DUT STID FA 2017 – Théorie des sondages

Martin Chevalier – Thomas Merly-Alpa

Devoir Maison

A rendre pour le 31/05/2017

Ce devoir maison est à réaliser par groupe de deux (gare au plagiat entre les groupes) et à rendre pour le 31 mai 2017 (par mail aux adresses martin.chevalier@insee.fr et thomas.merly-alpa@insee.fr). La notation tiendra compte de la présentation et de l'orthographe. Le sujet comporte 4 pages.

Une commune d'Île de France de 20 000 habitants souhaite mener une étude pour savoir s'il est pertinent ou non de remplacer certaines de ses lignes de bus par un bus en « site propre » (*i.e.* avec un espace réservé sur la chaussée).

Pour la maire de la commune, le principal objectif de cette nouvelle ligne est d'inciter davantage d'habitants à utiliser les transports en commun pour les trajets domicile-travail (ou domicile-lieu d'études) plutôt que la voiture en diminuant le temps de trajet vers la gare RER située dans la ville voisine.

1 Évaluation du besoin

La maire veut tout d'abord estimer, parmi ses administrés, combien ont besoin d'aller à Paris régulièrement, que ce soit pour leur travail, pour leurs études ou pour d'autres raisons. Pour cela, elle engage un institut de sondage dont vous êtes le représentant. Vous lui suggérez tout d'abord de réaliser un sondage aléatoire simple pour évaluer le besoin.

1.1 Indiquer pour chaque affirmation si elle est vérifiée ou non par un sondage aléatoire simple et le justifier rapidement :

1. On sait à l'avance combien de personnes vont être enquêtées ;
2. Le poids de sondage est différent selon l'âge des individus ;
3. Si deux invididus ont le même prénom, il y a moins de chances qu'ils soient sélectionnés conjointement ;
4. Les probabilités d'inclusion simples sont égales pour tous les individus.

Une étude sur la fréquentation des lignes de bus menée en 2014 avait montré que 32 % des habitants de la commune se rendaient régulièrement à Paris ; la maire pense que cela n'a pas beaucoup évolué.

1.2 On note z_i l'indicatrice qui vaut 1 si l'individu i se rend régulièrement à Paris, et 0 sinon. Rappelez la formule de l'estimateur d'Horvitz-Thompson de la proportion d'habitants de la commune se rendant régulièrement à Paris.

1.3 Combien de personnes faut-il interroger pour obtenir un coefficient de variation (CV) de 5 % pour cette estimation ? On notera ce nombre n_0.

1.4 On suppose que l'on a finalement, pour des raisons de budget, réalisé ce sondage auprès de $\frac{n_0}{2}$ personnes. Quel coefficient de variation peut-on obtenir ?

2 Nombre moyen de trajets en bus par semaine

Le principal objectif de l'étude est d'évaluer le nombre de trajets hebdomadaires supplémentaires qu'effectueraient les habitants de la commune avec le bus en site propre. Pour ce faire, le questionnaire adressé aux habitants comporte deux questions-clés :

— le nombre moyen de trajets en bus par semaine effectués avec les lignes actuelles ;
— le nombre moyen de trajets en bus par semaine qui seraient effectués avec la nouvelle ligne (après présentation de ses caractéristiques : fréquence, durée du trajet, etc).

À nouveau, vous disposez des résultats de l'étude menée en 2014 sur la fréquentation des lignes actuelles, que la maire vous communique ventilés selon les trois quartiers de la commune :

Quartier	**Habitants**	**Nombre moyen de trajets en bus par semaine**	**Écart-type**	**Proportion d'habitants ayant une voiture**
Est	5 000	3,6	3,6	45 %
Centre	10 000	5,9	6,4	5 %
Ouest	5 000	12,4	4,9	10 %

TABLE 1 – Résultats de l'étude de 2014

2.1 Comment justifier auprès de la maire d'utiliser un sondage aléatoire stratifié selon le quartier de résidence plutôt qu'un sondage aléatoire simple pour estimer le nombre moyen de trajets en bus par semaine en 2017 ?

2.2 Calculer l'allocation de Neyman associée à la variable « Nombre moyen de trajets en bus par semaine » pour $n = 500$.

2.3 Calculez le coefficient de variation de l'estimateur « Nombre moyen de trajets en bus par semaine » (variable Y) obtenu avec l'allocation déterminée à la question **2.2**.

2.4 Calculez le coefficient de variation de l'estimateur « Proportion d'habitants ayant une voiture » (variable Z) obtenu avec l'allocation déterminée à la question **2.2**. Comment expliquez-vous la différence avec le résultat de la question précédente ?

2.5 Quelle allocation serait optimale pour estimer efficacement la proportion d'habitants ayant une voiture (variable Z) ? La calculer pour un échantillon de 500 personnes.

2.6 On appelle *allocation mixte* la moyenne arithmétique des allocations calculées aux questions **2.2** et **2.5**. Calculez l'allocation mixte associée à chaque strate pour un échantillon de 500 personnes. Quelle est la précision obtenue sur les variables Y et Z pour un sondage aléatoire stratifié avec cette allocation mixte ? Commentez le résultat obtenu.

3 Correction de la non-réponse et redressements

En dépit des avantages de la statification, vous ne parvenez pas à convaincre la maire de mettre en œuvre ce protocole et revenez donc à un **sondage aléatoire simple de taille** $n = 100$.

Identifiant	Classe d'âge	Nombre moyen de trajets actuellement	Nombre moyen de trajets avec la nouvelle ligne
1	Moins de 30 ans	16	18
2	Moins de 30 ans	16	19
3	Moins de 30 ans	15	12
4	30-60 ans	2	3
5	30-60 ans	13	16
6	30-60 ans	10	13
7	30-60 ans	1	3
8	30-60 ans	11	14
9	30-60 ans	14	18
10	30-60 ans	0	3
11	30-60 ans	9	
12	30-60 ans	3	7
13	60 ans et plus	0	3
14	60 ans et plus	0	
15	60 ans et plus	0	1
16	60 ans et plus	0	
17	60 ans et plus	0	6
18	60 ans et plus	0	0
19	Moins de 30 ans		
20	30-60 ans		
21	30-60 ans		
22	30-60 ans		
23	30-60 ans		
24	30-60 ans		
25	60 ans et plus		

TABLE 2 – Résultats de la première semaine de collecte

La première semaine de collecte de l'enquête, 25 personnes sont interrogées par téléphone. Le tableau 2 synthétise ces premiers résultats (les cases vides indiquent une absence d'information).

3.1 Comment désigne-t-on la non-réponse correspondant à la situation des individus 19 à 25 ? Quels effets cette non-réponse peut-elle induire sur les estimateurs ? Décrivez précisément la ou les méthodes que vous mettriez-vous en œuvre pour la corriger ici.

3.2 Comment désigne-t-on la non-réponse correspondant à la situation des individus 11, 14, 16 ? Indiquez succinctement comment vous pourriez la corriger ici.

On poursuit la collecte sur plusieurs semaines en recontactant les personnes pour lesquelles certaines valeurs étaient manquantes, si bien qu'**on n'a en définitive aucune non-réponse au moment de mener à bien l'estimation**. Les estimations peuvent donc porter sur l'ensemble des $n = 100$ individus échantillonnés.

3.3 Vous cherchez à estimer de la façon la plus précise possible le nombre moyen de trajets qui seraient réalisés chaque semaine avec la nouvelle ligne (variable Y). Pour ce faire, vous recourez à une estimation par le ratio en utilisant comme variable auxiliaire la variable X = « Nombre moyen de trajets actuellement » :

— la moyenne de la variable X dans l'échantillon est 6,41 ;
— la moyenne de la variable X dans la population est connue *via* les données de la compagnie de bus et vaut 6,93 ;

— sur l'échantillon, vous calculez les quantités : $\bar{y} = 8,66$, $s_Y^2 = 36,47$, $s_X^2 = 34,95$ et $s_{X,Y} = 33,76$.

Calculez la valeur de l'estimateur par le ratio de la moyenne de Y et son intervalle de confiance à 95 %. Comparez-le à celui de l'estimateur d'Horvitz-Thompson de la moyenne (sans redressement).

3.4 Alternativement, vous testez une méthode de post-stratification par classe d'âge en utilisant les données du tableau 3. Calculez la valeur de l'estimateur post-stratifié de la moyenne de Y et son intervalle de confiance à 95 %. Comparez l'intervalle de confiance obtenu à cette question à ceux obtenus à la question précédente.

Classe d'âge	N	n	$\bar{y}$	s_Y^2
Moins de 30 ans	3 495	16	15,7	13,3
30-60 ans	9 990	50	10,2	25,9
60 ans et plus	6 515	34	3,1	4,9

TABLE 3 – Données relatives aux post-strates

3.5 Quelle technique pourrait-on utiliser pour combiner les avantages des estimateurs calculés aux questions **3.3** et **3.4** ?