

Travail Pratique 2 : Méthode de Monte-Carlo

LUC WACHTER

28 mai 2019

1 Introduction

Dans ce travail pratique, nous nous intéressons au calcul de la distance moyenne entre deux points du carré unité. Il s'agit donc de calculer l'espérance de la distance euclidienne séparant deux points indépendants et identiquement distribués dans le carré unité $[0; 1] \times [0; 1]$.

Si on note $P(x_1; y_1)$ et $Q(x_2; y_2)$ deux points du carré unité et D la distance euclidienne qui les sépare, l'espérance de cette distance est

$$E(D) = \int_0^1 \int_0^1 \int_0^1 \int_0^1 \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} dy_2 dy_1 dx_2 dx_1.$$

C'est cette espérance que l'on veut calculer.

Pour ce faire, nous allons réaliser une simulation de Monte-Carlo afin de générer des estimateurs \hat{D} de $E(D)$ à partir d'échantillons de populations générées aléatoirement.

Ce document a pour but de documenter les choix faits durant la conception du programme, mais surtout de présenter les résultats obtenus, de les interpréter et de les discuter.

Symbole mathématique	Symbole dans le code
N_{init}	<code>initialNumberOfRuns</code>
N_{add}	<code>additionalNumberOfRuns</code>
N	<code>estimatedNumberOfRunsNeeded</code>
Δ_{max}	<code>maxHalfWidth</code>
s	<code>standardDeviation</code>

TABLE 1 – Correspondance des symboles

2 Correspondance des symboles

Les symboles mathématiques proposés dans la donnée du travail ne sont bien sûr pas propices à une utilisation en Java. Le tableau 1 permet donc de faire le lien entre les symboles utilisés dans ce document et ceux utilisés dans le programme.

3 Choix du nombre d'expériences à effectuer

Après avoir lancé la simulation pour exécuter N_{init} expériences, il nous faut estimer le nombre d'expériences supplémentaires nécessaires pour arriver à un intervalle de confiance à 95% dont la demi-largeur ne dépasse pas Δ_{max} .

Ce nombre d'expériences n peut être estimé en utilisant l'équation pour la largeur de l'intervalle de confiance :

$$\Delta_{I_c} = 2 \cdot z_{1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$$

Adaptons cette équation pour notre cas pratique en remplaçant la largeur de l'intervalle de confiance Δ_{I_c} par la demi-largeur maximale Δ_{max} (valeur fixée par le code appelant). Ce changement nous permet de faire fi du facteur deux du côté droit de l'équation, puisque nous nous intéressons à la demi-largeur et non à la largeur complète.

$$\Delta_{max} = z_{1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$$

Il nous suffit alors d'isoler n pour déterminer la formule à utiliser dans notre programme.

$$\Delta_{max} = z_{1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$$

$$\Rightarrow \sqrt{n} = \frac{z_{1-\frac{\alpha}{2}} \cdot s}{\Delta_{I_c}}$$

$$\Rightarrow n = \frac{z_{1-\frac{\alpha}{2}}^2 \cdot s^2}{\Delta_{I_c}^2}$$

En utilisant 1.959964 comme quantile de la loi normale et avec l'estimateur s de l'écart-type calculé empiriquement à l'aide de nos expériences initiales, il nous est donc possible d'estimer le nombre d'expériences nécessaires pour arriver à la précision voulue.

4 Approche utilisée

Afin d'obtenir les résultats qui seront discutés dans la section suivantes, 1500 simulations ont été lancées en séquence. Pour chacune d'entre elles, les paramètres suivants ont été fixés :

$$\Delta_{max} = 5 \cdot 10^{-5}, \quad N_{init} = 10^6, \quad N_{add} = 10^5$$

En outre, la séquence de nombres créés par le générateur pseudo-aléatoire dans le programme a été fixée à l'aide de la graine 20190528.

Les valeurs ayant été enregistrées à la fin de chaque simulation sont le nombre d'observations effectuées, la valeur de l'estimateur calculé (la moyenne), l'écart-type, la variance et la demi-largeur de l'intervalle de confiance à 95%.

5 Résultats

5.1 Valeur exacte de l'espérance

Afin d'analyser les résultats de la simulation, la valeur exacte de $E(D)$ nous est donnée.

$$\begin{aligned} E(D) &= \int_0^1 \int_0^1 \int_0^1 \int_0^1 \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} dy_2 dy_1 dx_2 dx_1 \\ &= \frac{2 + \sqrt{2} + 5 \ln \sqrt{2} + 1}{15} \\ &\approx 0,5214054 \end{aligned}$$

Dans toutes les figures suivantes, cette valeur exacte sera mise en évidence par une droite de couleur verte.

5.2 Distribution des estimateurs ponctuels \hat{D}

En observant l'histogramme de la figure 1, nous remarquons la forme caractéristique de la loi normale. L'apparition d'une telle distribution n'est pas surprenante, puisque nos \hat{D} sont calculés comme la moyenne des résultats d'expériences aléatoires indépendantes, et que le théorème central limite nous permet d'établir la convergence de la loi de ces expériences vers la loi normale.

Il y donc une tendance vers une distribution symétrique autour de la valeur exacte de l'espérance recherchée. On observe plus facilement, à l'aide de la boîte à moustache de la figure 2 que la moyenne (représentée par un segment noir épais au centre de la boîte) se rapproche de la valeur exacte, et que les observations sont distribuées symétriquement et uniformément autour de cette valeur.

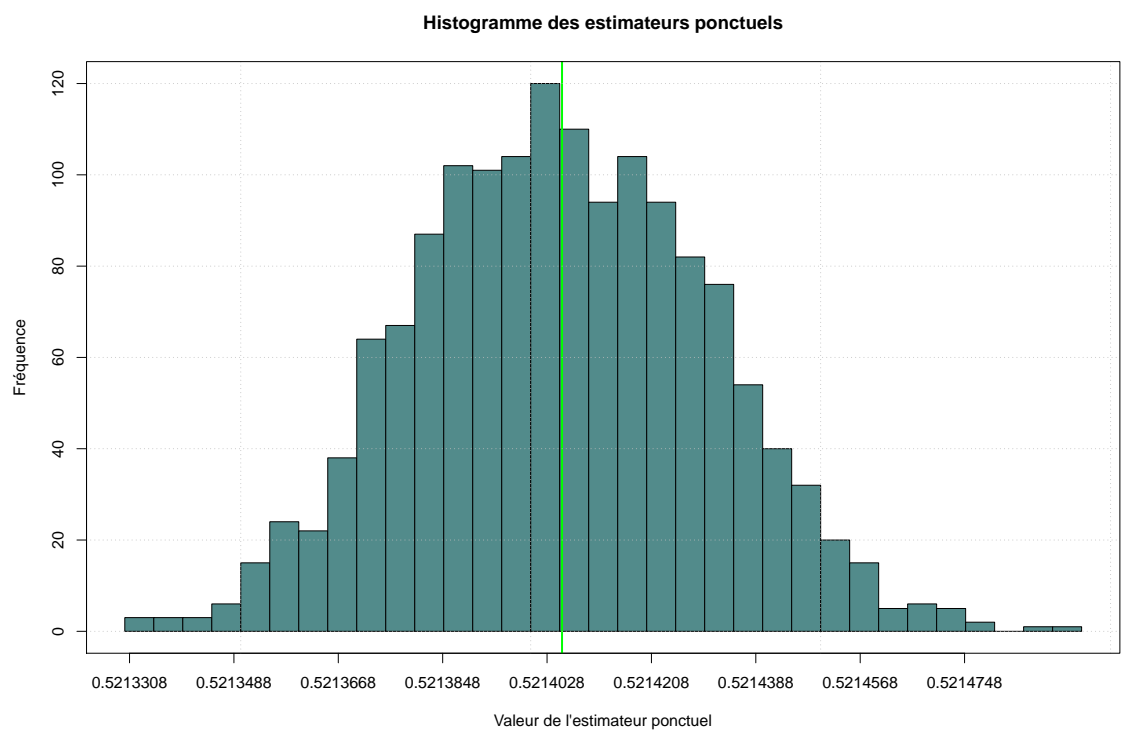


FIGURE 1 – Histogramme des estimateurs \hat{D}

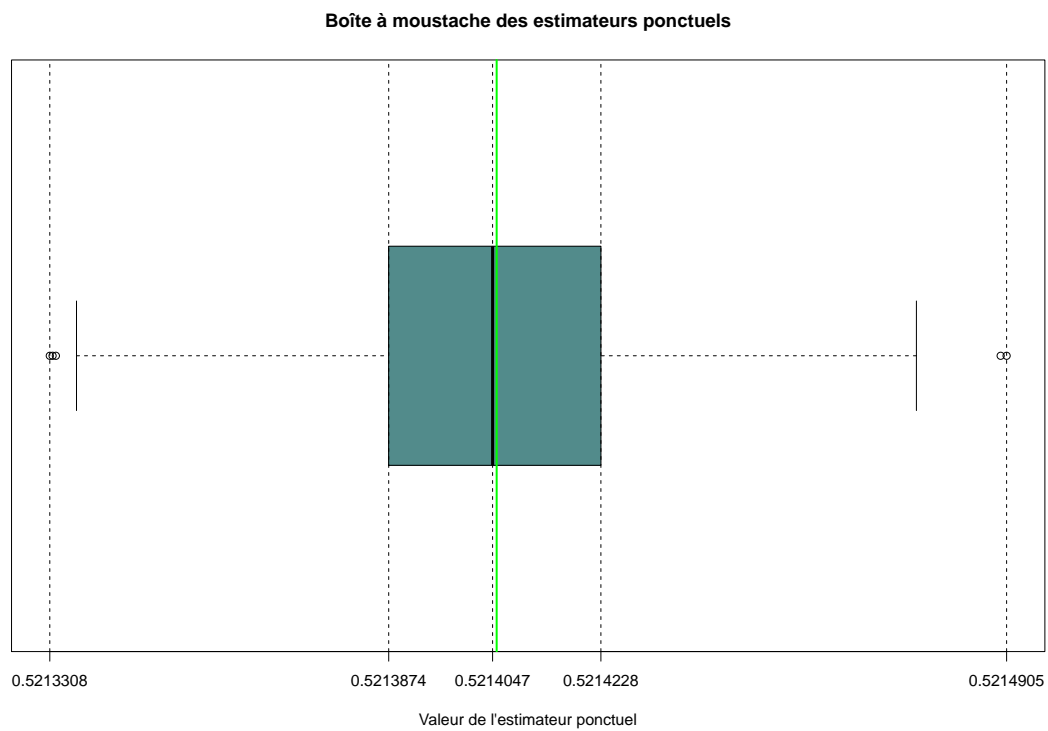


FIGURE 2 – Boîte à moustache des estimateurs \hat{D}

5.3 Couverture empirique des intervalles de confiance

À l'aide du graphique des intervalles de confiance de la figure 3, nous pouvons visualiser les intervalles de confiance des 100 premières simulations effectuées. Sur cette figure, les points représentent la valeur de l'estimateur \hat{D} , tandis que les barres montrent l'I.C. pour chaque estimateur. Les éléments en rouge sont ceux dont l'I.C. ne contient pas la valeur exacte de l'espérance recherchée.

On observe une distribution cohérente par rapport à l'histogramme de la figure 1. On peut également remarquer que 3 estimateurs sont si éloignés que leur I.C. ne contient pas la valeur exacte. Cela signifie que le seuil de confiance empirique de cet échantillon des données s'élève à 97%.

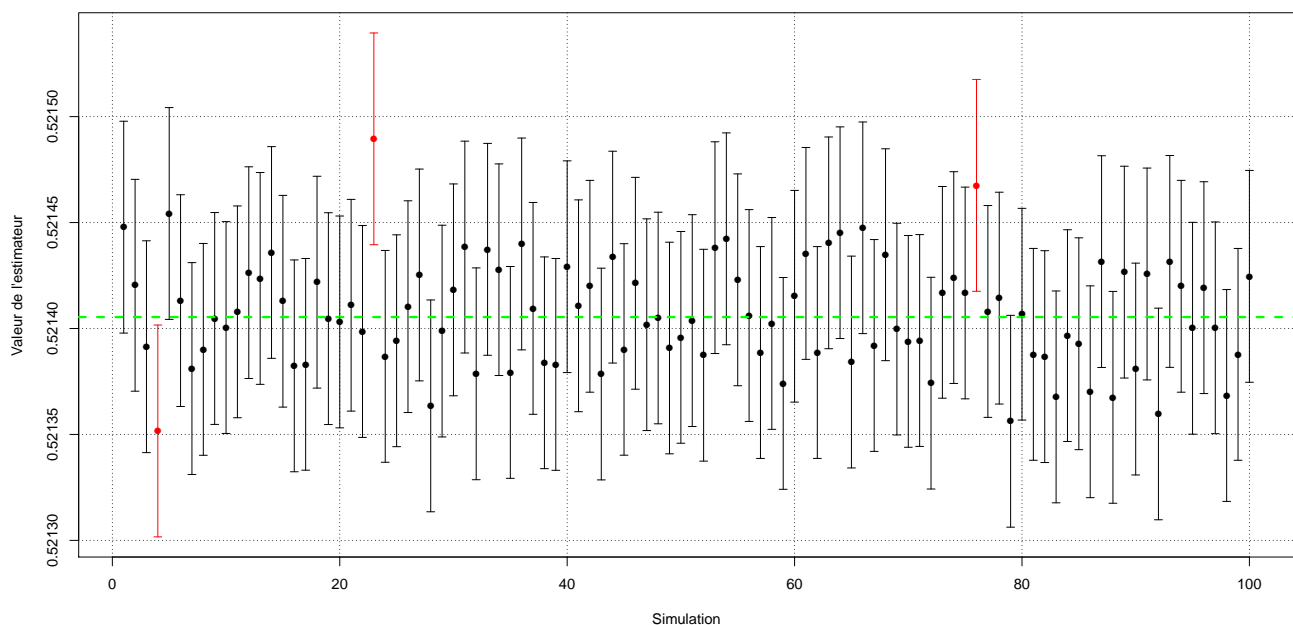


FIGURE 3 – 100 intervalles de confiance au seuil de 95%

Augmentons maintenant la portée de notre analyse en regardant la totalité des simulations, à l'aide du nuage de points de la figure 4. Ici, les deux droites rouges valent respectivement $E(D) \pm \Delta_{max}$. On compte alors 62 estimateurs en dehors de cet intervalle (4.133% des estimateurs générés), ce qui donne un seuil de confiance empirique de 95.867%.

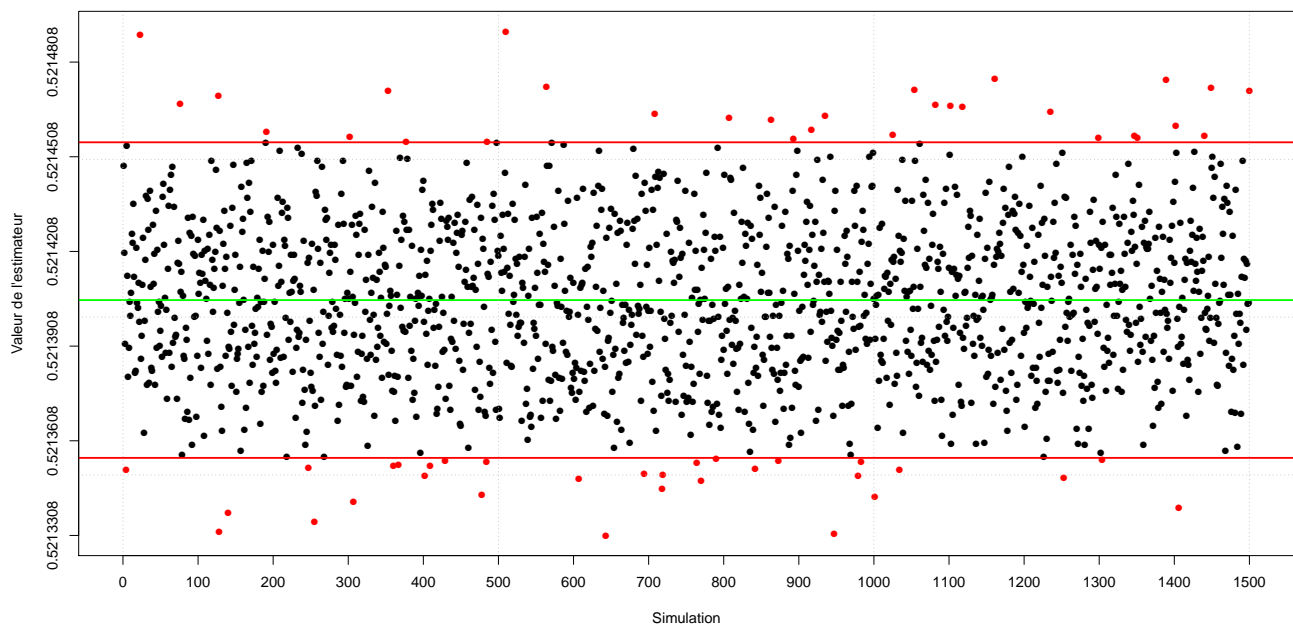


FIGURE 4 – Nuage de points

6 Conclusion

Au fur et à mesure de ce travail pratique, nous nous sommes efforcés d'estimer la valeur de $E(D)$ empiriquement. Il est clair cependant, que ce n'était pas le vrai but de ce travail. En effet, il est plus important de nous sensibiliser à la juste utilisation des méthodes approchées.

Grâce à ce travail, nous avons pu remarquer empiriquement qu'une simulation ne vaut pas grand chose si l'on ne quantifie pas l'erreur à l'aide de l'intervalle de confiance. En outre, même avec l'assurance d'atteindre un intervalle de confiance précis, il est toujours possible de se retrouver bien plus éloigné de la valeur exacte qu'on le voudrait en pratique. Il faut donc toujours avoir un regard critique sur les résultats des méthodes approchées.

Dans ce travail, cependant, nos observations empiriques nous ont permis de confirmer les intuitions acquises durant le cours. Nous avons vu qu'il est difficile de réduire la largeur de l'intervalle de confiance et nous avons confirmé par des observations les valeurs des intervalles de confiance, et la distribution des estimateurs.