Bootcamp Data Science
Zajęcia 4

Przemysław Spurek

Symulacje komputerowe

Symulacje komputerowe - to naśladowanie procesu przy pomocy programu komputerowego

Symulacje komputerowe

Symulacje komputerowe - to naśladowanie procesu przy pomocy programu komputerowego

Metody Monte Carlo

Metody Monte Carlo - metody symulacji wykorzystujące liczby losowe. Zwykle stosowane do obliczania prawdopodobieństw, wartości oczekiwanych i innych charakterystyk rozkładów (long-run proportion)

Symulacje komputerowe

Symulacje komputerowe - to naśladowanie procesu przy pomocy programu komputerowego

Metody Monte Carlo

Metody Monte Carlo - metody symulacji wykorzystujące liczby losowe. Zwykle stosowane do obliczania prawdopodobieństw, wartości oczekiwanych i innych charakterystyk rozkładów (long-run proportion)

Cel symulacji

Cel symulacji - estymacja wartości trudnych lub kosztownych do wyliczenia analitycznie

Liczby Pseudolosowe

Chociaż liczby pseudolosowe są generowane przez algorytm deterministyczny, możemy je traktować tak, jakby były to prawdziwe liczby losowe.

Liczby Pseudolosowe

Chociaż liczby pseudolosowe są generowane przez algorytm deterministyczny, możemy je traktować tak, jakby były to prawdziwe liczby losowe.

Zasadniczo algorytm generuje liczby całkowite, które są następnie znormalizowane, aby uzyskać dane zmiennoprzecinkowe ze standardowego rozkładu jednostajnego.

Liczby Pseudolosowe

Chociaż liczby pseudolosowe są generowane przez algorytm deterministyczny, możemy je traktować tak, jakby były to prawdziwe liczby losowe.

Zasadniczo algorytm generuje liczby całkowite, które są następnie znormalizowane, aby uzyskać dane zmiennoprzecinkowe ze standardowego rozkładu jednostajnego.

Próbki z innych rozkładów są z kolei generowane przy użyciu danych z rozkładu jednostajnego.

Liczby Pseudolosowe

- Zacznijmy od rozkładu dyskretnego, który jak pokażemy pozwala generować wszystkie inne.

Liczby Pseudolosowe

- Zacznijmy od rozkładu dyskretnego, który jak pokażemy pozwala generować wszystkie inne.
- Otóż chodzi nam o skonstruowanie zmiennej losowej X, która przyjmuje z jednolitym prawdopodobieństwem wartości w zbiorze $\mathbb{Z}_M = \{0, ..., M-1\}$, gdzie M jest bardzo dużą liczbą (z powodów numerycznych M jest często potęgą dwójki).

Liczby Pseudolosowe

- Zacznijmy od rozkładu dyskretnego, który jak pokażemy pozwala generować wszystkie inne.
- Otóż chodzi nam o skonstruowanie zmiennej losowej X, która przyjmuje z jednolitym prawdopodobieństwem wartości w zbiorze $\mathbb{Z}_M = \{0, ..., M-1\}$, gdzie M jest bardzo dużą liczbą (z powodów numerycznych M jest często potęgą dwójki).
- Czyli chcemy by każde $i \in \mathbb{Z}_M$ było losowane z jednakowym prawdopodobieństwem $1/M$.

Liczby Pseudolosowe

Uwaga

- Ponieważ programistycznie nie da się na komputerze zaimplementować generatora prawdziwych liczb losowych, w związku z czym istnieją sprzętowe generatory liczb losowych, które używają fizycznych zjawisk i posiadają własności losowe typu zjawiska kwantowe (korzystają z nich głównie banki dla bezpieczeństwa).

Liczby Pseudolosowe

Uwaga

- Ponieważ programistycznie nie da się na komputerze zaimplementować generatora prawdziwych liczb losowych, w związku z czym istnieją sprzętowe generatory liczb losowych, które używają fizycznych zjawisk i posiadają własności losowe typu zjawiska kwantowe (korzystają z nich głównie banki dla bezpieczeństwa).
- To co się robi najczęściej w praktyce z powodu szybkości i potencjalnej powtarzalności doświadczeń, to generowanie liczb pseudolosowych (pseudo-random number generator).

Liczby Pseudolosowe

Często stosowane generatory liczb losowych polegają na określeniu wartości startowej x (SEED) i funkcji $f : \mathbb{Z}_M \to \mathbb{Z}_M$ tak, że nasz ciąg pseudolosowy jest dany przez:

$$x_{n+1} = f(x_n).$$

Liczby Pseudolosowe

Często stosowane generatory liczb losowych polegają na określeniu wartości startowej x (SEED) i funkcji $f : \mathbb{Z}_M \to \mathbb{Z}_M$ tak, że nasz ciąg pseudolosowy jest dany przez:

$$x_{n+1} = f(x_n).$$

Funkcja f musi być tak dobrana aby nie było oczywistego związku pomiędzy poprzednimi wartościami, a następnymi oraz by była szybka w obliczaniu. Najprostsze generatory powyższego typu to LCG (linear congruential generator):

$$x_{n+1} = (ax_n + c) \ mod \ m.$$

Uwaga

Zły dobór parametrów może mieć tragiczne skutki – niesławny tu jest RANDU zaprojektowany w latach 60-tych przez IBM-a:

$$x_{j+1} = 65539 x_j \ mod \ 2^{31}.$$

Otóż $x_{k+2} = (2^16 + 3)x_{k+1} = (2^16 + 3)^2 x_k$ co oznacza, że:

$$x_{k+2} = (2^{16} + 6 \cdot 2^{16} + 9)x_k = [6 \cdot (2^{16} + 3) - 9]x_k = 6x_{k+1} - 9x_k \ mod \ 2^{31}.$$

W wyniku tego punkty (x_k, x_{k+1}, x_{k+2}) leżą w przestrzeni $\mathbb{R}^3$ na niewielkiej liczbie płaszczyzn (jest silna korelacja, nie ma niezależności). W konsekwencji wiele prac fizycznych bazujących na symulacjach losowych używających tego generatora okazało się być nieprawdziwych.

Liczby Pseudolosowe

Twierdzenie Hull-Dobell

LCG będzie generować "liczby pseudolosowe" gdy:

- c i m są względnie pierwsze,
- $a - 1$ jest podzielny przez wszystkie dzielniki pierwsze m,
- $a - 1$ jest wielokrotnością 4, jeśli m jest wielokrotnością 4.

LCG jest zazwyczaj tworzony tak, aby zwrócić z/m, która jest liczbą zmiennoprzecinkową z przedziału $(0, 1)$.

Liczby Pseudolosowe

https://github.com/przem85/bootcamp/blob/master/statistics/D14_Z01.ipynb

Zadanie 1.

Wylosuj próbkę 1000-elementową liczb pseudolosowych zgodnie z modelem LCG.

Zadanie 2.

Wylosuj próbkę 1000-elementową z rozkładu jednostajnego na odcinku (wykorzystując kod z Zadanie 1.).

Zadanie 3.

Wylosuj próbkę 1000-elementową z rozkładu Bernoulliego.

$$U \sim Uniform(0,1)$$

$$X = \begin{cases} 1 & \text{gdy } U < p \\ 0 & \text{gdy } U \leq p \end{cases}$$

Zadanie 4.

Wylosuj próbkę 1000-elementową z rozkładu dwumianowego.

$$U \sim Uniform(0, 1)$$

$$X = \sum_i X_i \sim Bernoulli(p)$$

Zadanie 5.

Wylosuj próbkę 1000-elementową z rozkładu geometrycznego.

$$U \sim Uniform(0, 1)$$

$$X =\sim Bernoulli(p)$$

```
p = ...
X = 1
while (runif(1) > p) {
    X = X+1
}
X
```

Liczby Pseudolosowe

Metoda losowania z dowolnego rozkładu dyskretnego.

$$p_i = P(X = x_i),\ i = 1, \ldots, n,\ \sum_{i=1}^{n} p_i = 1$$

Liczby Pseudolosowe

Metoda losowania z dowolnego rozkładu dyskretnego.

$$p_i = P(X = x_i),\ i = 1, \ldots, n,\ \sum_{i=1}^{n} p_i = 1$$

- Podziel odcinek $[0, 1]$ na pod-odcinki:

$$\begin{aligned} A_1 &= [0, p_1) \\ A_1 &= [p_1, p_1 + p_2) \\ A_3 &= [p_1 + p_2, p_1 + p_2 + p_3) \\ &\ldots \\ A_3 &= [p_1 + \ldots + p_{n-1}, 1) \end{aligned}$$

Liczby Pseudolosowe

Metoda losowania z dowolnego rozkładu dyskretnego.

$$p_i = P(X = x_i),\ i = 1, \ldots, n,\ \sum_{i=1}^{n} p_i = 1$$

- Podziel odcinek $[0, 1]$ na pod-odcinki:

$$\begin{aligned} A_1 &= [0, p_1) \\ A_1 &= [p_1, p_1 + p_2) \\ A_3 &= [p_1 + p_2, p_1 + p_2 + p_3) \\ &\ldots \\ A_3 &= [p_1 + \ldots + p_{n-1}, 1) \end{aligned}$$

- Wylosuj $U(0, 1)$

Liczby Pseudolosowe

Metoda losowania z dowolnego rozkładu dyskretnego.

$$p_i = P(X = x_i),\ i = 1, \ldots, n,\ \sum_{i=1}^{n} p_i = 1$$

- Podziel odcinek $[0, 1]$ na pod-odcinki:

$$\begin{aligned} A_1 &= [0, p_1) \\ A_1 &= [p_1, p_1 + p_2) \\ A_3 &= [p_1 + p_2, p_1 + p_2 + p_3) \\ &\ldots \\ A_3 &= [p_1 + \ldots + p_{n-1}, 1) \end{aligned}$$

- Wylosuj $U(0, 1)$
- Jeżeli U należy do A_i, to $X = x_i$

$$P(X = x_i) = P(U \in A_i) = p_i$$

Zadanie 6.

Wylosuj próbkę 1000-elementową z rozkładu X, takiego, że:

- $P(X = 1) = 0.2$,
- $P(X = 2) = 0.3$,
- $P(X = 3) = 0.25$,
- $P(X = 4) = 0.25$.

Metoda odwróconej dystrybuanty

Twierdzenia

Niech X będzie ciągłą zmienną losową o dystrybuancie $F_X(x)$. Niech $U = F_X(X)$.
Wtedy U ma rozkład jednostajny na $(0, 1)$.

Dowód jest prosty, zachęcam do pomyślenia nad nim.

Metoda odwróconej dystrybuanty

https://github.com/przem85/bootcamp/blob/master/statistics/D14_Z02.ipynb

Aby wygenerować zmienną losową X o dystrybuancie F, należy odwrócić formułę:

$$U = F(X).$$

Wtedy X może być uzyskane ze standaryzowanego rozkładu jednostajnego U jako:

$$X = F^{-1}(U)$$

Metoda Boxa-Mullera

`https://github.com/przem85/bootcamp/blob/master/statistics/D14_Z03.ipynb`

Dla rozkładu normalnego istnieje tzw. transformacja Boxa-Mullera, gdzie $(Z1, Z2)$ jest parą niezależnych zmiennych o standaryzowanych rozkładach normalnych:

$$Z_1 = \sqrt{-2\ln(U_1)}cos(2\pi U_2)$$

$$Z_2 = \sqrt{-2\ln(U_2)}sin(2\pi U_2)$$

gdzie:

$$U_1, U_2 \sim Uniform(0, 1).$$

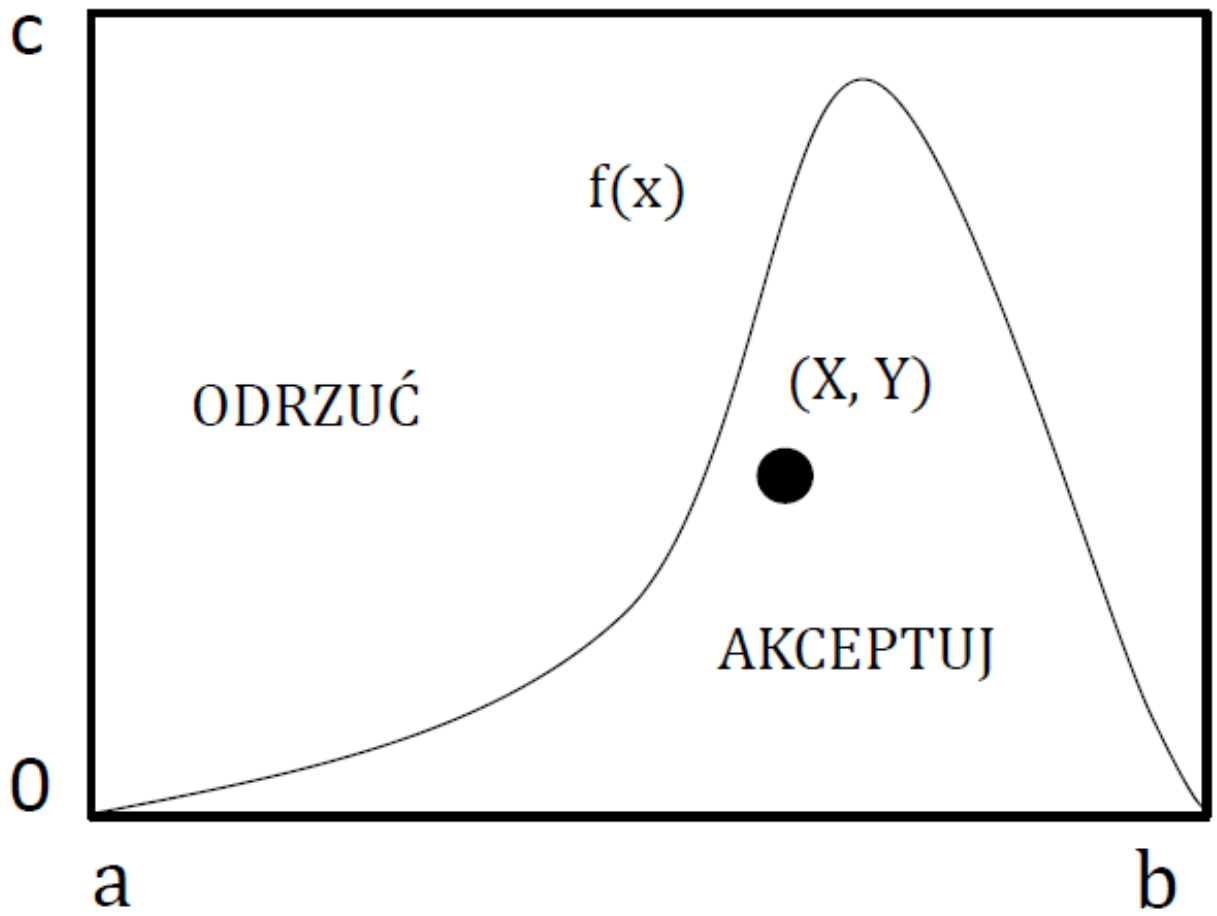
Metoda odrzucania

Można stosować, gdy postać odwróconej dystrybuanty jest zbyt skomplikowana, ale dana jest funkcja gęstości f.

Twierdzenie

Niech (X, Y) ma rozkład jednostajny w obszarze $A = (x, y) : 0 \leq y \leq f(x)$, dla pewnej funkcji gęstości f. Wtedy f jest gęstością X.

Metoda odrzucania



Metoda odrzucania

https://github.com/przem85/bootcamp/blob/master/statistics/D14_Z04.ipynb

1. Znajdź a, b, c takie, że:

$$0 \leq f(x) \leq c \text{ dla } a \leq x \leq b$$

2. $U \sim U(0,1)$, $V \sim U(0,1)$
3. Weźmy:

$$X = a + (b-a)U, \quad Y = cV,$$

wtedy: $X \sim Unif(a,b)$, $Y \sim Unif(0,c)$ i (X,Y) ma rozkład jednostajny na: $[a,b] \times [0,c]$

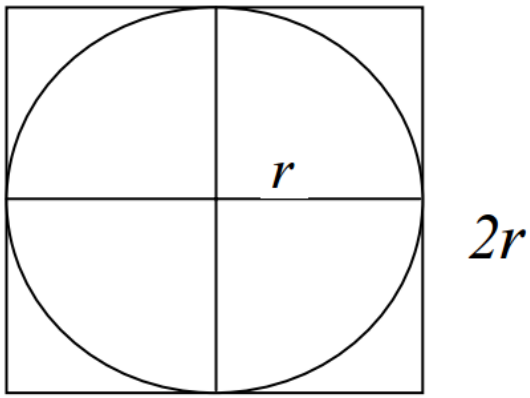
- Jeśli $Y > f(X)$ odrzuć i wróć do 2.
- Jeśli $Y \leq f(X)$, to X jest szukaną zmienną losową o gęstości $f(x)$

https://github.com/przem85/bootcamp/blob/master/statistics/D14_Z05.ipynb

METODY MONTE CARLO

Przykład wartość π

https://github.com/przem85/bootcamp/blob/master/statistics/D14_Z06.ipynb



Pole kwadratu to $4r^2$, a pole koła wynosi πr^2. W takim razie stosunek:

$$\frac{P_{kola}}{P_{kwadratu}} = \frac{\pi r^2}{4r^2} = \frac{\pi}{4}.$$

W konsekwencji:

$$\pi = 4\frac{P_{kola}}{P_{kwadratu}}.$$

Przykład wartość π

Mamy:

$$\pi = 4\frac{P_{kola}}{P_{kwadratu}}.$$

Jeżeli będziemy losować punkty o współrzędnych od $-2r$ do $2r$, to stosunek liczby punktów zawierających się w kole o środku w punkcie $(0, 0)$ i promieniu r do wszystkich wylosowanych punktów, będzie dążył w nieskończoności (z pewnym prawdopodobieństwem) do stosunku tego pola koła do koła kwadratu o boku $2r$.

Przykład wartość π

Mamy:

$$\pi = 4\frac{P_{kola}}{P_{kwadratu}}.$$

Jeżeli będziemy losować punkty o współrzędnych od $-2r$ do $2r$, to stosunek liczby punktów zawierających się w kole o środku w punkcie $(0, 0)$ i promieniu r do wszystkich wylosowanych punktów, będzie dążył w nieskończoności (z pewnym prawdopodobieństwem) do stosunku tego pola koła do koła kwadratu o boku $2r$.

Cała metoda sprowadza się więc do tego, by losować punkty i sprawdzać, czy mieszczą się w kole.

Przykład wartość π

Rozważmy troszkę ogólniejszą sytuację niech będzie dany zbiór A. Ponadto, niech prawdopodobieństwo, że zmienna losowa należy do A wynosi p:

$$p = P(X \in A)$$

Przykład wartość π

Rozważmy troszkę ogólniejszą sytuację niech będzie dany zbiór A. Ponadto, niech prawdopodobieństwo, że zmienna losowa należy do A wynosi p:

$$p = P(X \in A)$$

Zmienna losowa X przyjmuje:

- 1 (punkt leży wewnątrz obszaru A) z prawdopodobieństwem p,
- 0 (punkt leży poza obszarem A) z prawdopodobieństwem $1 - p$.

Przykład wartość π

$$p = P(X \in A)$$

Wtedy na podstawie próbki możemy określić estymator p:

$$\hat{p} = \hat{P}(X \in A) = \frac{|\{X_i : X_i \in A,\ i = 1, \ldots, N\}|}{N}$$

gdzie N - rozmiar eksperymentu Monte Carlo, X_i - zmienne generowane z tego samego rozkładu co X.

Przykład wartość π

$$\hat{p} = \hat{P}(X \in A) = \frac{|\{X_i : X_i \in A,\ i = 1, \ldots, N\}|}{N} = \frac{K}{N}.$$

Rozważmy zmienną losową K, która mówi ile punktów wpadło do zbioru A. Ma ona rozkład dwumianowy (suma zmiennych losowych o rozkładzie zero-jedynkowym):

$$K \sim Binomial(N, p).$$

Wiemy (`https://pl.wikipedia.org/wiki/Rozk%C5%82ad_dwumianowy`), że wartość oczekiwana i wariancja rozkładu dwumianowego to:

$$E(K) = Np, \quad Var(K) = Np(1 - p).$$

Przykład wartość π

$$\hat{p} = \hat{P}(X \in A) = \frac{|\{X_i : X_i \in A,\ i = 1, \ldots, N\}|}{N} = \frac{K}{N}.$$

Rozważmy zmienną losową K, która mówi ile punktów wpadło do zbioru A. Ma ona rozkład dwumianowy (suma zmiennych losowych o rozkładzie zero-jedynkowym):

$$K \sim Binomial(N, p).$$

Wiemy (`https://pl.wikipedia.org/wiki/Rozk%C5%82ad_dwumianowy`), że wartość oczekiwana i wariancja rozkładu dwumianowego to:

$$E(K) = Np, \quad Var(K) = Np(1-p).$$

W takim razie:

$$E(\hat{p}) = \frac{1}{N}(Np) = p, \qquad std(\hat{p}) = \frac{1}{N}\sqrt{Np(1-p)} = \sqrt{\frac{p(1-p)}{N}}$$

Przykład wartość π

Estymator p jest nieobciążony:

$$E(\hat{p}) = \frac{1}{N}(Np) = p$$

Odchylenie estymatora maleje wraz ze wzrostem N w tempie $1/\sqrt{N}$. Np. 100-krotne zwiększenie próby da 10-krotnie mniejsze odchylenie standardowe estymatora:

$$std(\hat{p}) = \frac{1}{N}\sqrt{Np(1-p)} = \sqrt{\frac{p(1-p)}{N}}$$

Przykład wartość π

Ponieważ w metodach Monte Carlo n jest zwykle duże, będziemy mogli skorzystać z centralnego twierdzenia granicznego (CTG), które mówi, że jeśli tylko $X_1, \ldots, X_n$ są niezależne i o jednakowym rozkładzie oraz $\sigma_X^2 = Var(X_i) < \infty$, to dla $n \to \infty$

$$P(\sum_{j=1}^{n} \frac{X_i - E(X_i)}{\sigma_X \sqrt{n}} \leq x) \to CDF(x)$$

gdzie $\sigma_X = \sqrt{Var(X_1)}$ oraz $CDF(x)$ jest dystrybuantą standardowego rozkładu normalnego.

Przykład wartość π

W konsekwencji możemy policzyć 95% przedział ufności dla naszego estymatora:

$$[\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}],$$

gdy nie znamy σ to możemy użyć estymatora:

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}).$$

Przykład wartość π

W przypadku naszej liczby π mamy:

$$p = \frac{P_{kola}}{P_{kwadratu}} = \frac{\pi r^2}{4r^2} = \frac{\pi}{4}.$$

$$E(\hat{p}) = \frac{\pi}{4}, \qquad std(\hat{p}) = \frac{\pi}{4}(1 - \frac{\pi}{4})$$

Przykład wartość π

W przypadku naszej liczby π mamy:

$$p = \frac{P_{kola}}{P_{kwadratu}} = \frac{\pi r^2}{4r^2} = \frac{\pi}{4}.$$

$$E(\hat{p}) = \frac{\pi}{4}, \qquad std(\hat{p}) = \frac{\pi}{4}(1 - \frac{\pi}{4})$$

- Tak więc aby przy zadanym poziomie istotności dostać jedną cyfrę wiodącą (długość przedziału ufności $[\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}]$, musi być równa jeden) tj. $\pi = 3$, należy zrobić $n \approx 1.96^2 \cdot 2.7^2/0.5^2 = 112$ replikacji.

Przykład wartość π

W przypadku naszej liczby π mamy:

$$p = \frac{P_{kola}}{P_{kwadratu}} = \frac{\pi r^2}{4r^2} = \frac{\pi}{4}.$$

$$E(\hat{p}) = \frac{\pi}{4}, \qquad std(\hat{p}) = \frac{\pi}{4}(1 - \frac{\pi}{4})$$

- Tak więc aby przy zadanym poziomie istotności dostać jedną cyfrę wiodącą (długość przedziału ufności $[\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}]$, musi być równa jeden) tj. $\pi = 3$, należy zrobić $n \approx 1.96^2 \cdot 2.7^2/0.5^2 = 112$ replikacji.
- Aby dostać drugą cyfrę wiodąca to musimy zrobić (długość przedziału ufności musi być równa 0.1) $n \approx 1.96^2 \cdot 2.7^2/0.05^2 = 11200$ replikacji, itd.
- Pamiętaj, że używamy $16 \cdot \frac{\pi}{4}(1 - \frac{\pi}{4})$ bo mamy $\hat{p} = \frac{\pi}{4}$

`https://github.com/przem85/bootcamp/blob/master/statistics/D14_Z06.ipynb`

Podsumowanie

Metody Monte Carlo służące do obliczenia wartość I polegają na stworzeniu estymatora nieobciążonego (w naszych rozważaniach) w postaci:

$$\hat{X}_n = \frac{1}{n} \sum_{j=1}^{n} X_j.$$

gdzie $X_1, \ldots, X_n$ są niezależnymi replikacjami zmiennej losowej X oraz mamy:

$$I = E(\hat{X}_n).$$

Podsumowanie

Metody Monte Carlo służące do obliczenia wartość I polegają na stworzeniu estymatora nieobciążonego (w naszych rozważaniach) w postaci:

$$\hat{X}_n = \frac{1}{n} \sum_{j=1}^{n} X_j.$$

gdzie $X_1, \ldots, X_n$ są niezależnymi replikacjami zmiennej losowej X oraz mamy:

$$I = E(\hat{X}_n).$$

Uwaga

W przypadku liczby π zmienna losowa X miała rozkład zero-jedynkowy (jeden gdy jest w kole i zero gdy poza nim), a:

$$p = \frac{1}{n} \sum_{j=1}^{n} X_j.$$

Całkowanie – metoda chybił-trafił

`https://github.com/przem85/bootcamp/blob/master/statistics/D14_Z07.ipynb`

Analogicznie jak przy liczeniu liczby π możemy postępować w ogólnym przypadku. Dla całki:

$$\int_a^b f(x)dx$$

Całkowanie – metoda chybił-trafił

`https://github.com/przem85/bootcamp/blob/master/statistics/D14_Z07.ipynb`

Analogicznie jak przy liczeniu liczby π możemy postępować w ogólnym przypadku. Dla całki:

$$\int_a^b f(x)dx$$

- Wylosuj dużą próbkę n ze zmiennej losowej o rozkładzie jednostajnym na kwadracie $[a, b] \times [0, max_{x \in [a,b]}\{f(x)\}]$

Całkowanie – metoda chybił-trafił

`https://github.com/przem85/bootcamp/blob/master/statistics/D14_Z07.ipynb`

Analogicznie jak przy liczeniu liczby π możemy postępować w ogólnym przypadku. Dla całki:

$$\int_a^b f(x)dx$$

- Wylosuj dużą próbkę n ze zmiennej losowej o rozkładzie jednostajnym na kwadracie $[a, b] \times [0, max_{x \in [a,b]}\{f(x)\}]$
- Policzyć liczbę punktów, które wpadają pod funkcję (tak jak w metodzie odrzuceń) A

Całkowanie – metoda chybił-trafił

`https://github.com/przem85/bootcamp/blob/master/statistics/D14_Z07.ipynb`

Analogicznie jak przy liczeniu liczby π możemy postępować w ogólnym przypadku. Dla całki:

$$\int_a^b f(x)dx$$

- Wylosuj dużą próbkę n ze zmiennej losowej o rozkładzie jednostajnym na kwadracie $[a, b] \times [0, max_{x \in [a,b]}\{f(x)\}]$
- Policzyć liczbę punktów, które wpadają pod funkcję (tak jak w metodzie odrzuceń) A
- Oszacuj wartość całki jako:

$$\frac{A}{n} \cdot Vol([a, b] \times [0, max_{x \in [a,b]}\{f(x)\}])$$

gdzie $Vol([a, b] \times [0, max_{x \in [a,b]}\{f(x)\}])$ objętość prostokąta, z którego próbkujemy (czynnik normalizujący).

Całkowanie

Możemy również stworzyć analogicznie jak wcześniej estymator nieobciążony całki i wykorzystać go w naszej metodzie.
Naszym celem jest obliczyć całkę:

$$I = \int_a^b \psi(x)dx.$$

Całkowanie

Możemy również stworzyć analogicznie jak wcześniej estymator nieobciążony całki i wykorzystać go w naszej metodzie.
Naszym celem jest obliczyć całkę:

$$I = \int_a^b \psi(x)dx.$$

Uwaga

Niech ψ będzie dowolną funkcją. Jeżeli zmienna losowa X ma rozkład ciągły o gęstości f i wartość oczekiwana $\psi(X)$ istnieje, to wyraża się wzorem:

$$E(\psi(X)) = \int_{-\infty}^{+\infty} \psi(x)f(x)dx$$

Całkowanie – metoda chybił-trafił

https://github.com/przem85/bootcamp/blob/master/statistics/D14_Z08.ipynb

$$I = \int_a^b \psi(x)dx.$$

Niech $X_1, \ldots, X_n$ będzie ciągiem niezależnych zmiennych losowych o rozkładzie jednostajnym na odcinku $[0, 1]$, wtedy:

$$\hat{I}_n = \frac{1}{n} \sum \psi(X_i)$$

jest estymatorem nieobciążonym I.

Całkowanie – metoda chybił-trafił

https://github.com/przem85/bootcamp/blob/master/statistics/D14_Z08.ipynb

$$I = \int_a^b \psi(x)dx.$$

Niech $X_1, \ldots, X_n$ będzie ciągiem niezależnych zmiennych losowych o rozkładzie jednostajnym na odcinku $[0, 1]$, wtedy:

$$\hat{I}_n = \frac{1}{n} \sum \psi(X_i)$$

jest estymatorem nieobciążonym I.

$$E(\hat{I}_n) = E(\tfrac{1}{n} \sum \psi(X_i)) = \tfrac{1}{n} \sum E(\psi(X_i)) =$$
$$= \tfrac{1}{n} \sum \int_{\mathbb{R}} \psi(x) U_{[0,1]}(x)dx = \tfrac{1}{n} \sum \int_0^1 \psi(x)dx = \tfrac{1}{n} n \int_0^1 \psi(x)dx = I.$$

Importance sampling

W poprzedniej procedurze możemy użyć innej gęstości $g(x)$ - tak zwanej "importance function". Można przekształcić:

$$\int_a^b \psi(x)dx = \int_a^b g(x)\frac{\psi(x)}{g(x)}dx,$$

gdzie gęstość g jest większa od f w każdym punkcie na $[a, b]$.

Takie podejście umożliwi też całkowanie na przedziale $[a, \infty)$ lub $(-\infty, b]$.

Wielowymiarowy rozkład normalny

https://github.com/przem85/bootcamp/blob/master/statistics/D14_Z09.ipynb

Funkcja gęstości n-wymiarowego rozkładu normalnego wektora losowego X, o wektorze wartości oczekiwanych: $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_n]^T$ i macierzy kowariancji: Σ, dana jest wzorem:

$$f_{\boldsymbol{\mu},\Sigma}(X) = \frac{1}{(2\pi)^{n/2}\,|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(X-\boldsymbol{\mu})^T\Sigma^{-1}(X-\boldsymbol{\mu})\right).$$

Oznacza się to w skrócie zapisem:

$$X \sim N(\boldsymbol{\mu}, \Sigma)$$

Losowanie z wielowymiarowego rozkładu normalnego

Dla dowolnego n i dla $X = (X_1, \ldots, X_n)$ oraz A takie:

$$\Sigma = A^T A$$

mamy:

Twierdzenie

Jeśli $Z = (Z_1, \ldots, Z_n)$ są niezależnymi zmiennymi losowymi o jednakowym rozkładzie normalnym $N(0, 1)$, to:

$$Z = ZA + m$$

ma rozkład $N(n, \Sigma)$.

Aby obliczyć "pierwiastek" A z macierzy kowariancji Σ można skorzystać z gotowych metod numerycznych - Rozkład Choleskiego.
`https://pl.wikipedia.org/wiki/Rozk%C5%82ad_Choleskiego`

Losowanie z wielowymiarowego rozkładu normalnego

https://github.com/przem85/bootcamp/blob/master/statistics/D14_Z10.ipynb

Zadanie

Wylosuj dane z rozkładu normalnego o parametrach: $m = [0, 0]$, $\Sigma = [[3.40, -2.75], [-2.75, 5.50]]$ przy użyciu próbki z rozkładu N([0,0], I), gdzie *I* to macierz identycznościowa.

Próbkowanie Gibbsa

Przypuśćmy, że chcemy losować dane z rozkładu dwuwymiarowego (trudnego):

$$p(X;\theta), \quad x = (x_1, x_2)$$

a umiemy losować z $p(x_1|x_2,\theta)$ oraz $p(x_2|x_1,\theta)$.
Wtedy, dla początkowych wartości: (x_1^0, x_2^0) wykonujemy próbkowanie Gibbsa:

- wylosuj
$$x_1^{(t)} \sim p(x_1|x_2^{(t)},\theta),$$
- wylosuj
$$x_2^{(t)} \sim p(x_2|x_1^{(t)},\theta).$$

Próbkowanie Gibbsa

https://github.com/przem85/bootcamp/blob/master/statistics/D14_Z11.ipynb

Zadanie

Wylosuj próbkę z rozkładu:

$$X \sim N(0, \Sigma), \qquad \Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}.$$

za pomocą próbkowanie Gibbsa wiedząc, że:

$$p(x_1|x_2) \sim N(\rho x_1, [1 - \rho^2])$$
$$p(x_2|x_1) \sim N(\rho x_2, [1 - \rho^2])$$

to są rozkłady warunkowe.