Dyskretny rozkład Bernoulliego, rozkład normalny, centralne twierdzenie graniczne

Metody Monte Carlo w fizyce
Nanoinżynieria materiałów

Łukasz Ruba

Akademia Górniczo-Hutnicza
im. Stanisława Staszica w Krakowie

29 lutego 2024

Spis treści

Cel ćwiczenia

Celem ćwiczenia było zapoznanie się z dyskretnym rozkładem Bernoulliego. W tym celu utworzono ciągi liczb $N = 10^7$ za pomocą algorytmu generowania liczb o rozkładzie Bernoulliego i przeanalizowano błąd estymowanych wartości oczekiwanej oraz wariancji z wartościami teoretycznymi.

1 Wstęp teoretyczny

Rozkład Bernoulliego opisuje proces, którego wynik może przyjmować jedynie dwie wartości [1]:

$$X_n = \{0, 1\}. \tag{1}$$

Prawdopodobieństwo uzyskania stanu 1 jest określane wartością p, zaś uzyskanie stanu 0 jest możliwe z prawdopodobieństwem q, gdzie:

$$q = 1 - p. \tag{2}$$

W rozkładzie Bernoulliego znając prawdopodobieństwo sukcesu można obliczyć analitycznie wartość oczekiwaną zmiennej X za pomocą równania (3):

$$E\langle X\rangle = q \cdot 0 + p \cdot 1 = p, \tag{3}$$

zaś jej wariancja wynosi wtedy:

$$\sigma_X^2 = E\langle X^2\rangle - E\langle X\rangle^2 = q \cdot 0^2 + p \cdot 1^2 - (q \cdot 0 + p \cdot 1)^2 = p - p^2 = p(1-p). \tag{4}$$

Te same wartości dla zmiennej X można estymować korzystając z technik numerycznych. Wartość oczekiwaną ciągu liczb z rozkładu Bernoulliego obliczymy jako średnią arytmetyczną z elementów tego ciągu:

$$E\langle X\rangle = \sum_{n=1}^{N} \frac{X_n}{N}. \tag{5}$$

Na podstawie wzoru (5) numerycznie można wyznaczyć również wariancję jako:

$$\sigma_X^2 = E\langle X^2\rangle - E\langle X\rangle^2 = \sum_{n=1}^{N} \frac{X^2}{N} - \left(\sum_{n=1}^{N} \frac{X}{N}\right)^2. \tag{6}$$

2 Metodyka

Obliczenia został wykonane za pomocą algorytmu napisanego w języku *C++*. Program utworzył trzy ciągi liczbowe o mocy $N = 10^7$ na podstawie warunku zależnego od prawdopodobieństwa wystąpienia sukcesu, który obrazuje pseudokod poniżej:

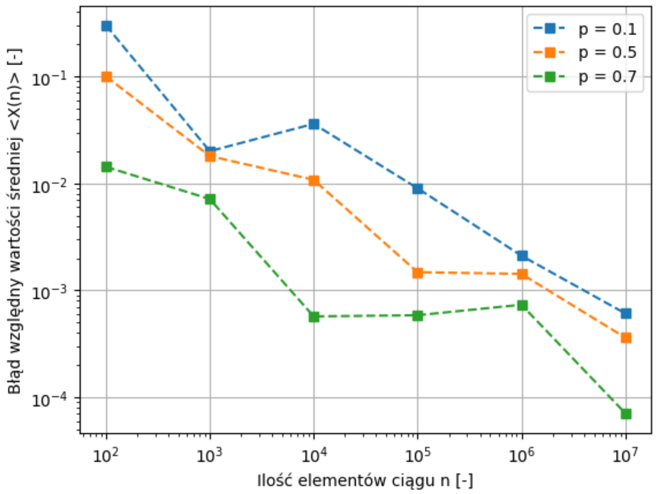
```
if [ RANDOM_NUMBER( value from 0 to 1) <= p ] then:
    X_n = 1
else
    X_n = 0
end
```

Kolejne liczby X_n sumowano w każdej iteracji, a następnie dla iteracji od 10^2 do 10^7 z krokiem co rząd wielkości liczono wartość średnią oraz wariancję powstałego ciągu na podstawie wzorów (5) i (6). Wartości te ostatecznie wykorzystano do porównania danych estymowanych z analitycznymi licząc moduł błędu względnego. Obliczenia wykonano dla trzech prawdopodobieństw p o wartościach kolejno 0.1, 0.5 oraz 0.7.

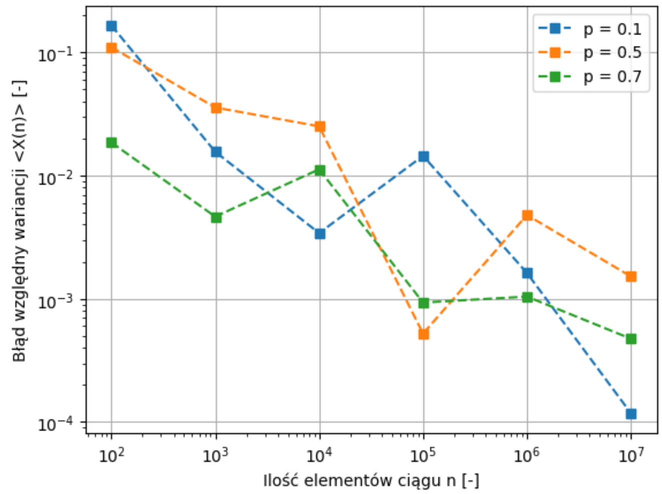
Wszystkie dane zapisano do pliku, a następnie przy użyciu języka programowania *Python* i paczki *matplotlib* utworzono na ich podstawie wykresy zależności błędów wartości oczekiwanej i wariancji od długości ciągu X.

3 Wyniki

Na podstawie obliczeń utworzono Rysunki 1 oraz 2.



Rysunek 1: Wartości błędów względnych wartości średniej dla prawdopodobieństw kolejno 0.1, 0.5 i 0.7 w zależności od ilości elementów ciągu. Obie skale przedstawiono logarytmicznie.



Rysunek 2: Wartości błędów względnych wariancji dla prawdopodobieństw kolejno 0.1, 0.5 i 0.7 w zależności od ilości elementów ciągu. Obie skale przedstawiono logarytmicznie.

Jak można zauważyć, w obu przypadkach wartości błędów zbiegają do zera wraz z coraz to większą ilością elementów w ciągu. Powołując się na równanie von Misesa prawdopodobieństwa P(A):

$$lim_{N\to\infty}\frac{k_N(A)}{N} = P(A), \tag{7}$$

gdzie $k_N(A)$ oznacza liczbę zdarzeń elementarnych sprzyjających zdarzeniu A, widzimy, że obserwowana tendencja jest zgodna z przewidywaniami. Zwiększając długość ciągu N coraz lepiej przybliżamy równanie (7). W naszym przypadku prawdopodobieństwo $P(A)$ jest równe p, czyli analitycznej wartości średniej, jak udowodniono we wzorze (3).

Należy jednak pamiętać, że obliczenia numeryczne nigdy w pełni nie przybliżą równania (7) oraz obarczone są fluktuacjami statystycznymi, które można zobaczyć na Rysunkach 1 i 2 w postaci chwilowych wzrostów wartości błędów.

4 Wnioski

Algorytm utworzony w trakcie pierwszego laboratorium, mimo swojej prostoty, świetnie pokazuje jak ilość obliczeń wpływa na dokładność oszacowania parametrów procesów stochastycznych. Pokazano, że wraz ze wzrostem liczebności ciągów zwiększa się dokładność estymacji wartości średniej i wariancji, co jest zgodne z przewidywaniami teoretycznymi.

Literatura

[1] Tomasz Chwiej, dr hab. inż., wykład pt. "*Wybrane aspekty podstaw rachunku prawdopodobieństwa i statystyki*", AGH, Kraków, 2024.