Université de Strasbourg
UFR de Mathématique et d'Informatique
Strasbourg
Année universitaire 2013/2014

Master de Mathématiques 1ère année - Biostatistiques et Statistiques Industrielles

Unité d'enseignement : Sondage
Session des examens d'automne 2013-2014
Durée : 2 heures
Enseignant responsable : M. Maumy-Bertrand

Tous les documents sont autorisés à l'exception des livres. Seules les calculatrices (autres que téléphone portable) et sans imprimante sont autorisées. Les téléphones mobiles et autres équipements communicants doivent être éteints et rangés dans les sacs pendant toute la durée de l'épreuve. Les trois exercices sont indépendants et peuvent être résolus dans n'importe quel ordre. Chaque réponse devra être justifiée précisément.

Exercice 1. (4 points)

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. Nous trouvons 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 genre conduit au tableau suivant :

	Masculin	Féminin
Opinion favorable	25%	50%
Base	600	400

Or nous savons 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. (10 points)

Nous désirons 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, nous notons Y_k la durée mensuelle que consacre le k^{ieme} étudiant aux réseaux sociaux. Le paramètre d'intérêt est noté : $\mu_Y = \frac{1}{N}\sum_{k\in U} Y_k$.

1. Nous menons une enquête auprès d'un échantillon $\mathcal{S}$ de 200 étudiants choisis selon un sondage aléatoire simple. Sur l'échantillon, nous observons :

$$\sum_{i\in\mathcal{S}} Y_i = 3\,000 \quad \text{et} \quad \sum_{i\in\mathcal{S}} Y_i^2 = 100\,000.$$

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

2. En fait, nous avions sélectionné l'échantillon d'étudiants selon un sondage aléatoire simple stratifié selon leur genre, avec allocation proportionnelle. Nous savons qu'il y a 9 000 filles dans l'ensemble universitaire considéré et nous relevons les résultats suivants :

$$\sum_{i\in\mathcal{S}_F} Y_i = 1\,400 \quad \sum_{i\in\mathcal{S}_G} Y_i = 1\,600 \quad \sum_{i\in\mathcal{S}_F} Y_i^2 = 60\,000 \quad \sum_{i\in\mathcal{S}_G} Y_i = 40\,000$$

   (a) Quelle est la taille de l'échantillon dans chaque strate ?
   (b) Que vaut l'estimateur de μ_Y ?
   (c) Estimer la variance de cet estimateur et donner 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 ?

3. 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 μ_Y ?

4. Qu'aurait-on obtenu pour μ_Y et pour l'effet de sondage si nous n'avions tenu compte du genre 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. (6 points)

Le but de cet exercice est de montrer que, lorsque la taille de l'échantillon est fixe, la précision d'un tirage sans remise avec probabilités inégales peut s'exprimer sous une forme « sympathique », dite de Sen-Yates-Grundy.

1. Si nous notons π_k la probabilité d'inclusion de l'individu k, N la taille de la population, et n la taille fixe de l'échantillon, montrer que : $\sum_{k\in U} \pi_k = n$.
2. Si nous notons π_{kl} la probabilité d'inclusion double de k et de l, montrer que, pour tout $k \in U$: $\sum_{l\in U, l\neq k} \pi_{kl} = (n-1)\pi_k$. Indication : utiliser les variables indicatrices.
3. Montrer que, pour tout $k \in U$: $\sum_{l\in U, l\neq k} \pi_k\pi_l = \pi_k(n-\pi_k)$. En déduire que, pour tout $k \in U$: $\sum_{l\in U, l\neq k} (\pi_k\pi_l - \pi_{kl}) = \pi_k(1-\pi_k)$.
4. Mettre la précision de l'estimateur de Horvitz-Thompson $\widehat{T}_Y$ (estimant sans biais le total Y) sous la forme :

$$\mathbb{V}ar\left(\widehat{T}_Y\right) = \left[\sum_{k\in U}\left(\frac{y_k}{\pi_k}\right)^2 \sum_{l\in U, l\neq k} (\pi_k\pi_l - \pi_{kl}) - \sum_{k\in U}\sum_{l\in U, l\neq k} \frac{y_k}{\pi_k}\frac{y_l}{\pi_l}(\pi_k\pi_l - \pi_{kl})\right].$$

   En déduire :

$$\mathbb{V}ar\left(\widehat{T}_Y\right) = \frac{1}{2}\sum_{k\in U}\sum_{l\in U, l\neq k} (\pi_k\pi_l - \pi_{kl})\left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2.$$

   Quel est l'intérêt de cette formulation ?