Livre de Travaux Dirigés
DUT STID 2017 - 2ème année

2nd semestre 2017-2018
Livre d'exercices

Théorie des Sondages

Martin Chevalier, Thomas Merly-Alpa

D'après les exercices d'A. Rebecq

1 Bases de la théorie des sondages

1.1 Estimateur pondéré

Un sondage de 15 individus parmi une population de 300 a été réalisé par un institut. Le sondage vise à mesurer la fréquentation mensuelle des cinémas par les individus de la population. L'institut nous transmet les résultats de son enquête *via* la table suivante :

Individu	Fréquentation des cinémas	Poids de sondage
A	1	24
B	2	20
C	4	18
D	15	18
E	2	20
F	3	14
G	1	22
H	0	20
I	0	17
J	3	20
K	6	22
L	0	21
M	0	22
N	1	22
O	1	20

1.1.1

On rappelle que l'estimateur naïf est l'estimateur qui ne prend pas en compte les poids de sondage. Quel est l'estimateur naïf de la fréquentation moyenne ?

1.1.2

Quel est l'estimateur pondéré (avec les poids de sondages donnés dans la table) de la fréquentation moyenne ?

1.2 Calcul de probabilités d'inclusion

On se donne une population $\mathcal{U} = \{1, 2, 3\}$ et un plan de sondage p défini par :

$$p(\{1\}) = 0 \quad p(\{1,2\}) = \frac{1}{2} \quad p(\{1,2,3\}) = 0$$

$$p(\{2\}) = 0 \quad p(\{1,3\}) = \frac{1}{3}$$

$$p(\{3\}) = 0 \quad p(\{2,3\}) = \frac{1}{6}$$

1.2.1

Calculer les probabilités d'inclusion simples et doubles pour toutes les unités de la population, ainsi que les $\Delta_{kl} = \pi_{kl} - \pi_k \pi_l$.

1.2.2

Quelle propriété intéressante possède le plan de sondage p ? Que peut-on dire de $\sum_{k \in \mathcal{U}} \pi_k$ dans ce cas ? Le vérifier.

1.3 Introduction à l'estimateur d'Horvitz-Thompson

On se donne la population et le plan de sondage de l'exercice 1.2. On suppose que le revenu des individus de la population est :

$$Y_1 = 1\ 000$$
$$Y_2 = 2\ 000$$
$$Y_3 = 3\ 000$$

1.3.1

Donner la loi de l'estimateur *plugin* du revenu total : pour chaque échantillon possible, rappelez sa probabilité de sélection et calculez la valeur de l'estimateur *plugin*. Calculer le biais et la variance de cet estimateur.

L'estimateur d'Horvitz-Thompson d'un total Y est un estimateur pondéré qui se définit par :

$$\hat{Y}_{HT} = \sum_{k \in s} \frac{y_k}{\pi_k}$$

1.3.2

Donner la loi de l'estimateur d'Horvitz-Thompson du revenu total. Calculer le biais et la variance de cet estimateur.

1.4 Enquête sur le patrimoine

Un sondeur réalise une enquête ayant pour but de mesurer le patrimoine moyen des ménages d'Île-de-France. Les individus sont tirés parmi la liste des titulaires d'une carte Navigo (carte d'abonnement aux transports en commun d'Île-de-France) pour l'année 2016. Le plan de sondage donne une probabilité de sélection plus forte aux individus habitant des communes aux revenus médians les plus élevés : Paris 16, Neuilly-sur-Seine, Paris 7, Versailles. L'estimateur utilisé est celui d'Horvitz-Thompson. 70 % des individus échantillonnés répondent au questionnaire, mais l'estimateur utilisé ne prend pas en compte la non-réponse.

Que penser des affirmations suivantes ?

1. Donner une probabilité de sélection plus forte pour certains individus conduit à un échantillon non représentatif de la population ;
2. Le poids de sondage d'un individu sélectionné qui vit à Paris 13 est supérieur à celui d'un individu sélectionné habitant à Paris 16 ;
3. L'estimateur d'Horvitz-Thompson est biaisé, il vaudrait mieux utiliser l'estimateur naïf ;
4. La base de sondage présente un défaut de couverture, l'estimation est potentiellement biaisée ;
5. Ne pas prendre en compte la non-réponse n'a pas d'impact sur l'estimation.

2 Sondage aléatoire simple

2.1 Émissions de CO_2

Une entreprise produit quatre modèles de voitures. On souhaite connaître le niveau moyen d'émissions de CO_2 des modèles de cette entreprise, mais le test de pollution est coûteux à réaliser. L'organisme en charge du contrôle décide donc de sélectionner au hasard deux des quatre modèles, en suivant un sondage aléatoire simple. Les informations sur les modèles sont disponibles ci-dessous :

Modèle	Émissions
1	1
2	2
3	3
4	10

TABLE 1 – Table de données pour l'exercice 2.1

2.1.1

On rappelle qu'on réalise un sondage aléatoire simple de 2 modèles parmi les 4. Donner les π_i et les $\pi_{i,j}$ associés à ce sondage.

2.1.2

Quels sont tous les échantillons possibles ? Donner pour chacun sa probabilité de sélection.

2.1.3

Pour chacun des échantillons obtenus à la question précédente, donner l'estimateur d'Horvitz-Thompson $\hat{Y}$ de la moyenne des émissions de gaz.

2.1.4

Retrouver que l'estimateur d'Horvitz-Thompson est sans biais.

2.1.5

Pour chacun des échantillons S obtenus, calculer l'estimateur de la variance de l'estimateur d'Horvitz-Thompson, qu'on notera $\hat{V}_S$.

2.1.6

Comparer les trois variances suivantes :

1. La moyenne des $\hat{V}_S$ obtenus à la question précédente ;
2. La vraie variance de l'estimateur d'Horvitz-Thompson mobilisant l'information sur toute la population ;
3. La dispersion empirique des estimations $\hat{Y}$ obtenues à la question 2.1.3.

Que remarquez-vous ?

2.2 Mesure de surfaces cultivées

D'après O. Sautory.

On veut estimer la surface moyenne cultivée dans les fermes d'un canton rural. Sur les $N = 2\ 010$ fermes que comprend ce canton, on en tire $n = 100$ par sondage aléatoire simple (sans remise). On mesure pour chaque ferme k de l'échantillon la surface cultivée y_k.

Station	mai	juin
1	5,82	5,89
2	5,33	5,34
3	5,76	5,92
4	5,98	6,05
5	6,2	6,2
6	5,89	6
7	5,68	5,79
8	5,55	5,63
9	5,69	5,78
10	5,81	5,84

TABLE 2 – Table de données pour l'exercice 2.4

On obtient : $\sum_{k \in s} y_k = 2\ 907\ ha$ et $\sum_{k \in s} y_k^2 = 154\ 593\ ha^2$.

2.2.1

Quelle est la valeur de l'estimateur de Horvitz-Thompson pour la moyenne de la surface cultivée $\bar{Y}$?

2.2.2

Proposer un intervalle de confiance pour cet estimateur.

2.3 Mesure d'audience

Un patron de chaîne veut connaître le nombre de personnes qui regardent l'émission de télévision qu'il diffuse en *access prime time*. Il commande ainsi une étude à un institut de sondages.

Celui-ci choisit d'échantillonner par sondage aléatoire simple n individus. Si la véritable audience (inconnue) de l'émission est de 1 %, combien faut-il tirer de personnes pour obtenir un coefficient de variation (CV) de 5% ?

2.4 Mesure de prix à la pompe

D'après Ardilly, 1992.

On veut estimer l'évolution du prix moyen du litre entre mai et juin *via* la différence des prix moyens, par sondage aléatoire simple (taille d'échantillon $n < 10$). Les prix pour toutes les stations sont donnés en Table 2.

On veut comparer deux méthodes d'échantillonnage :

1. On échantillonne n stations en mai ($n < 10$), puis n stations en juin, de manière totalement indépendante ;
2. On échantillonne n stations en mai, et on interroge de nouveau ces mêmes stations en juin.

2.4.1

Qualitativement, quelle est la meilleure stratégie d'échantillonnage à adopter ? Pourquoi ?

2.4.2

Que vaut $\sqrt{\frac{\mathbb{V}_1(\bar{p}_{juin} - \bar{p}_{mai})}{\mathbb{V}_2(\bar{p}_{juin} - \bar{p}_{mai})}}$ le rapport des écarts-types des estimateurs obtenus *via* les deux méthodes ?

3 Stratification et probabilités inégales

3.1 Dénombrement d'oiseaux

Un parc naturel souhaite savoir combien d'oiseaux se trouvent sur son espace protégé. Pour cela, il dispose de six sites d'observation, qui couvrent tout le territoire du parc. Chaque site couvre une partie différente du territoire du parc : quatre se trouvent près du chemin, et deux dans les hauteurs. Envoyer une équipe compter les animaux est coûteux, et le parc souhaite se limiter à des dénombrements sur deux sites uniquement. Les données sont résumées dans le tableau suivant :

Site	Position	Oiseaux
1	Chemin	1
2	Chemin	2
3	Chemin	1
4	Chemin	0
5	Hauteurs	8
6	Hauteurs	9

TABLE 3 – Table de données pour l'exercice 3.1

3.1.1

Le responsable du parc a décidé de tirer les deux sites selon un sondage aléatoire simple. Calculer la variance de l'estimateur d'Horvitz-Thompson du nombre total d'oiseaux présents dans le parc.

3.1.2

Quel type de sondage recommander ici pour améliorer la précision ? Pourquoi ?

3.1.3

Donner les π_i et les $\pi_{i,j}$ associés à ce sondage.

3.1.4

Quels sont tous les échantillons possibles ? Donner à chacun sa probabilité de sélection.

3.1.5

Pour chacun des échantillons obtenus à la question précédente, donner l'estimateur d'Horvitz-Thompson du nombre total d'oiseaux présents dans le parc.

3.1.6

Retrouver que, dans ce cas aussi, l'estimateur d'Horvitz-Thompson est sans biais.

3.1.7

Comparer la variance de l'estimateur recommandé à celle obtenue à la question 3.1.1. Le responsable du parc doit-il changer de méthode de sondage ?

3.2 Sondage stratifié selon le salaire

D'après O. Sautory.

Une entreprise emploie 7 500 salariés et souhaite connaître la proportion P d'entre eux qui possèdent au moins une voiture. Pour chaque individu de la base de sondage, on dispose de son salaire. On décide de stratifier la population selon 3 strates de salaires : salaires faibles (strate 1), salaires moyens (strate 2), salaires élevés (strate 3). Sauf indication contraire, on utilise dans l'exercice les notations du cours. On note également p_h l'estimateur de la proportion d'individus possédant au moins un véhicule dans la strate h.

Les résultats de l'enquête menée avec un sondage aléatoire simple (SAS) au sein de chaque strate sont donnés en table 4.

	Strate 1	Strate 2	Strate 3
N_h	3 500	2 000	2 000
n_h	500	300	200
p_h	0,13	0,45	0,50

TABLE 4 – Table de données pour l'exercice 3.2

3.2.1

Quelle est la valeur de l'estimateur d'Horvitz-Thompson pour ces données d'enquête ?

3.2.2

Proposer un intervalle de confiance associé à l'estimateur d'Horvitz-Thompson.

3.2.3

Que pensez-vous de la stratification choisie ?

3.2.4

Que pensez-vous de l'allocation utilisée ?

3.3 Choix d'allocation dans une enquête auprès des entreprises

On cherche à mener une enquête sur l'investissement des entreprises (nature, destination, etc.) d'un secteur donné par un SAS stratifié par tranches d'effectif de 300 unités parmi 1 060.

Grâce aux données fiscales, on dispose d'informations sur la moyenne ($\bar{y}_h$) et la dispersion (S_h^2) du montant total d'investissement pour chaque tranche d'effectif (en milliers d'euros). Ces informations sont indiquées en table 5.

Taille	N_h	$\bar{y}_h$	S_h^2
0-9	500	10	2
10-19	300	50	15
20-49	150	200	50
50-499	100	500	100
500 et plus	10	1 000	2 500

TABLE 5 – Table de données pour l'exercice 3.3

3.3.1

Déterminer les allocations proportionnelle et de Neyman en utilisant le montant total d'investissement comme variable auxiliaire.

3.3.2

Pour chaque cas, calculer la variance de l'estimateur du montant total d'investissement construit à partir de l'enquête.

4 Redressements et non-réponse

4.1 Traitements de la non-réponse

On souhaite mesurer les dépenses totales en électricité des occupants d'une résidence de 250 logements. Pour cela on tire un échantillon aléatoire simple de 25 logements et on demande aux résidents leur type de tarif (Classique ou Social [1]), la surface de l'appartement et leur facture mensuelle en électricité. Cinq résidents refusent de répondre mais on peut tout de même connaître leur orientation (Sud ou Nord). Parmi les autres, cinq n'ont pas voulu donner le montant de leur facture (voir table 6).

Logement	Orientation	Tarif	Surface	Facture électricité
a	N	S	70	60
b	N	S	70	
c	N	S	80	70
d	N	S	90	80
e	N	S	90	80
f	N	S	70	70
g	N	S	70	
h	N	C	80	80
i	N	C	80	80
j	N	C	80	80
k	N	C	80	
l	N	C	90	
m	N	C	90	90
n	S	S	50	40
o	S	S	60	50
p	S	S	70	60
q	S	C	50	40
r	S	C	60	50
s	S	C	70	60
t	S	C	80	
u	N			
v	N			
w	S			
x	S			
y	S			

TABLE 6 – Table de données pour l'exercice 4.1 – Déclarations

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.

4.1.2

Vous apprenez auprès du syndicat de copropriété que 130 logements ont un tarif classique (Tarif = "C") et 120 bénéficient de tarifs sociaux (Tarif = "S"). Modifiez-vous votre réponse ? Pourquoi ?

4.1.3

Parmi les 10 non-réponses sur les factures, on tire un échantillon aléatoire simple comprenant les résidents b, g, w et x. On contacte leur fournisseur d'électricité, qui nous indique combien ils ont payé le mois dernier (voir table 7). Que conclure ?

1. Certains consommateurs peuvent bénéficier pour leur résidence principale d'un tarif de première nécessité (ou tarif social) pour alléger le montant de leurs factures d'électricité.

b	90
g	100
w	70
x	60

TABLE 7 – Table de données pour l'exercice 4.1 – Factures

4.2 Mesure de surfaces cultivées 2

On considère une région agricole composée de $N = 2\ 010$ 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 $s_y^2 = 1\ 069$.

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 :

$$\begin{aligned} \bar{x} &= 132 \\ s_x^2 &= 7\ 619 \\ s_{xy} &= 1\ 453 \end{aligned}$$

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.

4.2.3

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).

En notant 1 et 2 les deux post-strates définies, on a dans l'échantillon :

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

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 %.