TP Statistiques 2

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

17 février 2023

Simulation et convergence

Tout type de **statistiques d'échantillon** telles que la moyenne ou les quantiles d'échantillon etc. présente une variabilité aléatoire. Nous explorerons cette propriété plus loin avec une expérience de simulation.

I. Variation sous-jacente et échantillonnage répété

La densité de la loi exponentielle avec le paramètre de taux $\lambda > 0$, notée $X \sim \mathcal{E}(\lambda)$, est $f(x) = \lambda e^{-\lambda x} 1_{\mathbb{R}^+}(x)$.

1. Si $X \sim \mathcal{E}(0.5)$, quelle est la probabilité qu'on observe une valeur supérieure à 3?
2. Simulez un échantillon de taille $n = 20$ d'un loi de $\mathcal{E}(0{,}5)$, créez un histogramme de votre échantillon et commentez la forme de votre histogramme. Superposer la vrai densité. Quelle est la probabilité empirique qu'on observe une valeur supérieure à 3 ?
3. Répétez cette opération 5 ou 6 fois et commentez les différences entre les histogrammes que vous obtenez à chaque fois. Utilisez la même limite sur les axes pour faciliter la comparaison. Notez également comment la probabilité empirique qu'on observe une valeur supérieure à 3 change.
4. Augmentez la taille de votre échantillon à 100 et répétez votre expérience. Que remarquez-vous?

II. Variabilité aléatoire du maximum de l'échantillon

1. Simuler un échantillon de taille $n = 10$ d'une loi $\mathcal{U}(-1, 1)$ et enregistrez le maximum de l'échantillon.
2. Répétez les deux étapes ci-dessus dix fois, en écrivant le maximum de l'échantillon à chaque fois. Commentez la variabilité des valeurs que vous obtenez pour les maxima de votre échantillon.
3. Répétez 100 fois et construisez un histogramme et une boîte à moustaches. Quelle est la loi du maximum, $M = \max_{1 \leq i \leq n} X_i$ où $X_i \sim \mathcal{U}(-1, 1)$ (TD1) ? Superposer la densité théoreique sur l'histogramme. Que remarquez-vous ?
4. Augmentez la taille de votre échantillon à 50 et répétez votre expérience. Que remarquez-vous? Sont-ils proches de la symétrie ?

Monte Carlo Methods

Nos expériences précédentes sont des exemples d'application des méthodes de Monte Carlo.

Dans de nombreux systèmes complexes, les quantités d'intérêt ne peuvent pas être résolues analytiquement, nous nous appuyons donc sur des méthodes numériques. Les méthodes de Monte Carlo sont des méthodes statistiques qui reposent sur un échantillonnage répété pour approximer les quantités d'intérêt.

Sous une forme basique, supposons que nous soyons intéressés par l'évaluation de l'espérance d'une variable aléatoire telle que: $\theta = E[\psi(X)]$ où $X \sim f(\cdot)$. S'il est possible de simuler $X_i \sim f, i = 1, \ldots, n$, nous pouvons

approximer l'intégrale par la **moyenne empirique**:

$$\theta = E[\psi(X)] = \int_{\mathbb{R}} \psi(x) f(x)\, dx \approx \frac{1}{n} \sum_{i=1}^{n} \psi(X_i) = \hat{\theta} \equiv \hat{\theta}(X_1, \ldots, X_n)\,.$$

Verifier que $E[\hat{\theta}] = \theta$. En raison de la variabilité de l'échantillon sous-jacent, nous n'avons pas de solution exacte. Néanmoins, nous pouvons donner une certaine garantie probabiliste pour l'erreur comme:

$$P(|\hat{\theta} - \theta| \geq \delta)\,, \quad \delta > 0\,.$$

Moyenne et phénomène de concentration.

1. Supposons que la variance $\sigma^2 = V[\psi(X)] < \infty$. Donner une borne de cette quantité en utilisant l'inégalité de Bienaymé Chebychev.
2. En supposant que $a \leq \psi(X_i) \leq b$, donner une borne en utilisant l'inégalité de Hoeffding.
3. De combien d'échantillons auriez-vous besoin pour que la prabilitié pour $\delta = 2\sigma$ soit inférieure à 1% ?

Si nous pouvons répéter la procédure, le principe de Monte Carlo également s'applique à l'estimation de la probabilité d'erreur ou les autre charactéristiques de variabilité de l'estimateur comme la variance.

Application pour l'estimation de probabilité:

1. Pour la question I avec $\mathcal{E}(0.5)$, identifier le paramètre d'intérêt θ et donner un estimateur $\hat{\theta}$.
2. Exprimer $\eta = P(|\hat{\theta} - \theta| \geq \delta)$ comme l'espérance d'une certaine variable aléatoire Z de sorte que $\eta = E[Z]$. Déterminer une estimation de η par la méthode de Monte Carlo pour $n = 20$ et $n = 100$.
3. Comparer avec les bornes obtenues par Bienaymé Chebychev.

Théorème Central Limite et Estimation Monte Carlo

La densité de la loi Pareto, $\mathcal{P}(a, \alpha)$, est $f(x; a, \alpha) = \alpha \frac{a^\alpha}{x^{\alpha+1}} 1_{[a,+\infty[}$

1. 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 $\alpha \geq 2$).
2. Simuler $N = 1000$ échantillons i.i.d de loi commune Pareto $\mathcal{P}(a, \alpha)$ (avec votre choix de paramètres) de taille $n = 5, 30, 100$ et calculer les moyennes et variances empiriques $\bar{X}_{n,i}$ et $S_{n,i}, i = 1, \ldots, N$.
3. Tracer l'histogramme des moyennes empiriques.
4. 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 de les moyennes empiriques normalisées, $U_{n,i}$, et *distribution théorique approchée*. Quelle est l'influence de la taille de l'échantillon n sur la qualité de cette approximation?

Quand le théorème de central limite ne s'applique pas

La densité de la loi de Cauchy $\mathcal{C}(\theta)$ est $f(x, \theta) = \frac{1}{\pi} \frac{1}{1 + (x - \theta)^2}$ pour tout $x \in \mathbb{R}$.

1. Simuler un échantillon de taille $n = 20$ d'une loi de $\mathcal{C}(2)$ et calculer la moyenne empirique $\bar{X}_n$.
2. Faites varier la taille de l'échantillon $n = 20, 100, 1000$ et 10000. Qu'en déduire ?
3. Expliquer ce comportement. On se rappellera notamment que la fonction caractéristique s'écrit $\phi_\theta(t) = \exp(i\theta t - |t|)$.
4. Quelle est la médiane d'une loi de Cauchy $\mathcal{C}(\theta)$?
5. En déduire un estimateur de θ et evaluer la performance de ce estimateur pour $n = 20, 100, 1000$.