

Zadanie domowe 2

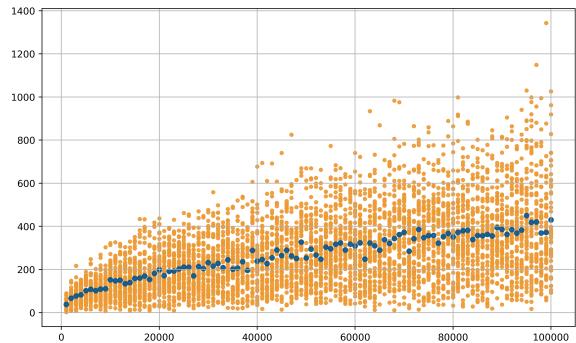
Paweł Smolnicki 283999

1 Wprowadzenie

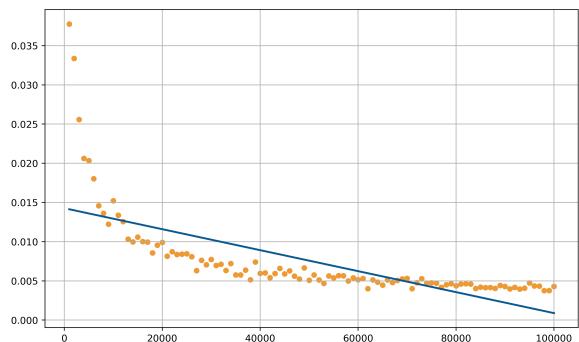
Rozwiążanie obejmuje 2 programy, jeden (main.cpp) uruchamia eksperymenty oraz zapisuje je do plików .csv. Drugi (main.py) nakreśla wykresy oraz zapisuje je do plików.

Na wykresach niebieskim kolorem zaznaczono wartość średnią, natomiast pomarańczowym - pojedyńcze wyniki eksperymentu. Dodatkowo, na wykresach aproksymujących asymptotykę danych funkcji, nałożona jest prosta dopasowania kolorem niebieskim.

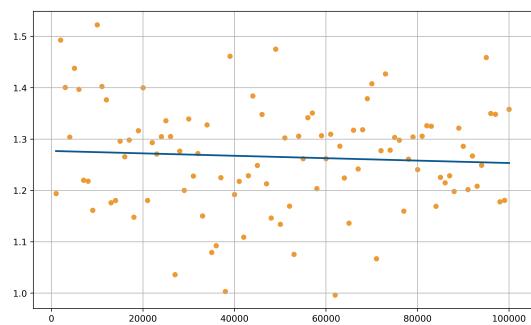
2 Paradoks urodzinowy (B_n)



(a) Wyniki eksperymentu dla pierwszej kolizji



(b) $\frac{B_n}{n}$



(c) $\frac{B_n}{\sqrt{n}}$

2.1 Wnioski

Funkcja wyznaczająca moment kolizji, ma asymptotykę $O(\sqrt{n})$ - stosunek $\frac{B_n}{\sqrt{n}}$ ma wtedy stałą wartość między 1,4 a 1,2 (co jest słuszne, gdyż przybliżenie wyniku to ok. $\sqrt{\frac{\pi}{2}}$). W przypadku $\frac{B_n}{n}$ wartość dąży do 0, co oznacza, że nie jest to asymptotyka tej funkcji

Sam paradoks dotyczy hipotetycznego problemu: Jakie jest prawdopodobieństwo, że spośród n wybranych osób co najmniej 2 będą miały urodziny w ten sam dzień. Rozwiążanie wskazuje na nieintuicyjny fakt, że wymagane są tylko 23 osoby, by przekroczyć szansę 50%, a mając 70 osób, prawdopodobieństwo to wynosi 99,9% - znacznie mniej niż można by się spodziewać.

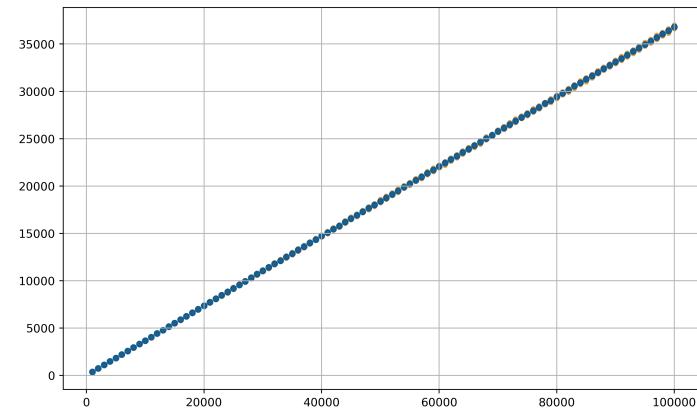
W kontekście funkcji hashujących, chociażby implementując tzw. hash mapę, należy mieć na uwadze, że kolizje bardzo szybko się pojawią - przy ok. \sqrt{n} wejściach, gdzie n to rozmiar tej mapy. Dlatego musi ona mieć sposób rozwiązywania kolizji - np. poprzez kubelkowanie (buckets). W kryptograficznych funkcjach hashujących problem ten staje się jeszcze bardziej dotkliwy, gdyż głównym ich zadaniem jest zapewnienie nieodwracalności wejścia. Dlatego przyjęto standard, aby zapewnić 2^n bitów bezpieczeństwa, należy użyć 2^{2n} bitów do wygenerowania takiego ciągu.

Niech X będzie zmienną losową oznaczającą liczbę rzutów do pierwszej kolizji, wówczas

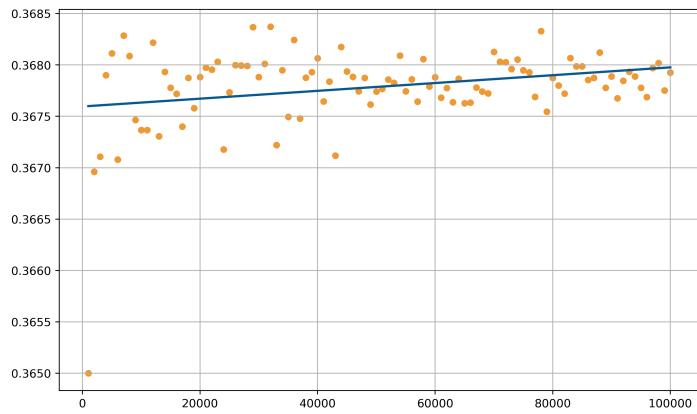
$$\begin{aligned}
 B_n &= E[X] = \sum_{k=0}^{\infty} P(X > k) \\
 P(X > k) &= \frac{n}{n} \times \frac{n-1}{n} \times \dots \times \frac{n-k+1}{n} = \prod_{i=0}^{k-1} \frac{n-i}{n} = \\
 &= \prod_{i=1}^{k-1} \left(1 - \frac{i}{n}\right) \quad \text{oraz } 1-x \approx e^{-x} \quad \text{dla } 0 \leq x \ll 1 \\
 &\approx \prod_{i=1}^{k-1} e^{-\frac{i}{n}} = e^{-\sum_{i=1}^{k-1} \frac{i}{n}} = e^{-\frac{k(k-1)}{2n}} \approx e^{-\frac{k^2}{2n}} \\
 \sum_{k=0}^{\infty} 1 \times e^{-\frac{k^2}{2n}} &\approx \int_0^{\infty} e^{-\frac{x^2}{2n}} dx \quad \text{niech } u = \frac{x}{\sqrt{2n}} \rightarrow du \sqrt{2n} = dx \\
 \int_0^{\infty} e^{-u^2} \sqrt{2n} du &= \sqrt{2n} \int_0^{\infty} e^{-u^2} du = \sqrt{2n} \sqrt{\frac{\pi}{2}} = \sqrt{\frac{\pi n}{2}}
 \end{aligned}$$

a zatem $U_n \approx \sqrt{\frac{\pi n}{2}}$

3 Puste urny po n kulach (U_n)



(a) Wyniki eksperymentu U_n - ilość pustych urn po n kulach



(b) $\frac{U_n}{n}$

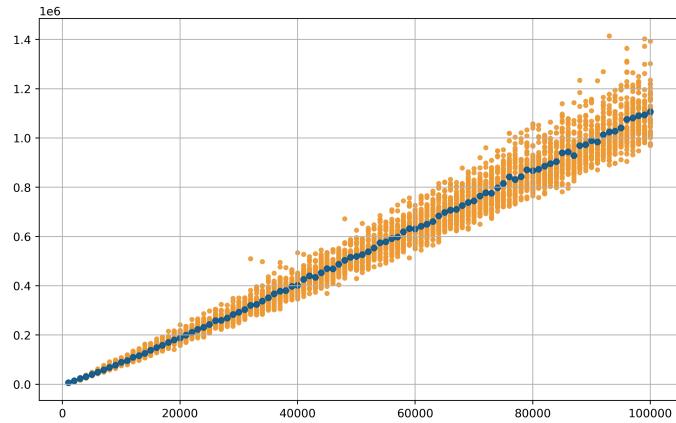
3.1 Wnioski

Na wykresie z eksperimentu pojedyńcze wykonania są mało widoczne, ponieważ wyniki oscylowały z małym odchyleniem od średniej. Widoczna jest zależność liniowa, z współczynnikiem między 0.367 a 0.368, co się zgadza z teoretycznym modelem ($U_n = \frac{n}{e}$). Oznacza to, że średnio po wrzuceniu n kul, ok. $\frac{1}{3}$ z nich dalej jest pusta, a zatem jest całkiem sporo kolizji.

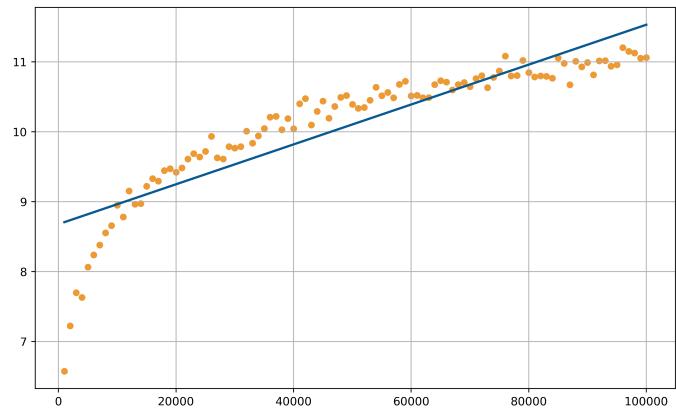
Do n urn wkładamy n kul, wtedy p_i - prawdopodobieństwo, że i-ta urna jest pusta wynosi $(1 - \frac{1}{n})^n$, bo każde włożenie kuli jest niezależne. Zdefiniuję zmienną losową X, tak że $X_i = 0$ jeśli i-ta urna ma kulę, $X_i = 1$ jeśli jest pusta. Wtedy:

$$U_n = E[X] = E[X_1 + X_2 + \dots + X_n] = \sum_{i=1}^n E[X_i] = \sum_{i=1}^n (1 - \frac{1}{n})^n = n(1 - \frac{1}{n})^n$$
$$\text{dla } n \rightarrow \infty \quad (1 + \frac{-1}{n})^n = e^{-1} \quad \text{zatem} \quad U_n \approx \frac{n}{e} \approx 0.3679 n$$

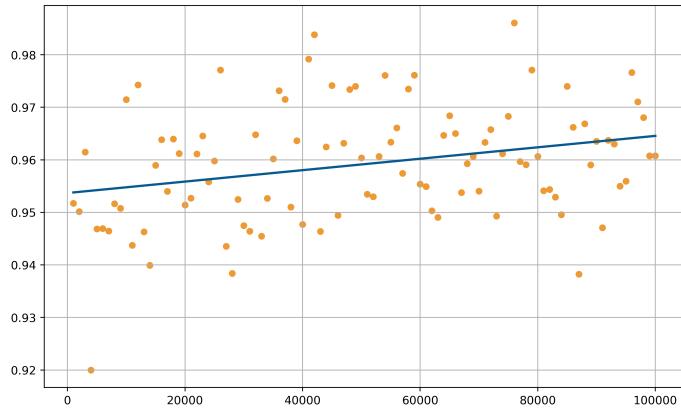
4 Problem kolekcjonera (C_n)



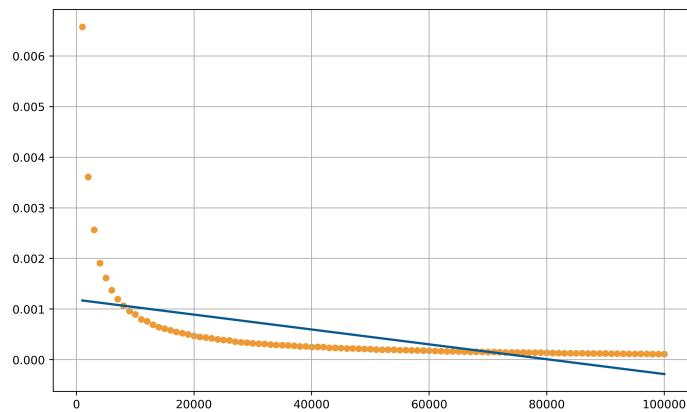
(a) Wyniki eksperymentu dla C_n



(b) $\frac{C_n}{n}$



(a) $\frac{C_n}{n \ln(n)}$



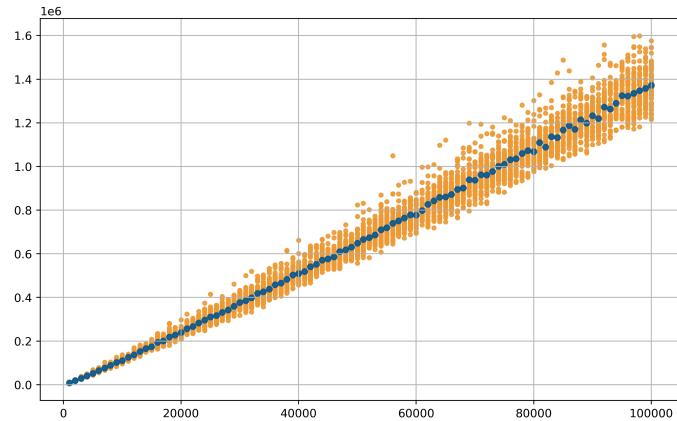
(b) $\frac{C_n}{n^2}$

4.1 Wnioski

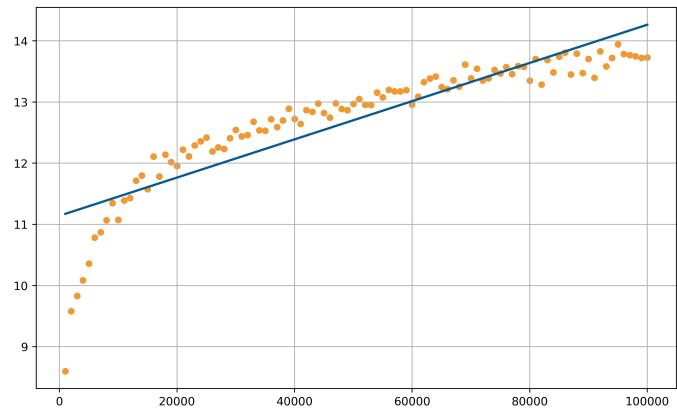
Asymptotyka tej funkcji wynosi $O(n \ln(n))$, dodatkowo współczynnik przy tym wyrazie wynosi 1, co można zauważyc dzięki zbiegającym się wartościom średniej na wykresie $\frac{C_n}{n \ln(n)}$. W przypadku $\frac{C_n}{n}$ widoczny jest logarytmiczny wzrost wartości, a dla $\frac{C_n}{n^2}$ wartości zbiegają do 0.

Problem kolekcjonera kuponów zadaje następujące pytanie: Mając n różnych kuponów, ile razy oczekujesz, że będziesz musiał wylosować (z wymianą) kuponów, by zdobyć wszystkie kupyony? Podobnie jak w paradkosie urodzinowym, wartość ta jest znacznie większa niż mogłoby się wydawać. Dodatkowo, już po n dobraniach, średnio, będziemy mieli ok. $\frac{2}{3}$ wszystkich kuponów. Jednak dobranie tych pozostałych zajmie nam jeszcze dużo czasu, ponieważ prawdopodobieństwo wylasowania kuponu, którego nie mamy jeszcze, maleje z każdym kolejnym dobraniem (nowego kuponu).

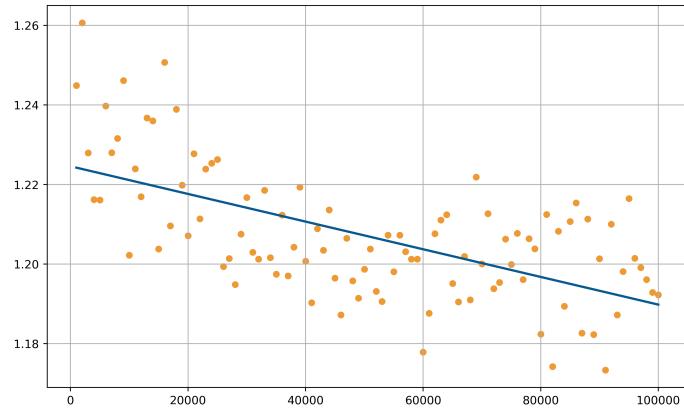
5 Podwójny problem kolekcjonera kuponów (D_n)



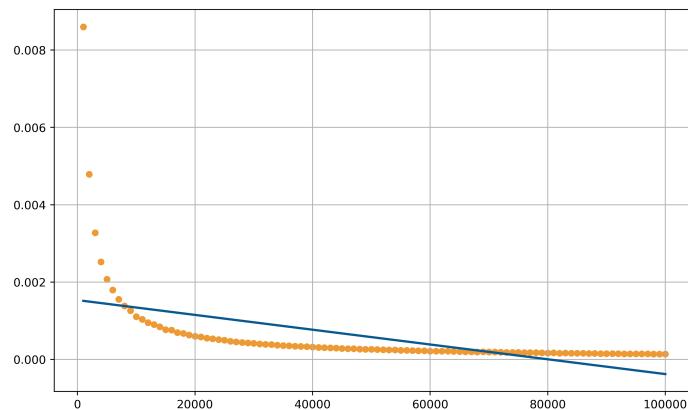
(a) Wyniki eksperymentu dla D_n



(b) $\frac{D_n}{n}$



(a) $\frac{D_n}{n \ln(n)}$



(b) $\frac{D_n}{n^2}$

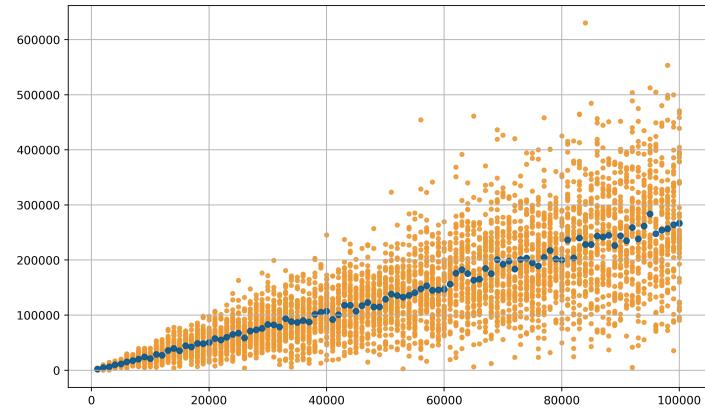
5.1 Wnioski

Zebranie po 2 kupony jest niemalże tak samo czasochłonne jak zebranie tylko 1 pełnego zestawu. Asymptotyka tej funkcji jest taka sama jak dla C_n , a stosunek $n \rightarrow \infty$ $\frac{D_n}{n \ln(n)} = \frac{C_n}{n \ln(n)} = 1$.

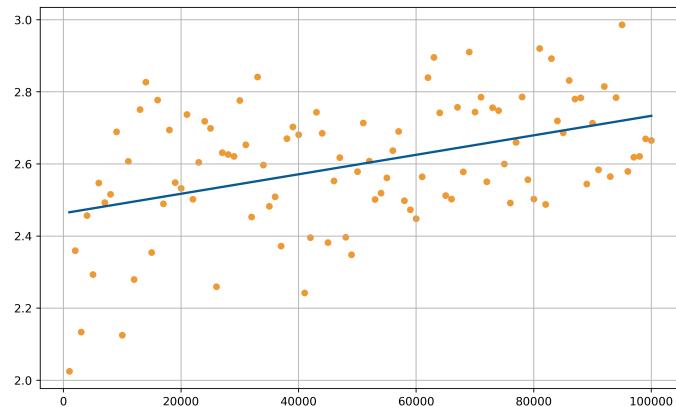
Można by myśleć, że skoro zebranie 1 całego kompletu zajmuje średnio C_n to zebranie podwójnego kompletu będzie wynosiło $2C_n$, jednak nie jest tak, ponieważ podczas zbierania kuponów do pełnego kompletu, natrafimy na wiele kuponów, które już mamy (średnio będziemy mieć po $\ln(n)$ po każdym kuponie). Wówczas dobranie do podwójnego zestawu, kończy się "polowaniem" na małą, nieskompletowaną ilość kuponów.

Na wykresie $\frac{D_n}{n \ln(n)}$ widoczny jest znaczący spadek wartości, ponieważ dokładniejsze przybliżenie D_n ma też wyraz $O(n \ln(\ln(n)))$ co dla małych n ma większe znaczenie. Zwiększa to wynik w tym ilorazie dla mniejszych n .

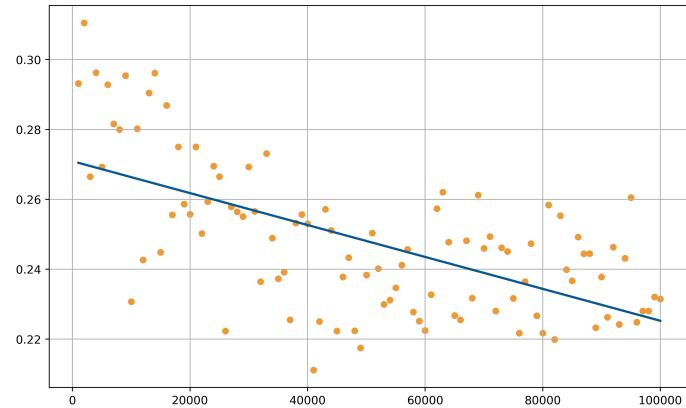
6 Różnica $(D_n - C_n)$



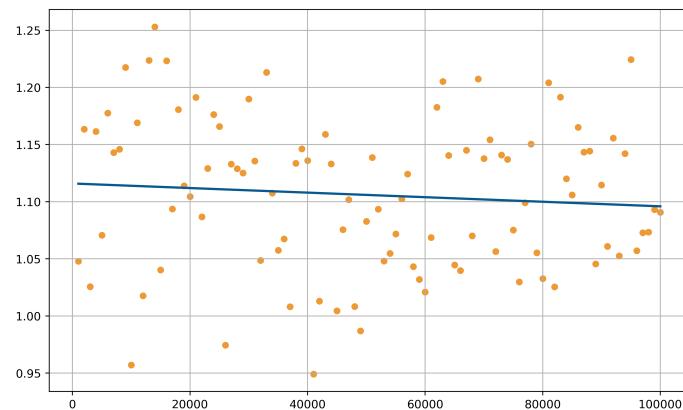
(a) Wyniki eksperymentu dla $D_n - C_n$



(b) $\frac{D_n - C_n}{n}$



(a) $\frac{D_n - C_n}{n \ln(n)}$



(b) $\frac{D_n - C_n}{n \ln(\ln(n))}$

6.1 Wnioski

Liczba kul wymaganych, by przejść z co najmniej 1 kuli we wszystkich urnach do 2, rośnie z asymptotyką $n \ln(\ln(n))$, co ciekawe nie tak jak C_n i D_n . Jednak jest to zrozumiałe, gdyż po wypełnieniu urn co najmniej 1 kulą, w wielu z nich wylądowało więcej niż 1 kula, z tą też ta funkcja rośnie "wolniej". Tzn potrzeba mniej kul by przejść z niepustych urn do każdej urny wypełnionej co najmniej dwoma kulami, niż zaczynając od 0 do niepustych.