Principes et Méthodes Statistiques
TP 2018

Le travail sera conduit par groupes de 2 ou 3 personnes, ces groupes étant constitués au hasard. Le délivrable de ce TP est une archive contenant deux fichiers. Le premier, au format pdf, est le compte-rendu du TP. Le second, au format texte, contiendra l'intégralité du code `R` développé. L'archive devra être déposée sur Teide avant le vendredi 13 avril 2018 à 22h. Tout retard sera pénalisé.

Le compte-rendu comprendra, suivant la nature des questions posées, des calculs mathématiques et/ou des sorties numériques et graphiques de `R`. Une grande importance sera accordée aux commentaires, visant à interpréter les résultats et mettre en valeur votre analyse du problème. Des conseils et des directives obligatoires pour la rédaction du compte-rendu sont disponibles sur Chamilo ; les enseignants pourront y faire référence dans leur correction.

Le but du TP est d'étudier deux méthodes élémentaires d'estimation de la taille d'une population, utilisées en écologie, dites de *capture-recapture*.

Voir `http://fr.wikipedia.org/wiki/Capture-marquage-recapture`.

Un étang contient un nombre inconnu θ de poissons. On souhaite estimer au mieux θ en évitant de vider l'étang de tous ses poissons. Pour cela, on fait une première pêche de n_0 poissons. On les bague et on les rejette à l'eau. Deux stratégies sont alors possibles pour estimer θ.

La première stratégie consiste à pêcher les uns après les autres un nombre fixé n de poissons. Pour chaque poisson pêché, on regarde s'il est bagué et on le rejette à l'eau. On note alors X_i la variable aléatoire qui vaut 1 si le $i^{\text{ème}}$ poisson pêché est bagué et 0 sinon. L'observation est donc le vecteur $(x_1, \ldots, x_n)$, réalisation de $(X_1, \ldots, X_n)$.

La deuxième stratégie consiste à pêcher des poissons les uns après les autres jusqu'à ce que l'on ait obtenu un nombre m fixé de poissons bagués. Chaque poisson pêché est rejeté à l'eau. On note Y_j le nombre aléatoire de poissons pêchés entre l'obtention du $(j-1)^{\text{ème}}$ et du $j^{\text{ème}}$ poisson bagué. L'observation est donc le vecteur $(y_1, \ldots, y_m)$, réalisation de $(Y_1, \ldots, Y_m)$. Par exemple, $y_1 = 4$ si le premier poisson bagué est le quatrième poisson péché.

On suppose que les conditions de pêche et de remise à l'eau sont telles que chacun des θ poissons de l'étang a la même probabilité d'être pêché à chaque tentative, et que les résultats des pêches successives sont indépendants, de sorte qu'à chaque pêche, la probabilité que le poisson pêché soit bagué est $\dfrac{n_0}{\theta}$.

1 Première stratégie

1.1. Montrer que $X_1, \ldots, X_n$ sont des variables aléatoires indépendantes et de même loi de Bernoulli de paramètre $\dfrac{n_0}{\theta}$.

Pour un étang contenant $\theta = 1000$ poissons, on a bagué $n_0 = 50$ poissons lors de la première pêche. Choisir une valeur du nombre n de poissons pêchés lors de la seconde pêche et simuler un échantillon $x_1, \ldots, x_n$. Comparer la moyenne et la variance empirique de cet échantillon à l'espérance et la variance théorique de la loi des X_i.

1.2. Donner la loi du nombre total T de poissons bagués parmi les n poissons pêchés. Donner sa réalisation t sur l'exemple simulé.

1.3. Montrer que l'estimateur des moments $\tilde{\theta}_n$ et l'estimateur de maximum de vraisemblance $\hat{\theta}_n$ de θ, calculés à partir de $X_1, \ldots, X_n$, sont confondus. Donner la valeur de l'estimation correspondante pour l'exemple.

1.4. En utilisant les résultats du cours, donner un intervalle de confiance exact et un intervalle de confiance asymptotique de seuil α pour θ. Donner ces intervalles de confiance aux seuils 1%, 5%, 10% et 20% pour les données simulées. Que constatez-vous ?

1.5. Calculer $P(\hat{\theta}_n = +\infty)$. Que peut-on en déduire sur le biais de cet estimateur ? Que vaut cette probabilité dans l'exemple ?

1.6. Pour quelles valeurs de n a-t-on $P(\hat{\theta}_n = +\infty) > \dfrac{1}{2}$? Faites des simulations en `R` qui mettent ce fait en évidence.

2 Deuxième stratégie

2.1. Montrer que $Y_1, \ldots, Y_m$ sont des variables aléatoires indépendantes et de même loi géométrique de paramètre $\frac{n_0}{\theta}$. Choisir une valeur de m et simuler un échantillon $y_1, \ldots, y_m$. Attention, en `R`, `rgeom` simule une variable aléatoire de même loi que $Y - 1$, où Y est de loi géométrique. Comparer la moyenne et la variance empirique de cet échantillon à l'espérance et la variance théorique de la loi des Y_j.

2.2. Donner la loi du nombre total N de poissons pêchés. Donner sa réalisation n sur l'exemple simulé.

2.3. Montrer que l'estimateur des moments $\tilde{\theta}'_m$ et l'estimateur de maximum de vraisemblance $\hat{\theta}'_m$ de θ, calculés à partir de $Y_1, \ldots, Y_n$, sont confondus. Donner la valeur de l'estimation correspondante pour l'exemple.

2.4. Calculer la quantité d'information de Fisher $\mathcal{I}_m(\theta)$. Montrer que l'estimateur $\hat{\theta}'_m$ est sans biais et de variance minimale.

2.5. On admettra qu'un intervalle de confiance asymptotique de seuil α pour θ est donné par :

$$\left[\hat{\theta}'_m - \frac{u_\alpha}{\sqrt{\mathcal{I}_m(\hat{\theta}'_m)}}, \hat{\theta}'_m + \frac{u_\alpha}{\sqrt{\mathcal{I}_m(\hat{\theta}'_m)}}\right]$$

Donner ces intervalles de confiance aux seuils 1%, 5%, 10% et 20% pour les données simulées. Que constatez-vous ?

3 Application et comparaison des stratégies

3.1. Le fichier `Peche.txt` contient les résultats de 1000 pêches successives sous la forme du vecteur $x_1, \ldots, x_{1000}$, pour lequel le nombre de poissons bagués est $n_0 = 100$. La commande `scan` permet de créer un vecteur dans `R` à partir du contenu de ce fichier. Calculer l'estimation de θ et les intervalles de confiance exact et asymptotique de seuil 5% selon la première stratégie.

3.2. Ecrire une fonction `R` permettant de créer le vecteur $y_1, \ldots, y_m$ à partir du vecteur $x_1, \ldots, x_n$. Calculer l'estimation de θ et un intervalle de confiance asymptotique de seuil 5% selon la deuxième stratégie.

3.3. À l'aide d'un graphe de probabilités, vérifier si l'hypothèse de loi géométrique est pertinente pour le vecteur $y_1, \ldots, y_m$.

3.4. Quelle est selon vous la meilleure stratégie pour estimer θ ?

4 Vérifications expérimentales à base de simulations

4.1. Choisir θ, n_0, n, m et α. Simuler m échantillons de taille n de la loi de Bernoulli $\mathcal{B}(n_0/\theta)$. Pour chacun d'entre eux, calculer les intervalles de confiance de seuil α pour θ obtenus dans la question 1.4. Comparer la proportion d'intervalles contenant la vraie valeur de θ à $1-\alpha$. Quel est l'impact du choix des valeurs de θ, n_0, n, m et α ?

4.2. Vérification de la loi faible des grands nombres. Simuler m échantillons de taille n de la loi $\mathcal{B}(p)$. Calculer le nombre de fois où l'écart en valeur absolue entre la moyenne empirique et l'espérance de la loi simulée est supérieure à un ϵ à choisir. Faire varier n et conclure.

4.3. Vérification du théorème central-limite. Simuler m échantillons de taille n de la loi $\mathcal{B}(p)$. Sur l'échantillon des m moyennes empiriques, tracer un histogramme et un graphe de probabilités pour la loi normale. Faire varier n en partant de $n = 5$ et conclure.