Podstawowe fakty

Wartość oczekiwaną zmiennej losowej X przyjmującej wartości $y_1, y_2, \ldots$ określamy wzorem

$$E[X] = \sum_{i=1}^{\infty} P(X = y_i) \cdot y_i.$$

(zazwyczaj będziemy to stosować dla zmiennych o wartościach całkowitych, nieujemnych, czyli gdy ciąg $y_1, y_2, \ldots$ jest ciągiem $0, 1, 2, \ldots$). Wartość oczekiwana opisuje średnią wartość zmiennej X i może być nieokreślona lub nieskończona.

Wartość oczekiwana ma własności podobne do całki.

Fakt 1. Wartość oczekiwana jest liniowa, tzn.

$$E[X + Y] = E[X] + E[Y]$$

oraz, dla $\alpha \in \mathbb{R}$,

$$E[\alpha X] = \alpha\ E[X].$$

Fakt 2. Jeżeli zmienne losowe X oraz Y są niezależne, tzn. $P(X \in A \wedge Y \in B) = P(X \in A)P(Y \in B)$ dla dowolnych zbiorów A, B, to zachodzi równość

$$E[X \cdot Y] = E[X]E[Y].$$

Ćwiczenie 1. Dla jakich zmiennych zachodzi równość $E[X \cdot X] = E[X]E[X]$?

Ćwiczenie 2. Jeżeli X jest nieujemną zmienną losową oraz $E[X] = 0$, to $X = 0$.

Wariancja zmiennej losowej X definiowana jest wzorem

$$Var[X] = E[(X - EX)^2].$$

Zgrubsza, mierzy ona, jak przeciętnie X odbiega od swojej średniej. Np. jeżeli wariancja jest zerowa, to znaczy, że X w ogóle nie odbiega od swojej średniej, więc jest stała.

Ćwiczenie 3. Wykaż, że

$$Var[X] = E[X^2] - E[X]^2.$$

Fakt 3. Wariancja sumy *niezależnych* zmiennych losowych jest sumą ich wariancji

$$\begin{aligned}
Var(X+Y) &= E[((X+Y) - E[X+Y])^2] \\
&= E[X^2 + 2XY + Y^2 - (E[X]^2 + 2E[X]E[Y] + E[Y]^2)] \\
&= E[X^2 - E[X]^2] + E[X^2 - E[Y]^2] + \underbrace{2E[XY] - 2E[X]E[Y]}_{\text{0 z niezeależnośći}} \\
&= Var[X] + Var[Y]
\end{aligned}$$

Problem kolekcjonera nakrętek

Kolekcjoner nakrętek od soczków Tarczyn ma na celu przeczytać każdą z 300 możliwych informacji umieszczonych na spodzie nakrętki soczku. Ile średnio musi kupić butelek, by ten cel osiągnąć?

Rozwiązanie Przyjmijmy dla uproszczenia opisu, że kolekcjoner kupuje nowy soczek każdego dnia począwszy od dnia numer 1 i jeżeli widzi nakrętkę, którą już posiada w kolekcji, to ją wyrzuca a w przeciwnym wypadku zachowuje. Oznaczmy liczbę 300 symbolem n. Dla $i = 1, \ldots, n$, niech T_i będzie momentem (numerem dnia) w którym kolekcjoner celebrował i-tą zakrętkę w kolekcji. Tak więc, $T_1 = 1$ oraz T_2 jest numerem dnia, w którym kolekcjoner otrzymał zakrętkę inną od tej pierwszej.

Interesuje nas wartość oczekiwana zmiennej $T := T_n$. Oznaczmy sobie przez S_i zmienną $T_i - T_{i-1}$. Zmienna ta mówi, ile dni czekał kolekcjoner od momentu, kiedy miał $i-1$ zakrętek, do momentu, w którym otrzymał zakrętkę, której wcześniej nie miał. Wówczas, jeśli przyjmiemy, że $T_0 = 0$, to

$$T = (T_1 - T_0) + (T_2 - T_1) + \cdots + (T_n - T_{n-1}) = S_1 + \cdots + S_n.$$

Łatwo zobaczyć, że zmienna S_i ma rozkład geometryczny, w którym prawdopodobieństwo sukcesu wynosi $\frac{n-i+1}{n}$, gdyż sukcesem kolekcjonera jest wylosowanie którejbądź nakrętki spośród $n-i+1$ pożądanych.

Poniżej obliczymy wartość oczekiwaną oraz wariancję zmiennej o rozkładzie geometrycznym za pomocą funkcji tworzących.

Fakt 4. Wartość oczekiwana oraz wariancja zmiennej losowej X o funkcji tworzącej $f_X(t)$ wyrażają się wzorami

$$\begin{aligned} E[X] &= f'_X(1) \\ Var[X] &= f''_X(1) + f'_X(1) - \left(f'_X(1)\right)^2 \end{aligned}$$

Dowodzi się tego poprzez zróżniczkowanie szeregów wyraz po wyrazie.

Dla zmiennej G_p o rozkładzie geometrycznym z parametrem sukcesu p, funkcja tworząca to

$$\begin{aligned} f_{G_p}(t) = \sum_{i=0}^{\infty} P(G_p = i)t^i &= \sum_{i=1}^{\infty} p(1-p)^{i-1}t^i \\ = tp\sum_{i=1}^{\infty} \left(t\ (1-p)\right)^{i-1} &= \frac{t\ p}{1 - t(1-p)} \end{aligned}$$

Wartości dwóch pierwszych pochodnych f_{G_p} w jedynce to $\frac{1}{p}$ oraz $\frac{2(1-p)}{p^2}$, skąd wnioskujemy, że

$$\begin{aligned} E[G_p] &= 1/p \\ Var[G_p] &= \frac{1-p}{p^2} \end{aligned}$$

A zatem, jako że zmienna S_i ma rozkład geometryczny z prawdopodobieństwem sukcesu $\frac{n-i+1}{n}$, to $E[S_i] = \frac{n}{n-i+1}$.

Tak więc, z liniowości wartości oczekiwanej,

$$E[T_n] = E[S_1] + \cdots + E[S_n] = \frac{n}{n} + \frac{n}{n-1} + \cdots + \frac{n}{1} = n \cdot H_n,$$

gdzie $H_n = 1 + \frac{1}{2} + \frac{1}{3} + \cdots \frac{1}{n}$ jest n-tą liczbą harmoniczną, na którą mamy oszacowanie $H_n = \ln n + \gamma + \frac{1}{2n} + o(1)$, dla stałej Eulera $\gamma \approx 0.577$. Wobec tego,

$E[T] = n \cdot \ln n + \gamma \cdot n + \frac{1}{2} + O(1) = \Theta(n \ln n)$ i tyle średnio trzeba czekać na zebranie wszystkich nakrętek. W praktyce, w przypadku Tarczyna, oznacza to, że średnio musimy kupić 1885 butelek zanim zbierzemy 300 różnych nakrętek.

Teraz policzymy wariancję zmiennej T, czyli miarę tego, jak bardzo brutalna rzeczywistość średnio może odbiec od oczekiwań kolekcjonera nakrętek.

Przypomnijmy, że wariancja zmiennej o rozkładzie geometrycznym z parametrem p wynosi $\frac{1-p}{p^2}$. Tak więc,

$$\begin{aligned} Var[S_i] &= \frac{1 - \frac{n-i+1}{n}}{(\frac{n-i+1}{n})^2} \\ &= \frac{\frac{i-1}{n}}{(\frac{n-i+1}{n})^2} \\ &= \frac{i-1}{n} \cdot \left(\frac{n}{n-i+1}\right)^2 \leq \left(\frac{n}{n-i+1}\right)^2 \end{aligned}$$

Ponieważ zmienne $S_1, \ldots S_n$ są *niezależne*, to stosuje się wzór na wariancję sumy zmiennych *niezależnych*.

$$\begin{aligned} Var[T] &= Var[S_1 + \cdots + S_n] \\ &= Var[S_1] + \cdots Var[S_n] \\ &\leq \left(\frac{n}{n}\right)^2 + \left(\frac{n-1}{n}\right)^2 + \left(\frac{n-2}{n}\right)^2 + \cdots + \left(\frac{1}{n}\right)^2 \\ &\leq n^2(\frac{1}{1^2} + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \cdots) \\ &= n^2 \sum_{i=1}^{\infty} \frac{1}{i^2} = n^2 \cdot \frac{\pi^2}{6} < 1.65\ n^2. \end{aligned}$$

Co z tego wynika w praktyce, mówią nam nierówności Markowa (słabsze oszacowanie, korzystające tylko z wartości oczekiwanej) i Czebyszewa (silniejsze, korzystające z wariancji)

Fakt 5. Nierówność Markowa. Dla $\alpha > 0$ oraz nieujemnej zmiennej losowej X,

$$P[X \geq \alpha] \leq \frac{1}{\alpha} E[X].$$

Dowód. Niech $y_1, y_2, \ldots$ będzie rosnącym ciągiem wartościami zmiennej X. Załóżmy, że $y_n \geq \alpha$ dla $n \geq N$ (jeżeli takie N nie istnieje, to lewa strona nierówności jest równa 0 i jest OK).
Wtedy $E[X] = \sum_{i=1}^{\infty} P(X = y_i) \cdot y_i \geq \sum_{i=N}^{\infty} P(X = y_i) \cdot \alpha = P(X \geq \alpha) \cdot \alpha$.

Stosując to dla $\alpha = c \cdot nH_n$ oraz zmiennej T dostajemy, że

$$P[T \geq c \cdot nH_n] \leq \frac{1}{c}.$$

A zatem, prawdopodobieństwo, że kolekcjoner nakrętek będzie musiał kupić przynajmniej 3770 soczków by zebrać całą kolekcję jest mniejsze niż 50%, a że będzie musiał kupić przynajmniej 18850 nakrętek – 10%.

Dużo bardziej optymistyczne oszacowania daje nam poniższa nierówność.

Fakt 6. Nierówność Czebyszewa. Dla $\alpha > 0$ oraz nieujemnej zmiennej losowej X,

$$P[|X - E[X]| \geq \alpha] \leq \frac{Var[X]}{\alpha^2}.$$

Wynika ona z nierówności Markowa dla zmiennej $(X - E[X])^2$.

Stosując to do problemu kolekcjonera nakrętek, otrzymujemy, że

$$P(|T - nH_n| \geq cn) \leq \frac{1.65}{c^2} < \frac{2}{c^2}.$$

W naszym przypadku $n = 300$, biorąc $c \approx 6$ otrzymujemy, że prawdopodobieństwo, że kolekcjoner nakrętek będzie musiał kupić przynajmniej 3770 soczków jest mniejsze niż 5%, a prawdopodobieństwo, że kolekcjoner zakończy po kupnie między 1280 do 2485 soczków jest większe niż 60%.