Model kul i urn w kontekście zagadnień algorytmicznych

Ewa Kasprzak

3 grudnia 2023

1. Wstęp

Model kul i urn jest klasyfikowany jako jeden z kluczowych modeli probabilistycznych, szczególnie w kontekście algorytmiki probabilistycznej. Ten model abstrahuje różne problemy, takie jak paradoks urodzinowy czy problem kolekcjonera kuponów.

Celem tego sprawozdania jest eksperymentalne zbadanie różnych aspektów związanych z modelem kul i urn. Poprzez symulacje komputerowe, będziemy analizować wartości takie jak moment pierwszej kolizji, liczba pustych urn, minimalna liczba rzutów do uzyskania co najmniej jednej kuli w każdej urnie, minimalna liczba rzutów do uzyskania co najmniej dwóch kul w każdej urnie, oraz różnica między tymi dwoma ostatnimi wielkościami.

2. Opis modelu

W modelu tym, m kul jest wrzucane kolejno do n ponumerowanych urn. Każda kula jest wrzucana niezależnie z jednakowym prawdopodobieństwem równym $\frac{1}{n}$ do jednej z urn. Wrzucenie kuli do urny możemy utożsamiać z losową funkcją f ze zbioru $\{1, \ldots, m\}$ do zbioru $\{1, \ldots, n\}$. Formalnie, przestrzeń zdarzeń elementarnych jest wówczas zbiorem $\Omega_{n,m} = \{1, \ldots, n\}^{1,\ldots,m}$.

Celem tego zadania jest eksperymentalne wyznaczenie następujących wielkości:

a) B_n – moment pierwszej kolizji; $B_n = k$, jeśli k-ta z wrzucanych kul jest pierwszą, która trafiła do niepustej urny (paradoks urodzinowy, ang. birthday paradox),

b) U_n – liczba pustych urn po wrzuceniu n kul,

c) C_n – minimalna liczba rzutów, po której w każdej z urn jest co najmniej jedna kula (pierwszy moment, w którym nie ma już pustych urn; problem kolekcjonera kuponów, ang. coupon collector's problem),

d) D_n – minimalna liczba rzutów, po której w każdej z urn są co najmniej dwie kule (the siblings of the coupon collector / coupon collector's brother),

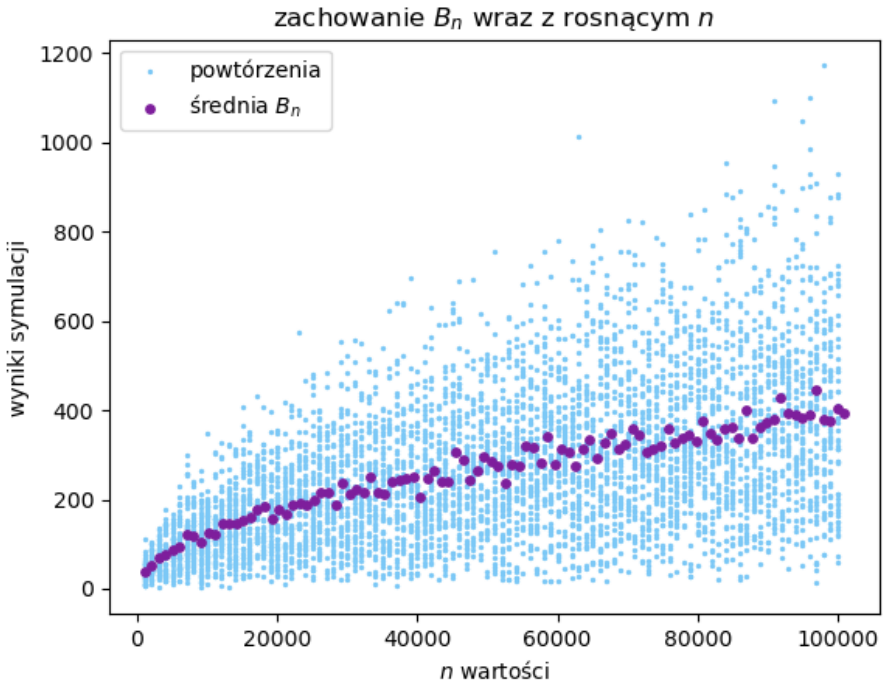
e) $D_n - C_n$ – liczba rzutów od momentu C_n, potrzebna do tego, żeby w każdej urnie były co najmniej dwie kule.

3. Opis symulacji

Wykonano dla każdej wartości n z zakresu $\{1000, 2000, \ldots, 100000\}$ 50 niezależnych powtórzeń eksperymentu wrzucania kul do urn. Każde pojedyncze powtórzenie polegało na wrzucaniu kul aż do pierwszego momentu, w którym w każdej z urn znajdowały się co najmniej dwie kule. W trakcie tego procesu zbierane były różne statystyki.

Do generacji liczb pseudolosowych został użyty generator SecureRandom z języka Java, co zapewniło dobre własności statystyczne. W celu prezentacji wyników użyto narzędzia NumPy z języka Python. Na wykresach przedstawiono wyniki poszczególnych powtórzeń, zaznaczając punkty danych dla każdego n, oraz dodano średnie wartości, co umożliwiło łatwą ocenę koncentracji wyników wokół wartości średniej.

4. Wyniki symulacji i asymptotyka wartości oczekiwanych zmiennych losowych



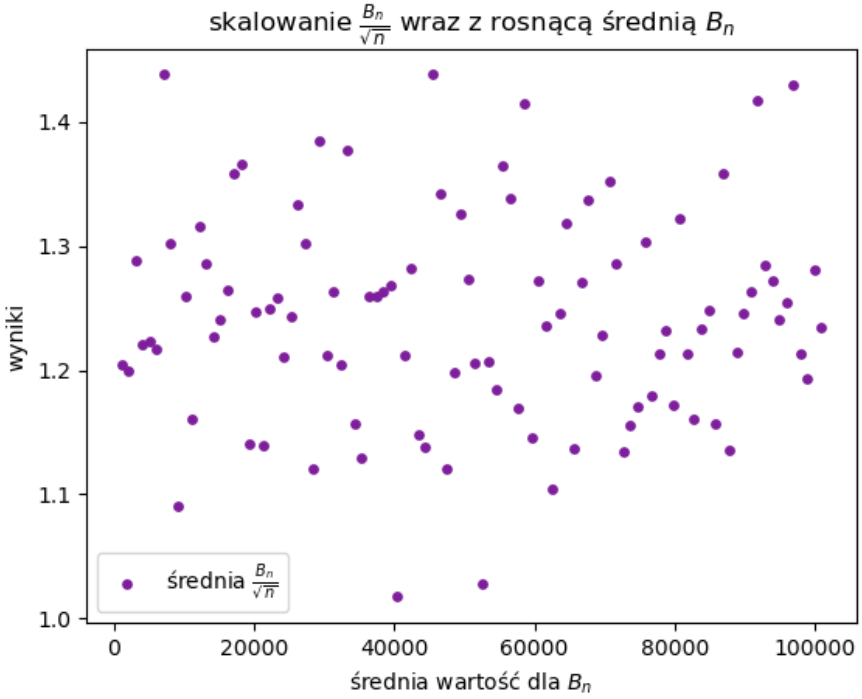
Rysunek 1: Wyniki eksperymentów dla wartości Bn.

Powyższy wykres ilustruje zmiany wartości B_n w miarę wzrostu parametru n podczas przeprowadzanej symulacji. Wraz ze zwiększaniem się wartości n, obserwuje się większe rozproszenie wyników na wykresie, co świadczy o rosnącej wartości odchylenia standardowego prób.

Paradoks urodzinowy to badanie, które skupia się na określeniu minimalnej liczby osób, jakie trzeba zgromadzić, aby prawdopodobieństwo, że przynajmniej dwie z nich obchodzą swoje urodziny tego samego dnia, przekroczyło 50%. Choć na pierwszy rzut oka mogłoby się wydawać, że konieczne jest zorganizowanie znacznej liczby uczestników, analiza matematyczna ujawnia, że minimalna liczba ta wynosi 23, co stanowi istotny paradoks w świetle intuicji.

Według Michaela Mitzenmachera i Eli Upfala w książce *Probability and Computing: Randomization and Probabilistic Techniques in Algorithms and Data Analysis* (Cambridge University Press, USA, 2nd edition, 2017), jeśli rozłożymy m kul w n urnach, gdzie $m = \Omega(\sqrt{n})$, to z prawdopodobieństwem bliskim jedności przynajmniej w jednej z urn znajdą się dwie kule. Na podstawie tej obserwacji możemy wywnioskować, że wartość oczekiwana $E(B_n) = O(\sqrt{n})$.

Potwierdzeniem tego stwierdzenia jest zamieszczony poniżej wykres, przedstawiający średnie wartości zmiennej losowej B_n przeskalowane przez $\sqrt{n}$. Jak można zauważyć, wartości te wydają się koncentrować w wąskim przedziale. W rezultacie otrzymujemy funkcję, która przybliża się do funkcji stałej.



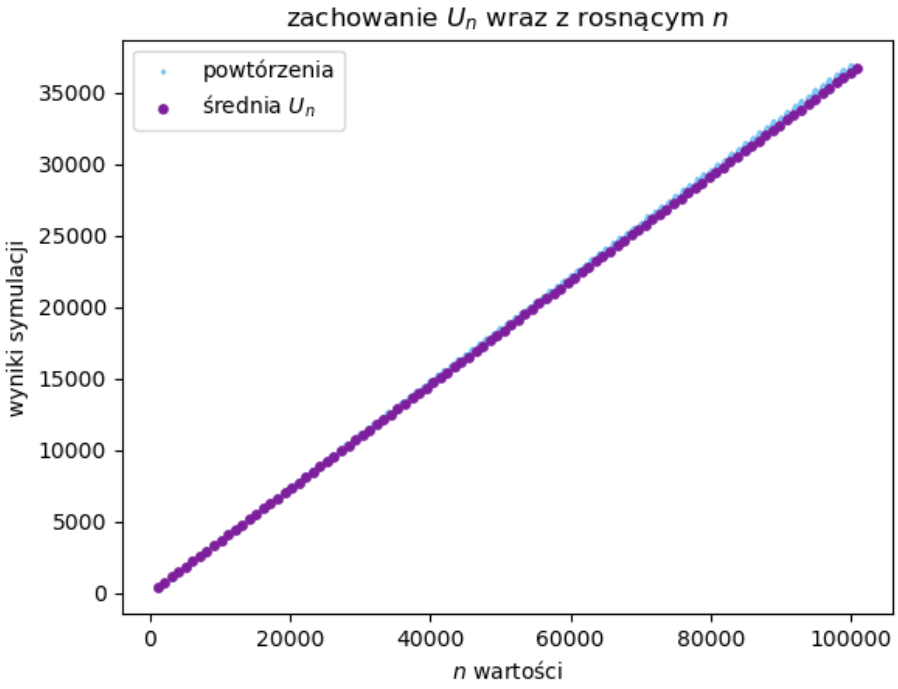
Rysunek 2: Wyniki eksperymentów dla wartości Bn.

Funkcja haszująca jest narzędziem używanym do przypisania unikalnego klucza, zwanej skrótem, do danej wiadomości, co umożliwia późniejsze rozpoznawanie. Atak urodzinowy stanowi potencjalne zagrożenie dla integralności tej funkcji, polegające na wytworzeniu kolizji, czyli sytuacji, w której dwie różne wiadomości generują ten sam skrót. Taki atak może prowadzić do zastąpienia oryginalnej wiadomości przez jej bliźniaczą kopię, tj. o identycznym skrócie,

co umożliwia cyberprzestępcy zamianę wiadomości bez wykrycia.

Paradoks urodzinowy w tym kontekście odnosi się do zjawiska, w którym kolizje mogą wystąpić znacznie szybciej niż można by przewidzieć na podstawie rozmiaru przestrzeni skrótów funkcji haszującej. W tym przypadku, "urodziny" są analogiczne do danych wejściowych, a "osoby w grupie" reprezentują ilość różnych możliwych danych, które są analizowane. Liczba potrzebnych prób atakującego rośnie wraz z pierwiastkiem kwadratowym liczby wszystkich możliwości danych wejściowych.

Z tego wynika, że zwiększanie długości skrótu jest konieczne w celu zmniejszenia ryzyka kolizji, ponieważ większa przestrzeń skrótów przekłada się na mniejsze prawdopodobieństwo wystąpienia kolizji.



Rysunek 3: Wyniki eksperymentów dla wartości Un.

Powyższy wykres ilustruje zmiany wartości U_n w miarę wzrostu parametru n podczas przeprowadzanej symulacji. Wraz ze zwiększaniem się wartości n, obserwuje się większe rozproszenie wyników na wykresie, co świadczy o rosnącej wartości odchylenia standardowego prób. Warto jednak zauważyć, że odchylenia te są bardzo małe. Dla małych wartości n są praktycznie niewidoczne, a dla dużych n są znikome. Nieduże odchylenia standardowe świadczą o stabilności i małej różnorodności wyników.

Niech U_i dla $i \in \{1, \ldots, n\}$ będzie zmienną losową, przyjmującą wartość 0, gdy i-ta urna jest pełna, i 1, gdy i-ta urna jest pusta. Wówczas zmienna losowa $U_n = \sum_{i=1}^{n} U_i$. Ze względu na niezależność zmiennych losowych U_i, można obliczyć wartość oczekiwaną $E(U_n)$, wykorzystując

linearność wartości oczekiwanej:

$$E(U_n) = E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(U_i)$$

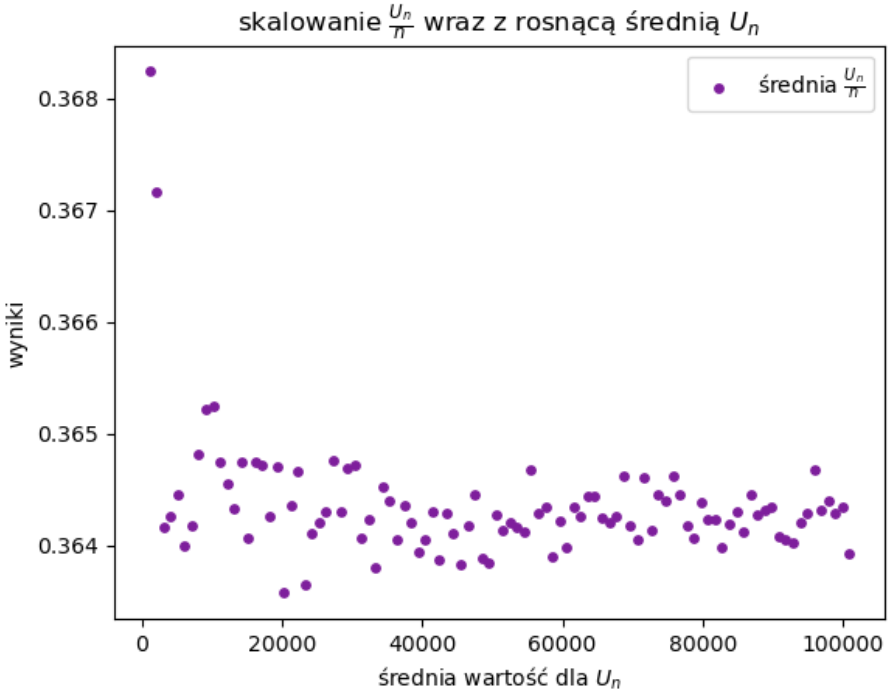
Ponieważ prawdopodobieństwo trafienia kuli do urny jest takie samo dla każdego $i \in \{1, \ldots, n\}$, można wyznaczyć $E(U_i)$ jako $(1 - 1/n)^m \approx e^{-m/n}$ Wynika stąd, że $E(U_n) = n(1 - 1/n)^m = ne^{-n/m}$.

Uzyskane informacje pozwalają na określenie asymptotycznego zachowania zmiennej U_n, ponieważ

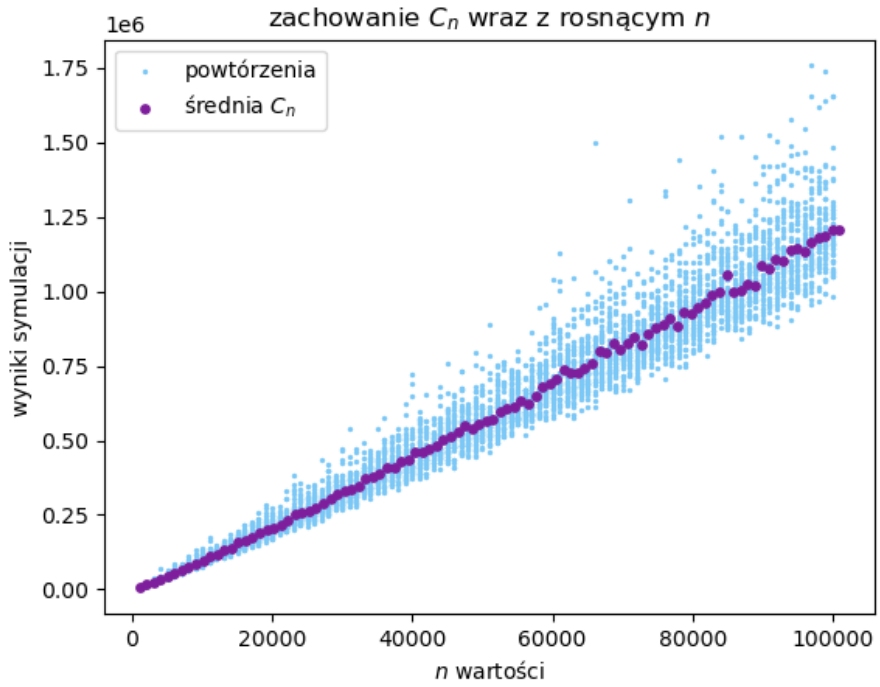
$$\lim_{n\to\infty} \frac{n(e^{-m/n})}{n} = \lim_{n\to\infty} e^{-m/n} = 1.$$

Stąd mamy $E(U_n) = O(n)$.

Potwierdzeniem tego stwierdzenia jest zamieszczony poniżej wykres, przedstawiający średnie wartości zmiennej losowej U_n przeskalowane przez n. Jak można zauważyć, wartości te wydają się koncentrować w wąskim przedziale. W rezultacie otrzymujemy funkcję, która przybliża się do funkcji stałej.



Rysunek 4: Wyniki eksperymentów dla wartości Cn.



Rysunek 5: Wyniki eksperymentów dla wartości Cn.

Powyższy wykres ilustruje zmiany wartości C_n w miarę wzrostu parametru n podczas przeprowadzanej symulacji. Wraz ze zwiększaniem się wartości n, obserwuje się większe rozproszenie wyników na wykresie, co świadczy o rosnącej wartości odchylenia standardowego prób.

Problem kolekcjonera kuponów stanowi wyzwanie matematyczne, którego celem jest analiza minimalnej liczby prób niezbędnych do zebrania kompletnego zbioru, w warunkach, gdzie każdy element występuje losowo i niezależnie od pozostałych. Problem ten znajduje swoje zastosowanie w sytuacji, gdy każda paczka płatków śniadaniowych zawiera jeden z n różnych kuponów, a uzyskanie pełnej kolekcji umożliwia zgłoszenie się po nagrodę.

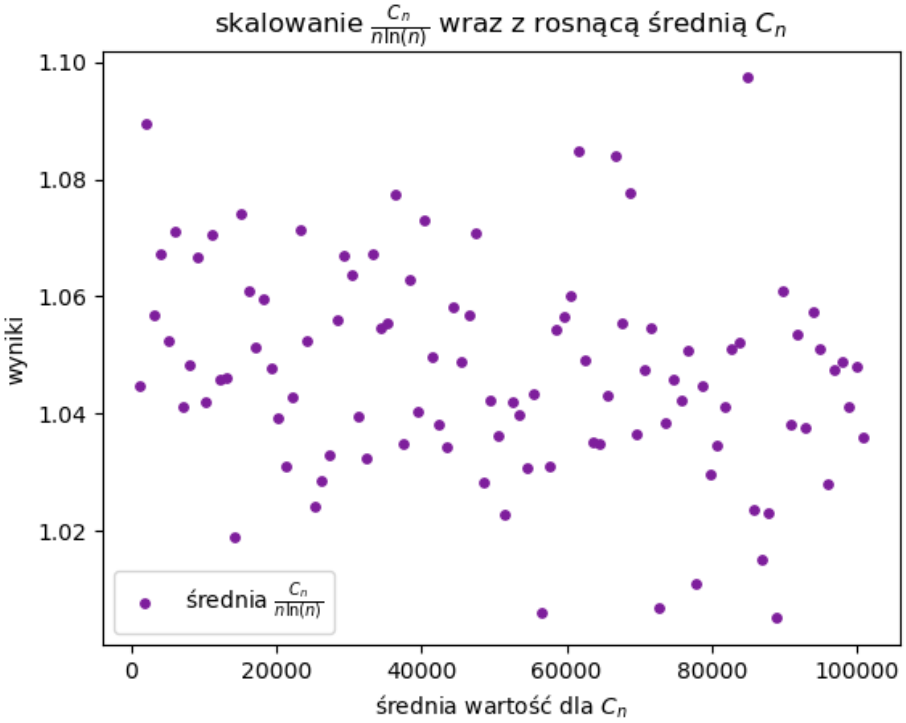
Okazuje się, że średnia liczba zakupów potrzebna do zebrania wszystkich kuponów wynosi $O(n \log n)$. Aby zrozumieć to zjawisko, przypatrzmy się szczegółowo analizie matematycznej tego problemu.

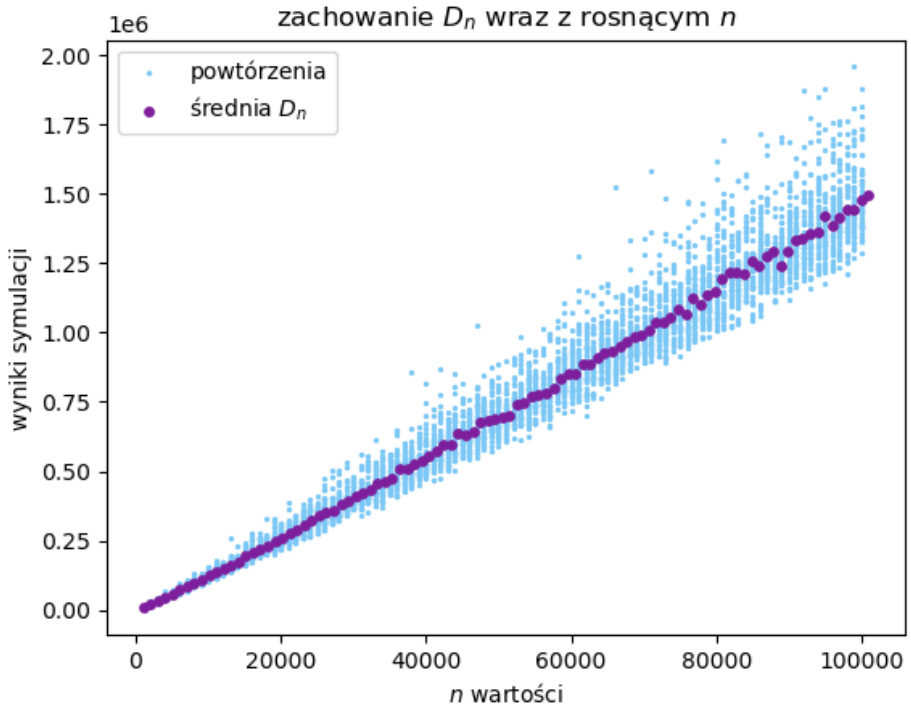
Niech C_i dla $i \in \{1, \ldots, n\}$ będzie zmienną losową oznaczającą liczbę wyrzuconych kul potrzebnych do wypełnienia i-tej urny. Wówczas zmienna losowa $C_n = \sum_{i=1}^{n} C_i$. Każda zmienna C_i ma rozkład geometryczny, co znacznie ułatwia dalsze obliczenia. Prawdopodobieństwo wypełnienia i-tej urny jest równe $p_i = 1 - \frac{i-1}{n}$. Wartość oczekiwana tych zmiennych można wyznaczyć korzystając z własności rozkładu geometrycznego: $E(C_i) = \frac{1}{p_i} = \frac{n}{n-i+1}$. Ze względu na niezależność zmiennych losowych C_i, można obliczyć wartość oczekiwaną $E(C_n)$, wykorzystując linearność wartości oczekiwanej:

$$
\begin{aligned}
E(C_n) &= E\left(\sum_{i=1}^{n} C_i\right) \\
&= \sum_{i=1}^{n} E(C_i) \\
&= \sum_{i=1}^{n} \frac{n}{n-i+1} \\
&= n \sum_{i=1}^{n} \frac{1}{n-i+1} \\
&= n \sum_{i=1}^{n} \frac{1}{i} \\
&= nH_n.
\end{aligned}
$$

Dzięki przeprowadzonym obliczeniom jesteśmy w stanie ustalić asymptotyczne tempo wzrostu wartości oczekiwanej zmiennej losowej C_n: $nH_n = O(n\ln(n)) \implies E(C_n) = O(n\ln(n))$

Potwierdzeniem tego stwierdzenia jest zamieszczony poniżej wykres, przedstawiający średnie wartości zmiennej losowej C_n przeskalowane przez $n\ln(n)$. Jak można zauważyć, wartości te wydają się koncentrować w wąskim przedziale. W rezultacie otrzymujemy funkcję, która przybliża się do funkcji stałej.





Rysunek 6: Wyniki eksperymentów dla wartości Dn.

Powyższy wykres ilustruje zmiany wartości D_n w miarę wzrostu parametru n podczas przeprowadzanej symulacji. Wraz ze zwiększaniem się wartości n, obserwuje się większe rozproszenie wyników na wykresie, co świadczy o rosnącej wartości odchylenia standardowego prób.

Zgodnie z wynikami przedstawionymi przez Erdősa i Rényiego w ich pracy pt. "On a classical problem of probability theory", możemy stwierdzić, że dla zmiennej losowej D_r, gdzie $r+1$ oznacza minimalną liczbę kul, które chcemy uzyskać we wszystkich urnach, otrzymujemy wyrażenie na wartość oczekiwaną:

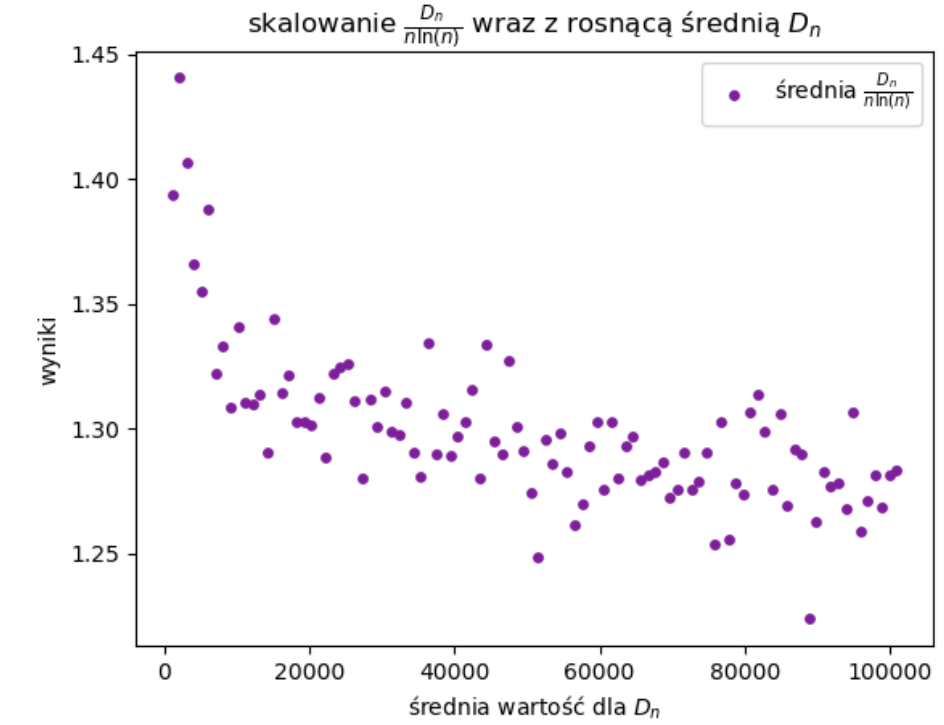
$$E(D_r) = n\ln n + rn\ln\ln n + (\gamma - \ln r!)n + O(n)$$

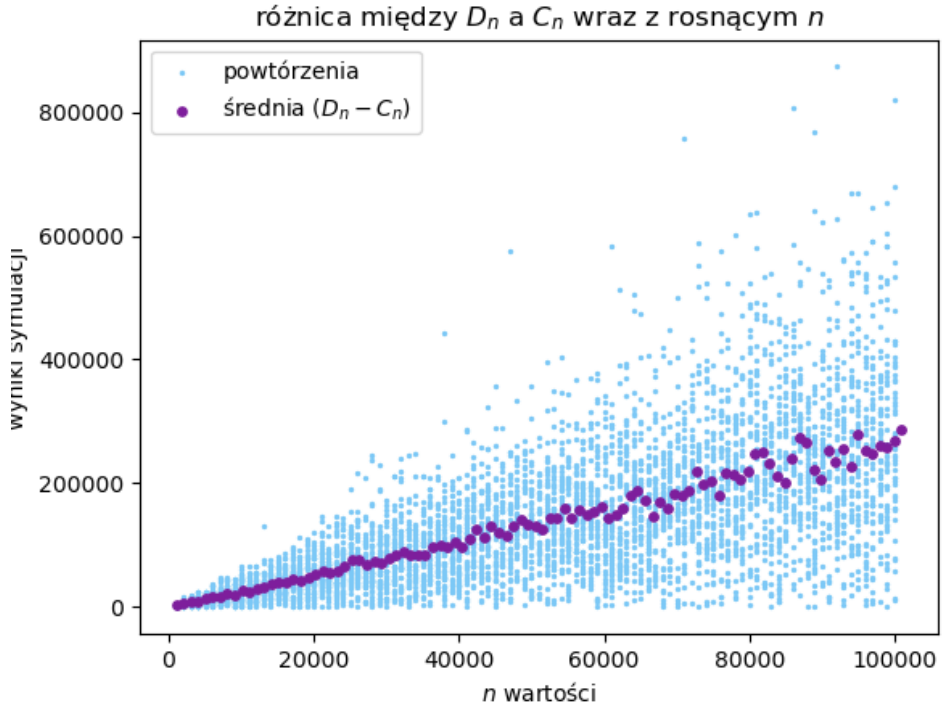
gdzie γ to stała Eulera-Mascheroniego. W wyniku przeprowadzonych obliczeń dla $r = 1$ otrzymujemy:

$$E(D_1) = n\ln n + n\ln\ln n$$

Następnie, obliczając granicę $\lim_{n\to\infty} \frac{n\ln n + n\ln\ln n}{n\ln n}$, uzyskujemy $\lim_{n\to\infty} 1 + \frac{\ln(\ln n)}{\ln n} = 1$. Na podstawie tego wyniku wnioskujemy, że $E(D_n) = O(n\ln n)$.

Potwierdzeniem tego stwierdzenia jest zamieszczony poniżej wykres, przedstawiający średnie wartości zmiennej losowej D_n przeskalowane przez $n\ln(n)$. Jak można zauważyć, wartości te wydają się koncentrować w wąskim przedziale. W rezultacie otrzymujemy funkcję, która przybliża się do funkcji stałej.





Rysunek 7: Wyniki eksperymentów dla wartości Dn - Cn.

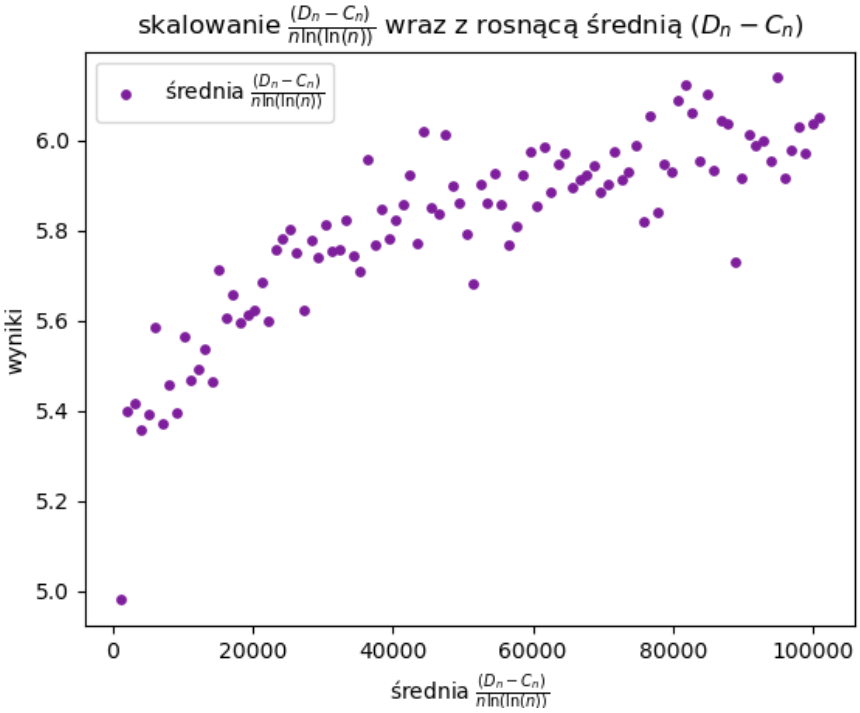
Powyższy wykres ilustruje zmiany wartości $D_n - C_n$ w miarę wzrostu parametru n podczas przeprowadzanej symulacji. Wraz ze zwiększaniem się wartości n, obserwuje się większe rozproszenie wyników na wykresie, co świadczy o rosnącej wartości odchylenia standardowego prób.

Z własności wartości oczekiwanej otrzymujemy:

$$E(D_n - C_n) = E(D_n) - E(C_n) = n \ln n + n \ln \ln n - n \ln n = n \ln \ln n.$$

Wynika z tego, że $E(D_n - C_n) = O(n \ln \ln n)$.

Potwierdzeniem tego stwierdzenia jest zamieszczony poniżej wykres, przedstawiający średnie wartości zmiennej losowej $D_n - C_n$ przeskalowane przez $n \ln(\ln(n))$. Jak można zauważyć, wartości te wydają się koncentrować w wąskim przedziale. W rezultacie otrzymujemy funkcję, która przybliża się do funkcji stałej.



5. Wnioski

Analiza modelu kul i urn w kontekście paradoksu urodzinowego czy problemu kolekcjonera kuponów ukazuje, że pozornie proste problemy mogą kryć w sobie głębokie rozważania matematyczne. Symulacje komputerowe pełnią kluczową rolę w eksperymentalnym potwierdzaniu teoretycznych wyników. Zastosowanie adekwatnych modeli i narzędzi matematycznych do analizy tych zagadnień jest istotne w obszarach takich jak kryptografia, teoria informacji oraz badania algorytmów o charakterze probabilistycznym. czy to podsumowanie ma sens