

4 Redressements et non-réponse

4.1 Traitements de la non-réponse

On souhaite mesurer le poids total des 250 passagers d'un vol charter. Pour cela on tire un échantillon aléatoire simple de 25 personnes à qui on a l'intention de demander leur age (en 2 classes), leur taille et leur poids. Cinq personnes refusent de répondre mais on peut connaître leur sexe. Parmi les autres, cinq n'ont pas voulu indiquer leur poids (voir table 6).

individu	sexe	age	taille	poids
a	1	20	170	60
b	1	20	170	
c	1	20	180	70
d	1	20	190	80
e	1	20	190	80
f	1	30	170	70
g	1	30	170	
h	1	30	180	80
i	1	30	180	80
j	1	30	180	80
k	1	30	180	
l	1	30	190	
m	1	30	190	90
n	2	20	150	40
o	2	20	160	50
p	2	20	170	60
q	2	30	150	40
r	2	30	160	50
s	2	30	170	60
t	2	30	180	
u	1			
v	1			
w	2			
x	2			
y	2			

TABLE 6 – Données échantillon des passagers de l'avion

4.1.1

Quelles méthodes adoptez-vous pour corriger les effets de la non-réponse ? Justifier vos décisions de façon précise en explicitant les modèles que vous utilisez.

b	90
g	100
w	70
x	60

TABLE 7 – Poids pour certains non-répondants sur la variable poids

4.1.2

Vous apprenez que 130 passagers sont des hommes (sexe=1) et 120 des femmes (sexe=2). Modifiez-vous votre réponse? Pourquoi?

4.1.3

Parmi les 10 non-réponses sur le poids, on tire un échantillon aléatoire simple comprenant les individus b, g, w et x. On arrive à leur faire avouer leur poids (voir table 7). Que conclure?

4.2 Mesure de surfaces cultivées 3

On considère une région agricole composée de $N = 2010$ fermes. On cherche à mesurer la moyenne sur la région de la surface cultivée **en céréales** (\bar{Y}). Pour cela, on réalise un sondage aléatoire simple de taille $n = 100$. On obtient dans l'échantillon : $\bar{y} = 29$ et $s_y^2 = 1069$

4.2.1

On ne dispose pas d'information auxiliaire. Calculer l'estimateur d'Horvitz-Thompson et donner un intervalle de confiance à 95% pour \bar{Y} .

4.2.2

On connaît la moyenne sur la région, notée \bar{X} , de la surface cultivée **totale**, i.e. toutes cultures confondues : $\bar{X} = 118$ On connaît la surface cultivée totale de chaque ferme de l'échantillon. On obtient, sur l'échantillon :

$$\begin{aligned}\bar{x} &= 132 \\ s_x^2 &= 7619 \\ s_{xy} &= 1453\end{aligned}$$

Justifier l'utilisation de la méthode d'estimation par le ratio. Donner l'estimateur et l'intervalle de confiance à 95% obtenu par cette méthode.

4.3

Dans cette question, on suppose que l'information dont on dispose sur la surface cultivée est la suivante : on sait qu'il y a 1 580 fermes de moins de 160 hectares (post-strate 1), et 430 fermes de 160 hectares et plus (post-strate 2).

Dans l'échantillon, (ayant noté avec des indices 1 et 2 les deux post-strates définies) on a :

$$\begin{array}{ll} n_1 = 70 & n_2 = 30 \\ \bar{y}_1 = 19 & \bar{y}_2 = 52 \\ s_1^2 = 312 & s_2^2 = 922 \end{array}$$

Quel est l'estimateur post-stratifié \bar{Y}_{post} ? Est-il différent de la moyenne simple \bar{y} ? donner un intervalle de confiance à 95 %.