T. D. n° 4
Stratification a posteriori

Exercice 1. Application des formules.

À un questionnaire mensuel sur la consommation des ménages, nous incorporons à la demande d'un producteur, une question sur le nombre de rosiers achetés l'année précédente. Nous avons 8 300 réponses qui donnent $\widehat{\mu} = 0,530$ et $s^2 = 2,0$. Comme nous avons beaucoup d'informations sur les enquêtes, nous allons essayer d'affiner l'estimation par une stratification a posteriori.

1. Nous ventilons les résultats de l'enquête selon la catégorie d'habitation : Rural contre Non-rural

	Effectifs	Moyenne	Variance corrigée
Rural : R	2573	$\widehat{\mu}_R = 0,882$	$s^2_{R,c} = 2,8^2$
Non rural : NR	5727	$\widehat{\mu}_{NR} = 0,372$	$s^2_{NR,c} = 1,4^2$

   La population totale est de 20 589 000 ménages dont 25, 9% de ménages ruraux. Calculer l'estimateur post stratifié $\widehat{\mu}_{post}$ et donner un intervalle de confiance à 95% pour la moyenne μ.

2. Nous avons également quelques raisons de penser que la taille du ménage est un facteur susceptible de jouer un rôle : grande maison donc grand jardin,... Nous ventilons les résultats de l'enquête selon les deux critères (la catégorie d'habitation et la taille du ménage).
   Or, en France, les ménages se répartissent, selon la taille, avec les fréquences suivantes :
   — 1 personne 22, 6 %
   — 2 personnes 30, 4 %
   — 3 personnes ou plus 47, 0 %

	1 personne	2 personnes	3 personnes ou plus	Total
Rural				
Effectif	171	882	1520	2573
Moyenne	0, 241	0, 861	0, 966	0, 882
Non rural				
Effectif	1082	1475	3170	5727
Moyenne	0, 122	0, 412	0, 438	0, 372
Total				
Effectif	1253	2357	4690	8300
Moyenne	0, 138	0, 580	0, 609	0, 530

   (i) Faire les deux premières étapes de l'algorithme RAS.

(ii) Calculer l'estimateur post-stratifié sur le tableau suivant qui provient du résultat de l'algorithme RAS :

	1 personne	2 personnes	3 personnes ou plus	Total
Rural	215	832	1103	2150
Non rural	1660	1691	2799	6150
Total	1875	2523	3902	8300

(iii) Commenter.

Exercice 2. Stratification a priori et a posteriori.

Dans la population des laboratoires d'examens biologiques ($N = 4000$) nous nous intéressons à Y_i, le nombre d'examens relatifs à une maladie dans un laboratoire i.

1. **Tirage aléatoire simple à PESR.**

(i) Rappeler les formules pour $\widehat{T}_n$, $\text{Var}\left(\widehat{T}_n\right)$ et $\widehat{\text{Var}}\left(\widehat{T}_n\right)$.

(ii) Sur un échantillon de taille 400, nous avons observé :

$$\widehat{\mu}_y = 3,56 \quad \text{et} \quad s_c = 4,5.$$

Calculer $\widehat{T}_n$ et $\widehat{\text{Var}}\left(\widehat{T}_n\right)$. Donner un intervalle de confiance à 95% pour le total T.

(iii) Critiquer et commenter.

2. **Stratification a priori avec allocation proportionnelle.**

(i) Nous faisons deux strates : $h = 1$ pour les 400 gros laboratoires spécialisés (hôpitaux,...) et $h = 2$ pour les 3600 autres laboratoires. Nous avons observé sur l'échantillon proportionnel, de taille $n = 400$:

$$\widehat{\mu}_1 = 10, \quad \widehat{\mu}_2 = 2, \quad s_{1,c} = 6,942 \quad \text{et} \quad s_{2,c} = 1.$$

Calculer $\widehat{T}_{stp}$ et $\widehat{\text{Var}}\left(\widehat{T}_{stp}\right)$. Donner un intervalle de confiance à 95% pour le total T.

(ii) Critiquer et commenter.

3. **Stratification a posteriori.**
Nous repartons du plan de sondage aléatoire simple à PESR et nous supposons $\widehat{\mu}_1, \widehat{\mu}_2, s_{1,c}$ et $s_{2,c}$ inchangés.

(i) Quelle était la répartition n_1, n_2 dans l'échantillon ? Commenter.

(ii) Vérifier que nous retrouvons $s_c = 4,5$.

(iii) Calculer $\widehat{T}_{post}$ et $\widehat{\text{Var}}\left(\widehat{T}_{post}\right)$. Donner un intervalle de confiance à 95% pour le total T.

(iv) Commenter.

Exercice 3. Redressement d'une proportion.

Avant d'envisager la construction d'un nouveau parking dans une université, nous faisons un sondage pour estimer la proportion d'étudiants utilisant un véhicule personnel pour venir sur le campus. Comme nous connaissons l'origine géographique des étudiants (ville, le reste du département, autre), nous utiliserons cette information pour faire éventuellement un redressement.

1. Nous avons tiré selon un SAS PESR un échantillon de 150 étudiants, les résultats sont donnés dans le tableau ci-dessous :

	Véhicule personnel	
Origine	Non	Oui
Ville	45	15
Département	25	25
Autre	10	30

   (i) Donner une estimation de la proportion d'étudiants utilisant un véhicule personnel.

   (ii) Quelle est la variance de cet estimateur ?

2. Sachant que sur ce campus la répartition des étudiants est la suivante

Ville	5000
Département	3000
Autre	2000

   (i) Faire un redressement de l'estimation.

   (ii) Donner une variance de ce nouveau estimateur.

   (iii) Ce redressement est-il justifié ?

Exercice 4. Redressement avec de mauvaises valeurs.

Dans un échantillon nous avons procédé à un redressement pour estimer $\overline{Y}$ par $\overline{Y}_{post}$ à partir des valeurs de $\overline{y}_h$ en utilisant N_h, les effectifs des classes dans la population de taille N et $\overline{y}_h$ les moyennes dans les classes de l'échantillon. Si les N_h et N sont entachés d'erreurs, nous avons introduit un biais que nous allons calculer.

Nous notons par N_h^*, N^* les vraies valeurs des effectifs dans la population et $\widehat{\overline{Y^*}}_{post}$ est la bonne estimation post-stratifiée.

1. Montrer que

$$\widehat{\overline{Y}}_{post} = \widehat{\overline{Y^*}}_{post} + \sum_h \left(\frac{N_h}{N} - \frac{N_h^*}{N^*} \right) \overline{y}_h.$$

2. Écrire $\mathbb{E}[\widehat{\overline{Y}}_{post}]$. Quel est le biais introduit ?

3. A.N. : nous estimons un revenu dans un échantillon et nous avons trouvé les valeurs suivantes ventilées par sexe : $n = 1000$, $n_H = 600$, $n_F = 400$, $\overline{y} = 7000$, $\overline{y}_H = 7500$ et $\overline{y}_F = 6250$.

(i) Si nous supposons que les proportions d'hommes et de femmes sont de 54 % et 46 %, en déduire $\widehat{\overline{Y}}_{post}$ par redressement.

(ii) Si les vraies valeurs dans la population sont de 51 % et 49 %, calculer $\widehat{\overline{Y^*}}_{post}$.

(iii) Vérifier la valeur du biais de la question 2).

Exercice 5. *Post-stratification.* **Extrait du livre « Exercices corrigés de méthodes de sondage » de P. Ardilly et de Y. Tillé**

Nous considérons une région agricole comportant $N = 2010$ fermes. Nous réalisons un SAS de fermes de taille $n = 100$. Nous possédons l'information sur la surface totale cultivée de chaque ferme. En particulier, nous savons que qu'il y a 1580 fermes de moins de 160 hectares (post-strate 1) et 430 fermes de plus de 160 hectares (post-strate 2). Nous cherchons à estimer la surface moyenne cultivée en céréales μ_Y. D'après l'échantillon simple et sans remise (ayant noté avec les indices 1 et 2 les deux post-strates ainsi définies), nous avons :

$$n_1 = 70, \quad n_2 = 30, \quad \widehat{\mu}_{1,Y} = 19,40, \quad \widehat{\mu}_{2,Y} = 51,63, \quad s^2_{1,y} = 312, \quad s^2_{2,y} = 922.$$

1. Quel est l'estimateur post-stratifié de la moyenne, noté $\widehat{\mu}_{post}$? Est-il différent de l'estimateur du SAS PESR de la moyenne ?
2. Quelle est la loi de n_1 ? Que vaut son espérance ? Que vaut sa variance ?
3. Donner l'estimateur sans biais de la variance, noté $\widehat{\mathrm{Var}}\,(\widehat{\mu}_{post})$ et un intervalle de confiance pour la moyenne μ à 95%.