TP Statistiques 3

Juhyun Park, Angie Pineda, Nicolas Brunel

11 mars 2022

Echantillon, Théorème Central Limite, Estimation Monte Carlo

Ceci est un brève illustration du théorème central limite et une sensibilisation à l'estimation "par Monte Carlo''.

1. Soit $\mu = 1$, $\sigma = 2$. Simuler $N = 1000$ échantillons i.i.d $S^i = (X_1^i, \dots, X_n^i), i = 1, \dots, N$, de taille $n = 5, 30, 100$, et dont la loi commune est une loi gaussienne de $\mathcal{N}(\mu, \sigma^2)$.
   Calculer les moyennes et variances empiriques $\bar{X}_{n,i}$ et $\sigma^2_{n,i} = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{X}_n)^2$ pour $i = 1, \dots, N$.
   Tracer l'histogramme des moyennes empiriques. Quelle est la loi théorique de la moyenne empirique ?
   A l'aide d'une renormalisation adéquate (a_n, b_n), montrer que $U_{n,i} = \frac{\bar{X}_{n,i} - a_n}{b_n}$ a une loi connue que vous expliciterez. Comparez histogramme des $U_{n,i}$ et distribution théorique. Quelle est l'influence de la taille de l'échantillon n ?

2. Simuler $N = 1000$ échantillons i.i.d de loi commune Pareto $\mathcal{P}(a, \alpha)$ (la densité est $f(x; a, \alpha) = \alpha \frac{a^\alpha}{x^{\alpha+1}} 1_{[a,+\infty[}$) de taille $n = 5, 30, 100$.
   Calculer les moyennes et variances empiriques $\bar{X}_{n,i}$ et $\sigma^2_{n,i}$.
   Vérifier que l'espérance théorique d'une loi de Pareto est $E[X] = \frac{\alpha a}{\alpha - 1}$ (avec la formule $\int_0^\infty P(X > t)dt$). On rappelle que la variance d'une Pareto est $V(X) = \left(\frac{\alpha a}{\alpha - 1}\right)^2 \frac{\alpha}{\alpha - 2}$ (pour $\alpha \geq 2$).
   Tracer l'histogramme des moyennes empiriques. Quelle est la loi théorique de la moyenne empirique ?
   A l'aide d'une renormalisation adéquate (a_n, b_n), montrer que $U_{n,i} = \frac{\bar{X}_{n,i} - a_n}{b_n}$ a une loi que vous pouvez approchez. Comparez histogramme et "distribution théorique approchée''. Quelle est l'influence de la taille de l'échantillon n sur la qualité de cette approximation?

3. Simuler $N = 1000$ échantillons i.i.d de loi de Poisson de taille $n = 5, 30, 100$.
   Calculer les moyennes et variances empiriques $\bar{X}_n$ et σ^2_n.
   Rappeler les expressions théoriques de la moyenne et variance d'une loi de Poisson.
   Tracer l'histogramme des moyennes empiriques. Quelle est la loi théorique de la moyenne empirique ?
   A l'aide d'une renormalisation adéquate (a_n, b_n), montrer que $U_{n,i} = \frac{\bar{X}_{n,i} - a_n}{b_n}$ a une loi que vous pouvez approchez. Comparez histogramme et "distribution théorique approchée''. Quelle est l'influence de la taille de l'échantillon n?

4. Déduire des expérimentations précédentes une méthodologie d'estimation de quantités de la forme $E[T(X_1, \dots, X_n)]$ où T est une statistique d'un échantillon $(X_1, \dots, X_n)$ que l'on est capable de simuler "facilement''. Vous expliciterez comment N influence la qualité de cette approximation.

Maximum de vraisemblance

Ceci a pour objectif de vous familiariser avec la méthode des estimateur de maximum de vraisemblance. Cette méthode est utilisée pour estimer, pour une loi donnée et un échantillon, les paramètres qui maximisent la probabilité que cette loi ait généré l'échantillon.

Ajuster une loi Bernoulli

Soit une loi Bernoulli (`rbinom`) avec $p = 0.7$. Simuler un échantillon i.i.d de taille $n = 10$.

5. Quelle est une façon simple d'estimer p?
6. Générez une fonction de vraisemblance, nommée `L_bern`, qui donne la vraisemblance d'un échantillon de Bernoulli pour une valeur donnée de p.
7. Pour votre échantillon, estimez 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?
8. En utilisant la fonction `optimize` de R, trouvez la valeur de p la plus probable d'avoir généré cet échantillon. (Attention à utiliser l'option maximum = TRUE et de borner les valeurs de p en utilisant l'option interval).
9. Testez avec des échantillons de taille allant de $n = 10$ à $n = 2000$ et comparez l'écart entre la valeur théorique attendue et la valeur obtenue. Que remarquez-vous? Comment combattre l'instabilité numérique due aux multiplications de probabilités?

Ajuster d'une loi normale d'écart type connu

Soit une loi Normale (`rnorm`) avec $\mu = 2$ et $\sigma = 1$. Simuler un échantillon i.i.d de taille $n = 10$.

10. Générez une fonction de vraisemblance, nommée `L_norm`, qui donne la vraisemblance d'un échantillon pour une valeur donnée de μ. (Voir la fonction `dnorm`)
11. Pour votre échantillon, estimez la vraisemblance de l'échantillon pour n lois normales de paramètres μ allant de 0 à 4. Tracez la courbe des valeurs calculées. Que remarquez-vous?
12. En utilisant la fonction `optimize` de R, trouvez la valeur de μ la plus probable d'avoir généré cet échantillon. (Attention à utiliser l'option maximum = TRUE et de borner les valeurs de μ en utilisant l'option interval).
13. Testez avec des échantillons de tailles allant de $n = 10$ à $n = 2000$ et comparez l'écart entre la valeur théorique attendue et la valeur obtenue. Utilisez une technique qui permet d'éviter les instabilités en caclulant des sommes.
14. Utilisez les données sur l'ozone ("summer_ozone.csv", "winter_ozone.csv") pour trouver l'estimateur de maximum de vraisemblance pour chaque site à chaque saison si on considére que c'est une loi normale.
    Est-ce que c'est raisonnable de fixer le paramètre d'écart type pour qu'il soit le même ?