Raport 3

Aleksandra Niedziela

2023-05-29

Problem kolekcjonera kuponów

Aby wygrać w konkursie potrzebujemy n kuponów. Interesuje nas liczba pudełek X_n, po których zakupie otrzymamy nagrodę. Liczba X_n to zmienna losowa. Przyjmijmy, że mamy już $k-1$ kuponów, niech $X_{n,k}$ będzie liczbą pudełek, które musimy kupić, aby posiadać k kuponów.

- $X_{n,k}$ to zmienna losowa, o rozkładzie geometrycznym z parametrem $\frac{n-k+1}{n}$
- $X_n = X_{n,1} + X_{n,2} + \ldots + X_{n,n} = \sum_{k=1}^{n} X_{n,k}$

Policzmy teraz wartość oczekiwaną zmiennej X_n

$$E[X_n] = \sum_{k=1}^{n} \frac{n}{n-k+1} = n \sum_{k=0}^{n-1} \frac{1}{n-k} = n \sum_{k=1}^{n} \frac{1}{k}$$

Symulacja

Przeprowadźmy teraz symulację, za pomocą poniższej funckji

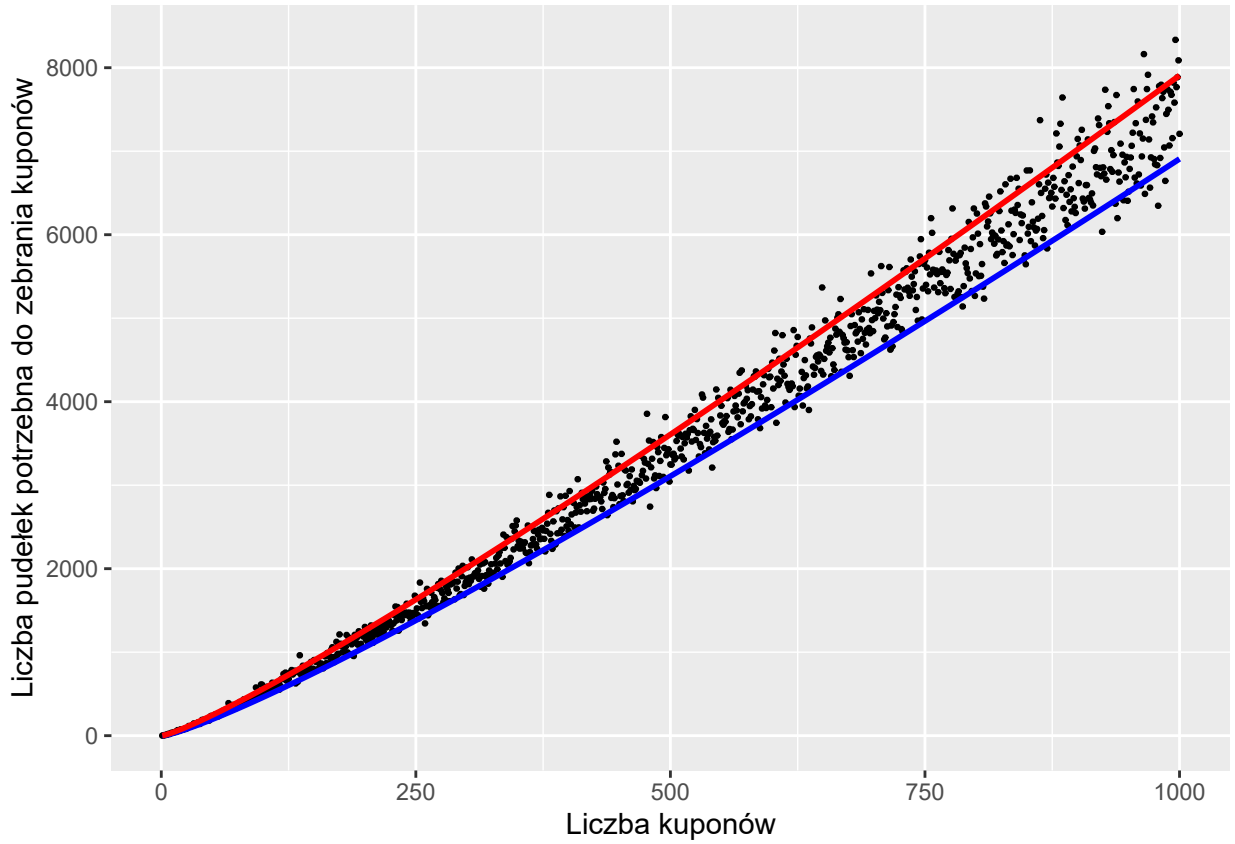
```
boxes <- function(n){
  coupons <- 1:n
  aquired_coupons <- numeric(n)
  prize <- sum(1:n)
  a <- 0

  while (sum(aquired_coupons) != prize){
    coupon <- sample(1:n, 1)
    aquired_coupons[coupon] = coupon
    a <- a + 1
  }
  return(a)
}
```

Wyniki możemy przedstawić w tabeli:

Table 1: Porównanie wartości eksperymentalnych i teoretycznych

ilość kuponów	Wartość teoretyczna	Wartość eksperymentalna
10	29	23
25	95	155
100	519	482
500	3396	3050
1000	7485	7236

Spójrzmy teraz na wykres pokazujący średnią liczbę pudełek (10 prób), po których zakupieniu zbierzemy wszystkie kupony



Czerwona i niebieska linia oznaczają ograniczenia wynikające z faktu:

$$\ln(n) < \sum_{k=1}^{n} \frac{1}{k} < \ln(n) + 1$$
$$n \cdot \ln(n) < E[X_n] < n \cdot (\ln(n) + 1)$$

Nierówność Markowa

Niech X będzie zmienną losową, która przyjmuje jedynie nieujemne wartości. Wtedy dla wszystkich a > 0,

$$P(X \geq a) \leq \frac{E[X]}{a}$$

Korzystając z nierówności Markowa możemy oszacować prawdopodobieństwo uzyskania co najmniej $\frac{3n}{4}$ orłów w n rzutach monetą. Przyjmijmy za zmienną losową X_n liczbę wyrzuconych orłów w n rzutach. Widzimy, iż jest to schemat Bernoulliego, gdzie prawdopodobieństwo sukcesu (wyrzucenia orła) wynosi $\frac{1}{2}$, stąd mamy:

$$E[X_n] = \frac{n}{2}$$

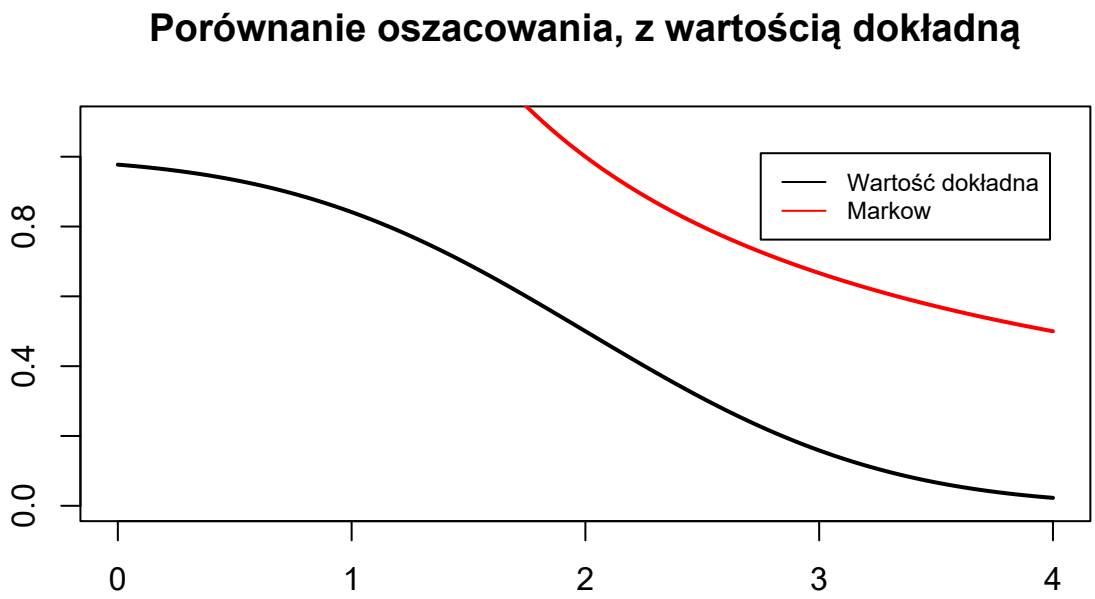
Teraz z nierówności Markowa otrzymujemy:

$$P(X_n \geq \frac{3n}{4}) \leq \frac{\frac{n}{2}}{\frac{3n}{4}} = \frac{2}{3}$$

Możemy wyliczyć dokładne prawdopodobieństwa dla ustalonej ilości rzutów, a następnie zauważyć, że za każdym razem są one ograniczone przez $\frac{2}{3}$

n	prawdopodobieństwo wyrzucenia 3n/4 orłów
4	0.31250000
10	0.17187500
20	0.02069473
50	0.00046811
100	0.00000028

Zobaczmy teraz jak nierówność Markowa sprawdza się dla rozkładu normalnego, dla różnych wartości a. Wiemy, że dla rozkładu normalnego $N(\mu, \sigma)$, $E[X] = \mu$. Weźmy $\mu = 2$ oraz $\sigma = 1$.



Nierówność Czebyszewa

Dla dowolnego $a > 0$ mamy:

$$P(|X - E[X]| \geq a) \leq \frac{Var[X]}{a^2}$$

Korzystając z powyższej nierówności możemy ponownie oszacować prawdopodobieństwo wyrzucenia $\frac{3n}{4}$ orłów w n rzutach symetryczną monetą. Wiemy, że $Var[X] = np(1-p)$. Mamy:

$$P(|X - E[X]| \geq a) = P(X \geq a + E[X]) + P(X \leq -a + E[X])$$

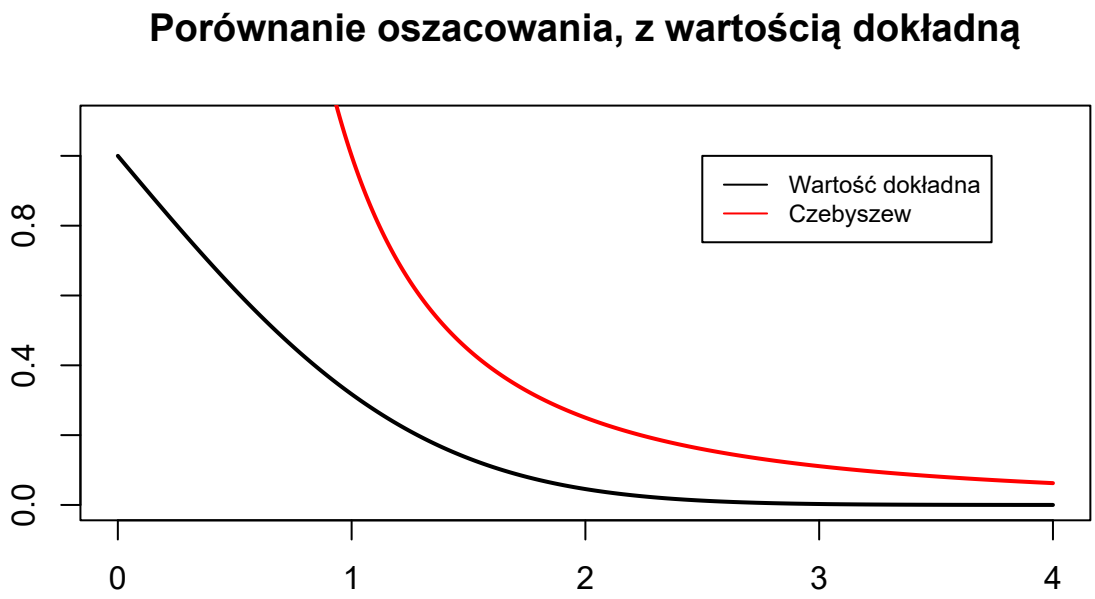
Chcemy, aby $a + E[X] = \frac{3n}{4}$, więc $a = \frac{n}{4}$. Ponieważ patrzymy tutaj na rozkład Bernoulliego, to mamy:

$$P(X \geq a + E[X]) + P(X \leq -a + E[X]) = 2P(X \geq \frac{3n}{4}) \leq \frac{Var[X]}{(\frac{n}{4})^2} = \frac{n \cdot 0.5 \cdot 0.5}{\frac{n^2}{16}} = \frac{4}{n}$$

Stąd prawdopodobieństwo wyrzucenia $\frac{3n}{4}$ orłów możemy oszacować przez $\frac{2}{n}$

W przeciwieństwie do ograniczenia wynikającego z nierówności Markowa wynik jest zależny od ilości powtórzeń eksperymentu.

Spójrzmy teraz na wykres, gdzie dla rozkładu normalnego porównamy oszacowanie z wartością dokładną ogona dystrybuanty. Dla rozkładu normalnego mamy $Var[X] = \sigma^2$.



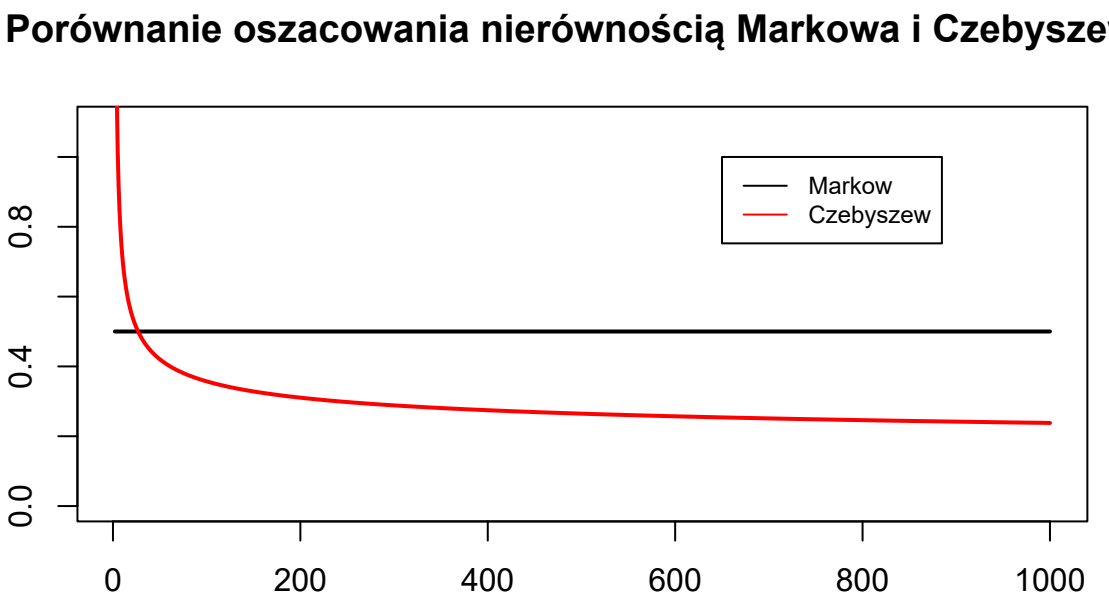
Korzystatjąc z nierówności Markowa, możemy oszacować wartość $P(X \geq 2nH_n)$, gdzie X to zmienna losowa oznaczająca ilość pudełek, które musimy kupić by wygrać nagrodę (problem kolekcjonera kuponów), natomiast $H_n = \sum_{i=1}^{n} \frac{1}{i}$. Mamy $E[X] = n \sum_{i=1}^{n} \frac{1}{i}$

$$P(X \geq 2nH_n) \leq \frac{E[X]}{a} = \frac{n\sum_{i=1}^{n}\frac{1}{i}}{2n\sum_{i=1}^{n}\frac{1}{i}} = \frac{1}{2}$$

Teraz korzystając z nierówności Czebyszewa oszacujmy $P(|X - nH_n| \geq nH_n)$ Aby policzyć $Var[X]$ skorzystamy z faktu, iż $Var[X] = \sum_{i=1}^{n} Var[X_i]$. Wiemy, iż zmienne X_i mają rozkład geometryczny, stąd $Var[X_i] = \frac{(1-p)}{p^2}$

$$P(|X - nH_n| \geq nH_n) \leq \frac{Var[X]}{a^2} = \frac{\sum_{i=1}^{n} Var[X_i]}{(nH_n)^2} \leq \frac{\pi^2 n^2}{6n^2 H_n^2} \leq \frac{\pi^2}{6(\ln n)^2}$$

Porównajmy otrzymane oszacowania



Widzimy, że dla małych wartości n lepsze oszacowanie daje nam nierówność Markowa, natomiast wraz ze wzrostem n nierówność Czebyszewa daje dokładniejsze przybliżenia.

Średnia, dyspersja, mediana

Jeśli X jest zmienną losową o średniej μ odchyleniu standardowym σ i medianie m wtedy:

$$|\mu - m| \leq \sigma$$

Aby sprawdzić, czy nierówność ta jest prawdziwa dla estymatorów mediany, średniej i odchylenia przeprowadźmy prostą symulację. Wybierzmy próbę 200 mężczyzn z populacji i przyjrzyjmy się estymatorom. Losujemy wartości z rozkładu normalnegoz parametrami $\mu = 178$, $\sigma = 5$, jako że tak możemy modelować zmienną losową opisującą wzrost w populacji.

Otrzymujemy:

- $\mu = 177.830$
- $m = 178.000$
- $\sigma = 5.018$

$$|\mu - m| = 0.17 \leq 5.02 = \sigma$$