Master Statistique et Econométrie de Toulouse

Cours de théorie des sondages
Examen 2009/2010

Lundi 2 Novembre de 9h45 à 11h45

Toutes vos réponses doivent être justifiées.

1 Questions de cours (6 points - 35 minutes)

Répondez aux questions suivantes par des réponses courtes mais sans oublier de préciser toutes les notations que vous introduirez.

1. Donnez 4 types d'erreurs possibles lors d'un enquête par sondage. Quel est le type d'erreur principalement étudiée dans le cours de théorie des sondages ?
2. Expliquez ce qu'est la non-réponse totale (respectivement partielle) et décrivez (en 2 ou 3 lignes) une méthode pour corriger la non-réponse totale (respectivement partielle) dans une enquête.
3. Si on ne tient pas compte de la non-réponse, à quel problème majeur peut-on être confronté ? Donnez un exemple (4 ou 5 lignes maximum) pour illustrer votre propos.
4. On considère un plan SI
   (a) Rappelez la définition d'un plan SI.
   (b) Précisez les probabilités d'inclusion du premier et du second ordre de ce plan.
   (c) Donnez, pour ce plan, l'estimateur de Horvitz-Thompson d'une proportion et la variance de cet estimateur.
5. On considère un plan BE
   (a) Rappelez ce qu'est un plan BE.
   (b) Précisez les probabilités d'inclusion du premier et du second ordre de ce plan.
   (c) Donnez, pour ce plan, l'estimateur de Horvitz-Thompson d'une proportion et la variance de cet estimateur.
6. On considère un plan avec remise et probabilités proportionnelles à la taille (la taille est définie par une variable auxiliaire x).
   (a) Donnez, pour ce plan, un estimateur du total d'une variable d'intérêt y.
   (b) Donnez, pour ce plan, un estimateur de la variance de cet estimateur du total.

2 Exercice (6 points - 35 minutes)

La population d'étude est l'ensemble des 554 communes de Haute-Garonne de moins de 10000 habitants au recensement 1999. Le statisticien chargé de l'étude dispose pour ces communes du nombre de logements (variable auxiliaire x) et son objectif est d'estimer le nombre total de logements vacants (variable d'intérêt y). Pour cela, il propose de stratifier la population des 554 en communes en 2 strates :

- strate 1 : les 390 communes de moins de 300 logements,
- strate 2 : les 164 communes de plus de 300 logements.

On suppose qu'il dispose de l'information suivante sur chacune des strates $h = 1, 2$:

	$\sum_{U_h} x_k$	$S^2_{yU_h}$
$h = 1$	41584	37.5
$h = 2$	155730	2393.1
U	197314	1104.5

Le statisticien propose de tirer $n = 55$ communes selon un plan STSI. Il propose trois méthodes d'allocation des $n = 55$ communes aux deux strates. Les trois allocations proposées sont les suivantes avec n_1 (respectivement n_2) la taille de l'échantillon tiré dans la première (respectivement 2ème) strate.

	Allocation 1	Allocation 2	Allocation 3
n_1	39	12	13
n_2	16	43	42

1. Expliquez en détaillant les calculs à quelle méthode d'allocation correspond chacune des trois colonnes ci-dessus.
2. Pour l'allocation optimale au sens de Neyman, retrouvez la formule dans le cas particulier de deux strates en utilisant la méthode du multiplicateur de Lagrange.
3. Pour chacune des trois allocations, calculez l'effet plan du plan stratifié considéré.
4. Parmi les 3 allocations, laquelle conduit à la meilleure précision de l'estimateur du total ? Est-ce étonnant ? L'information auxiliaire que cette allocation utilise vous paraît elle facilement accessible en pratique ? Quelle allocation préconiseriez-vous en pratique ?

3 Problème (8 points - 50 minutes)

On considère une population U de $N = 86$ cantons américains avec pour variable d'intérêt y_k : le nombre de médecins généralistes dans le canton $k \in U$. On cherche à estimer le nombre total de médecins $t_{yU} = \sum_U y_k$ à partir d'un échantillon s de taille $n = 20$ tiré selon un plan SI. On suppose que l'on dispose de la variable auxiliaire x_k : le nombre d'habitants dans le canton $k \in U$. On propose deux estimateurs différents : l'estimateur de Horvitz-Thompson et l'estimateur par le ratio utilisant l'information auxiliaire. On note R le ratio $\frac{\sum_U y_k}{\sum_U x_k}$ et $\hat{R}$ l'estimateur usuel de R.

Application numérique : on donne $\sum_U x_k = 2\,398\,534$, $\sum_s x_k = 514\,616$, $\sum_s y_k = 384$, $S^2_{ys} = 530$, $S^2_{xs} = 463\,850\,120$, $S_{xy,s} = 409\,119$.

1. **Cas de l'estimateur d'Horvitz-Thompson**
   (a) Rappelez la définition de l'estimateur de Horvitz-Thompson $\hat{t}_{y\pi}$ pour un plan SI ainsi que la variance et un estimateur de la variance de $\hat{t}_{y\pi}$.
   (b) *Application numérique :* calculez l'estimation par les valeurs dilatées de t_{yU}, précisez le CVE de cet estimateur et donnez un intervalle de confiance à 95% pour t_{yU}.
2. **Cas de l'estimateur par le ratio**
   (a) A partir de la figure ci-dessous qui donne un diagramme de dispersion de y en fonction de x pour les 86 cantons, justifiez l'utilisation de l'estimateur par le ratio $\hat{t}_{y,\text{ratio}}$.
   (b) Rappelez la définition de l'estimateur par le ratio.
   (c) Montrez en utilisant la linéarisation de Taylor que :

$$\hat{R} \simeq R + \frac{1}{\sum_U x_k} \sum_s \frac{y_k - R x_k}{\pi_k}.$$

   (d) Montrez à partir de l'approximation précédente que l'estimateur par le ratio est approximativement sans biais pour estimer t_{yU} et donnez une variance approchée de $\hat{t}_{y,\text{ratio}}$.
   (e) Justifiez l'utilisation de l'estimateur suivant pour estimer la variance approchée de $\hat{t}_{y,\text{ratio}}$ calculée précédemment :

$$N^2 \left(\frac{1}{n} - \frac{1}{N}\right) \left[S_{ys}^2 - 2\widehat{R} S_{xy,s} + \widehat{R}^2 S_{xs}^2\right] \tag{1}$$

   (f) *Application numérique :* donnez une estimation par le ratio du total t_{yU} ainsi que le CVE de cet estimateur en utilisant (1). Donnez un intervalle de confiance à 95% pour t_{yU}.
3. Comparez les résultats obtenus en 1 et 2.
4. On sait par ailleurs que $t_{yU} = 2160$. Pouvez-vous en conclure un résultat de portée générale ?