

SUJET UE STA108 :
(Enquêtes et sondages)

durée de l'examen : 2 heures
sujet de 4 pages, celle-ci comprise
responsable : G.Saporta

Tous documents et calculatrices autorisés

Les exercices sont indépendants et peuvent être résolus dans n'importe quel ordre.

Les téléphones mobiles et autres équipements communicants doivent être éteints et rangés dans les sacs pendant toute la durée de l'épreuve.

Vérifiez que vous disposez bien de la totalité des pages du sujet en début d'épreuve et signalez tout problème de reprographie le cas échéant.

Exercice 1

Un institut d'études mesure la popularité d'un homme politique H en interrogeant 1000 personnes selon une méthode que l'on considère équivalente à un tirage équiprobable sans remise.

1. On trouve 350 personnes ayant une opinion favorable de H. Donner un intervalle de confiance à 95% sur la cote de popularité de H
2. La ventilation des résultats selon le sexe conduit au tableau suivant :

	Masculin	Féminin
Opinion favorable	25 %	50%
Base	600	400

Or on sait que la proportion de femmes dans la population étudiée est de 52%. Comment s'appelle la méthode de redressement à utiliser ? Redresser les résultats en conséquence et commenter.

Exercice 2

On désire estimer le nombre d'heures passées sur les réseaux sociaux chaque mois par les 30 000 étudiants d'un ensemble universitaire U . Dans la suite, on note Y_k la durée mensuelle que consacre le $k^{\text{ème}}$ étudiant aux réseaux sociaux. Le paramètre d'intérêt est noté :

$$\bar{Y} = \frac{1}{N} \sum_{k \in U} Y_k$$

1. On mène une enquête auprès d'un échantillon S de 200 étudiants choisis selon un sondage aléatoire simple.

Sur l'échantillon, on observe $\sum_{k \in S} Y_k = 3000$ et $\sum_{k \in S} Y_k^2 = 100000$

Estimer \bar{Y} nombre d'heures passées sur les réseaux sociaux chaque mois par l'ensemble des étudiants et en donner un intervalle de confiance à 95 %.

- En fait, on avait sélectionné l'échantillon d'étudiants selon un sondage aléatoire simple stratifié selon leur sexe, avec allocation proportionnelle. On sait qu'il y a 9 000 filles dans l'ensemble universitaire considéré et on relève les résultats suivants :

$$\begin{aligned} \sum_{k \in S_{\text{filles}}} Y_k &= 1400 & \sum_{k \in S_{\text{garçons}}} Y_k &= 1600 \\ \text{et } \sum_{k \in S_{\text{filles}}} Y_k^2 &= 60000 & \sum_{k \in S_{\text{garçons}}} Y_k^2 &= 40000 \end{aligned}$$

- (a) Quelle est la taille de l'échantillon dans chaque strate ?
- (b) Que vaut l'estimateur de \bar{Y} ?
- (c) Estimer la variance de cet estimateur et fournir un intervalle de confiance à 95 %. Que vaut l'effet de sondage ? Commenter.
- (d) En utilisant les données de la question 1, quelle taille d'échantillon aurait-il fallu pour obtenir la même précision estimée à partir d'un sondage aléatoire simple ?
- En faisant l'hypothèse que les dispersions observées dans chaque strate données dans la question 2 sont les "vraies" dispersions par strate dans la population, quelle est l'allocation par strate qui permettrait d'obtenir :
- (a) la même précision pour la durée moyenne d'intérêt estimée chez les filles et la durée moyenne estimée pour les garçons ? Commenter.
- (b) la meilleure précision possible sur l'estimateur de la durée moyenne globale \bar{Y} ?
- Qu'aurait-on obtenu pour \bar{Y} et pour l'effet de sondage si on n'avait tenu compte du sexe qu'une fois l'enquête réalisée ? Comparer l'information auxiliaire qui est utile ici par rapport à celle requise à la question 2. Dans quel cas peut-il être intéressant de tenir compte du sexe après avoir déjà stratifié sur cette variable ?

Exercice 3 : Méthode des quotas

La méthode des quotas, consiste en l'absence de base de sondage à constituer un échantillon de personnes qui ressemblent à la population visée selon un certain nombre de caractéristiques. Du fait des contraintes légales qui limitent la disponibilité de bases de sondage nominatives, c'est la méthode la plus utilisée par les instituts privés. Dans les études de marché, les critères utilisés pour les quotas sont généralement le sexe, l'âge, la catégorie socio-professionnelle, le type de commune, la région...

Cet exercice a pour but de passer en revue un certain nombre de propriétés de la méthode et voir sous quelles conditions l'appliquer correctement. Mise à part la première partie il n'y a pas de calcul à réaliser ou de formule à établir, seulement poser des raisonnements.

1 Joignabilité des interviewés aux différentes heures d'appel

On veut réaliser une enquête téléphonique avec la méthode des quotas. On a constaté par le passé que les profils sociodémographiques n'étaient pas joignables de la même façon aux heures d'appels des enquêtes. On a pu établir le tableau suivant de probabilités de joindre une personne lors d'un coup de téléphone fixe au hasard aux heures d'appels des enquêtes. L'enquête se fait par téléphone à partir de numéros de téléphones fixes générés au hasard.

	Hommes	Femmes	Hommes	Femmes
	actifs		inactifs	
18-24	0.1	0.2	0.2	0.3
25-34	0.2	0.25	0.3	0.4
35-49	0.4	0.5	0.6	0.7
50+	0.7	0.8	0.8	0.9

- a) En notant p la probabilité d'être joint au téléphone et $(1-p)$ la probabilité de ne pas être joint, et en supposant l'indépendance des appels, calculer le nombre d'appels qu'il faudra en moyenne pour joindre un homme actif de 18 à 24 ans avec une probabilité de 80%
- b) Quelles conséquences sur le rythme de l'enquête auront les différences de joignabilité ? Que peut-on proposer pour améliorer la situation ?
- c) Supposons que l'enquête ne fixe des quotas que sur le sexe et l'âge, sans tenir compte du fait que la personne soit active ou non :
 - a. Que va-t-il se passer si on donne comme consigne à l'enquêteur de passer à un nouveau numéro au bout de 2 appels infructueux sur un numéro donné. (pas de calculs, donnez les conséquences sans les évaluer)
 - b. Quelles conséquences cela peut-il avoir sur les estimations ?
 - c. Comment les éviter ? Commenter

2 Recevabilité de la méthode des quotas

A partir de cet exemple, dites les principaux reproches que l'on fait à la méthode de sélection par quotas versus une sélection aléatoire ?

Quelle est la nature de l'erreur faite ?

Au final qu'est ce qui différencie dans les conditions de réalisation, une bonne d'une mauvaise enquête par quotas ?

3 Modèle théorique de la méthode des quotas

Les données suivantes, issues du recensement de la population détaillent en 96 groupes, les 51 millions de résidents en France métropolitaine âgés de 15 ans et plus en 2013.

On veut réaliser une enquête par quotas pour étudier les opinions des français. Pour cela on se sert du tableau suivant pour répartir l'échantillon en 96 cellules proportionnellement aux données INSEE.

- Quel est le modèle théorique le plus proche de cette approche ?
- A quelles conditions une enquête par quotas sera-t-elle plus précise qu'un sondage aléatoire simple ?
- Pourquoi déconseille-t-on la méthode des quotas pour des enquêtes de grande taille par rapport à un sondage aléatoire simple ? (Penser au comportement de l'erreur de la méthode des quotas en fonction de la taille de l'échantillon)

(source INSEE, recensement de la population, 1^{er} janvier 2013)

Age regroupé - HOMMES		15 à 19 ans	20 à 24 ans	25 à 39 ans	40 à 54 ans	55 à 64 ans	65 ans ou plus
CS regroupée	Agriculteurs exploitants	1 029	8 886	89 074	182 431	65 048	8 194
	Artisans, commerçants, chefs entreprise	6 310	25 898	342 035	584 139	216 124	22 096
	Cadres, professions intellectuelles sup.	1 803	77 009	1 022 521	1 131 737	447 401	35 039
	Professions intermédiaires	25 814	268 994	1 377 558	1 337 612	348 399	19 271
	Employés	70 196	280 878	822 443	641 804	147 358	7 694
	Ouvriers	307 407	696 751	2 047 779	2 042 302	445 130	23 163
	Retraités	55	76	2 088	60 942	1 826 565	4 174 216
	Autres sans activité professionnelle	1 579 069	584 343	320 955	348 897	258 361	50 760

Age regroupé - FEMMES		15 à 19 ans	20 à 24 ans	25 à 39 ans	40 à 54 ans	55 à 64 ans	65 ans ou plus
CS regroupée	Agriculteurs exploitants	224	1 276	22 418	70 920	37 775	8 781
	Artisans, commerçants, chefs entreprise	4 757	15 161	140 223	218 477	82 881	16 999
	Cadres, professions intellectuelles sup.	1 670	61 258	743 274	673 478	228 442	12 428
	Professions intermédiaires	22 338	294 765	1 598 048	1 395 609	367 172	16 280
	Employés	174 247	702 153	2 308 409	2 567 372	709 409	36 671
	Ouvriers	35 213	119 153	435 072	595 878	144 126	10 591
	Retraités	90	255	2 784	71 831	1 609 473	5 616 275
	Autres sans activité professionnelle	1 662 559	734 893	860 331	962 908	781 250	454 515