Metody probabilistyczne i statystyka
Zadanie domowe 2
Illia Azler 239537

Celem zadania była implementacja symulacji polegającej na wykonaniu dla każdego $n \in \{1000, 2000, \ldots., 100000\}$ po $k = 50$ niezależnych powtórzeń eksperymentu wrzucania kul do urn i zapisaniu wszystkich powyższych statystyk. Dla eksperymentu wyznaczyliśmy następujące wielkości:

(a) B_{n} – moment pierwszej kolizji; $B_{\mathrm{n}} = k$, jeśli $k - ta$ z wrzucanych kul jest pierwszą, która trafiła do niepustej urny.

(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.

(d) D_{n} – minimalna liczba rzutów, po której w każdej z urn są co najmniej dwie kule.

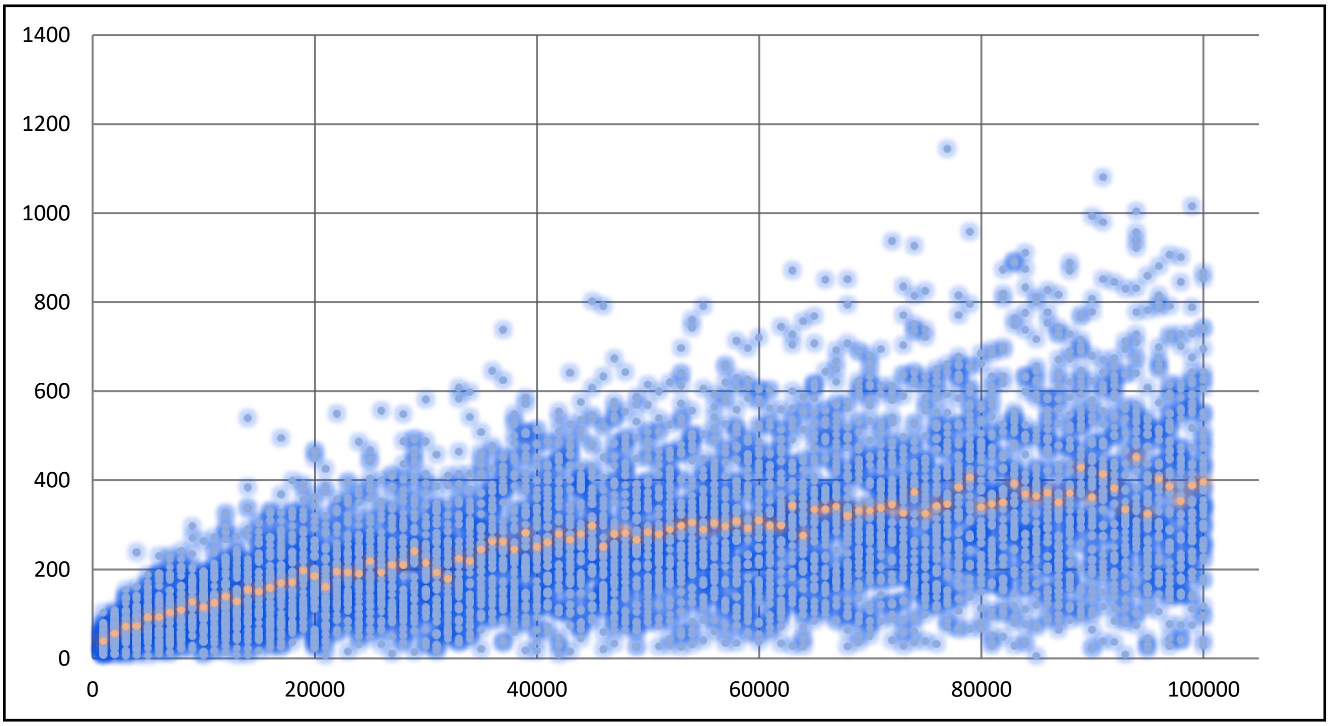
(e) $D_{\mathrm{n}} - C_{\mathrm{n}}$ – liczba rzutów od momentu C_{n} potrzeba do tego, żeby w każdej urnie były co najmniej dwie kule.

Implementacja zadania 1 za pomocą językų ***Python*** przedstawiona w załączonym pliku ***ppb_z2_kod_239537.py***, w programie użyty generator liczb pseudolosowych ***SecureRandom***.

Test 2(a,b):

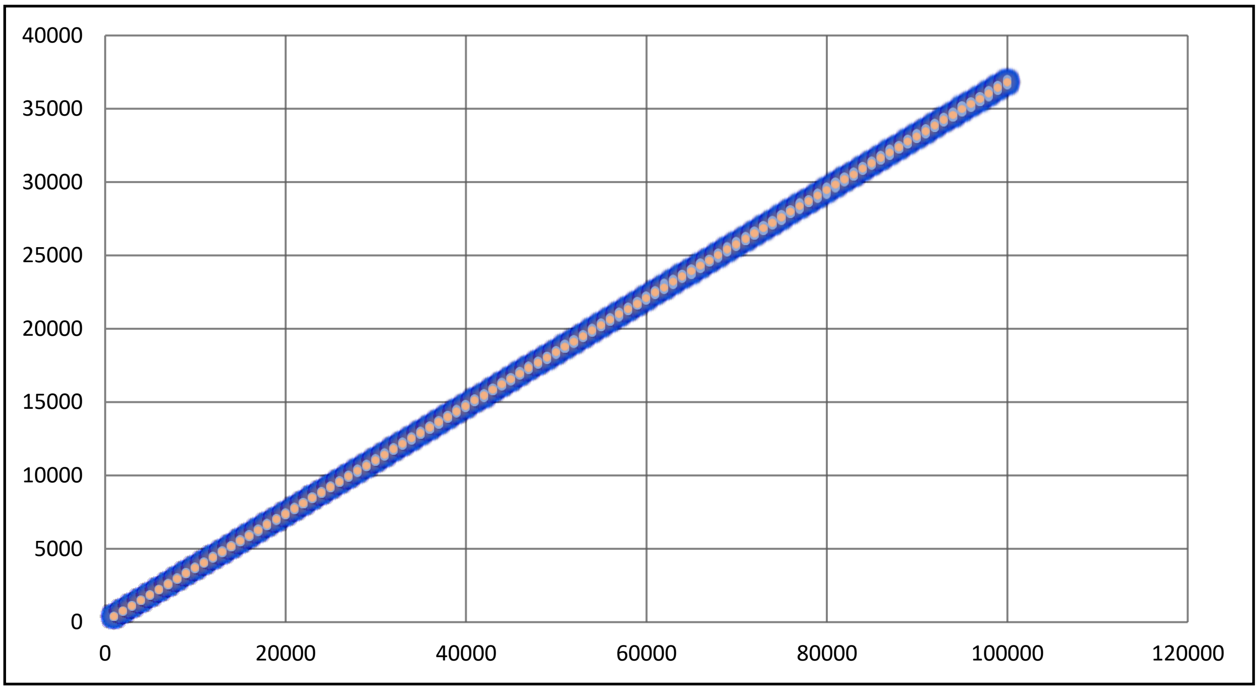
Na postawie uzyskanych wyników dostałem następne wykresy dla wielkości. Wyniki poszczególnych powtórzeń są reprezentowane przez niebieskie punkty, natomiast pomarańczowe punkty odzwierciedlają średnią wartość dla każdej próby.

B_{n}:



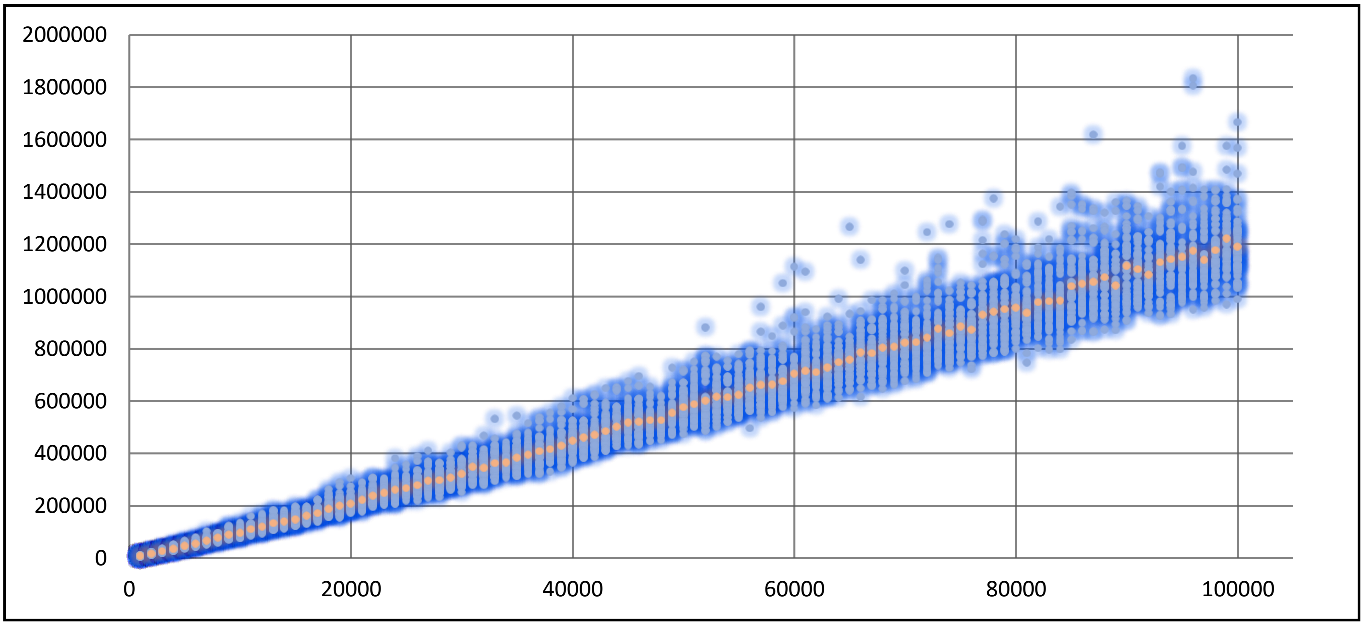
__Uwagi:__ Zauważalne są rozbieżności w wartościach momentów kolizji w różnych próbach.

U_n:



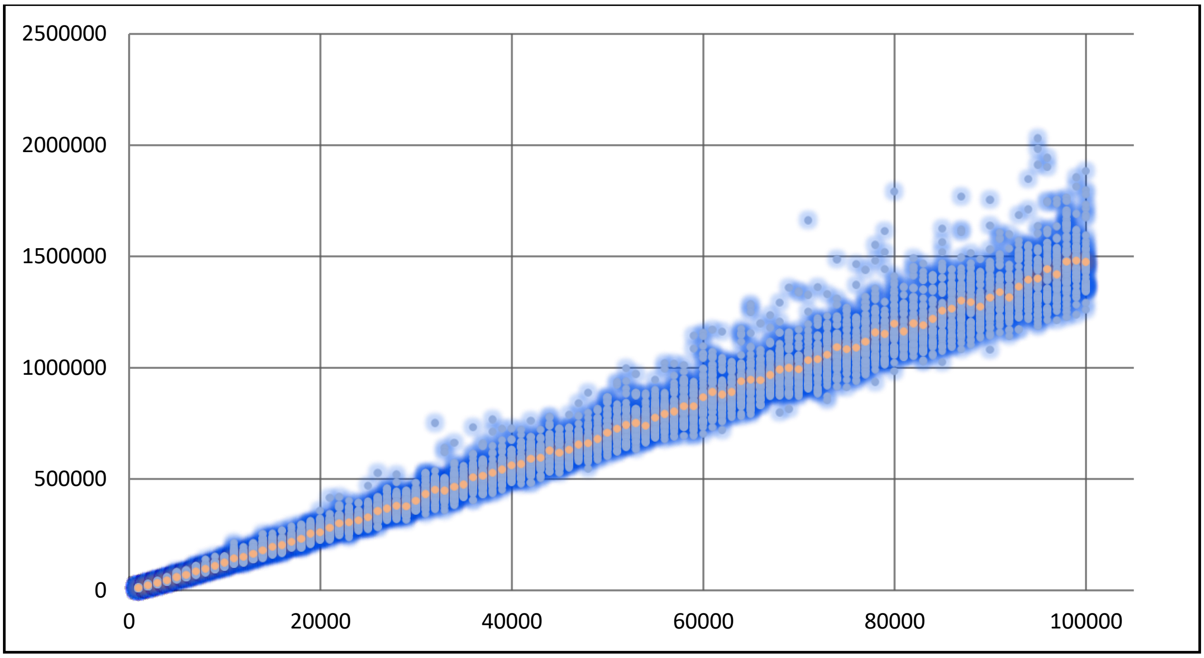
__Uwagi:__ Ilość pustych urn rośnie w sposób liniowy po wrzuceniu n kul, a wyniki poszczególnych iteracji zbliżają się do wartości średniej, czyli mamy minimalne odchylenia.

C_n:



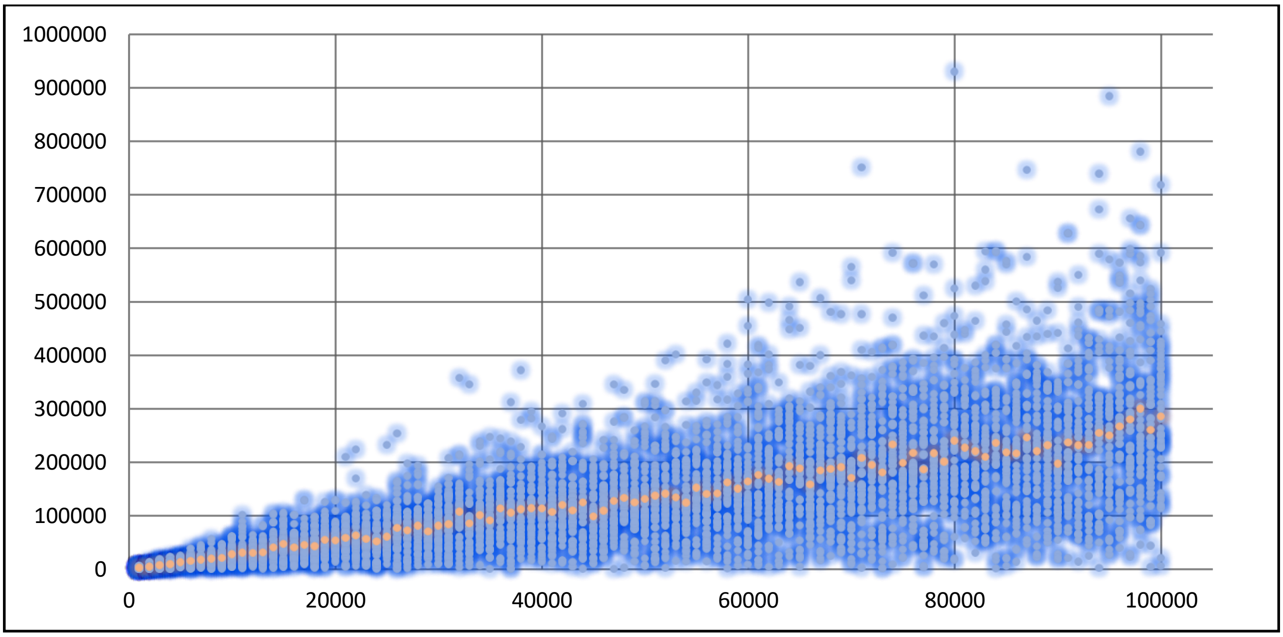
__Uwagi:__ Wyniki symulacji różnią się między sobą w stopniu ich wzrostu.

D_n:



__Uwagi:__ Rosną symetrycznie liniowo i nieco rozbieżne.

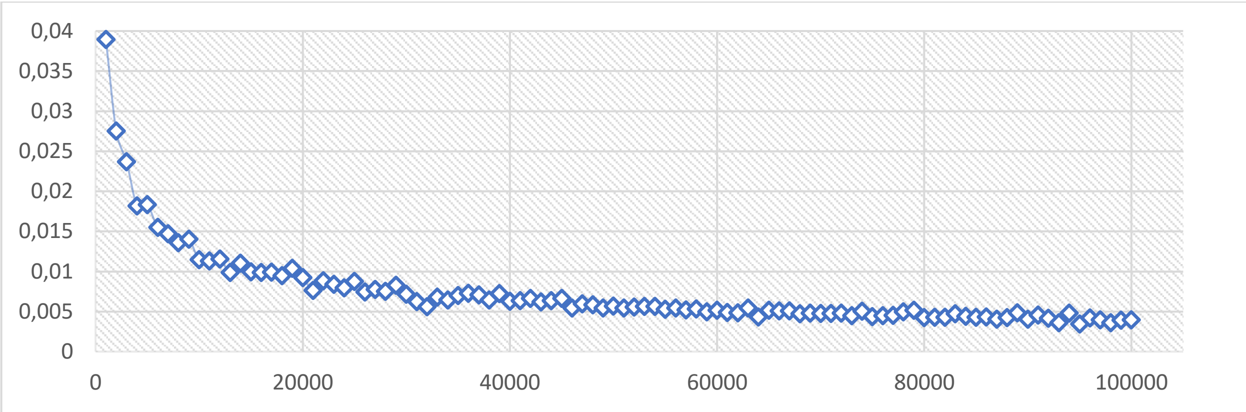
$D_n - C_n$:



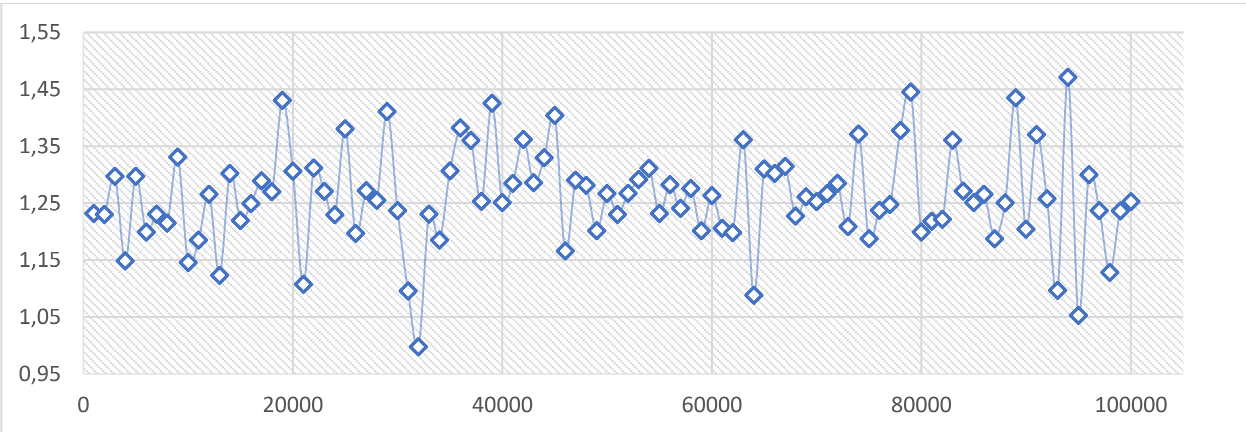
__Uwagi:__ Wyniki średniej w pełnym wielkim stopniu różnią się pod względem tempa wzrostu, przy czym obserwuje się symetryczny wzorzec wzrostu linowego.

Test 2(c):

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

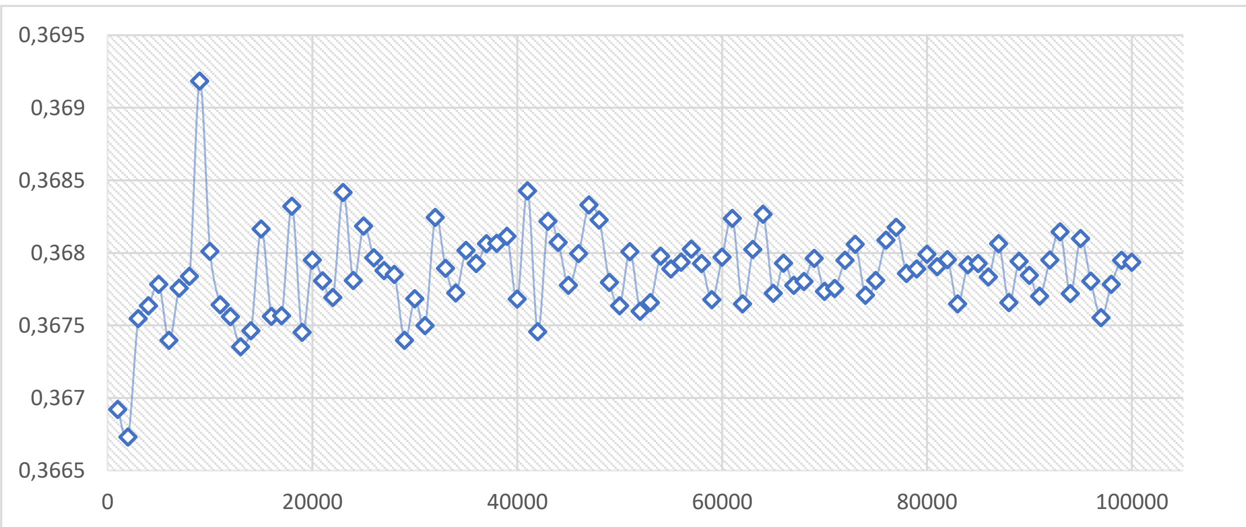


$$\frac{b(n)}{\sqrt{n}}$$



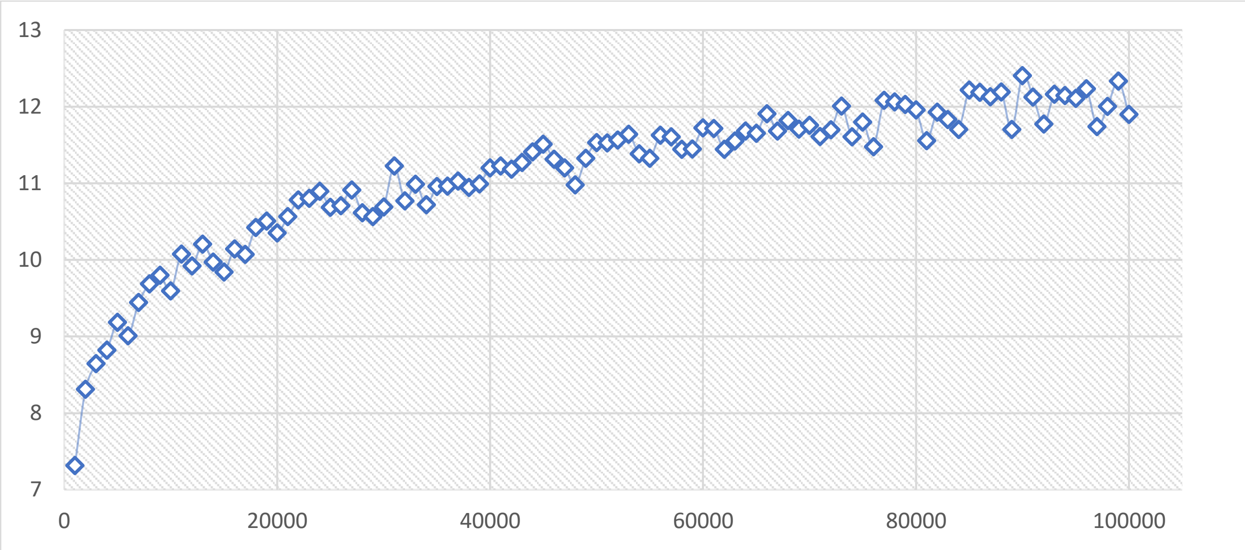
Asymptota wartości średniej: $B_{\mathrm{n}} = O(\sqrt{n})$

$$\frac{u(n)}{n}$$

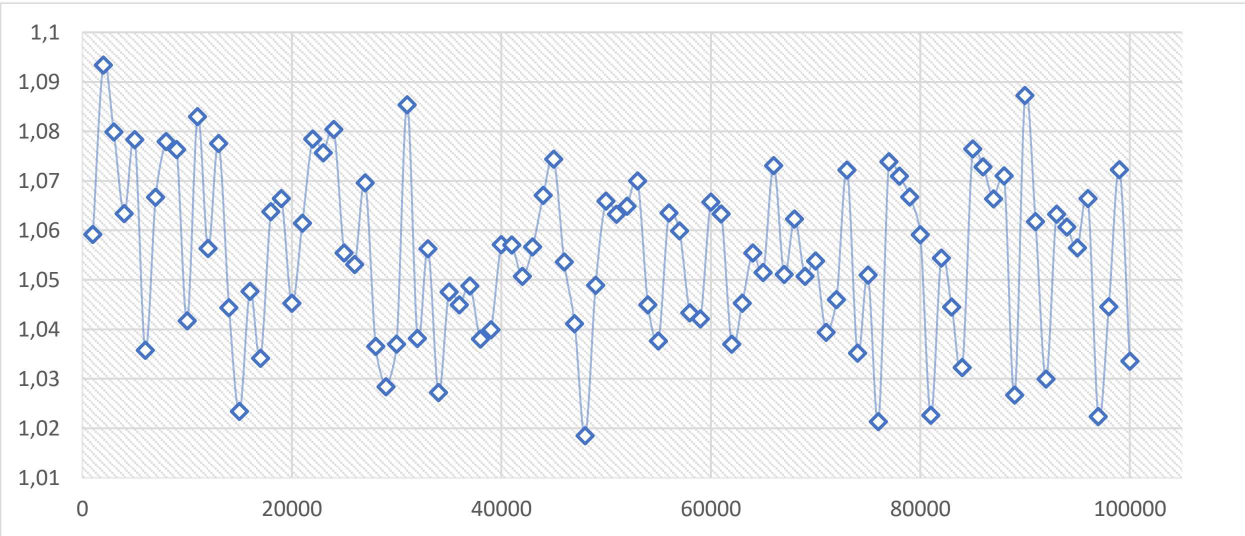


Asymptota wartości średniej: $\mathrm{U_n} = O(\sqrt{n})$

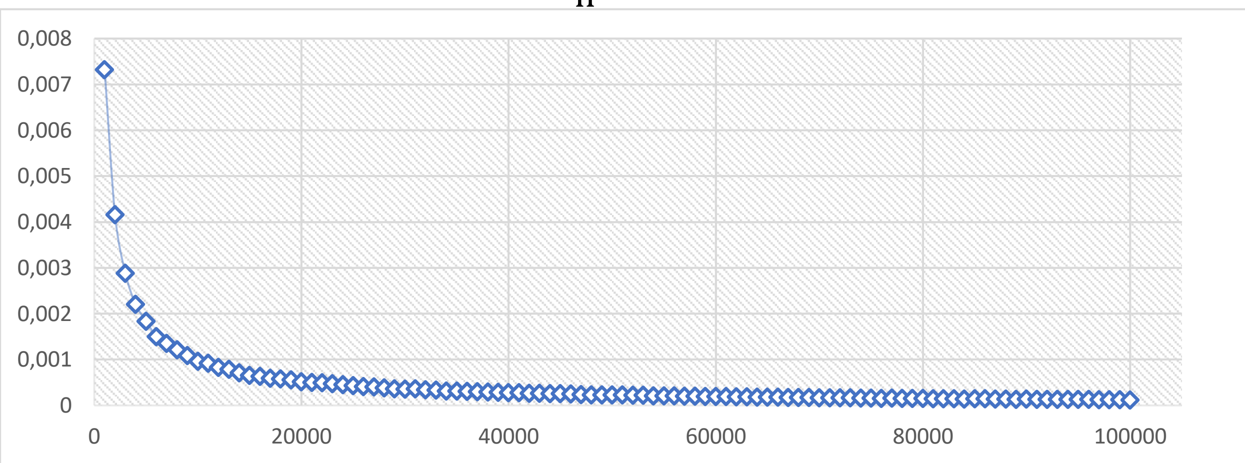
$$\frac{\mathrm{c(n)}}{\mathrm{n}}$$



$$\frac{\mathrm{c(n)}}{\mathrm{n*ln(n)}}$$

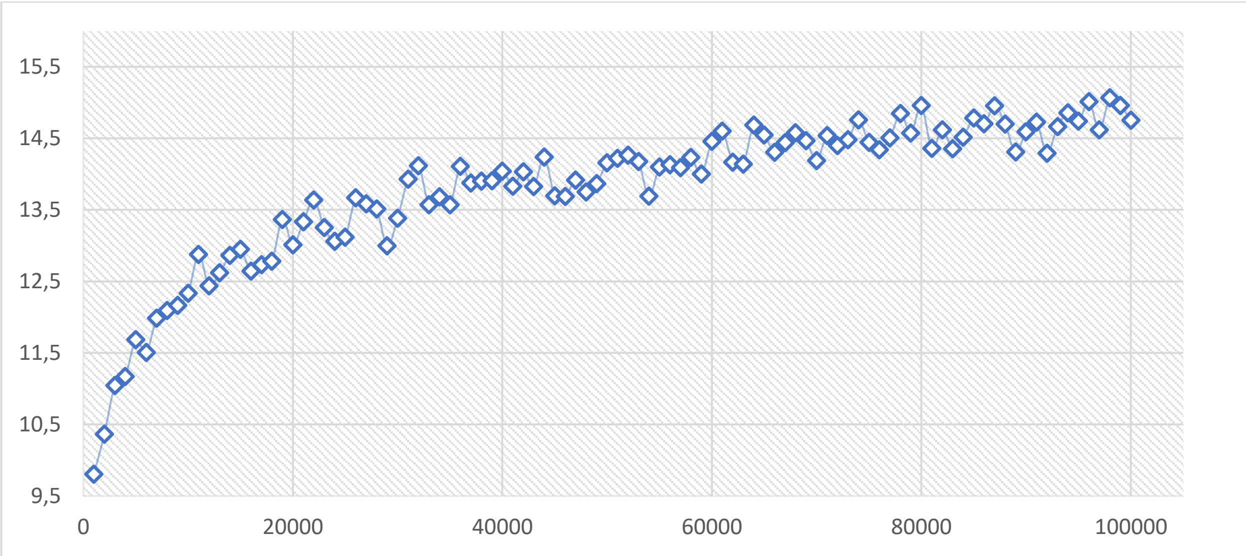


$$\frac{\mathrm{c(n)}}{\mathrm{n^2}}$$

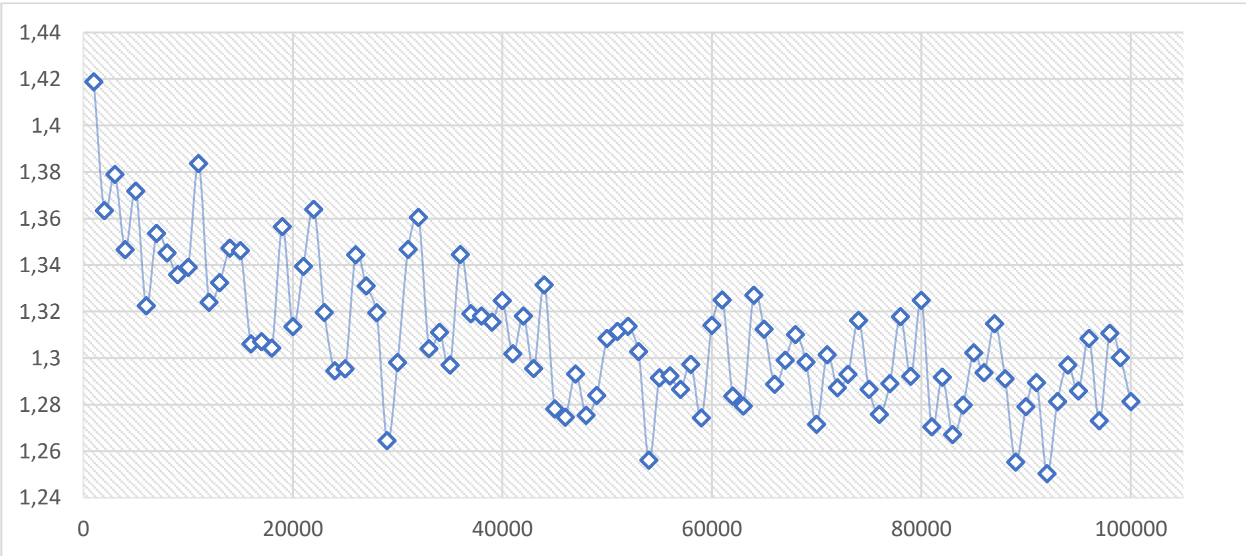


Asymptota wartości średniej: $C_n = O(n * \ln(n))$

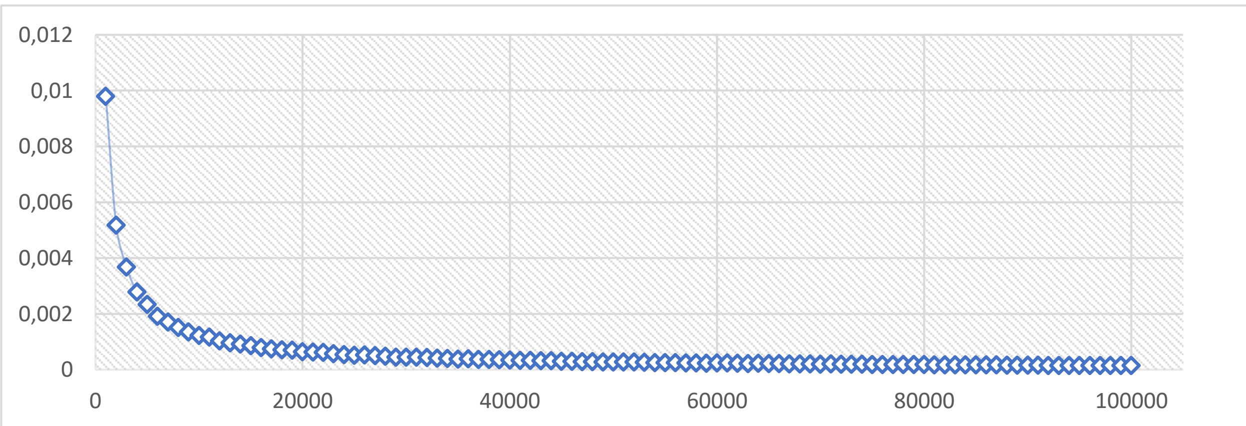
$$\frac{d(n)}{n}$$



$$\frac{d(n)}{n*\ln(n)}$$

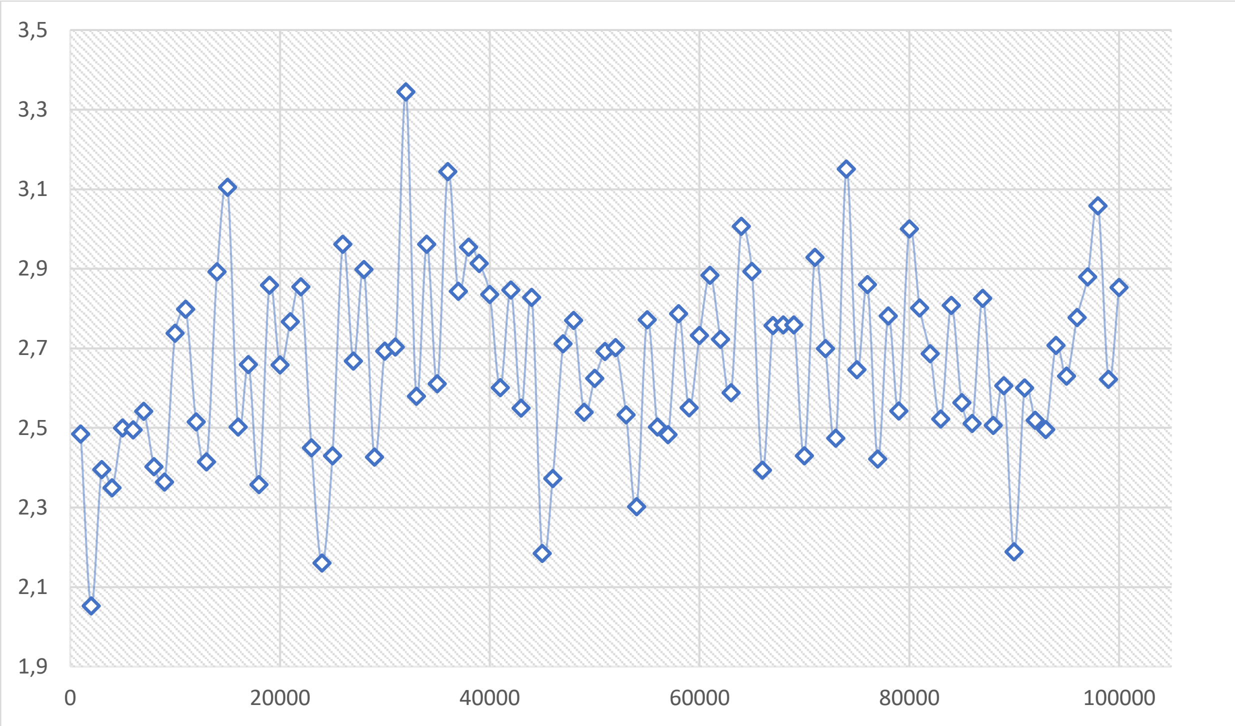


$$\frac{d(n)}{n^2}$$

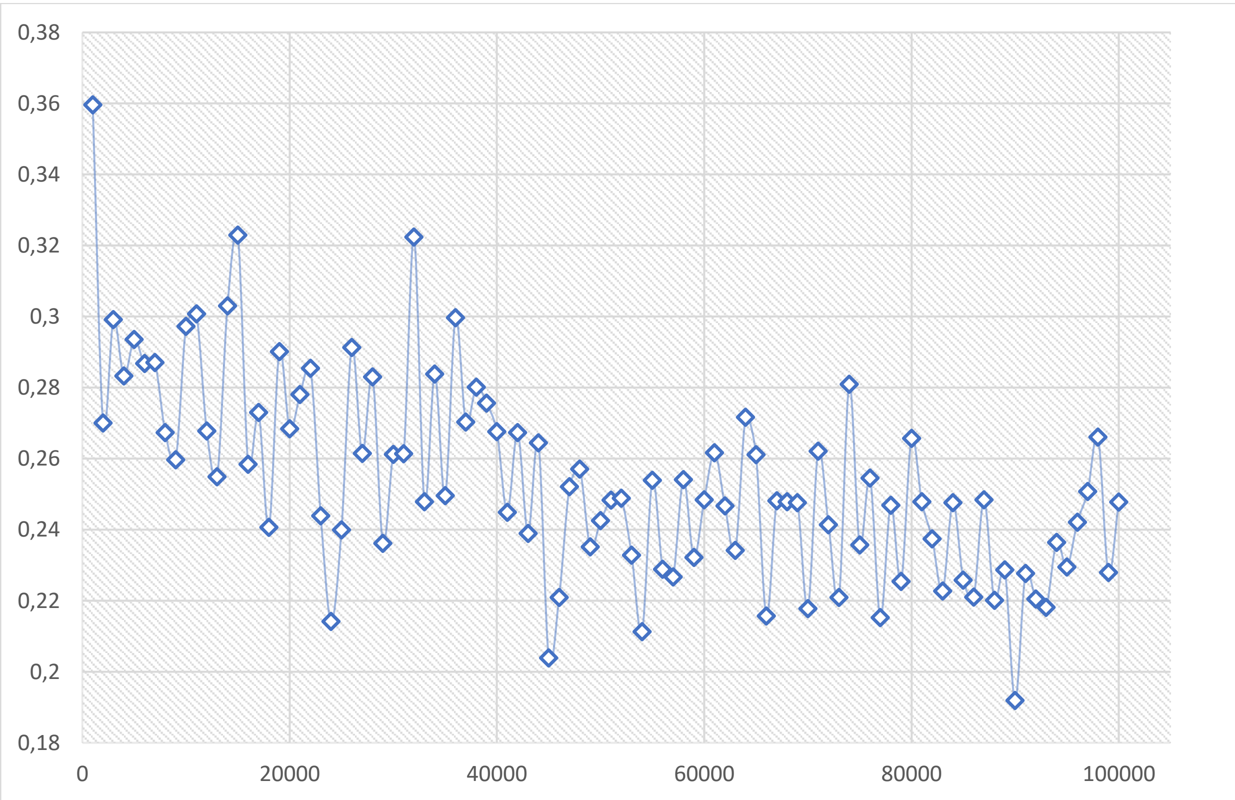


Asymptota wartości średniej: $D_n = O(n * \ln(n))$

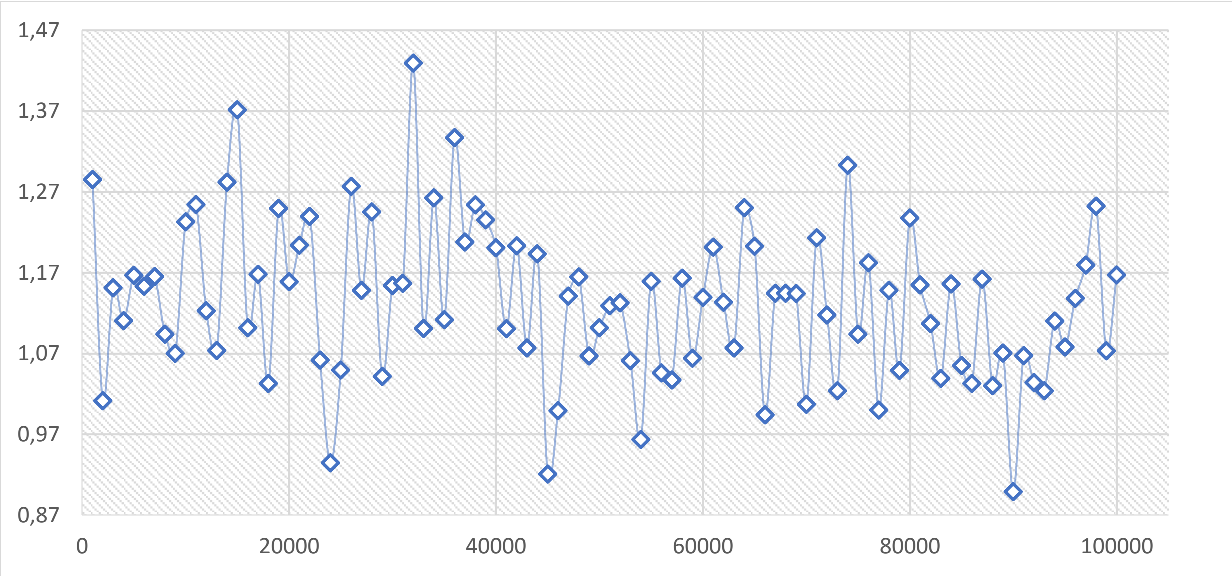
$$\frac{d(n) - c(n)}{n}$$



$$\frac{d(n) - c(n)}{n*\ln(n)}$$



$$\frac{\mathrm{d(n)\ -\ c(n)}}{\mathrm{n*ln(ln(n))}}$$



Asymptota wartości średniej: $D_{\mathrm{n}} - C_{\mathrm{n}} = O(\mathrm{n} * \ln(\ln(\mathrm{n})))$

<u>Test 2(d):</u>

Nazwa "Birthday Paradox" w kontekście tego zadania może być uzasadniona poprzez moment B_{n}, który reprezentuje pierwszą kolizję. W przypadku paradoksu urodzinowego, czyli "Birthday Paradox", chodzi o to, że w grupie stosunkowo niewielkiej liczby osób istnieje duże prawdopodobieństwo, że dwie z nich obchodzą urodziny tego samego dnia. W zadaniu, moment Bn oznacza pierwszą kolizję, czyli pierwszy raz, gdy dwie kule trafiają do tej samej urny. Analogicznie do paradoksu urodzinowego, moment Bn wskazuje, że już przy stosunkowo niewielkiej liczbie urn istnieje duże prawdopodobieństwo, że dwie kule trafią do tej samej urny.

Natomiast nazwa "Coupon Collector's Problem" może być uzasadniona poprzez zmienne $\mathrm{C_n}$ i $\mathrm{D_n}$. W przypadku problemu kolekcjonera kuponów, kolekcjoner dąży do zebrania pełnej kolekcji, a zmienna $\mathrm{C_n}$ oznacza minimalną liczbę rzutów potrzebną do uzyskania co najmniej jednej kuli w każdej urnie. To jest pierwszy moment, w którym każda urna zawiera przynajmniej jedną kulę, co można porównać do momentu, w którym kolekcjoner zdobywa swoje pierwsze unikalne karty w zbiorze. Zmienna $\mathrm{D_n}$ reprezentuje minimalną liczbę rzutów potrzebną do uzyskania co najmniej dwóch kul w każdej urnie, co można porównać do momentu, w którym kolekcjoner zdobywa dwie karty każdego typu w swojej kolekcji.

Intuicja za tymi nazwami polega na analogii między eksperymentem z kulami i urnami, a znanymi problemami matematycznymi, które mają podobne struktury i cechy. Wybór tych nazw pomaga zrozumieć istotę problemów i intuicyjnie podejść do analizy wyników symulacji.

Test 2(e):

W kontekście funkcji haszujących i kryptograficznych funkcji haszujących, problem "Birthday Paradox" ma istotne znaczenie ze względu na zastosowanie funkcji haszujących i prawdopodobieństwo kolizji.

"Paradoks urodzinowy" w kontekście funkcji haszujących oznacza sytuację, w której, pomimo dużej przestrzeni haszowej, istnieje znaczne prawdopodobieństwo, że dwie różne dane wejściowe spowodują wygenerowanie tego samego wartości skrótu (kolizji).

W przypadku funkcji haszujących używanych w kryptografii, takie kolizje są niepożądane, ponieważ prowadzą do utraty integralności i bezpieczeństwa systemów. Odporność na kolizje jest jednym z kluczowych kryteriów, które kryptograficzne funkcje haszujące muszą spełniać.

Algorytmy hashujące używane w kryptografii starają się zapewnić, że nawet minimalne zmiany w danych wejściowych powodują znaczące i trudne do przewidzenia zmiany w wartości skrótu. Paradoks urodzinowy wymaga większej przestrzeni haszowej, aby zminimalizować ryzyko kolizji, co jest jednym z powodów, dla których nowoczesne algorytmy hashujące posiadają długie długości skrótu.

Ogólnie rzecz biorąc, birthday paradox przypomina nam o tym, że przyjęcie bezpiecznych algorytmów haszujących wymaga staranności w doborze odpowiednich długości skrótu oraz regularnej aktualizacji algorytmów, aby utrzymać ich odporność na coraz bardziej zaawansowane ataki.