TP Statistiques 3

Juhyun Park, Phuong Thuy Vo, Lassaad Mchri, Nicolas Brunel

31 mars 2023

Estimation du Maximum de vraisemblance et L'intervalle de confiance

Vraisemblance: La loi Bernoulli

Soit X une variable aléatoire de Bernoulli (`rbinom`) avec $p = 0.6$.

1. Simuler un échantillon i.i.d de taille $n = 10$. Quelle est une façon simple d'estimer p?
2. Générer une fonction de vraisemblance, nommée `L_bern`, en fonction de (p, x), qui donne la vraisemblance d'un échantillon $x = (x_1, \ldots, x_n)$ pour une valeur donnée de p.
3. Pour votre échantillon, estimer la vraisemblance de l'échantillon pour n lois Bernoulli de paramètres p allant de 0 à 1. Tracez la courbe des valeurs calculées. Que remarquez-vous?
4. En utilisant la fonction `optim` de R, trouvez la valeur de p la plus probable d'avoir généré cet échantillon.

Attention : `optim` est par défaut une routine de **minimization**. Remarque : Avec la méthode de *L-BFGS-B* dans la fonction `optim`, vous pouvez traiter des contraintes sur le(s) paramètre(s), lorsque c'est nécessaire.

5. Tester avec des échantillons de taille allant de $n = 10$ à $n = 2000$ et comparer l'écart entre la valeur théorique attendue et la valeur obtenue. Que remarquez-vous? Comment combattre l'instabilité numérique due aux multiplications de probabilités?

6. Trouver deux intervalles de confiance de niveau 0.90 pour le paramètre p, d'après (i) l'inégalité de Bienaymé-Chebycheff et (ii) l'inégalité de Hoeffding et les comparer. Incluent-ils la valeur réelle ?

Vraisemblance pour plusieurs paramètres: La Loi Gamma

Soit $X_1, \ldots, X_n$ un échantillon de n variables indépendantes de loi de Gamma(α, β) où $\theta = (\alpha, \beta)$ est inconnue. Simuler un échantillon i.i.d de taille $n = 25$ avec $\theta_0 = (2.5, 1.5)$.

7. Présentez l'histogramme des données simulées. Choisir trois paramètres candidats, disons, θ_0 (vrai) θ_1, θ_2. Comparer l'histogramme avec les densités candidates. Que remarquez-vous?

8. Ecrire la log vraisemblance `logL_gamma`. Générez une fonction de log-vraisemblance avec les arguments (θ, x), qui donne la log vraisemblance d'un échantillon pour une valeur donnée de $\theta = (\alpha, \beta)$ et les donnéé $x = (x_1, \ldots, x_n)$. Pour votre échantillon, estimer la log-vraisemblance de paramètre $\theta = (\alpha, \beta)$, en faissant varier un paramètre à la fois.

9. Tracer la courbe des valeurs ainsi calculées. Comme la fonction, $\ell(\theta)$, a deux arguments, vous pouvez réaliser des tracés de contour (`contour`). Pour plus de simplicité, il suffit de tracer comme une fonction unidimensionnelle en supposant que l'autre est fixe: $\ell(\alpha|\beta = \beta_0)$ avec quelque β_0 de votre choix et $\ell(\beta|\alpha = \alpha_0)$ pour quelque α_0 de votre choix. Que remarquez-vous?

10. Donner l'expression mathématique du vecteur Score (les dérivées premières) à laquelle l'EMV répond. En utilisant la fonction `optim`, trouver la valeur de θ la plus probable pour votre l'échantillon.

11. Répéter l'estimation 100 fois avec de nouveaux ensembles de données (échantillons) et tracer les estimations $\hat{\alpha}$ vs $\hat{\beta}$. Sont-elles indépendantes ? Trouver l'intervalle où 95% des estimations sont incluses pour chaque paramètre. Vous venez de trouver un intervalle de confiance empirique à 95% ! Visualisez vos résultats à l'aide d'un histogramme. Que remarquez-vous?

12. Tester avec des échantillons de taille $n = 15$ et $n = 100$ et comparer avec les résultats précédents. Quel est l'effet de la taille de l'échantillon ?

Normalité asymptotique de l'EMV et l'intervalle de confidence

En pratique, nous n'avons qu'un seul ensemble de données avec des paramètres $\theta = (\theta_1, \ldots, \theta_p)$ inconnus, la simulation de multiples échantillons permettant de construire la distribution de l'estimateur n'est donc plus possible. Dans ce cas, on peut construire des intervalles basés sur une approximation asymptotique de la distribution, en utilisant le fait que $\hat{\theta} \approx \mathcal{N}(\theta, I_n(\theta)^{-1})$.

L'évaluation de l'information de Fisher directe nécessite de calculer le hessien H et son espérance analytiquement en différenciant l'opposé de la log-vraisemblance deux fois par rapport aux paramètres (pour obtenir la matrice complète), puis d'inverser explicitement la matrice, et enfin de remplacer θ par $\hat{\theta}$.

13. Pour revenir à l'exemple du Gamma ($n = 25$), trouver l'information de Fisher et estimer la covariance asymptotique. A partir de là, construire un intervalle de confiance asymptotique de niveau 0.95. Comparer avec la solution de 11.

En général, il n'est pas possible d'évaluer l'information de Fisher de manière analytique, ou cela peut prendre trop de temps. Dans ce cas, on utilise l' "information observée" (sans ésperance), $I_O(\hat{\theta})^{-1} = -H(\hat{\theta})^{-1}$.

Nous pouvons utiliser l'option 'hessian=TRUE' dans `optim` pour obtenir la matrice hessienne et estimer la fonction de covariance par son inverse.

14. Estimer la covariance asymptotique avec l'information observée et construire un intervalle de confiance asymptotique de niveau 0.95.

15. Nous utilisons la simulation pour comparer la performance de ces deux estimateurs. Simuler plusieurs fois un nouvel ensemble de données, construiser les deux intervalles de confiance et compter combien de fois cet intervalle contient les vraies valeurs. Quelle est votre conclusion?