TP Statistiques 2

Nawal Belaid, Nicolas Brunel et Salim Amoukou

28 Fevrier 2020

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 $\sigma_{n,i}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X}_n)^2$ 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 $\sigma_{n,i}^2$. 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_n^2. 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.

Moyenne et phénomène de concentration

Nous allons montrer que la moyenne d'une variable aléatoire est un résumé déterministe d'une v.a., dont la qualité est contrÃ´lée par la variance.

1. Rappeler l'inégalité de Bienaymé-Chebychef dans les cas Gaussien et Poisson.
2. Estimer par Monte Carlo les probabilités de déviation d'une variable aléatoire de sa moyenne.
   (a) Exprimer $P(|X - \mu| \geq \delta)$ comme l'espérance d'une certaine variable aléatoire Z.
   (b) Simuler un échantillon de taille N $Z_1, Z_2, \dots, Z_N$ de meme loi que Z (dans le cas Gaussien, Pareto et Poisson) - on prendra N grand. Déterminer une estimation de $P(|X - \mu| \geq \delta)$. Pouvez vous déterminer la précision de cette estimation?

(c) Comparer avec les bornes obtenues par Bienaymé Chebychev pour plusieurs δ. Faites varier σ.

(d) Dans le cas Gaussien et Poisson, comparer les estimations Monte-Carlo de $P(X - \mu \geq \delta)$ avec les bornes données par les inégalités Chernoff pour plusieurs δ et σ (cf. cours).

3. Simuler un échantillon de taille $n = 20$ pour les lois de Gauss et de Poisson *(choisir σ, λ approprié)*

   (a) Calculer les bornes de Chernoff dans le cas échantillon pour $\bar{X}_n$. Faites varier $n = 20$, 100, 1000.

   (b) En déduire un estimateur de μ et λ respectivement.

4. Simuler un échantillon de taille $n = 20$ d'une loi de Cauchy $\mathcal{C}(\theta)$ de densit? $f(x, \theta) = \frac{1}{\pi} \frac{1}{1+(x-\theta)^2}$.

   (a) Calculer la moyenne empirique $\bar{X}_n$. Faites varier la taille de l'échantillon $n = 20$, 100, 1000, 10000. Qu'en déduire ?

   (b) Expliquer ce comportement. On se rappellera notamment que la fonction caractéristique s'écrit $\phi_\theta(t) = \exp(i\theta t - |t|)$.

   (c) Quelle est la médiane d'une loi de Cauchy $\mathcal{C}(\theta)$? En déduire un estimateur de θ pour $n = 20, 100, 1000$.