Données Manquantes
TP 1 : correction de la non-réponse totale

Exercice 1

Dans une population contenant $N = 800$ individus dont $N_1 = 400$ femmes et $N_2 = 400$ hommes, nous sélectionnons un échantillon S de $n = 15$ individus. Pour chacun des individus échantillonnés, son sexe et sa classe d'âge (20-39 ans, 40 ans et plus) sont connus.
Parmi ces n individus, 9 acceptent de participer à l'enquête. Pour chacun d'entre eux, nous obtenons son poids (en kilogrammes) et sa taille (en mètres). Nous supposons qu'il n'y a pas de problème de non-réponse partielle. Les résultats obtenus sont donnés dans le tableau suivant :

Identifiant	Sexe	Classe d'âge	Réponse r_k	Poids y_k (kg)	Taille x_k (m)
01	F	20-39	1	60	1.50
02	F	20-39	1	60	1.65
03	F	40-59	1	70	1.70
04	F	40-59	1	65	1.60
05	F	40-59	1	65	1.55
06	F	40-59	1	55	1.50
07	H	20-39	1	80	1.80
08	H	40-59	1	70	1.70
09	H	40-59	1	60	1.70
10	F	20-39	0		
11	F	20-39	0		
12	F	40-59	0		
13	F	40-59	0		
14	H	20-39	0		
15	H	40-59	0		

Partie 1

Nous supposons que l'échantillon S est sélectionné selon un sondage aléatoire simple stratifié selon le sexe.

1) Donner les probabilités d'inclusion et les poids de sondage pour les individus de l'échantillon.
2) Donner une estimation du nombre total de personnes de chaque classe d'âge.

Partie 2

3) Nous notons $p_k \equiv Pr(r_k = 1|S)$ la probabilité de réponse à l'enquête. Pour chacun des modèles de non-réponse suivant, indiquer s'il s'agit d'un dispositif MAR ou NMAR, en justifiant très brièvement :

Modèle 1 : $\text{logit}(p_k) = 0.5 + 2 \times 1\,(k \text{ est une femme}) + 1\,(k \text{ a entre 20 et 39 ans})$

Modèle 2 : $\text{logit}(p_k) = 0.5 + 2 \times 1\,(k \text{ est une femme}) + 1\,(k \text{ pèse plus de 75 kg})$

Modèle 3 : $p_k = \begin{cases} 0.5 & \text{si la taille de } k \text{ est} \leq \text{à } 1.60 \\ 0.8 & \text{si la taille de } k \text{ est} > \text{à } 1.60 \end{cases}$

Modèle 4 : $\text{logit}(p_k) = 0.5 + 2 \times 1\,(k \text{ est une femme}) \times 1\,(k \text{ a entre 40 et 59 ans})\,.$

Partie 3

Nous supposons que le modèle de réponse est homogène par classe d'âge, tous sexes confondus.

4) Donner les probabilités de réponse estimées pour les individus répondants.
5) Donner les poids corrigés de la non-réponse pour les individus répondants.
6) En déduire une estimation du poids moyen $\mu_y = \frac{1}{N} \sum_{k \in U} y_k$ des individus de la population.

Exercice 2

Nous nous intéressons à une enquête réalisée par l'Unedic, ayant pour objet de connaître la situation vis-à-vis de l'emploi des personnes concernées. La population-cible comprend 329 374 individus. Un échantillon de 2006 personnes a été sélectionné par sondage aléatoire simple stratifié. Les 9 strates ont été obtenues en croisant deux critères. Le poids de sondage est donné par la variable POIDSINIT. L'échantillon est de plus affecté d'un problème de non réponse totale (fichier `CHOMEURS_REP`).

Nous nous intéressons à la variable MOTIF décrivant les raisons de sortie des listes d'allocataires :

- MOTIF=1 : l'individu a trouvé un emploi ou créé son entreprise.
- MOTIF=2 : l'individu a repris une formation.
- MOTIF=3 : l'individu est au chômage mais ne s'est pas ré-inscrit.
- MOTIF=4 : autre situation (retraite, maladie, abandon de la recherche d'emploi).

Nous souhaitons estimer le nombre total d'individus dans chacune des 4 modalités.

Partie 1

1) Vérifier que la variable AGE est bien disponible sur tous les individus de l'échantillon. Créer dans l'échantillon une variable contenant les probabilités d'inclusion.

2) Utiliser le package R `survey` (voir Annexe B) pour obtenir une estimation de la proportion de personnes par classe d'âge (moins de 30 ans, 31-35 ans, 36 ans ou plus), avec un intervalle de confiance à 95 % . Expliquer à quoi sert le paramètre `fpc`, et la conséquence sur l'estimation de variance si ce paramètre n'est pas renseigné.

Partie 2

3) En dehors des variables identifiantes, quelles sont les variables disponibles sur l'ensemble de l'échantillon ? Analyser le mécanisme de non-réponse totale en fonction de l'information disponible. Vous utiliserez un test de Wald pour

évaluer la significativité globale de chaque variable auxiliaire.

4) Réaliser une correction de la NR totale par Groupes Homogènes de Réponse, avec la méthode par croisement. Quels problèmes rencontre-t-on ?

5) Réaliser une correction de la NR totale par Groupes Homogènes de Réponse, avec la méthode des scores.

Partie 3

Nous supposons dans cette partie que le comportement de réponse est indépendant d'une strate à l'autre, et que le mécanisme de réponse est homogène à l'intérieur des strates.

6) Sous ce modèle de réponse, donnez les poids corrigés de la non-réponse totale. Donner une estimation du nombre total d'individus qui ont créé un emploi ou créé une entreprise.

7) En vous appuyant sur les résultats vus en cours, montrer que les estimateurs de variance associés à chaque phase de tirage peuvent être réécrits sous la forme

$$
\begin{aligned}
v_p(\hat{t}_{yr}) &= \sum_{h=1}^{H} (N_h)^2 \left(\frac{1}{n_h} - \frac{1}{N_h} \right) s^2_{yhr}, \qquad (1) \\
v_{nr}(\hat{t}_{yr}) &= \sum_{h=1}^{H} (N_h)^2 \left(\frac{1}{n_{rh}} - \frac{1}{n_h} \right) s^2_{yhr}, \qquad (2) \\
\text{avec } s^2_{yhr} &= \frac{1}{n_{rh}} \sum_{k \in S_{hr}} (y_k - \bar{y}_{hr})^2 \text{ et } \bar{y}_{hr} = \frac{1}{n_{rh}} \sum_{k \in S_{hr}} y_k.
\end{aligned}
$$

En déduire que l'estimateur de variance peut s'écrire sous forme simplifiée

$$
v(\hat{t}_{yr}) = \sum_{h=1}^{H} (N_h)^2 \left(\frac{1}{n_{rh}} - \frac{1}{N_h} \right) s^2_{yhr}. \qquad (3)
$$

8) Donner un intervalle de confiance à 95 % pour les estimateurs demandés.

A Variables présentes dans la table

STRATE
1 création d'entreprise
2 entrée en formation et pop RAC ('0' et '2')
3 autre entrée en formation
4 non réponse à convocation et pop RAC ('0' et '2')
5 autre non réponse à convocation
6 reprise emploi et pop RAC ('0' et '2')
7 autre reprise emploi
8 Inconnu
9 maladie, retraite, service militaire, reprise d'étude

SIT
1 absent fichier chômage en avril
2 chômeur non-indemnisé en avril
3 exclu pour activité réduite en avril
4 inscrit au chômage en avril avec un droit mais non-indemnisé
5 indemnisé mandaté en avril
6 indemnisé bénéficiaire fin avril
7 formation en avril
8 CES en avril

SEXE
1 homme
2 femme

AGE
1 30 ans ou moins
2 31 à 35 ans
3 36 ans ou plus

NATION
1 Etranger
2 Français

SITFAM
1 en couple
2 autre

QUALIF
1 Cadre
2 prof inter
3 employé non qual
4 employé qual
5 ouvrier non qualifié
6 ouvrier qualifié
7 Autre

RMI
1 bénéficiaire du RMI
2 non bénéficiaire du RMI

REGION
1 Ile de France, Provence-Alpes-Côte d'Azur et Corse
2 Autres régions

B Utilisation du package SURVEY

Le package SURVEY permet de réaliser un grand nombre d'analyses statistiques (statistique descriptives, tests, modèles de régression) sur des données d'enquête. Elle permet de prendre en compte le plan de sondage de l'enquête, et certains ajustements réalisés par des méthodes de calage.

La documentation complète du package SURVEY est disponible ici : `https://cran.r-project.org/web/packages/survey/index.html`.

Nous décrivons ci-dessous les principaux paramètres de la fonction `svydesign`, permettant de spécifier le plan de sondage utilisé (extrait de l'aide de SURVEY) :

- `ids` : formula or data frame specifying cluster ids from largest level to smallest level, 0 or 1 is a formula for no clusters.
- `probs` : formula or data frame specifying cluster sampling probabilities
- `strata` : formula or vector specifying strata, use NULL for no strata
- `variables` : formula or data frame specifying the variables measured in the survey. If NULL, the data argument is used.
- `fpc` : finite population correction.
- `weights` : formula or vector specifying sampling weights as an alternative to prob
- `data` : data frame to look up variables in the formula arguments, or database table name
- `calibrate.formula` : model formula specifying how the weights are "already" calibrated (raked, poststratified).
- `variance` : for pps without replacement, use variance="YG" for the Yates-Grundy estimator