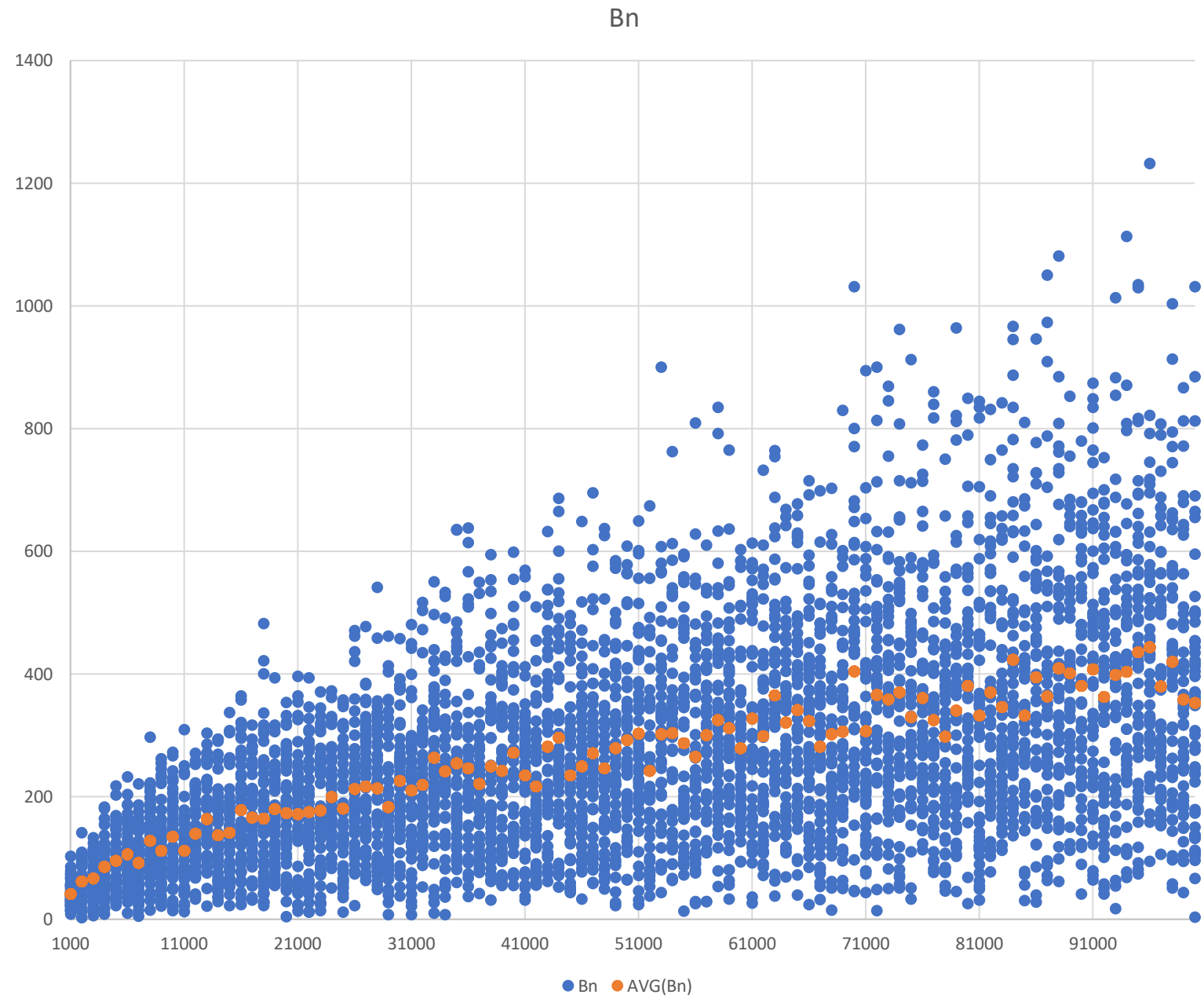
Łukasz Machnik

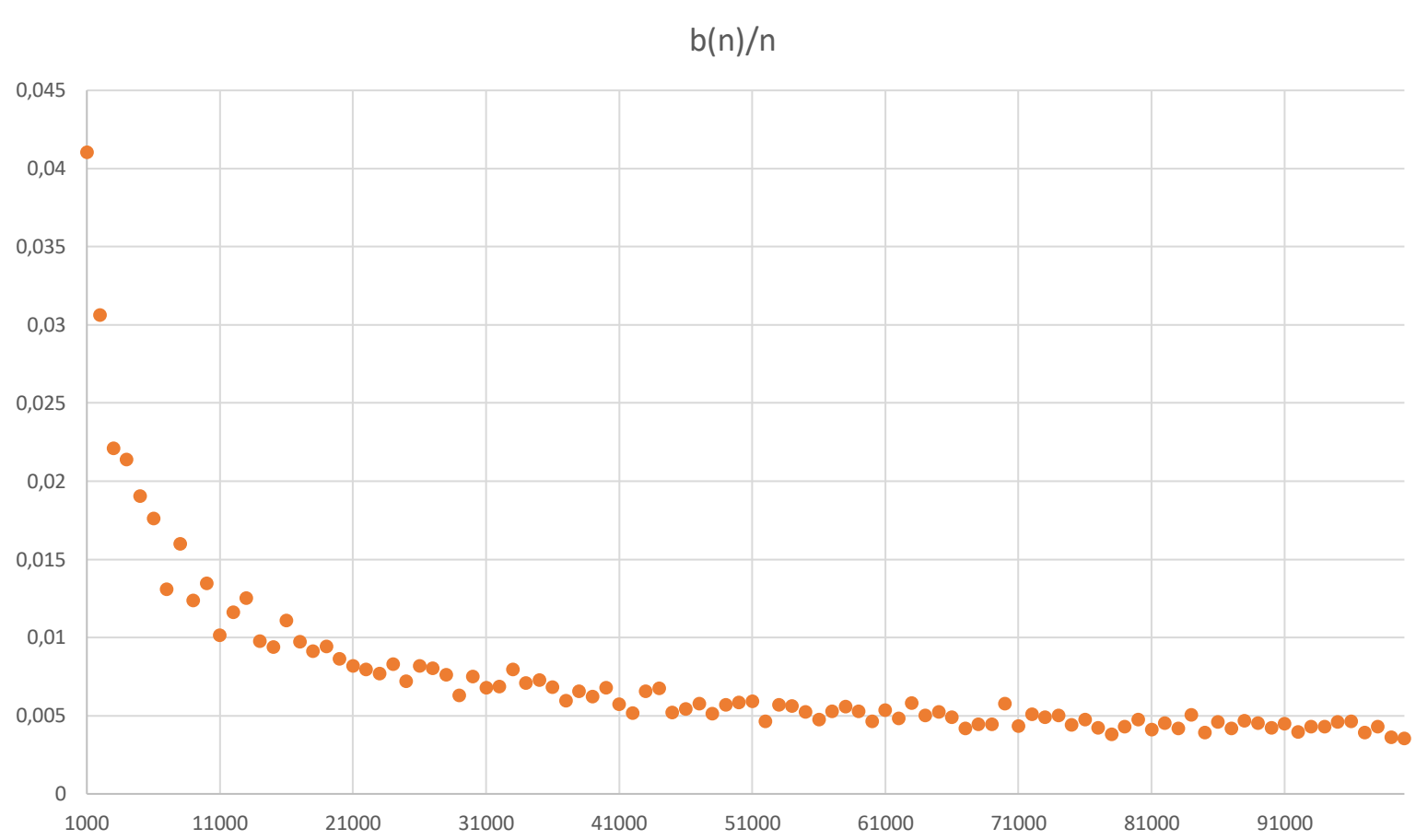
Metody probabilistyczne i statystyka

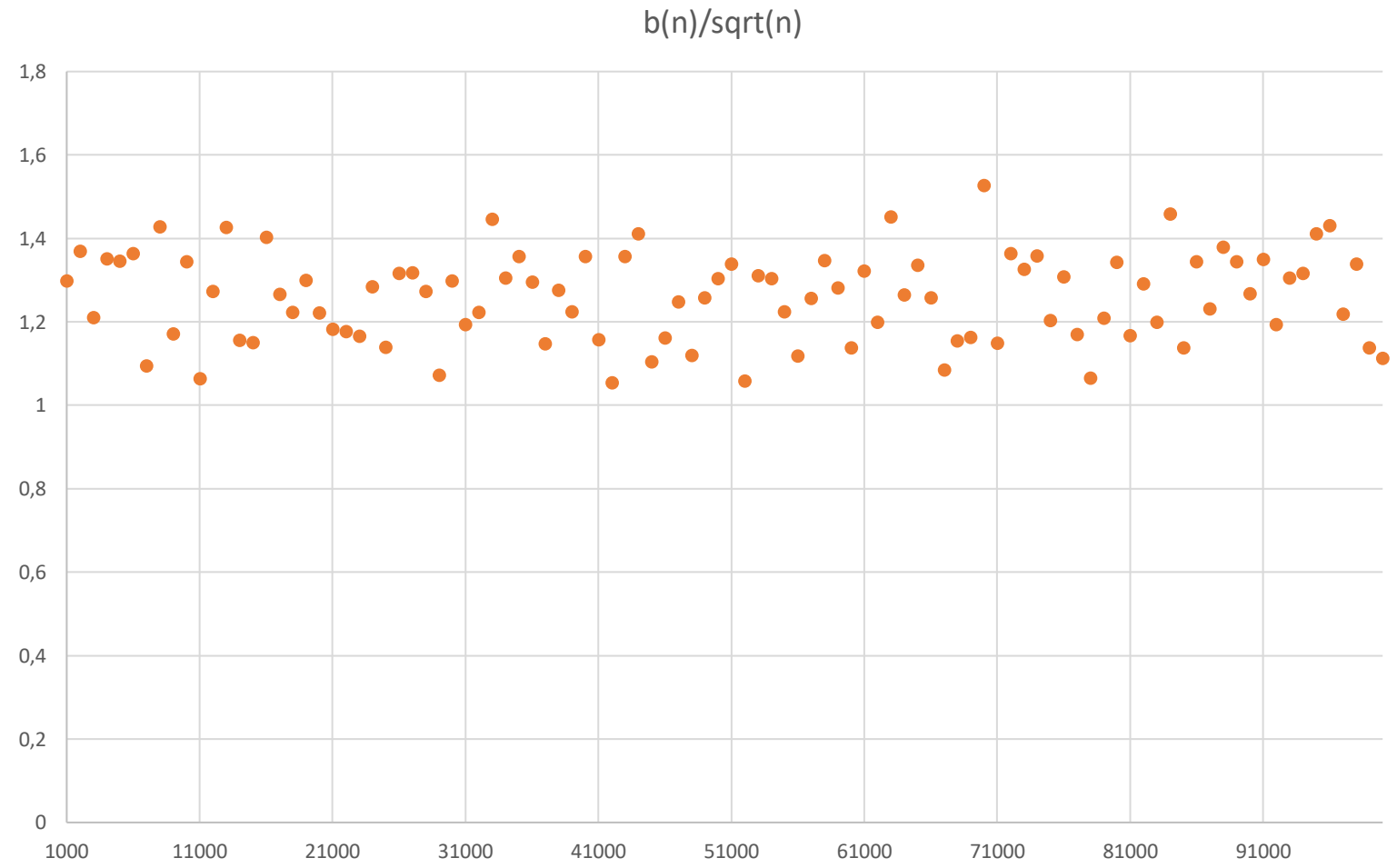
Homework 2



B_n – moment pierwszej kolizji, numer pierwszej kuli wrzuconej do niepustej urny.

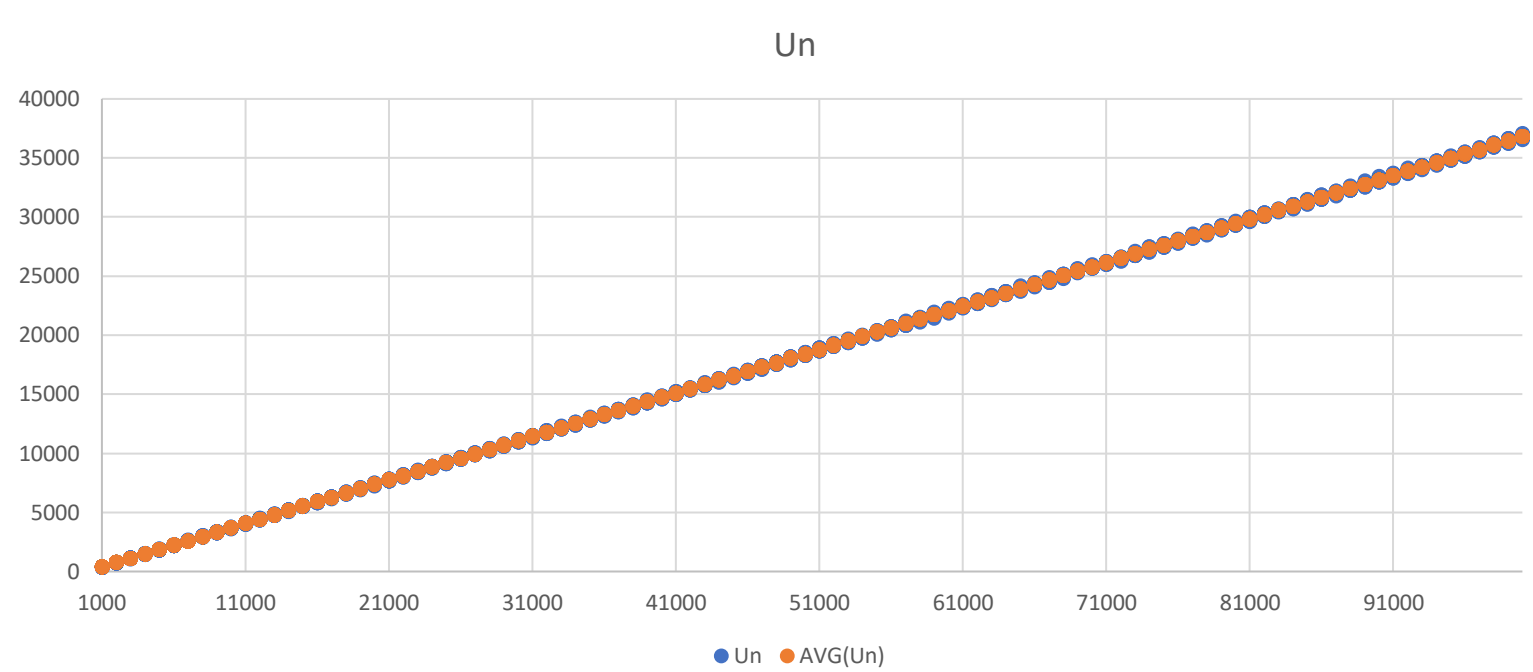
Widać że moment B_n następuje stosunkowo szybko bo przy 1000 urn średnio już 41. kula jest wrzucana do niepustej urny. Zgodnie z intuicją widać też że im więcej urn tym później średnio ten moment następuje, jednak wzrost ten jest bardzo powolny, wykładniczy i nawet dla 100000 urn w jednym z wygenerowanych przeze mnie przypadków taka sytuacja nastąpiła już przy 3. kuli. Wartości poszczególnych powtórzeń są tym bardziej rozbieżne od średniej czym więcej urn rozważamy: dla 1000 urn wartości B_n są bardzo skoncentrowane i różnica między największą a najmniejszą z nich to tylko 94, natomiast dla 96000 różnica między skrajnymi wartościami to aż 1138.





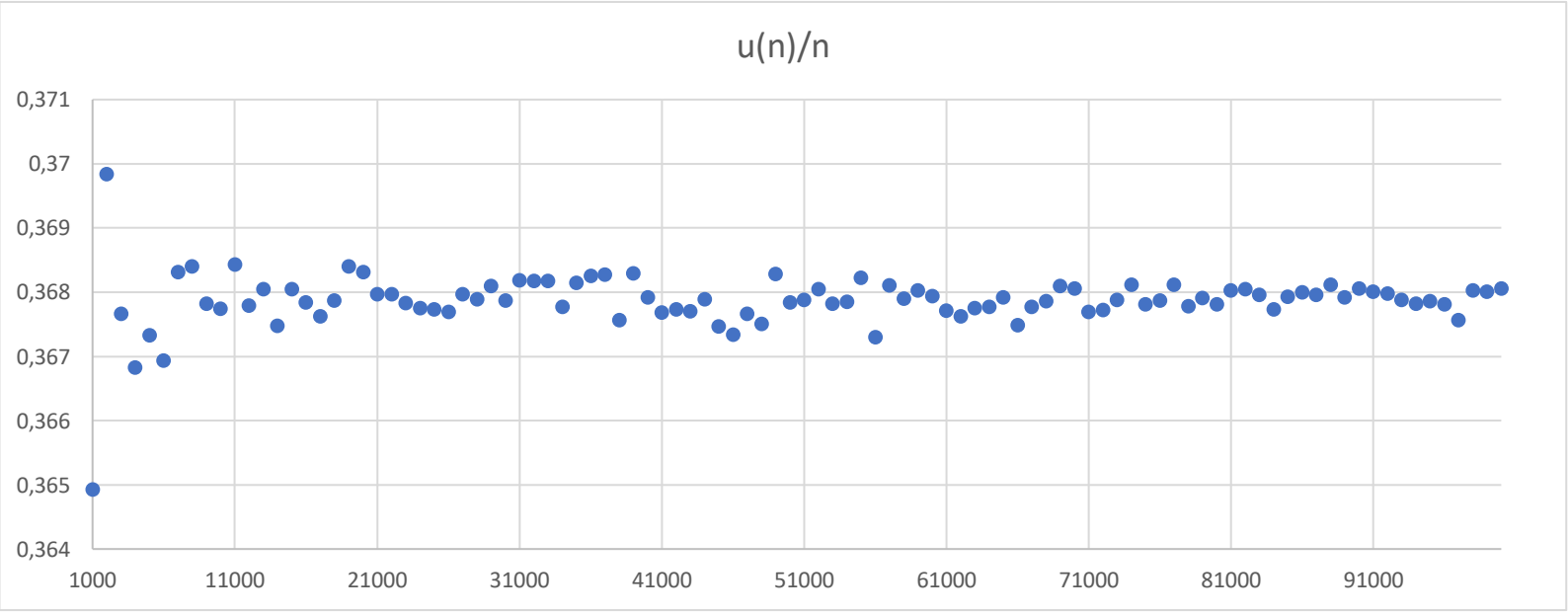
Stosunek momentu pierwszej kolizji do ilości urn to funkcja malejąca wykładniczo i dążąca do 0.

Stosunek momentu pierwszej kolizji do pierwiastka ilości urn to funkcja stała na poziomie ok. 1,25.

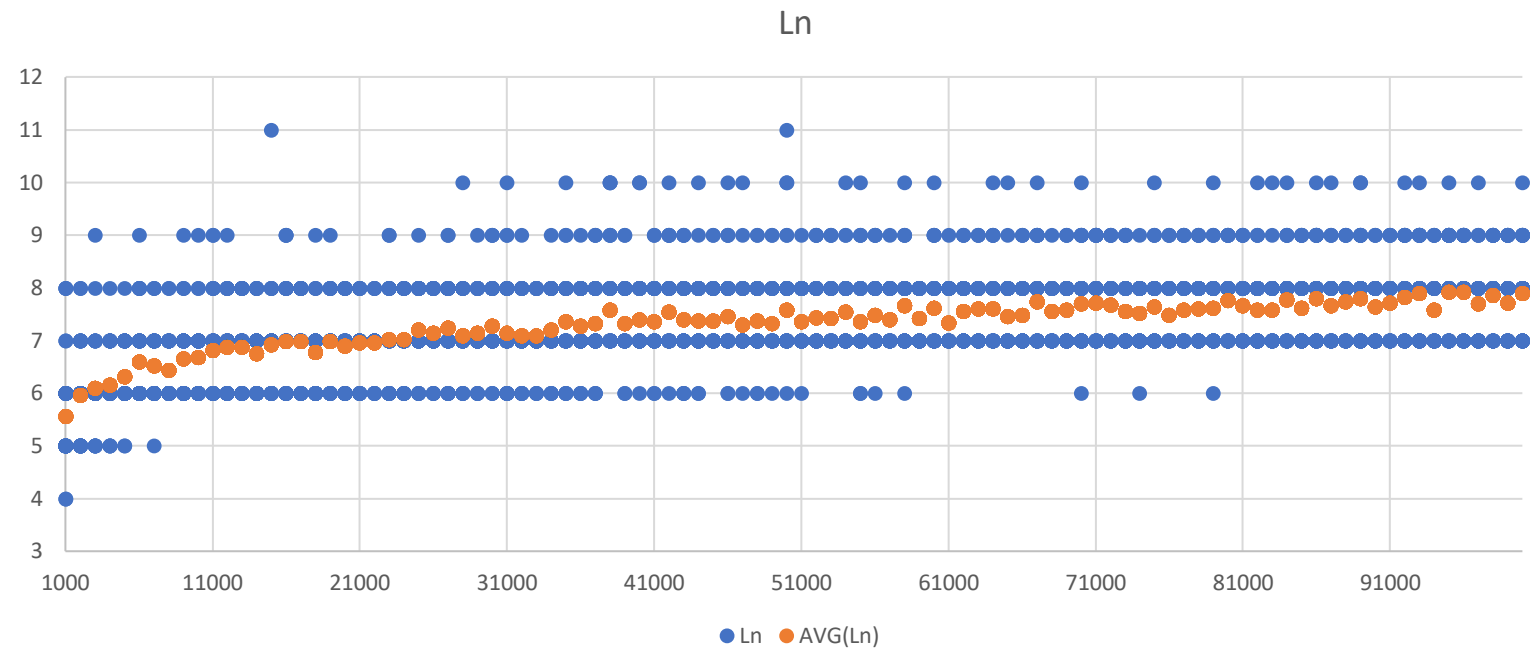


U_n – liczba pustych urn po wrzuceniu n kul

Widzimy że po wrzuceniu n kul do n urn dla danego n liczba pustych urn jest zazwyczaj podobna – poszczególne wyniki są mocno skoncentrowane wokół wartości średniej która to rośnie wprost proporcjonalnie do ilości urn. Przyrost wartości średniej wydaje się być liniowy, można go przybliżyć następującym równaniem: $U_n = \frac{4049}{11000}n - \frac{45}{11}$

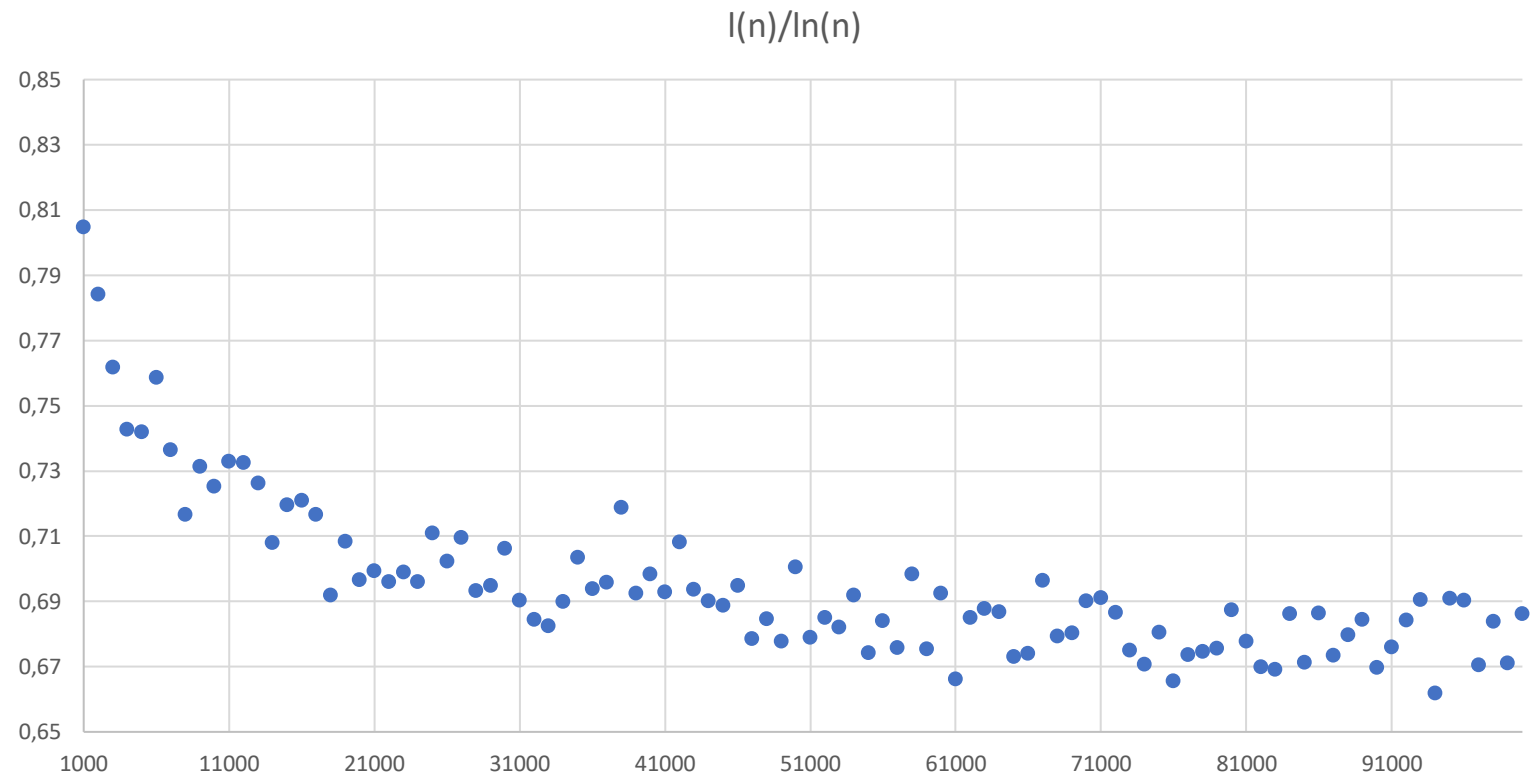


Stosunek liczby pustych urn po n rzutach do liczby urn (rzutów) to funkcja dążąca do 0,368

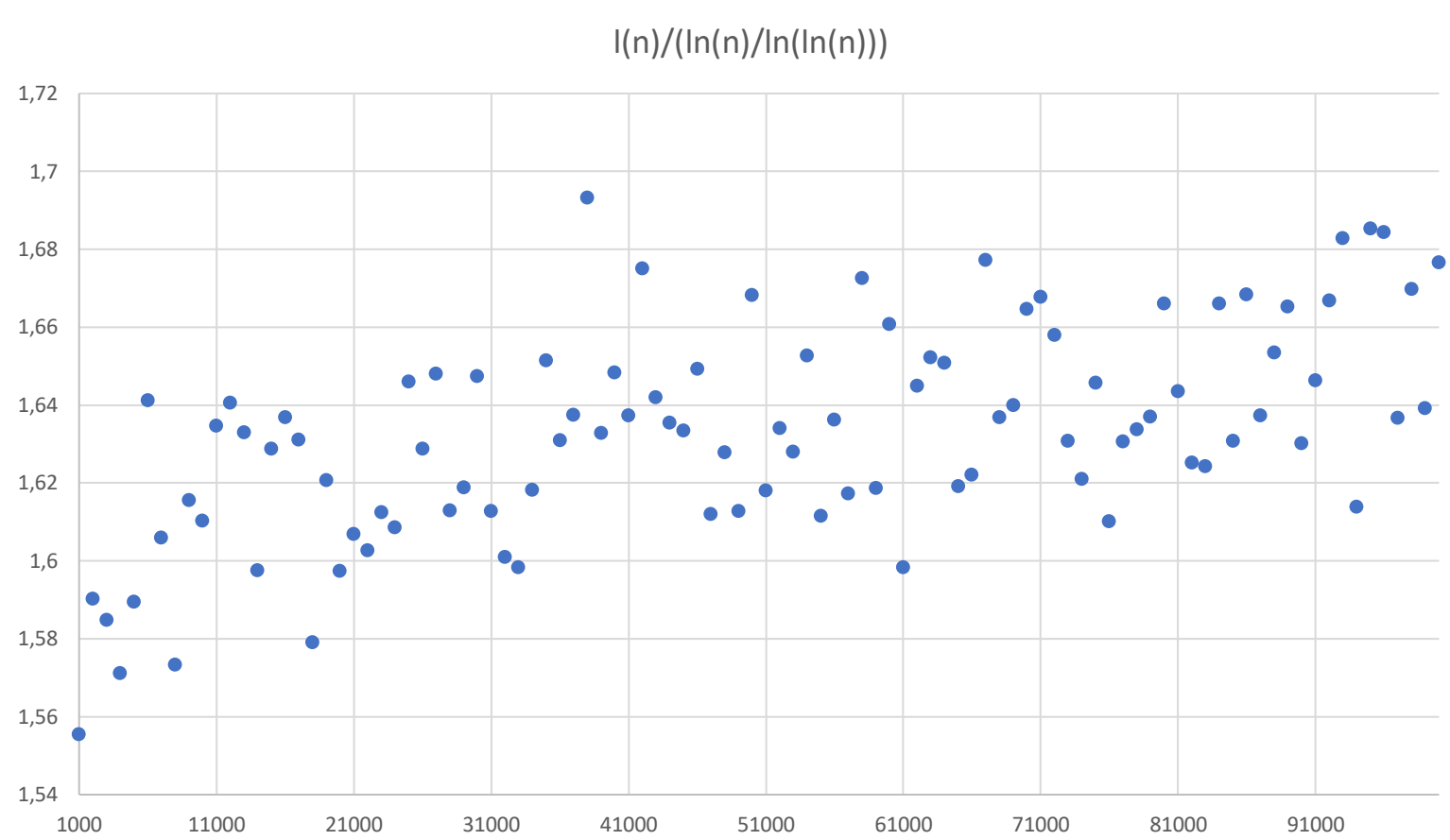


L_n – maksymalna liczba kul w urnie po wrzuceniu n kul

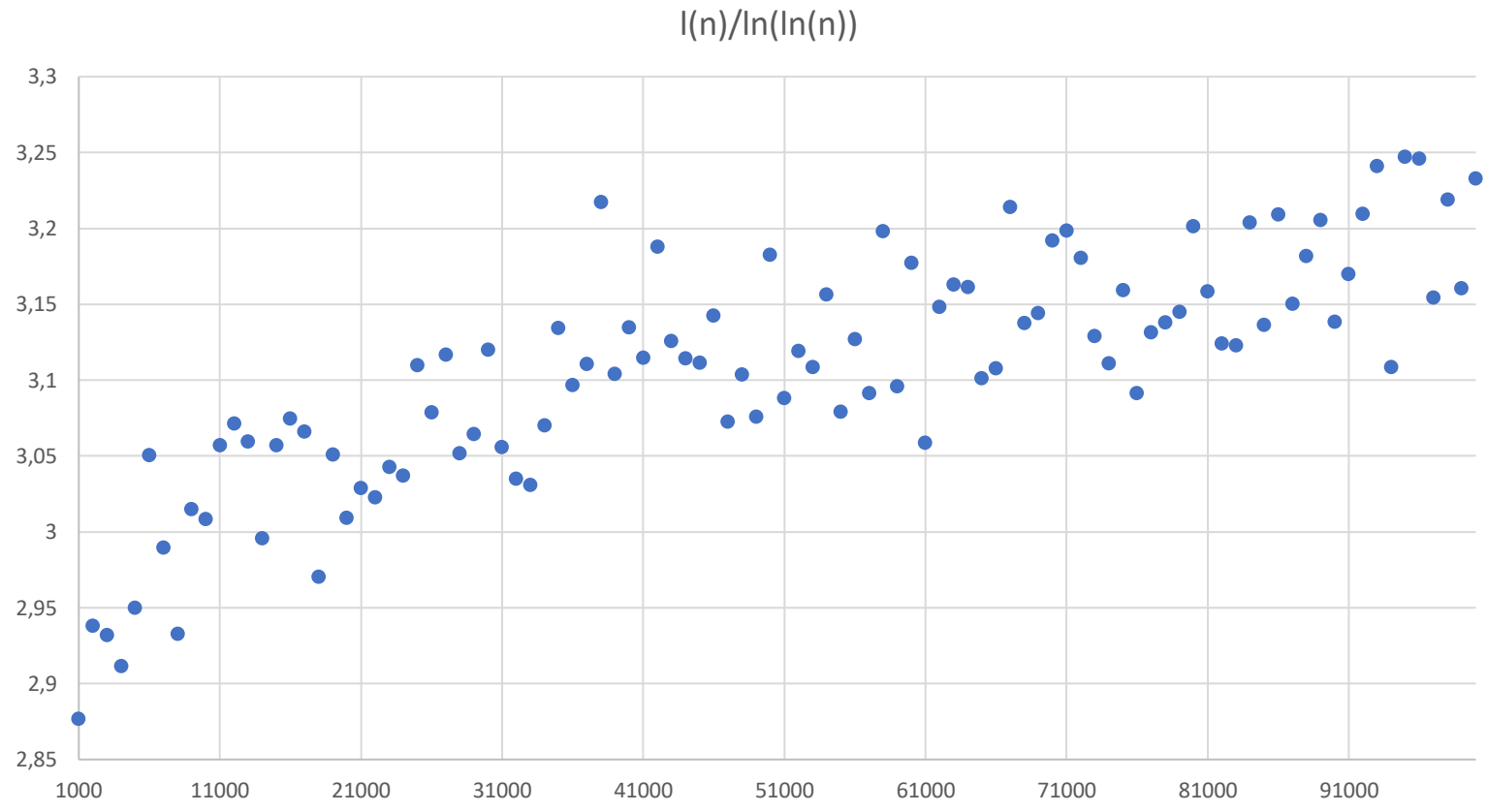
Jak widać niezależnie od liczby urn wartość ta jest niewielka - dla 1000 zdarza się że w jednej urnie są tylko 4 kule, przy mniej niż 8000 urn zdarzały się urny z 5 kulami, jeśli mamy co najmniej 8000 urn to największa ilość kul w jednej to co najmniej 6 itd. Zatem wartość średnia L_n rośnie przy rosnącej liczbie urn, jednak bardzo powoli, wykładniczo. Dla poszczególnych wartości L_n widzimy również pewien efekt – np. 9 dla niskich n pojawia się rzadko mimo aż 50 powtórzeń eksperymentu dla każdego n nie dla każdego wygenerowała się choć raz wartość 9. Jednak im większy n tym większe prawdopodobieństwo otrzymania wartości 9. Sądząc jednak po innych wartościach dla odpowiednio dużych n prawdopodobieństwo otrzymania 9 znowu zaczęłoby spadać.



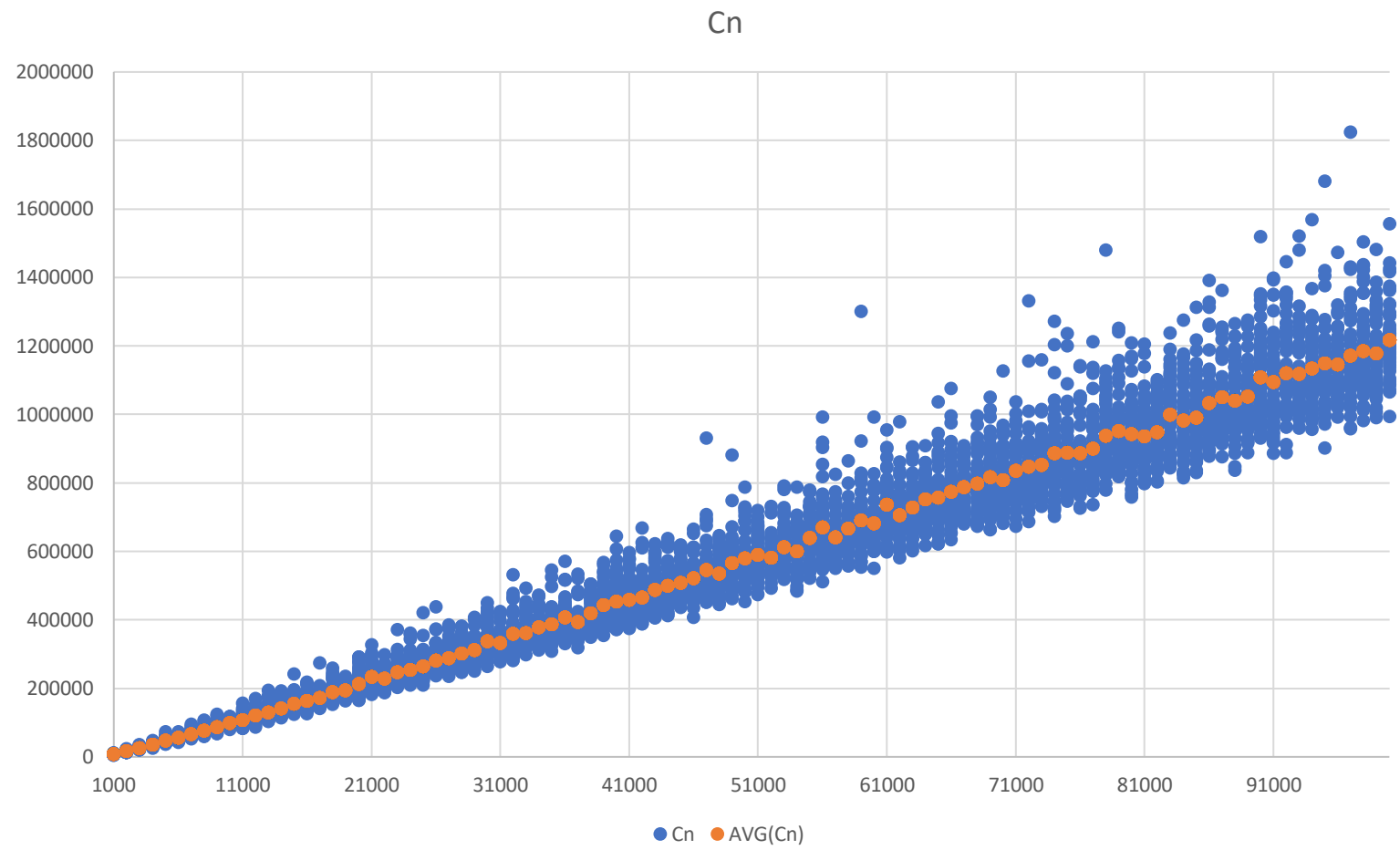
Stosunek maksymalnej liczby kul w urnie po wrzuceniu n kul do logarytmu naturalnego z n to funkcja malejąca wykładniczo.



$\frac{l(n)}{\frac{\ln n}{\ln\ln n}}$ to funkcja rosnąca wykładniczo.

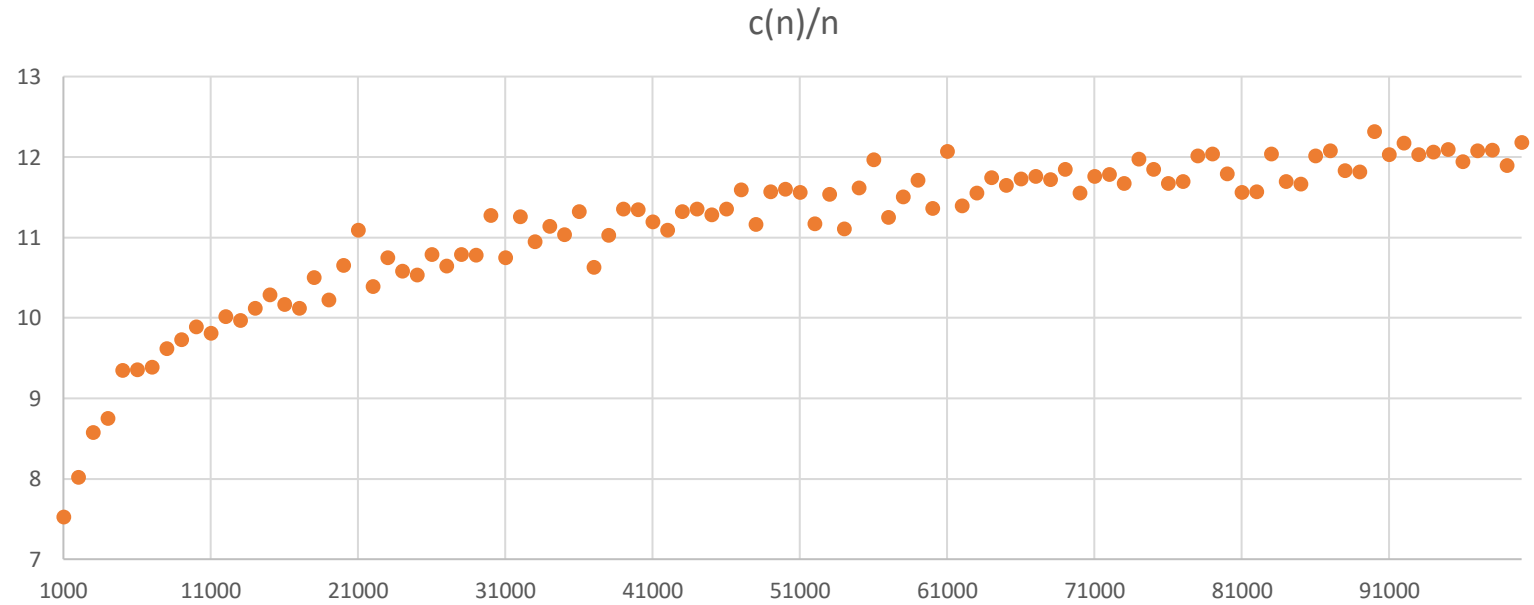


$\frac{l(n)}{\ln\ln n}$ to funkcja rosnąca wykładniczo.

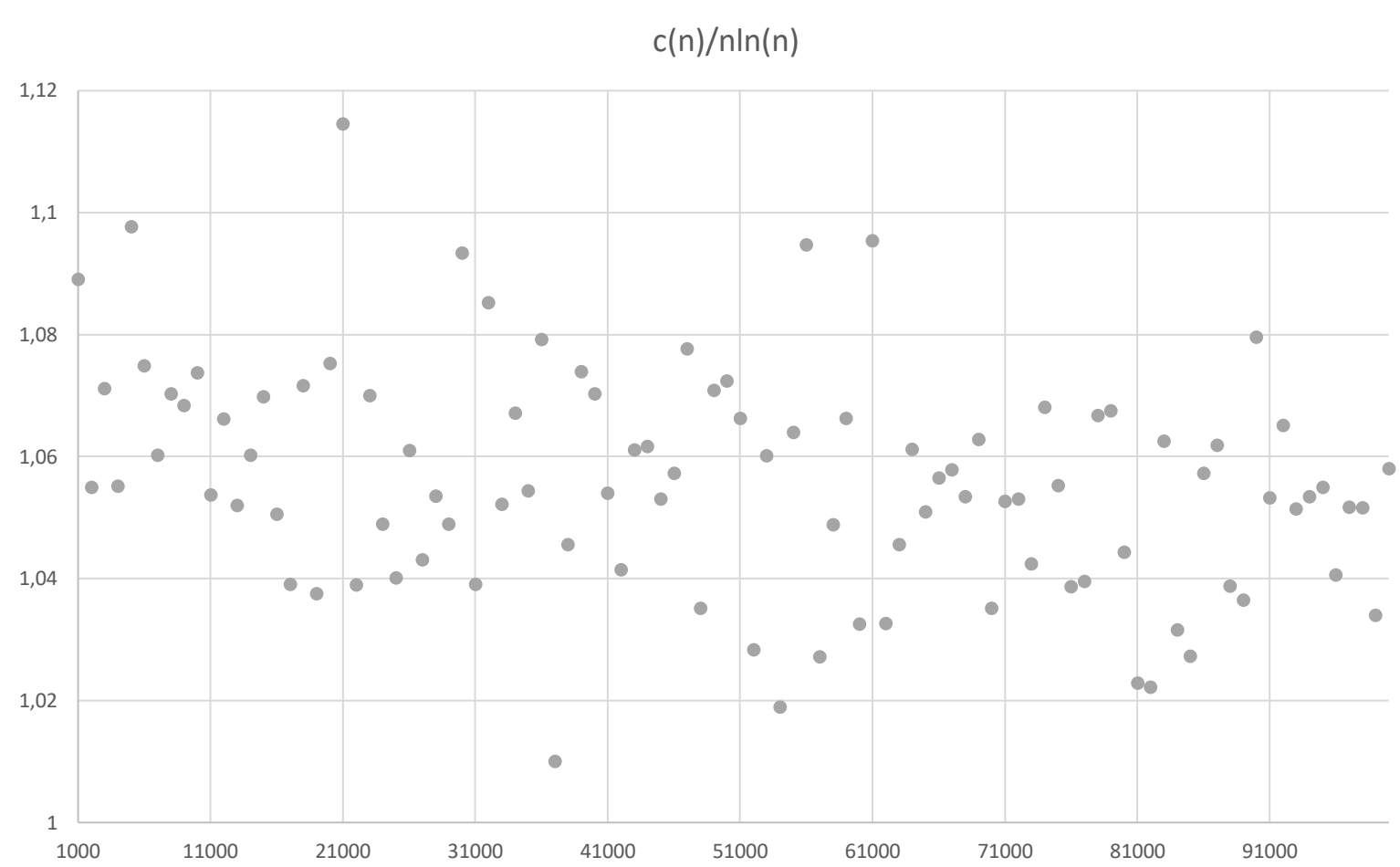


C_n – liczba rzutów po której żadna urna nie jest pusta

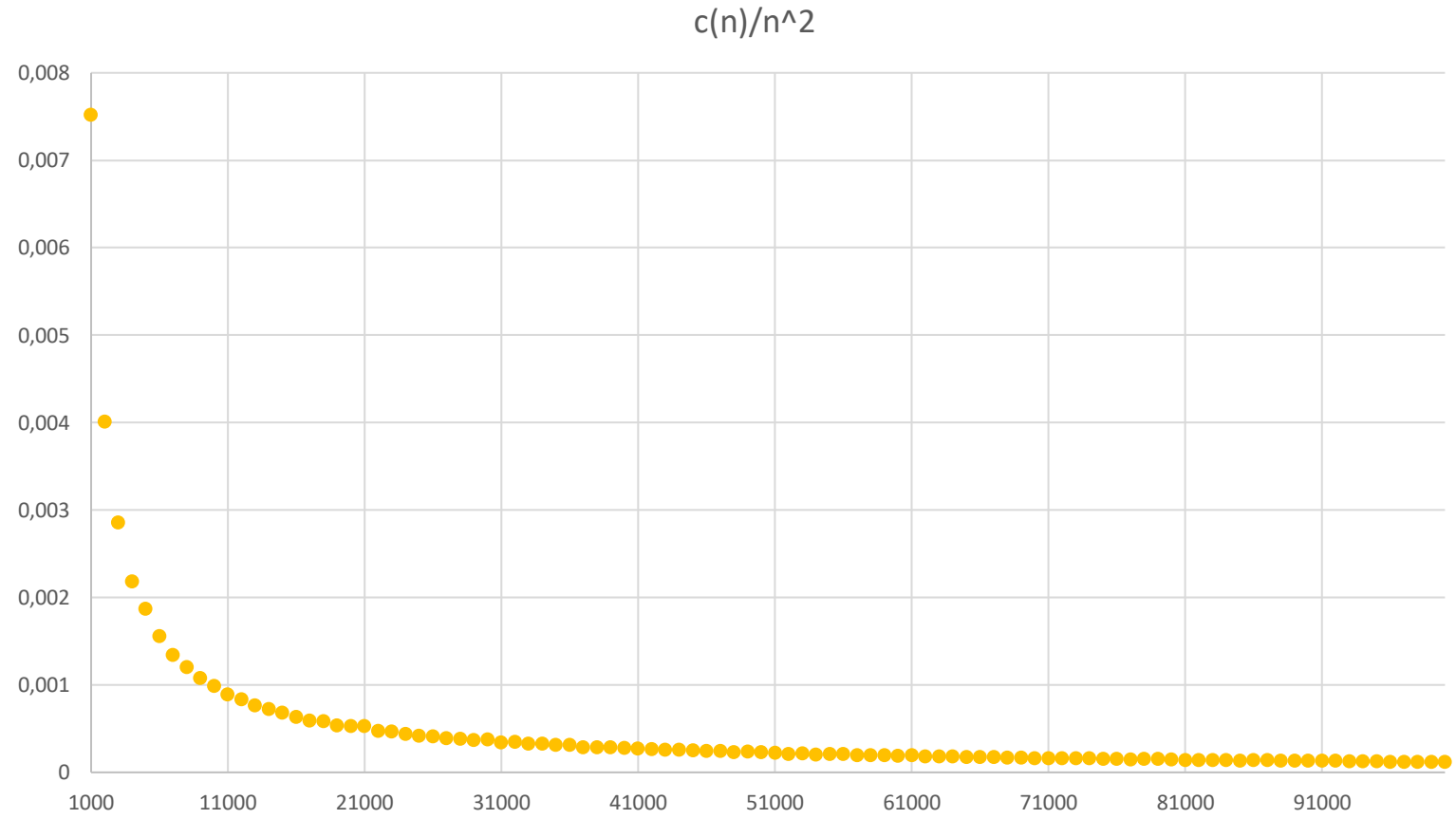
Tutaj raczej bez zaskoczenia – im więcej urn tym więcej kul potrzeba aby je wszystkie wypełnić. Wartość ta rośnie prawie liniowo, z minimalnym zakrzywieniem ku górze. Dla niewielu urn jest mocno skoncentrowana dookoła wartości średniej, a im więcej urn tym rozrzut jest większy.



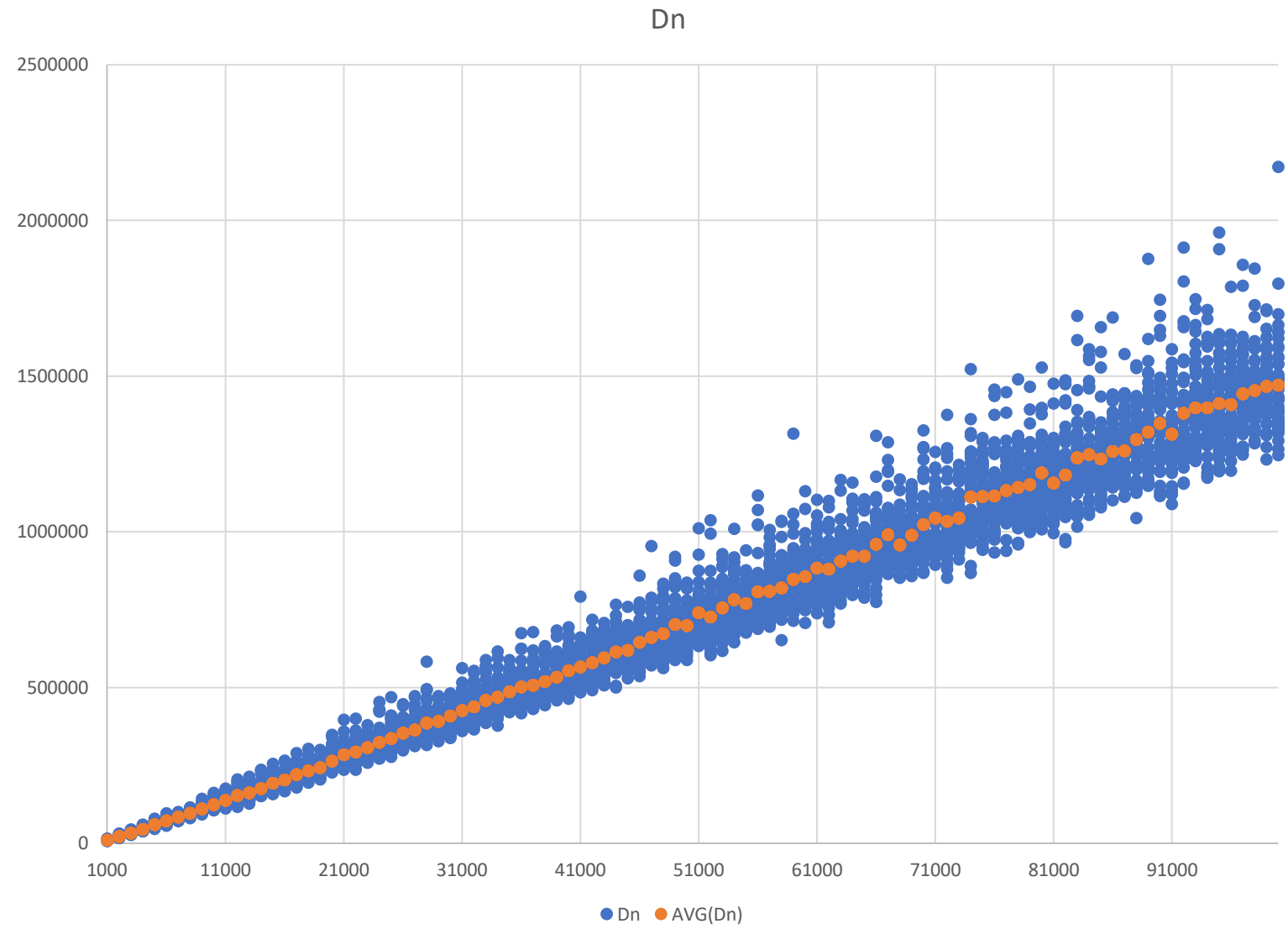
Stosunek liczby rzutów potrzebnych do wypełnienia wszystkich urn do liczby urn to funkcja rosnąca wykładniczo.



Ta funkcja wydaje się być malejącą wykładniczo.

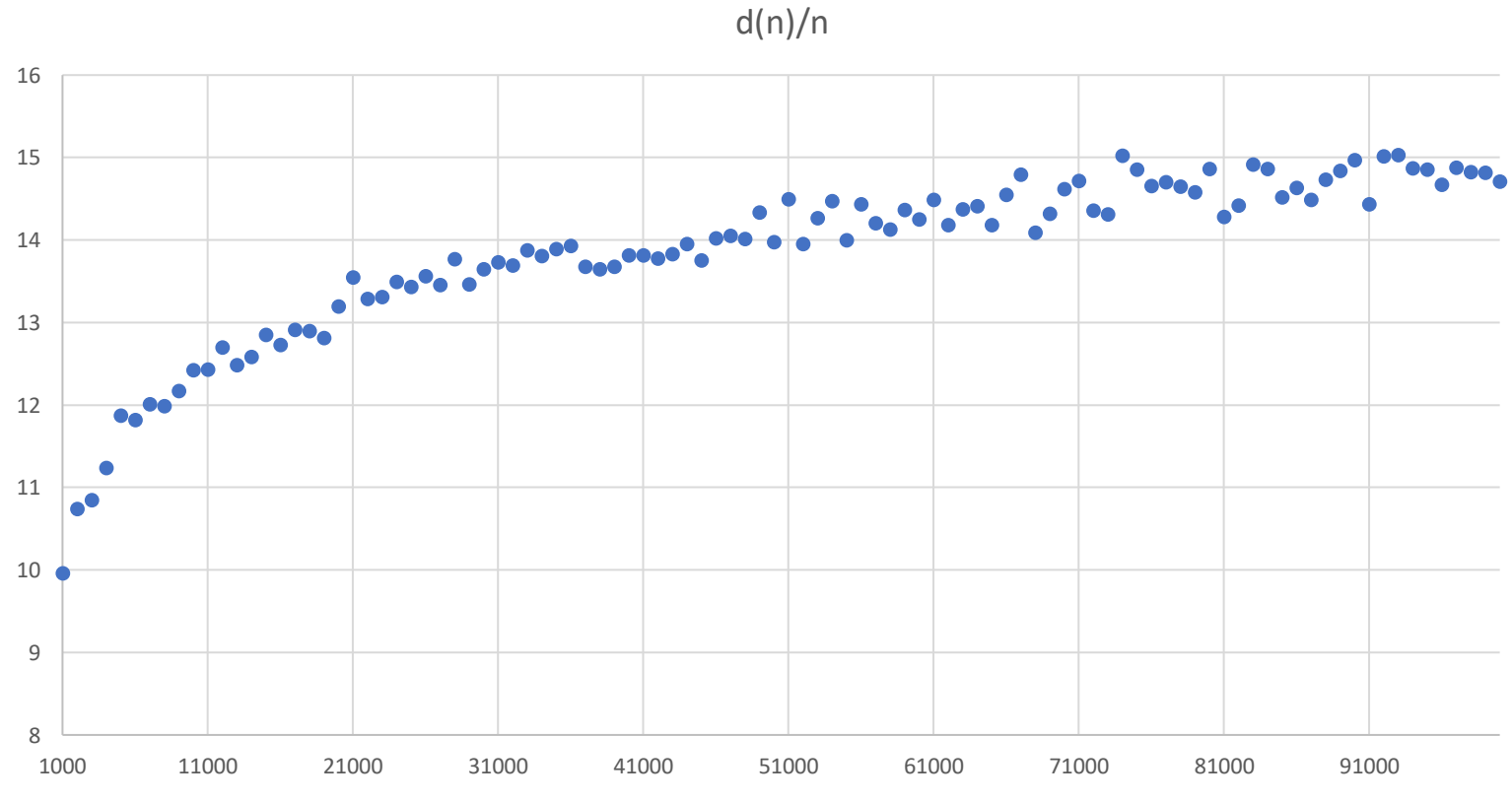


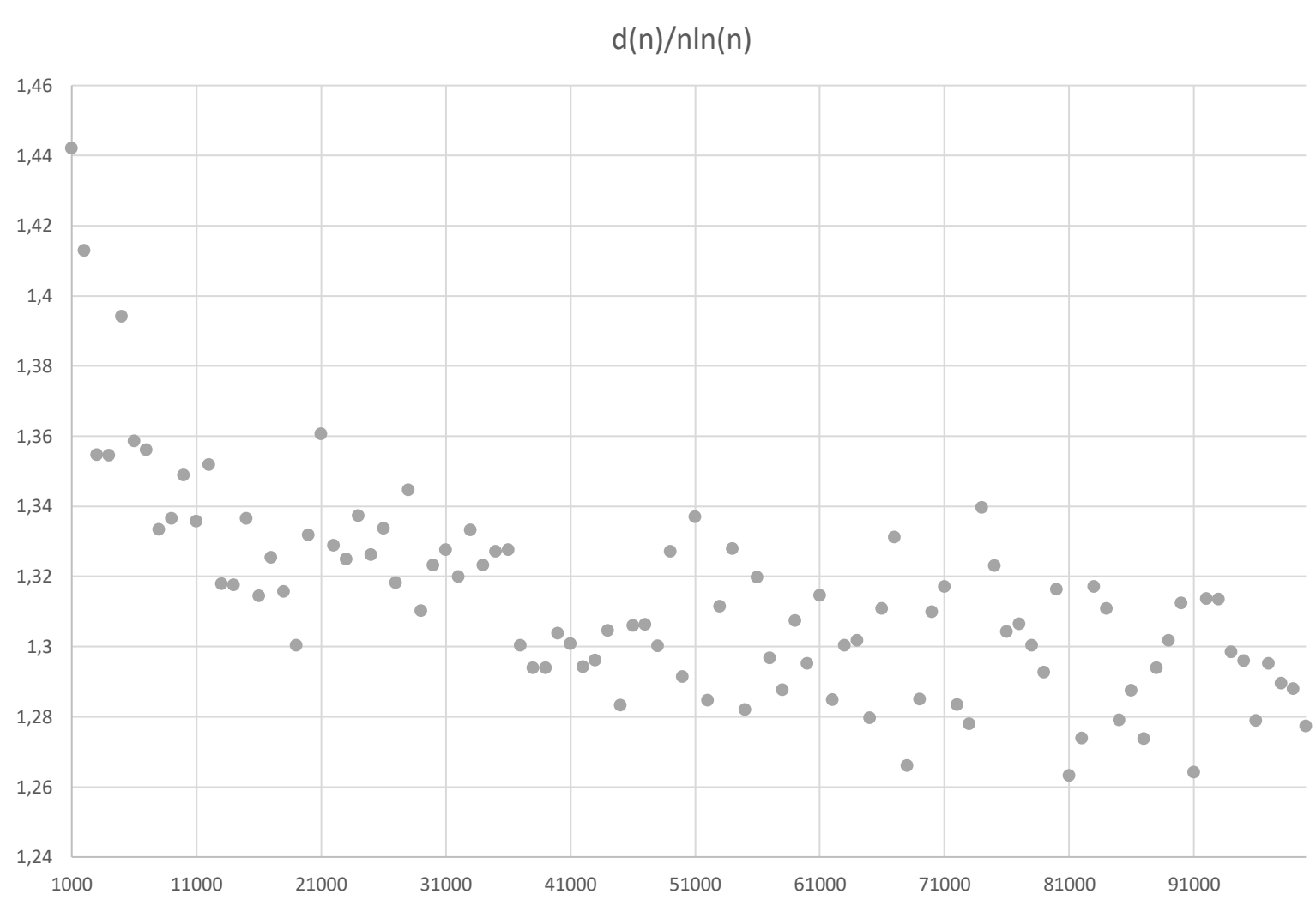
Powyższa funkcja jest pięknym przykładem funkcji malejącej wykładniczo.

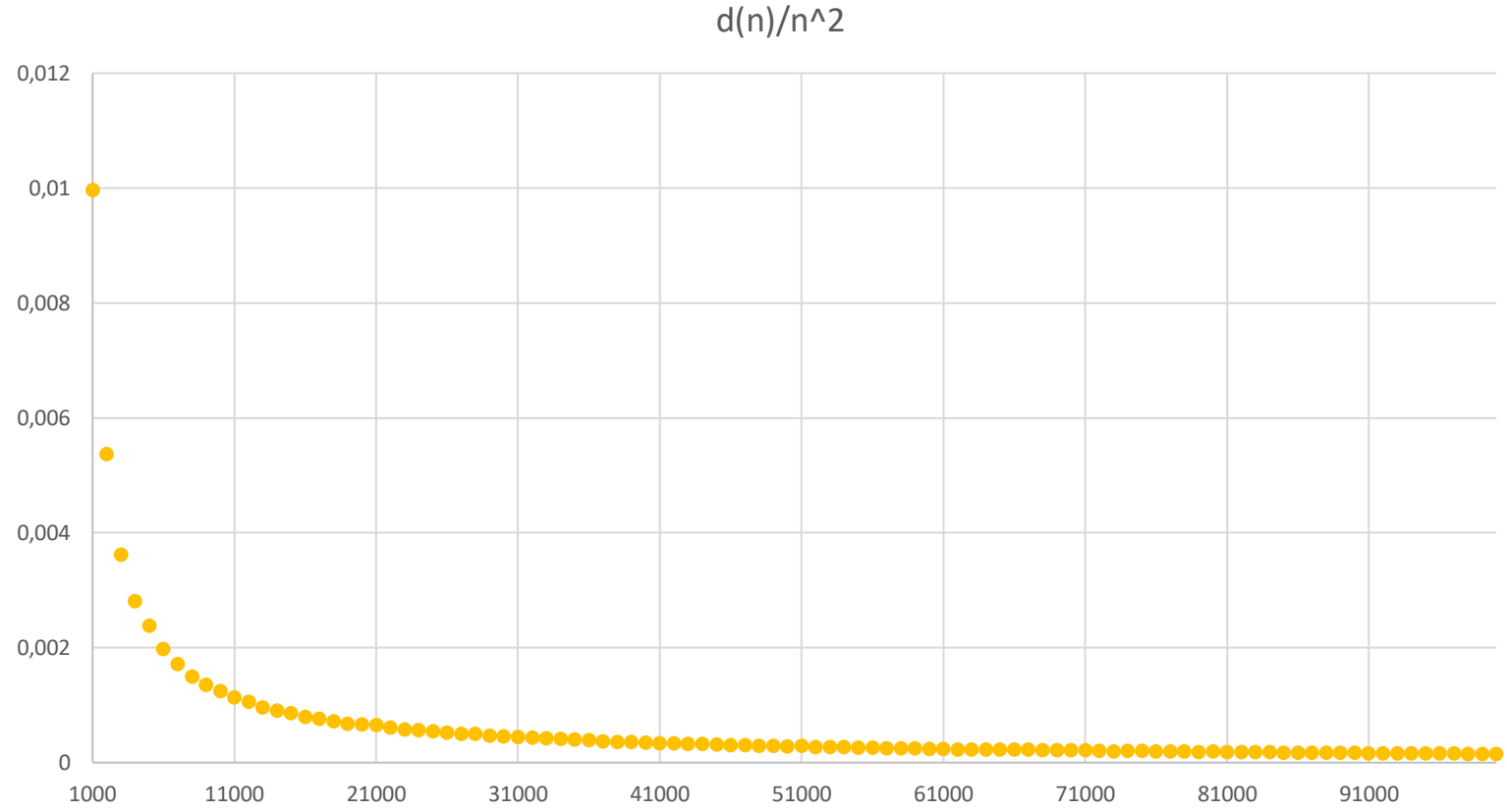


D_n – liczba rzutów po których w każdej urnie są co najmniej 2 kule (koniec wykonywania programu)

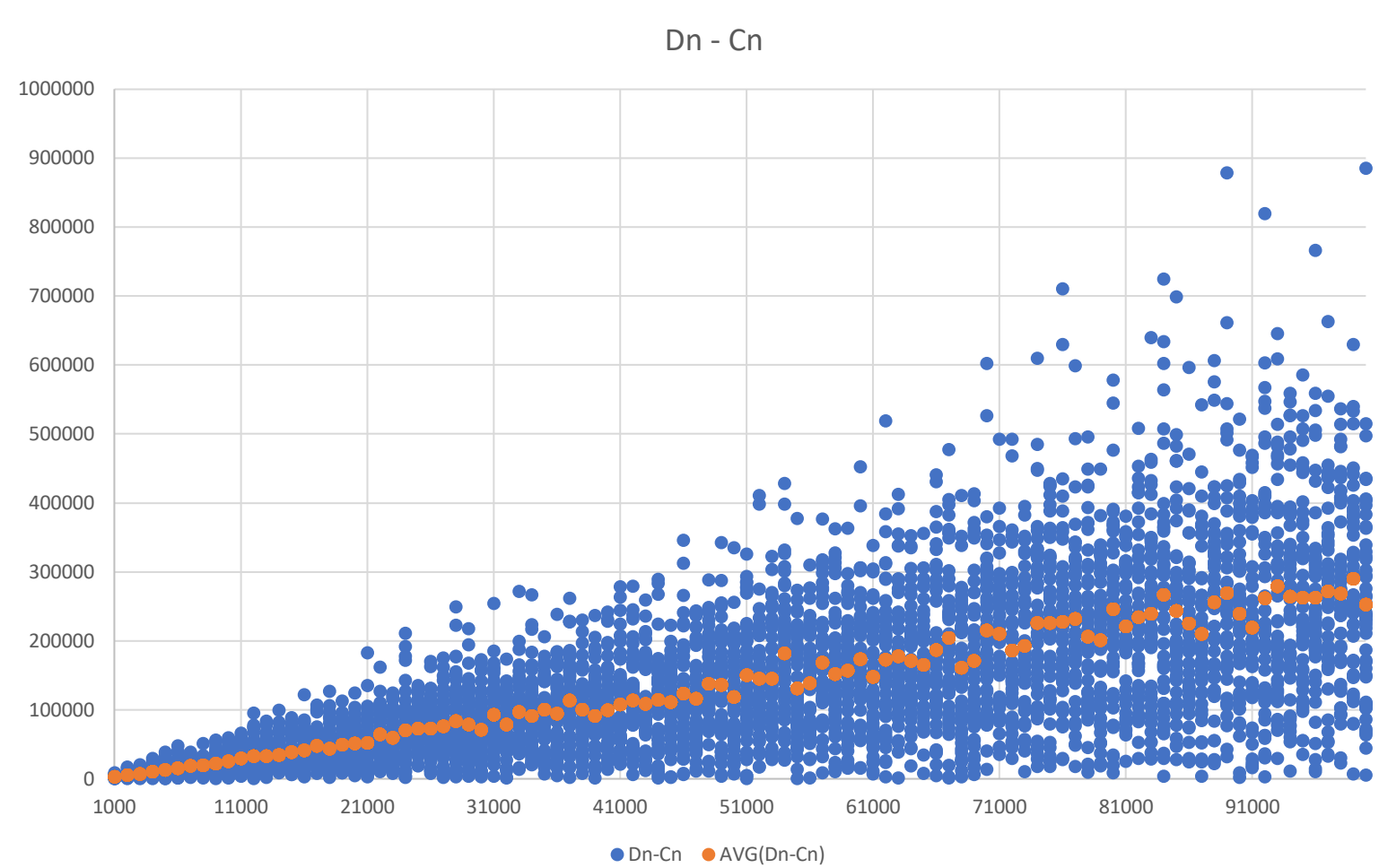
Wykres jest bardzo podobny do wykresu wartości C_n. Tutaj również średnie wartości rosną na pierwszy rzut oka liniowo i im większe n tym większa rozbieżność między wartościami minimalnymi i maksymalnymi.





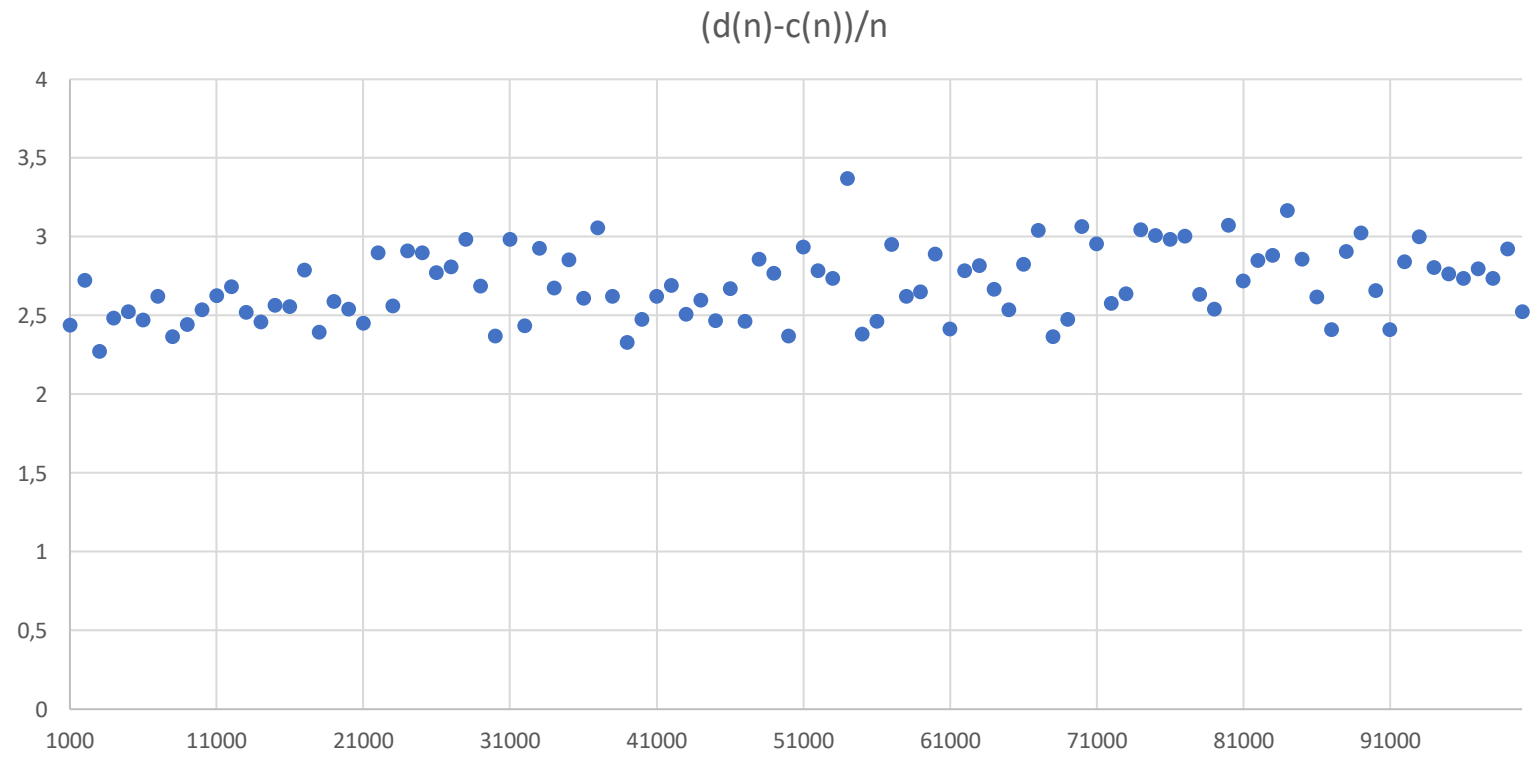


Wykresy bardzo podobne do tych analogicznych dla c(n) z tą różnicą że na drugim wykresie lepiej widać tutaj że funkcja maleje logarytmicznie.

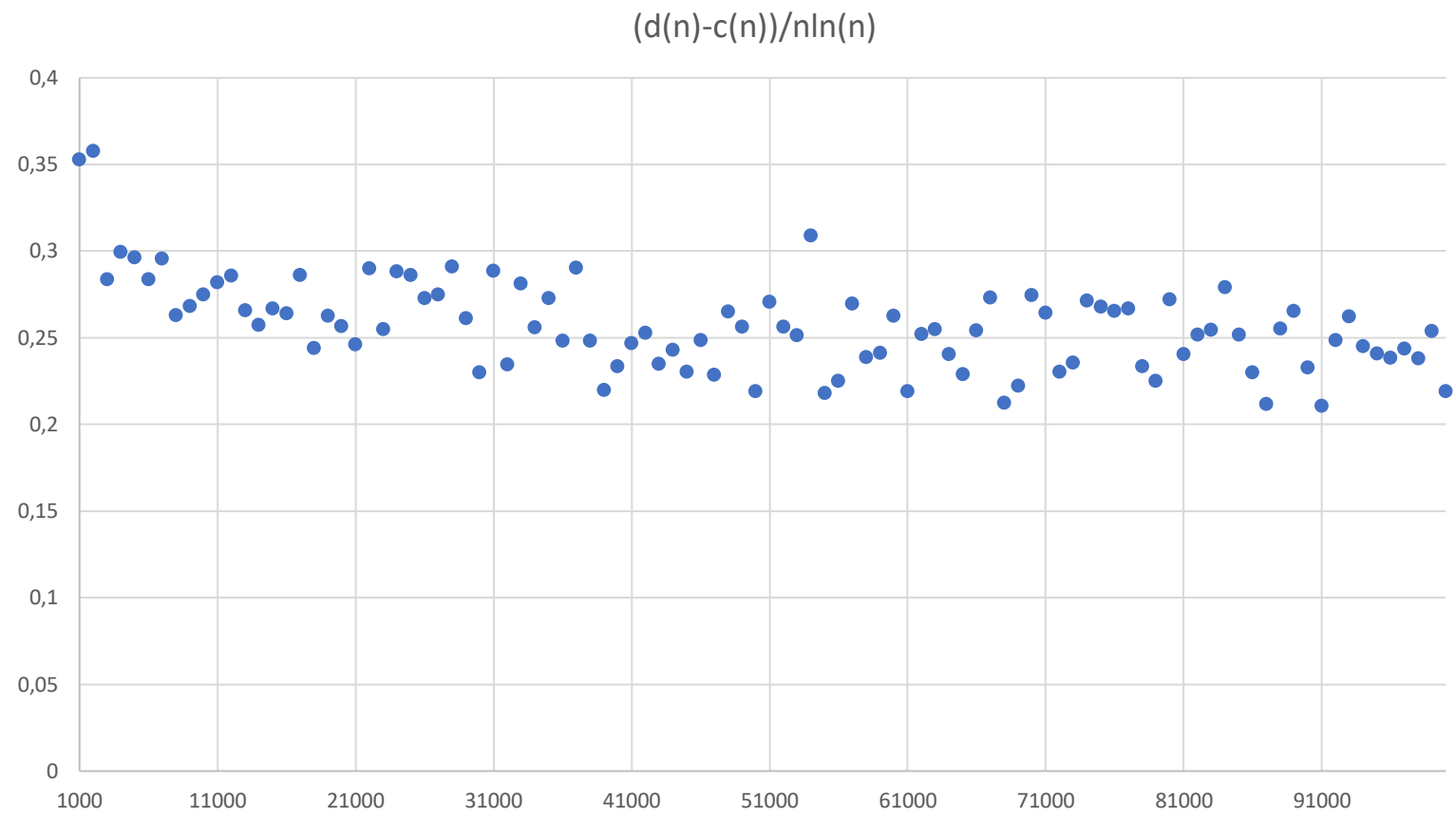


D_n – C_n – ilość rzutów między zdarzeniami C_n a D_n

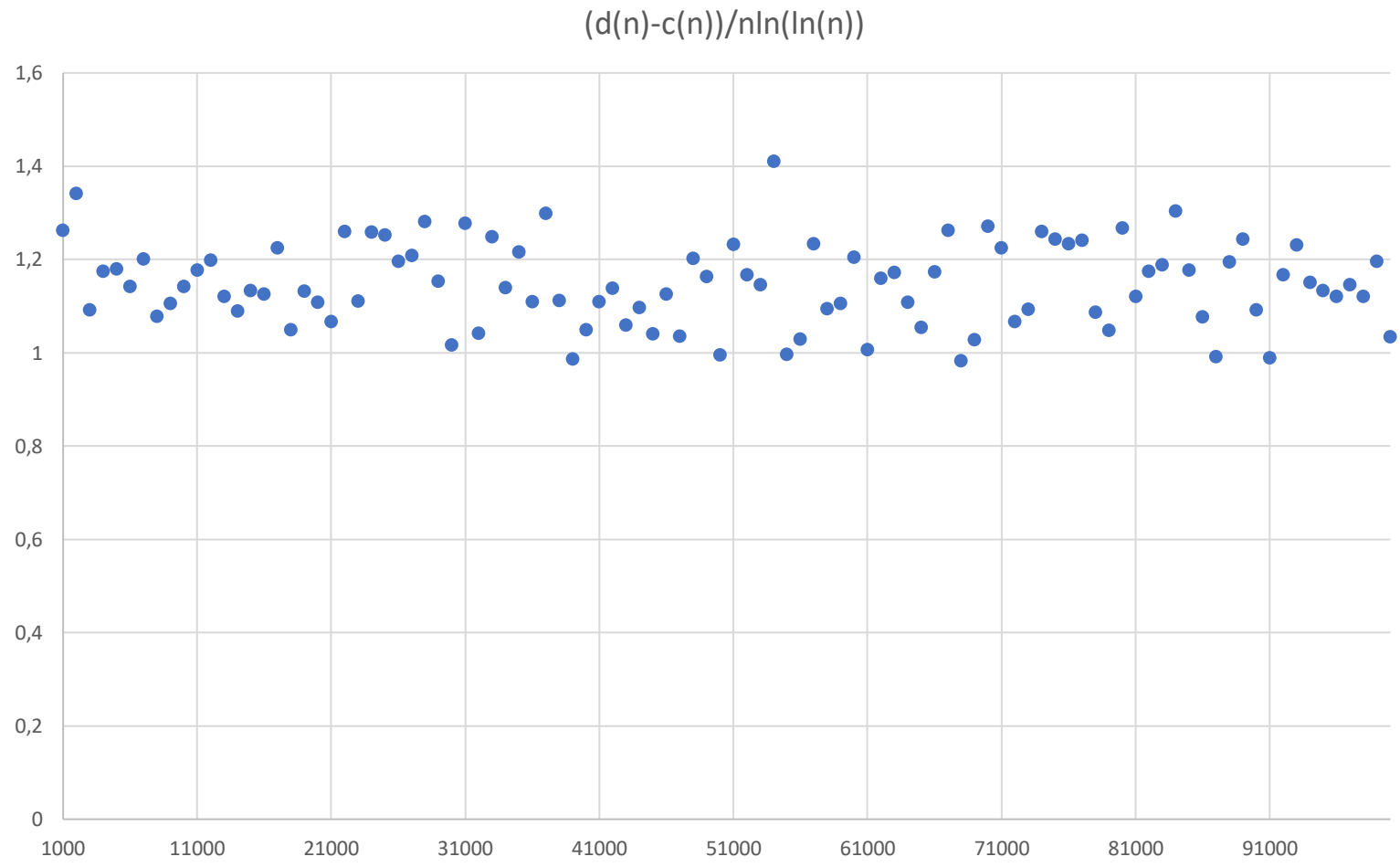
I ten wykres jest analogiczny do dwóch poprzednich – im większe n tym większa średnia ale i większe rozbieżności. W tym przypadku jednak średnia rośnie dużo wolniej ale nadal liniowo.



Z wykresu trudno odczytać jego monotoniczność lub asymptotykę. Na pierwszy rzut oka wydaje się to być w przybliżeniu funkcja stała jednak średnio wartości minimalnie rosną.



Wykres przedstawia funkcję malejącą wykładniczo.



Ten wykres również zdaje się przedstawiać funkcję stałą.

Birthday paradox

Funkcja B_n przedstawia paradoks urodzinowy. W jego podstawowej wersji mamy do czynienia z grupą n ludzi i sprawdzamy prawdopodobieństwo tego że dwoje z nich ma te same urodziny. Mimo tego że jest 365 dni w roku to już dla 23 osób to prawdopodobieństwo wynosi ponad 50%. W naszym przykładzie podobnie: pytamy o to ile kul (zamiast ludzi) musimy wrzucić do losowych spośród n urn (losowe dni w roku) aby dwie trafiły do tej samej urny (miały urodziny tego samego dnia). Okazuje się że prawdopodobieństwo tego że i-ta kula trafi do niepustej urny jest zaskakująco wysokie i nie maleje wcale tak bardzo ze wzrostem ilości urn jak by się można było tego spodziewać. To dlatego że kul jest niewiele, ale my porównujemy pary kul i urnę do której one trafiły a zatem dla k kul jest $\frac{k*(k-1)}{2}$ par do porównania.

Paradoks urodzinowy jest istotny w tematyce funkcji hashujących. Gdy kodujemy jakiś zestaw danych jako hash to chcemy aby każdy bit składową hashu miał takie samo prawdopodobieństwo na bycie zakodowanym przez dany bit wejścia. Jednak podobnie jak w w/w paradoksie – nawet jeśli wejście jest mniejsze niż wyjście to pewna ilość kolizji jest nieunikniona.

W przypadku kryptograficznych funkcji hashujących staje się to problemem ponieważ można znaleźć dwa takie różne zestawy danych zwracające ten sam hash aby temu zapobiec hash musiałby być bardzo długi. Można to wykorzystać przeprowadzając „atak urodzinowy” szukający dwóch różnych wejść zwracających ten sam hash i podmieniający część danych przesyłanych do atakowanego na inne lecz o takim samym hashu przez co atakowany nie jest świadom że otrzymywane dane są skorumpowane.

Coupon collector’s problem

Funkcja C_n przedstawia ten problem – ile razy musimy losować spośród n kuponów (w tym wypadku urn) aby każdą wylosować co najmniej 1 raz. Logicznym i zgodnym z intuicją jest fakt że funkcja ta rośnie wraz ze wzrostem n. Zaskoczeniem może być fakt że mimo że na wykresie wydaje się że funkcja jest liniowa jest ona $\Theta(n \ln n)$ co jednak również jest raczej spodziewane.