Examen de sondage : Sujet 1

Tous les documents sont autorisés. Les exercices sont indépendants.
Traiter au choix l'un des deux sujets proposés.

Exercice 1. Strates de médecins
Dans une grande ville, on considère le nombre moyen de clients que peut avoir un médecin pendant une journée de travail. On part de l'idée *a priori* que plus le médecin a d'expérience, plus il a de clients. Cela nous amène à classer la population de médecins en 3 groupes : les "débutants" (classe 1), les "confirmés" (classe 2) et les "très expérimentés" (classe 3). Par ailleurs, on suppose que l'on connaît, dans la base de sondage de médecins, la classe de chacun d'entre eux (1 ou 2 ou 3). Ainsi, on dénombre 500 médecins en classe 1, 1000 en classe 2 et 2500 en classe 3. Par sondage aléatoire simple, on tire 200 médecins dans chaque classe. On calcule alors, dans chaque classe, le nombre moyen de clients par jour et par médecin échantillonné : 10 en classe 1, puis 15 en classe 2 et 20 en classe 3. On calcule enfin les dispersions des nombres de clients par médecin dans chacun des trois échantillons et on trouve respectivement 4 (classe 1), 7 (classe 2), et 10 (classe 3).

1. Comment s'appelle ce plan de sondage ? Justifier *a priori* sa mise en œuvre.
2. Comment estimer le nombre moyen de clients soignés par jour et par médecin ?
3. Donner un intervalle de confiance à 95 % pour le "vrai" nombre moyen de clients soignés par médecin et par jour.
4. En n'ayant comme contrainte que le nombre total de médecin à enquêter (soit 600), procéderait-on comme ci-dessus ?
5. Quel est le gain de variance estimée obtenu avec une allocation proportionnelle par rapport au sondage aléatoire simple (de taille 600) ?
6. Ce gain aurait-il été différent si on avait naïvement estimé la dispersion vraie S_y^2 par la dispersion simple s_y^2 calculée sur l'ensemble de l'échantillon ?

Exercice 2. Espérance de la dispersion
Soit la dispersion non corrigée dans l'échantillon :

$$v_y^2 = \frac{1}{n}\sum_{k\in S}\left(y_k - \widehat{\overline{Y}}\right)^2, \quad \text{où} \quad \widehat{\overline{Y}} = \frac{1}{n}\sum_{k\in S} y_k.$$

1. Donnez l'espérance de v_y^2 pour un plan stratifié avec allocation proportionnelle (on néglige les problèmes d'arrondis qui se posent pour calculer les $n_h = nN_h/N$).
2. Si v_y^2 est utilisée pour estimer

$$\sigma_y^2 = \frac{1}{N}\sum_{k\in U}\left(y_k - \overline{Y}\right)^2,$$

quel est le biais de cet estimateur ? A-t-on tendance à surestimer ou à sous-estimer σ_y^2 ?
3. Quel est l'intérêt pratique du résultat précédent ?

Examen de sondage : Sujet 2

Tous les documents sont autorisés. Les exercices sont indépendants.
Traiter au choix l'un des deux sujets proposés.

Exercice 3. Stratification et estimateur par la différence
Soit un plan stratifié composé de H strates de taille N_h. L'objectif est d'estimer la moyenne de la population $\overline{Y}$ d'un caractère y. Notons $\overline{X}_h$, $h = 1, \ldots, H$ les moyennes dans les strates (dans la population) d'un caractère auxilliaire x. Les $\overline{X}_h$ sont supposées connues et on se propose d'estimer $\overline{Y}$ au moyen de l'estimateur suivant :

$$\widehat{\overline{Y}}_D = \widehat{\overline{Y}}_\pi + \overline{X} - \widehat{\overline{X}}_\pi.$$

On réalise un sondage aléatoire simple dans chaque strate.

1. Montrer que $\widehat{\overline{Y}}_D$ estime $\overline{Y}$ sans biais.
2. Donnez la variance de $\widehat{\overline{Y}}_D$.
3. Quelle est l'allocation optimale des n_h pour minimiser la variance de $\widehat{\overline{Y}}_D$? On considère que le coût unitaire d'enquête ne dépend pas de la strate.
4. Dans quel cas favorable $\widehat{\overline{Y}}_D$ est-il indiscutablement préférable à $\widehat{\overline{Y}}_\pi$?

Exercice 4. Effet de sondage
Lorsqu'on met en oeuvre des plans de sondage complexes et que l'on cherche à calculer des précisions en utilisant un logiciel, on obtient en général le calcul d'un rapport appelé "design effect" ou "effet de sondage". Ce rapport est défini comme le rapport de la variance de l'estimateur du total $\widehat{Y}$ sur la variance de l'estimateur que l'on obtiendrait si on effectuait un sondage aléatoire simple de même taille n. On note $\widehat{\overline{Y}}$ la moyenne simple des y_k pour k dans S.

1. En notant $\mathrm{Var}_p\left[\widehat{Y}\right]$ la variance vraie (éventuellement très compliquée) obtenue sous le plan complexe (noté p), donner l'expression du design-effet (noté désormais DEFF).
2. Comment va-t-on naturellement estimer DEFF (on note $\widehat{\mathrm{DEFF}}$ l'estimateur) ?
   On se restreint désormais à des plans complexes p à probabilités égales et de taille fixe.
3. Dans ces conditions, comment estime-t-on sans biais n'importe quel "vrai" total Y ?
4. Calculer l'espérance de la dispersion s_y^2 dans l'échantillon, sous le plan p (on la note $\mathbb{E}\left[s_y^2\right]$). On l'exprimera en fonction de $\mathrm{Var}_p\left[\widehat{\overline{Y}}\right]$, S_y^2, n et N.
5. Considérant le dénominateur de $\widehat{\mathrm{DEFF}}$, montrer que son utilisation introduit un biais que l'on exprime en fonction de n, N et $\mathrm{Var}_p\left[\widehat{\overline{Y}}\right]$. Pour cette question, on considère que n est "grand".

6. En déduire que le dénominateur de $\widehat{\text{DEFF}}$ a une espérance égale à la valeur souhaitée multipliée par le facteur :
$$1 - \frac{1-f}{n}\text{DEFF}.$$
Conclure dans le cas où n est "grand".