Principes et Méthodes Statistiques
TP 2022

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 sera le compte-rendu du TP au format Rmd, qui comprendra le code R et vos réponses détaillées aux questions selon les règles présentées ci-dessous. Le second sera le fichier pdf issu directement de la compilation (knit) du fichier Rmd. L'archive devra être déposée sur Teide avant le vendredi 15 avril 2022 à 22h. Tout retard sera pénalisé.

Le compte-rendu Rmd 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.

---

1 Estimation de la densité foliaire

Couvrant à peine 7% de la surface terrestre de la Terre, les forêts tropicales jouent un rôle disproportionné dans la biosphère : elles stockent environ 25% du carbone terrestre et contribuent à plus d'un tiers des émissions mondiales d'oxygène. Elles recyclent également environ un tiers des précipitations par évapotranspiration et contribuent ainsi à générer et à maintenir un climat humide au niveau régional, avec des effets positifs s'étendant bien au-delà des tropiques.

Pour mieux comprendre ces phénomènes et prédire les cycles biogéochimiques mondiaux, de nombreuses mesures sont effectuées sur les forêts tropicales. Parmi elles, l'indice de surface foliaire (https ://fr.wikipedia.org/wiki/Indice_de_surface_foliaire) représente la surface des feuilles d'un arbre par unité de surface de sol. C'est un indicateur clé contrôlant l'écoulement d'eau et l'apport de carbone. Pour le mesurer, on a recours à la télédétection par laser (lidar, https ://fr.wikipedia.org/wiki/Lidar).

On s'intéresse à la détection de feuilles sur une zone déterminée. Le lidar lance un rayon qui atteint une feuille avec une probabilité p. Une expérience consiste à faire k tirs de laser et à relever le nombre X de tirs ayant touché des feuilles. On suppose que les tirs sont des expériences identiques et indépendantes, de sorte que X est une variable aléatoire de loi binomiale $\mathcal{B}(k,p)$. On fait cette expérience dans n zones différentes. On suppose que les nombres relevés $X_1,\ldots,X_n$ sont indépendants et de même loi $\mathcal{B}(k,p)$.

1. On considère que $p=\lambda/k$, où λ est une constante. Montrer que quand k tend vers l'infini, $P(X=x)$ tend vers $e^{-\lambda}\,\dfrac{\lambda^x}{x!}$. On pourra utiliser la formule de Stirling, qui dit que, quand k tend vers l'infini, $k!$ est équivalent à $\sqrt{2\pi k}\,(k/e)^k$.

Ce résultat est utilisé en pratique pour dire que, quand k est grand et p petit, la loi binomiale $\mathcal{B}(k,p)$ peut être approchée par la loi de Poisson $\mathcal{P}(kp)$.

2. Proposer une expérimentation en `R` permettant de comparer les lois $\mathcal{B}(k,p)$ et $\mathcal{P}(kp)$ pour diverses valeurs de k et p. Pour quelles valeurs de k et p peut-on considérer que l'approximation ci-dessus est satisfaisante ?

On considère que les feuilles sont peu présentes dans la zone de recherche et que le nombre de mesures laser est important, de sorte que les variables aléatoires $X_1,\ldots,X_n$ définies plus haut peuvent être considérées comme indépendantes et de même loi de Poisson $\mathcal{P}(\lambda)$. Le paramètre λ caractérise la densité foliaire, c'est lui que l'on souhaite estimer avec précision au vu d'une réalisation $x_1,\ldots,x_n$ de $X_1,\ldots,X_n$. On rappelle que l'estimateur de maximum de vraisemblance et l'estimateur des moments sont identiques et valent $\hat{\lambda}_n=\bar{X}_n$. On rappelle également que $\sum_{i=1}^{n} X_i$ est de loi $\mathcal{P}(n\lambda)$.

3. La fonction de répartition de la loi de Poisson n'a pas d'expression simple, ce qui fait qu'on ne peut pas utiliser la méthode habituelle pour construire un graphe de probabilités pour la loi de Poisson.
   (a) Pour X de loi $\mathcal{P}(\lambda)$ et x entier, calculer $\ln[x!P(X=x)]$.
   (b) En déduire une méthode de type graphe de probabilités pour évaluer visuellement si la loi de Poisson pourrait être un modèle approprié pour un échantillon $x_1,\ldots,x_n$ observé. Proposer alors une procédure pour estimer graphiquement λ. Commandes utiles de `R` : `table` et `unique`.
   (c) Vérifier que la méthode fonctionne sur des échantillons simulés de loi $\mathcal{P}(3)$ et de loi géométrique $\mathcal{G}(0.5)$, de taille $n=100$.

4. Calculer la quantité d'information de Fisher $\mathcal{I}_n(\lambda)$.
5. Montrer que $\hat{\lambda}_n$ est l'estimateur sans biais et de variance minimale de λ.
6. En appliquant le théorème central-limite, construire un intervalle de confiance bilatéral asymptotique de seuil α pour λ.
7. Construire un test de H_0 : "$\lambda \leq \lambda_0$" contre H_1 : "$\lambda > \lambda_0$".

Le fichier `lidar.csv` contient 100 mesures indépendantes de la variable aléatoire X. Charger le tableau de données dans `R` en utilisant la commande `lidar<-scan("lidar.csv")`.

8. Utiliser la méthode de la question 3 pour évaluer si la loi de Poisson pourrait être un modèle approprié pour ces données.
9. Estimer λ.
10. Donner un intervalle de confiance asymptotique de seuil 5% pour λ.
11. Les mesures effectuées permettent de comparer des zones géographiques différentes, par exemple les forêts équatoriales et tropicales. On admet qu'une zone de référence est caractérisée par une densité foliaire $\lambda_0 = 5$. Au vu de ces données, peut-on admettre avec moins de 5% de chances de se tromper que la zone sur laquelle les données lidar ont été récoltées a une plus grande densité foliaire que la zone de référence ?

2 Vérifications expérimentales à base de simulations

1. Vérifiez expérimentalement que, quand on simule un grand nombre m d'échantillons de taille n de la loi $\mathcal{P}(\lambda)$, alors une proportion approximativement égale à $1-\alpha$ des intervalles de confiance de seuil α obtenus contient la vraie valeur du paramètre λ. Prendre plusieurs valeurs de α, λ, n et m et commenter les résultats.

2. Vérification de la convergence faible de l'estimateur.

   Un estimateur T_n de λ est dit faiblement convergent (ou convergent en probabilité) si et seulement si

   $$\forall \varepsilon > 0, \lim_{n \to +\infty} P(|T_n - \lambda| > \varepsilon) = 0.$$

   Vérifier la convergence faible de $\hat{\lambda}_n$ en choisissant plusieurs valeurs adaptées de ε et de n, puis en simulant m échantillons de taille n de la loi $\mathcal{P}(\lambda)$ et enfin, en calculant le pourcentage de fois où l'écart en valeur absolue entre l'estimation et le vrai paramètre est supérieur à ε (vous pouvez tracer des courbes). Conclure.

3. Vérification de la normalité asymptotique de l'estimateur.

   Simuler m échantillons de taille n de la loi $\mathcal{P}(\lambda)$. Sur l'échantillon des m estimations, tracer un histogramme et un graphe de probabilités pour la loi normale. Faire varier n en partant de $n = 5$ et conclure.