Simulation en Biologie

Bruno Toupance

Université de Paris

SEP 2021

Modélisation

- Modèles déterministes : Analyses numériques
  - Récurrences
  - Équations différentielles

- Modèles stochastiques : Simulations
  - Génération de nombres aléatoires
  - Méthodes de Monte-Carlo

Méthodes de Monte-Carlo

Casino de Monte-Carlo

- Étude du comportement d'un système complexe
- Étude empirique des fluctuations d'échantillonnage
- Détermination des propriétés probabilistes d'une variable aléatoire de loi inconnue (bootstrap)
- Validation d'un modèle probabiliste
- Calcul d'intégrales, d'espérances, de vraisemblances
- Statistiques bayésiennes : Monte-Carlo Markov Chain (MCMC), Approximate Bayesian Computation (ABC)

Historique : Aiguille de Buffon

Georges-Louis Leclerc, comte de Buffon (1707-1788)

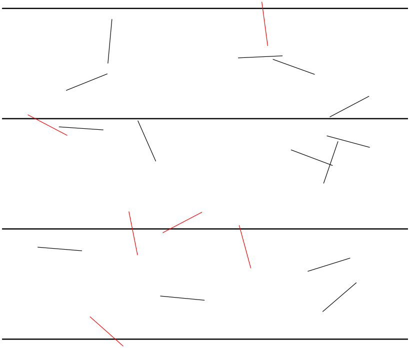
- Détermination expérimentale de la valeur de π
- Dispositif expérimental :
  - Lames de parquet de largeur D
  - Aiguille de longueur L
  - On lance l'aiguille n fois
  - On compte le nombre de lancers n_S pour lesquels l'aiguille chevauche deux lames
  - Estimation de la probabilité que l'aiguille chevauche deux lames :

$$\hat{p} = \frac{n_S}{n}$$

- On peut par ailleurs montrer que :

$$p = \frac{2}{\pi}\frac{L}{D} \qquad \text{donc} \qquad \pi \simeq 2\frac{1}{\hat{p}}\frac{L}{D}$$

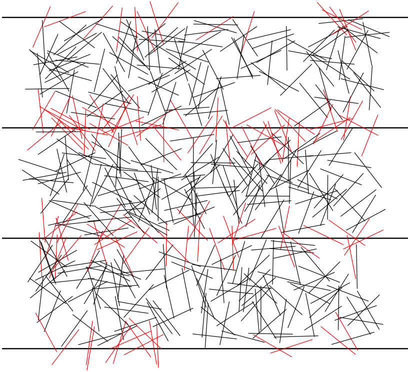
Aiguille de Buffon - Simulation 1



- Nombre de lancers : $n = 20$
- Rapport $\frac{L}{D} = 0.4$
- Proportion de succès : $p = \frac{6}{20} = 0.3$
- Estimation :

$$\hat{\pi} = 2.666667$$

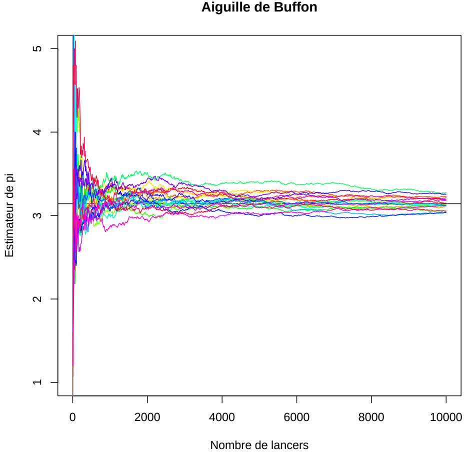
Aiguille de Buffon - Simulation 2



- Nombre de lancers : $n = 500$
- Rapport $\frac{L}{D} = 0.4$
- Proportion de succès :
  $p = \frac{123}{500} = 0.246$
- Estimation :

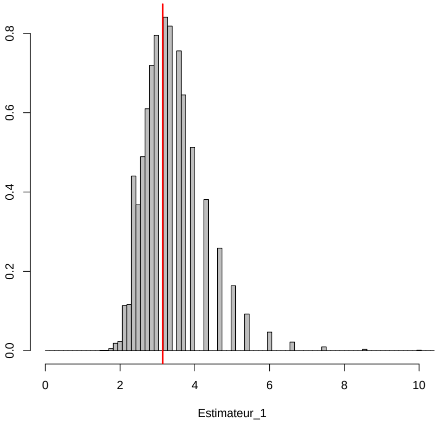
$$\hat{\pi} = 3.252033$$

Qualité de l'estimateur en fonction du nombre de lancers



- Rapport $\frac{L}{D} = 0.4$
- Nombre de lancers : $n = 1, 2, ..., 10000$
- 20 répétitions indépendantes

Aiguille de Buffon - Distribution de l'estimateur



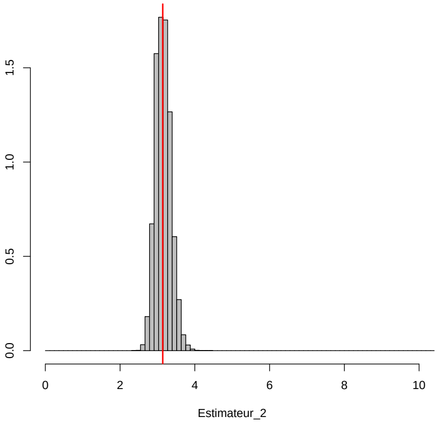
- Rapport $\frac{L}{D} = 0.3$
- Nombre de lancers : $n = 100$
- 1 000 000 répétitions indépendantes
- Moyenne :

$$m = 3.289791$$

- Variance :

$$s^2 = 0.5749353$$

Aiguille de Buffon - Distribution de l'estimateur



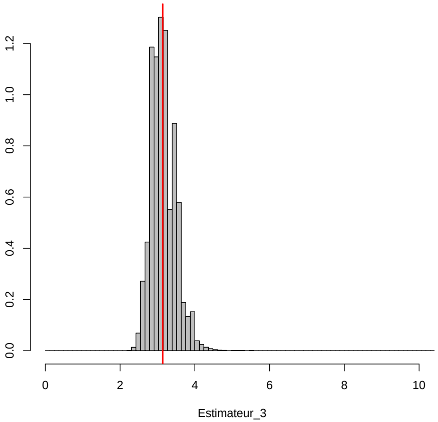
- Rapport $\frac{L}{D} = 0.3$
- Nombre de lancers : $n = 1000$
- 1 000 000 répétitions indépendantes
- Moyenne :

$$m = 3.154735$$

- Variance :

$$s^2 = 0.04304659$$

Aiguille de Buffon - Distribution de l'estimateur



- Rapport $\frac{L}{D} = 0.8$
- Nombre de lancers : $n = 100$
- 1 000 000 répétitions indépendantes
- Moyenne :

$$m = 3.172636$$

- Variance :

$$s^2 = 0.1033642$$

Soit une épreuve aléatoire définie par l'ensemble Ω des événements élémentaires (c'est-à-dire des résultats possibles). Soit $\mathcal{A}$ l'ensemble des événements. Une **probabilité** $\mathbb{P}$ est une application de l'ensemble des événements $\mathcal{A}$ dans l'ensemble des réels $\mathbb{R}$:

$$\begin{array}{rccc} \mathbb{P}: & \mathcal{A} & \rightarrow & \mathbb{R} \\ & A & \mapsto & \mathbb{P}[A] \end{array}$$

vérifiant :

1. $0 \leq \mathbb{P}[A] \leq 1$
2. $\mathbb{P}[\Omega] = 1$
3. Si $(A_1, A_2, ..., A_n)$ sont incompatibles deux à deux, alors :

$$\mathbb{P}[A_1 \cup A_2 \cup ... \cup A_n] = \mathbb{P}[A_1] + \mathbb{P}[A_2] + ... + \mathbb{P}[A_n]$$

- Événement impossible :

$$\mathbb{P}[\emptyset] = 0$$

- Complémentaire :

$$\mathbb{P}\left[\bar{A}\right] = 1 - \mathbb{P}[A] \qquad \text{et} \qquad \mathbb{P}[A] = 1 - \mathbb{P}\left[\bar{A}\right]$$

- Si $(B_1, B_2, ..., B_n)$ forment une partition de Ω, alors :

$$\mathbb{P}[A] = \mathbb{P}[A \cap B_1] + \mathbb{P}[A \cap B_2] + ... + \mathbb{P}[A \cap B_n]$$

- Si A et B sont deux événements quelconques :

$$\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B]$$

Si B est un événement pour lequel $\mathbb{P}[B] \neq 0$, on peut définir une nouvelle probabilité $\mathbb{P}_B$, appelée "**probabilité conditionnelle sachant** B" définie par :

$$\mathbb{P}_B[A] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}$$

Cette probabilité est souvent notée $\mathbb{P}[A \mid B]$ bien que "$A \mid B$" ne soit pas un événement.

La **probabilité conditionnelle** $\mathbb{P}[A \mid B]$ s'écrit en fonction de la **probabilité conjointe** $\mathbb{P}[A \cap B]$:

$$\mathbb{P}[A \mid B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}$$

Inversement, une probabilité conjointe peut s'écrire en fonction d'une probabilité conditionnelle (si cette probabilité peut être définie) :

$$\begin{aligned} \mathbb{P}[A \cap B] &= \mathbb{P}[B]\ \mathbb{P}[A \mid B] \qquad \text{si} \qquad \mathbb{P}[B] \neq 0 \\ \mathbb{P}[A \cap B] &= \mathbb{P}[A]\ \mathbb{P}[B \mid A] \qquad \text{si} \qquad \mathbb{P}[A] \neq 0 \end{aligned}$$

Si $(B_1, B_2, ..., B_n)$ forment une partition de Ω, alors :

$$\mathbb{P}[A] = \mathbb{P}[B_1]\,\mathbb{P}[A \mid B_1] + \mathbb{P}[B_2]\,\mathbb{P}[A \mid B_2] + ... + \mathbb{P}[B_n]\,\mathbb{P}[A \mid B_n]$$

En particulier, si B est un événement tel que $0 < \mathbb{P}[B] < 1$, alors :

$$\mathbb{P}[A] = \mathbb{P}[B]\,\mathbb{P}[A \mid B] + \mathbb{P}\left[\overline{B}\right]\mathbb{P}\left[A \mid \overline{B}\right]$$

Rappels de Probabilité - Indépendance

Deux événements A et B sont **indépendants** en probabilité si et seulement si :

$$\mathbb{P}[A \cap B] = \mathbb{P}[A] \times \mathbb{P}[B]$$

L'indépendance signifie que la réalisation d'un des événements ne modifie pas la probabilité de réalisation de l'autre événement :

$$\mathbb{P}[A \mid B] = \mathbb{P}[A] \qquad \text{ou} \qquad \mathbb{P}[B \mid A] = \mathbb{P}[B]$$

Rappels de Probabilité - Événements mutuellement indépendants

On dira que n événements $(A_1, A_2, ..., A_n)$ sont **mutuellement indépendants** si la probabilité conjointe de réalisation simultanée d'un nombre quelconque k de ces n événements est égale au produit des k probabilités :

$$\text{Quel que soit } k \text{ tel que } 2 \leq k \leq n,$$

$$\mathbb{P}[A_1 \cap A_2 \cap ... \cap A_k] = \mathbb{P}[A_1] \times \mathbb{P}[A_2] \times ... \times \mathbb{P}[A_k]$$

Dans le cas où les événements $(A_1, A_2, ..., A_n)$ ne sont pas mutuellement indépendants, la probabilité conjointe s'écrire :

$$\mathbb{P}[A_1 \cap A_2 \cap ... \cap A_n] =$$
$$\mathbb{P}[A_1] \times \mathbb{P}[A_2 \mid A_1] \times \mathbb{P}[A_3 \mid A_1 \cap A_2] \times ... \times \mathbb{P}[A_n \mid A_1 \cap A_2 \cap ... \cap A_{n-1}]$$

Cette formule n'est qu'une généralisation de :

$$\mathbb{P}[A \cap B] = \mathbb{P}[A] \times \mathbb{P}[B \mid A]$$

Succession d'événements/états $(A_1, A_2, ..., A_i, ...)$:
La loi conditionnelle de X_{n+1} sachant les états passés ne dépend que de X_n

$$\mathbb{P}\left[X_{n+1} = x \mid X_1 = x_1 \cap X_2 = x_2 \cap \ldots \cap X_n = x_n\right]$$
$$= \mathbb{P}\left[X_{n+1} = x \mid X_n = x_n\right]$$

Donc :

$$\mathbb{P}\left[A_1 \cap A_2 \cap \ldots \cap A_i \cap \ldots\right] =$$
$$\mathbb{P}\left[A_1\right] \times \mathbb{P}\left[A_2 \mid A_1\right] \times \mathbb{P}\left[A_3 \mid A_2\right] \times \ldots \times \mathbb{P}\left[A_i \mid A_{i-1}\right] \times \ldots$$

Le théorème de Bayes permet de calculer une probabilité *a posteriori* $\mathbb{P}[A \mid B]$ en fonction d'une probabilité *a priori* $\mathbb{P}[A]$:

$$\mathbb{P}[A \mid B] = \mathbb{P}[A] \times \frac{\mathbb{P}[B \mid A]}{\mathbb{P}[B]}$$

$$\mathbb{P}[A \mid B] = \mathbb{P}[A] \times \frac{\mathbb{P}[B \mid A]}{\mathbb{P}[A] \times \mathbb{P}[B \mid A] + \mathbb{P}\left[\bar{A}\right] \times \mathbb{P}\left[B \mid \bar{A}\right]}$$

Soit une épreuve aléatoire définie par l'ensemble Ω des événements élémentaires (c'est-à-dire des résultats possibles) et par une probabilité $\mathbb{P}$. Soit une application X qui associe à chaque événement élémentaire ω, une valeur réelle x :

$$X : \begin{array}{ccc} \Omega & \to & \mathbb{R} \\ \omega & \mapsto & x = X(\omega) \end{array}$$

On appellera X **variable aléatoire réelle**.

L'ensemble des valeurs que peut prendre la variable aléatoire X, c'est-à-dire $X(\Omega)$, est appelé **support**.

Une variable aléatoire réelle X est complètement définie par sa **fonction de répartition** F :

$$F(x) = \mathbb{P}\left[X \leq x\right] \qquad \text{pour} \qquad x \in \mathbb{R}$$

Toute fonction de répartition vérifie :

- $0 \leq F(x) \leq 1$ pour tout $x \in \mathbb{R}$
- F est une fonction croissante
- $\lim_{x \to -\infty} F(x) = 0$
- $\lim_{x \to +\infty} F(x) = 1$
- F est continue à droite

Une variable aléatoire réelle est **discrète** si son support $X(\Omega)$ est **fini** ou **dénombrable**.

Une variable aléatoire réelle est (absolument) **continue** si sa fonction de répartition F est **continue et dérivable** sur $\mathbb{R}$.

La **loi de probabilité** d'une variable aléatoire réelle discrète X est définie par l'ensemble des doublets (x_i, p_i) (pour $i = 1, 2, ..., n$ ou ∞) tel que :

- x_i est l'une des valeurs possibles de X et
- p_i est la probabilité que X soit égale à x_i, $\mathbb{P}[X = x_i] = p_i$, c'est-à-dire la probabilité d'observer les événements élémentaires ω associés à la valeur x_i.

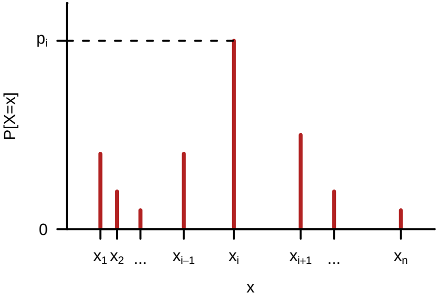
L'ensemble des probabilités p_i doit vérifier les conditions suivantes :

$$0 \leq p_i \leq 1 \qquad \text{et} \qquad \sum_{i=1}^{n \text{ ou } \infty} p_i = 1$$

Variable aléatoire discrète - Loi de probabilité

x_i	x_1	x_2	$\dots$	x_{i-1}	x_i	x_{i+1}	$\dots$	x_n
$\mathbb{P}[X = x_i]$	p_1	p_2	$\dots$	p_{i-1}	p_i	p_{i+1}	$\dots$	p_n

Diagramme en bâtons :



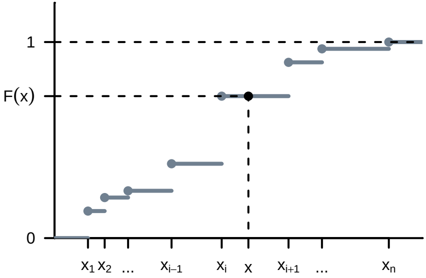
Soit X une variable aléatoire réelle discrète de support $X(\Omega) = \{x_1, x_2, x_3, ..., x_i, ...\}$ (éventuellement fini) où les valeurs x_i sont triées par ordre croissant. La **fonction de répartition** F de X pour une valeur x_i du support est :

$$F(x_i) \quad = \quad \mathbb{P}[X = x_1] + \mathbb{P}[X = x_2] + ... + \mathbb{P}[X = x_i]$$

Pour toute valeur réelle x comprise entre deux valeurs successives x_i et x_{i+1} du support $X(\Omega)$, la fonction de répartition est constante :

$$F(x) = F(x_i) \qquad \text{pour} \qquad x_i < x < x_{i+1}$$

La représentation graphique de la fonction de répartition $F(x)$ d'une variable aléatoire discrète X est en **forme d'escalier** :



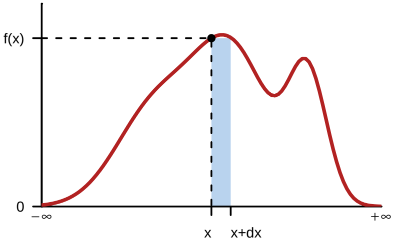
Discontinuité :

$$\mathbb{P}[X = x_i] = F(x_i) - F_g(x_i)$$

où $F_g(x_i)$ est la limite à gauche de F en x_i

La **densité de probabilité** (d.d.p.) d'une variable aléatoire réelle continue X est la fonction f définie par :

$$\mathbb{P}\left[x < X \leq x + dx\right] = f(x)\ dx \qquad \text{pour} \qquad x \in \mathbb{R}$$

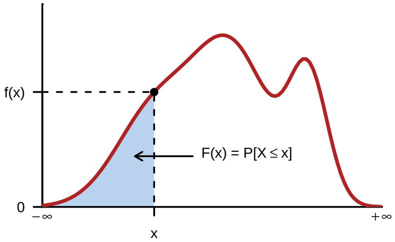


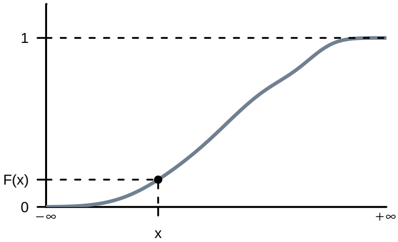
Toute fonction de densité de probabilité doit vérifier :

$$f(x) \geq 0 \text{ pour tout } x \in \mathbb{R} \qquad \text{et} \qquad \int_{-\infty}^{+\infty} f(u)\ du = 1$$

Fonction de répartition F de X :

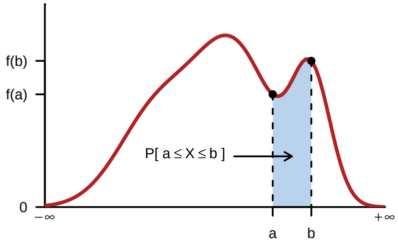
$$F(x) = \mathbb{P}[X \leq x] = \int_{-\infty}^{x} f(u)\, du \qquad \text{pour} \qquad x \in \mathbb{R}$$

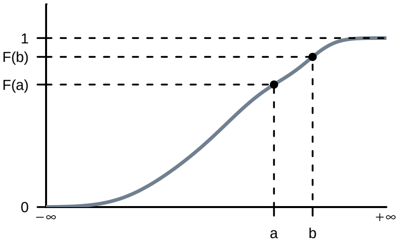




Probabilité pour une valeur x :

$$\mathbb{P}[X = x] = F(x) - F_g(x) = 0$$





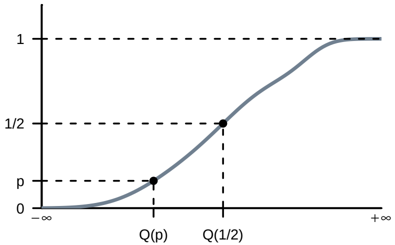
Calcul de probabilité :

$$\mathbb{P}\left[a \leq X \leq b\right] = \int_a^b f(u)\, du = \int_{-\infty}^b f(u)\, du - \int_{-\infty}^a f(u)\, du = F(b) - F(a)$$

$$\mathbb{P}\left[a < X < b\right] = \mathbb{P}\left[a \leq X < b\right] = \mathbb{P}\left[a < X \leq b\right] = \mathbb{P}\left[a \leq X \leq b\right]$$

Fonction quantile

La fonction de quantile Q de X est la **fonction réciproque** F^{-1} de la fonction de répartition.



Pour trouver la fonction quantile, il suffit de résoudre pour $F^{-1}(x)$:

$$F(F^{-1}(x)) = x \qquad \text{pour} \qquad 0 < x < 1$$

Variable aléatoire réelle - Espérance

L'**espérance** de X est le nombre réel noté $\mathbb{E}[X]$ et défini par :

- X discrète :

$$\mathbb{E}[X] = \sum_{i=1}^{n \text{ ou } \infty} x_i \; p_i$$

- X continue :

$$\mathbb{E}[X] = \int_{-\infty}^{+\infty} u \; f(u) \; du$$

La **variance** de X est le nombre réel noté $\mathbb{V}[X]$ et défini par :

- X discrète :

$$\mathbb{V}[X] = \sum_{i=1}^{n \text{ ou } \infty} (x_i - \mathbb{E}[X])^2 \, p_i$$

- X continue :

$$\mathbb{V}[X] = \int_{-\infty}^{+\infty} (u - \mathbb{E}[X])^2 \, f(u) \, du$$

On montre que :

$$\mathbb{V}[X] = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2$$

avec :

- X discrète :

$$\mathbb{E}\left[X^2\right] = \sum_{i=1}^{n \text{ ou } \infty} x_i^2 \; p_i$$

- X continue :

$$\mathbb{E}\left[X^2\right] = \int_{-\infty}^{+\infty} u^2 \; f(u) \; du$$

Variable aléatoire X quelconque. On crée la variable Y :

$$Y = \phi(X)$$

Comment caractériser la variable Y ?

- Support ?
- Loi ?
- Espérance ?
- Variance ?

- Cas général :

  Pas de résultat...

- Si X est continue et si la fonction ϕ est continue et strictement monotone, alors g, la fonction de densité de Y est définie par :

$$g(y) = \left| \frac{d\phi^{-1}}{dy}(y) \right| \ f(\phi^{-1}(y)) \qquad \text{pour} \qquad y \in \phi(X(\Omega))$$

  Calculer explicitement la densité de Y nécessite donc de pouvoir exprimer la fonction réciproque ϕ^{-1} et pouvoir dériver cette dernière.

Opération linéaire

Variable aléatoire X quelconque
Deux nombres réels a et b quelconques ($b \neq 0$)

Transformation linéaire de X

$$Y = a + b\,X$$

Densité de Y (cas continu) :

$$g(y) = \frac{1}{|b|}\, f\left(\frac{y-a}{b}\right) \qquad \text{pour} \qquad y \in \phi(\mathbb{R})$$

La forme de la loi est conservée par une transformation linéaire (avec une éventuelle symétrie si $b < 0$)

Espérance Y : Théorème du transport

$$\mathbb{E}\left[\phi(X)\right] = \sum_{x_i} \phi(x_i) \times \mathbb{P}\left[X = x_i\right] \qquad \text{(Cas discret)}$$

$$\mathbb{E}\left[\phi(X)\right] = \int_{-\infty}^{+\infty} \phi(u) \times f(u)\, du \qquad \text{(Cas continu)}$$

Estimation d'une espérance (cas continu)

Espérance X :

$$\mathbb{E}[X] = \int_{-\infty}^{+\infty} u \times f(u)\, du$$

On tire un n-échantillon dans la loi de X :

$$\{x_1, x_2, ..., x_n\}$$

Estimateur de $\mathbb{E}[X]$:

$$\mathbb{E}[X] \simeq \frac{1}{n} \sum_i x_i$$

Espérance $y = \phi(X)$:

$$\mathbb{E}[Y] = \int_{-\infty}^{+\infty} \phi(u) \times f(u)\, du$$

On tire un n-échantillon dans la loi de X :

$$\{x_1, x_2, ..., x_n\}$$

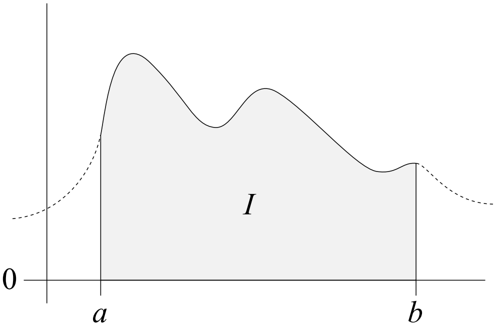
On applique la fonction ϕ sur le n-échantillon :

$$\{\phi(x_1), \phi(x_2), ..., \phi(x_n)\}$$

Estimateur de $\mathbb{E}[Y]$:

$$\mathbb{E}[Y] \simeq \frac{1}{n} \sum_i \phi(x_i)$$

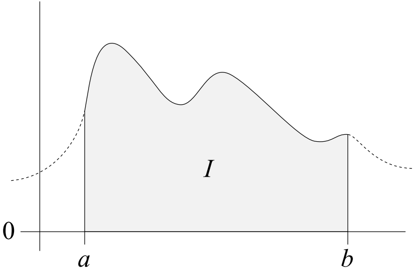
Problème général :



$$I = \int_a^b g(x)\, dx$$

Calcul d'intégrales

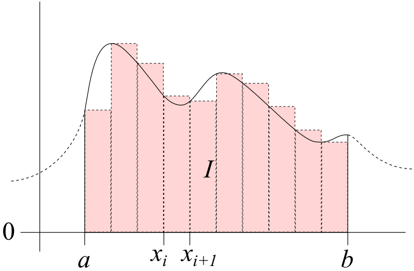
Méthodes classiques d'intégrations numériques :



$$I \simeq \sum_{i=1}^{n} \omega_i \; g(x_i)$$

- rectangles
- trapèzes
- Simpson

Méthodes des rectangles :

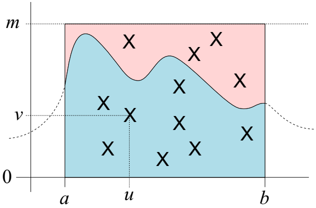


$$I \simeq I_1 = \sum_{i=1}^{n} \omega_i \, g(x_i)$$

avec

$$\omega_i = \frac{b-a}{n} \qquad \text{et} \qquad x_i = a + (i-1)\frac{b-a}{n}$$

Méthode de Monte-Carlo par « tirage noir ou blanc »



- m majorant de la fonction g
- U suit une loi uniforme sur $[a, b]$
- V suit une loi uniforme sur $[0, m]$

$$I \simeq m\,(b-a)\,\frac{n_S}{n}$$

avec

n_S Nombre de succès

Méthode de Monte-Carlo par « tirage noir ou blanc » - Preuve :
Rappels :

- Loi de U :

$$f_U(u) = \frac{1}{b-a} \qquad \text{pour} \qquad u \in [a,\ b]$$

$$F_U(u) = \frac{u-a}{b-a} \qquad \text{pour} \qquad u \in [a,\ b]$$

- Loi de V :

$$f_V(v) = \frac{1}{m} \qquad \text{pour} \qquad v \in [0,\ m]$$

$$F_V(v) = \frac{v}{m} \qquad \text{pour} \qquad v \in [0,\ m]$$

Méthode de Monte-Carlo par « tirage noir ou blanc » - Preuve :

$$\begin{aligned}
\mathbb{P}\left[V \leq g(U)\right] &= \int_{u=a}^{b} \mathbb{P}\left[V \leq g(u) \mid U = u\right] \; f_U(u)\, du \\
&= \int_{u=a}^{b} \mathbb{P}\left[V \leq g(u)\right] \; f_U(u)\, du \\
&= \frac{1}{b-a} \int_{u=a}^{b} F_V(g(u))\, du \\
&= \frac{1}{m(b-a)} \int_{u=a}^{b} g(u)\, du \\
&= \frac{1}{m(b-a)}\; I
\end{aligned}$$

Méthode de Monte-Carlo par « échantillonnage »
Si la fonction g peut s'écrire :

$$g(x) = h(x)\ f_X(x) \qquad \text{pour} \qquad x \in [a,\ b]$$

avec f_X densité d'une variable aléatoire X de support $[a,\ b]$

Alors le problème d'intégration revient à un calcul d'espérance :

$$I = \int_{u=a}^{b} g(x)\ dx = \int_{u=a}^{b} h(x)\ f_X(x)\ dx$$

$$I = \mathbb{E}\left[h(X)\right]$$

L'intégration peut donc se généraliser à $\mathbb{R}$

Calcul d'intégrales

$$I = \mathbb{E}\left[h(X)\right]$$

Donc : si $(x_1, x_2, ..., x_i, ..., x_n)$ est un n-échantillon de la loi de X

$$I \simeq \frac{1}{n}\sum_{i=1}^{n} h(x_i)$$

Le problème d'intégration revient donc à trouver une variable X adéquate et à savoir échantillonner dans sa loi.

Méthode de Monte-Carlo par « échantillonnage »

$$g(x) = h(x)\ f_X(x) \qquad \text{pour} \qquad x \in [a,\ b]$$

A priori, n'importe quelle variable X de support $[a,\ b]$ convient.
Pourvu qu'on sache générer un n-échantillon dans sa loi.

Il suffit de prendre :

$$h(x) = \frac{g(x)}{f_X(x)}$$

Méthode de Monte-Carlo par « échantillonnage simple »
Dans ce cas, la variable X est la loi uniforme sur $[a,\ b]$:

$$f_X(x) = \frac{1}{b-a} \qquad \text{pour} \qquad x \in [a,\ b]$$

Alors :

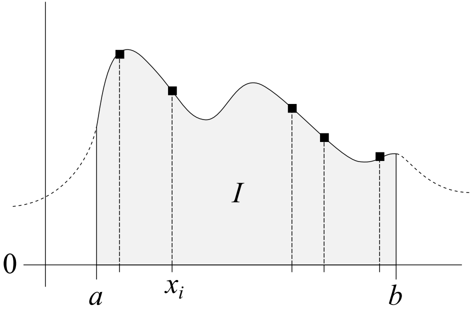
$$h(x) = (b-a)\ g(x) \qquad \text{pour} \qquad x \in [a,\ b]$$

et :

$$I = (b-a)\ \mathbb{E}\left[g(X)\right]$$

Calcul d'intégrales

Méthode de Monte-Carlo par « échantillonnage simple »



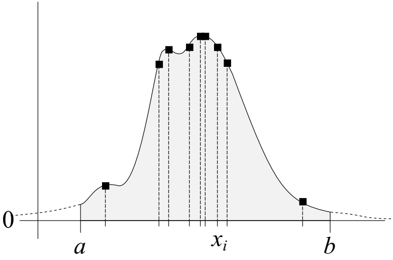
$$I \simeq \frac{b-a}{n} \sum_{i=1}^{n} g(x_i)$$

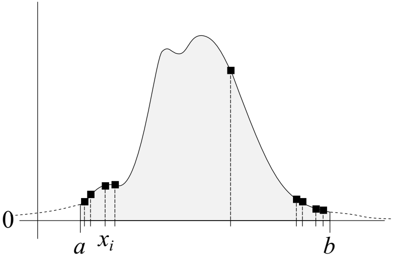
avec $(x_1, x_2, \ldots, x_i, \ldots, x_n)$ un échantillon de n valeurs tirées dans la loi uniforme sur $[a,\ b]$.

Ressemble à la méthode des rectangles.

Méthode de Monte-Carlo par « échantillonnage suivant l'importance »

Idée : Certaines régions ont plus d'importance que d'autres sur la quantité I à estimer. Mieux vaut concentrer l'échantillonnage sur ces régions (être précis).





Méthode de Monte-Carlo par « échantillonnage suivant l'importance »

Il « suffit » d'échantillonner en mimant la forme de la fonction g