0% ([0.6% ; 6.0%]) dans la cohorte C, représentant des femmes de 70 à 90 ans.

La prévalence est enfin estimée à 0.8% ([0.1% ; 6.0%]), dans la cohorte D, représentant des hommes de 70 à 90 ans.

Il s'agit donc d'estimations dans les cinq populations cibles, et tenant compte des pondérations sous l'hypothèse initiale. Les estimations ponctuelles semblent varier selon la cohorte, avec une amplitude de 0.8% à 10,6%. Les précisions de ces estimations varient largement en raison de certaines tailles d'échantillon plus faibles, de l'ordre d'une centaine de sujets pour une, contre plus de deux milles pour une autre.

Question 2

Sous l'hypothèse que les poids de sondage décrivent le nombre de sujets représentés dans leur population cible, alors les estimations des prévalences ne sont valables que pour les

populations cibles spécifiquement. L'estimation issue de l'échantillon A est uniquement généralisable à la population cible A et ainsi de suite jusqu'à l'échantillon et la population cible E.

Les tranches d'âge varient en fonction de la population cible et sont généralement différentes de celles de la population cible sur laquelle on cherche à estimer la prévalence, soit la population générale adulte en Île-De-France âgée de 20 à 90 ans.

En effet certaines des populations cibles ne couvrent pas tout l'intervalle de l'âge désiré. Par exemple les populations cibles C et D ne couvrent que les âges de 70 à 90 ans alors que l'on cherche à couvrir l'intervalle de 20 à 90 ans.

La population cible B présente les mêmes caractéristiques de tranche d'âge que la population cible (femmes et hommes de 20 à 90 ans), toutefois on ne connaît pas les autres caractéristiques qui pourraient conduire à une différence entre ces deux populations, par exemple la répartition femmes-hommes ou la répartition des catégories socio-professionnelles.

L'utilisation des poids indiqués permettent donc d'effectuer des estimations sur les cinq populations cibles respectives, sous l'hypothèse initiale, mais pas sur la population générale cible puisqu'elle n'est pas identique à ces cinq populations.

Finalement, nous avons besoin d'obtenir un échantillon représentatif de la population cible, soit une population adulte d'Île-De-France de 20 à 90 ans. Il faut donc connaitre les caractéristiques des 5 cohortes afin de déterminer des poids via des probabilités d'inclusion dans l'échantillon. Il manque par exemple des informations concernant la répartition femme-homme, ou bien le critère de provenance géographique. Ces informations sont manquantes à la fois dans la population cible (celle où l'on désire effectuer l'estimation) et à la fois dans les cinq autres populations cibles.

L'objectif serait de pondérer les observations de l'échantillon afin de constituer un échantillon plus représentatif de la population cible. Il serait possible par exemple de représenter différemment les individus issus des populations cibles dont la tranche d'âge est très restreinte, par rapport aux individus issus des populations cibles correspondant à la tranche d'âge souhaitée (si l'on ne considère que l'âge). Les autres caractéristiques pourraient être prises en compte afin de construire les pondérations, comme le critère géographique, la répartition femme-homme, la distribution de l'âge et les catégories socio-professionnelles entre autres.

Ceci devrait permettre que les strates, représentées par les cohortes, respectent les proportions de chacun des groupes au sein de la population étudiée en fonction de ses caractéristiques, pour finalement conduire à une réduction du biais d'estimation dans la population cible.

```
library(tidyverse)
library(survey)

df <- read_csv("C:/Users/Fanny/Desktop/Projets R/Méthodes avancées de
modélisation DEVOIR.Rproj/devoir_sample.csv")

summary(df)

df %>%
  group_by(cohorte) %>%
  summarise(sum(is.na(poids_sond)))



# Question 1 :

  # Division des tableaux de données

df_A <- df %>%
  filter(cohorte == "A") %>%
  filter(! is.na(poids_sond))


df_B <- df %>%
  filter(cohorte == "B")

df_C <- df %>%
  filter(cohorte == "C")

df_D <- df %>%
  filter(cohorte == "D")

df_E <- df %>%
  filter(cohorte == "E")

  # Création des designs


design_A <- svydesign(id=~ 1,
                      weights = ~ poids_sond,
                      data = df_A)

design_B <- svydesign(id=~ 1,
                      weights = ~ poids_sond,
                      data = df_B)

design_C <- svydesign(id=~ 1,
                      weights = ~ poids_sond,
                      data = df_C)

design_D <- svydesign(id=~ 1,
                      weights = ~ poids_sond,
                      data = df_D)

design_E <- svydesign(id=~ 1,
                      weights = ~ poids_sond,
                      data = df_E)
```

```
  # Calcul des prévalences (moyennes pondérées)


prev_A <- svymean(
  x = df_A$statut,
  design = design_A)


prev_B <- svymean(
  x = df_B$statut,
  design = design_B)


prev_C <- svymean(
  x = df_C$statut,
  design = design_C)


prev_D <- svymean(
  x = df_D$statut,
  design = design_D)



prev_E <- svymean(
  x = df_E$statut,
  design = design_E)


prev_A
prev_B
prev_C
prev_D
prev_E



  # Calcul des intervalles de confiance

ci_A <-svyciprop(
  ~statut,
  design = design_A)

ci_B <-svyciprop(
  ~statut,
  design = design_B)

ci_C <-svyciprop(
  ~statut,
  design = design_C)

ci_D <-svyciprop(
  ~statut,
  design = design_D)

ci_E <-svyciprop(
  ~statut,
  design = design_E)
```

ci_A
ci_B
ci_C
ci_D
ci_E