Wstęp do rachunku prawdopodobieństwa
Wykład 6: Parametry rozkładów dyskretnych.

Przypomnijmy na początek, że z liniowości wartości oczekiwanej wynika poniższe twierdzenie.

Twierdzenie 1. *Dla dowolnych zmiennych losowych $X_1, X_2, \ldots, X_n$ zachodzi:*

$$\mathbb{E}(X_1 + X_2 + \ldots + X_n) = \mathbb{E}X_1 + \mathbb{E}X_2 + \ldots + \mathbb{E}X_n.$$

Niestety w przypadku wyznaczania wariancji sumy zmiennych losowych $X = X_1 + X_2 + \ldots + X_n$ sytuacja się nieco komplikuje i w ogólnym przypadku oprócz znajomości wariancji poszczególnych składników sumy, musimy też wyznaczyć kowariancje par zmiennych losowych występujących w tej sumie.

Twierdzenie 2. *Niech $X_1, X_2, \ldots, X_n$ będzie ciągiem zmiennych losowych. Jeśli dla każdego $i = 1, 2, \ldots, n$ istnieje $VarX_i$, wówczas*

$$Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} VarX_i + 2\sum_{1\leqslant i<j\leqslant n} Cov(X_i, X_j).$$

W szczególności, jeśli zmienne losowe $X_1, X_2, \ldots, X_n$ są **parami nieskorelowane***, tzn. dla każdych $1 \leqslant i < j \leqslant n$ mamy $Cov(X_i, X_j) = \rho(X_i, X_j) = 0$, to*

$$Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} VarX_i.$$

Dowód. W dowodzie korzystamy głównie z liniowości wartości oczekiwanej. Mamy zatem:

$$\begin{aligned}
\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) &= \mathbb{E}\left(\sum_{i=1}^{n} X_i\right)^2 - \left(\mathbb{E}\left(\sum_{i=1}^{n} X_i\right)\right)^2 \\
&= \mathbb{E}\left(\sum_{i=1}^{n} X_i^2 + 2\sum_{1\leqslant i<j\leqslant n} X_iX_j\right) - \left(\sum_{i=1}^{n} \mathbb{E}X_i\right)^2 \\
&= \sum_{i=1}^{n} \mathbb{E}(X_i^2) + 2\sum_{1\leqslant i<j\leqslant n} \mathbb{E}(X_iX_j) - \left(\sum_{i=1}^{n} (\mathbb{E}X_i)^2 + 2\sum_{1\leqslant i<j\leqslant n} \mathbb{E}X_i\mathbb{E}X_j\right) \\
&= \sum_{i=1}^{n} \left(\mathbb{E}(X_i^2) - (\mathbb{E}X_i)^2\right) + 2\sum_{1\leqslant i<j\leqslant n} (\mathbb{E}(X_iX_j) - \mathbb{E}X_i\mathbb{E}X_j) \\
&= \sum_{i=1}^{n} \mathrm{Var}X_i + 2\sum_{1\leqslant i<j\leqslant n} \mathrm{Cov}(X_i, X_j).
\end{aligned}$$

□

Twierdzenie 3 (Własności wariancji)**.** *Jeśli wariancja $VarX$ zmiennej losowej X istnieje, to dla dowolnej stałej $a \in \mathbb{R}$ zachodzi*

$$Var(aX) = a^2\, VarX$$

oraz

$$Var(X + a) = VarX.$$

Dowód. Znów wykorzystujemy tylko i wyłącznie liniowość wartości oczekiwanej. Mamy zatem:

$$\mathrm{Var}(aX) = \mathbb{E}((aX)^2) - (\mathbb{E}(aX))^2 = \mathbb{E}(a^2X^2) - (a\mathbb{E}X)^2 = a^2\left(\mathbb{E}(X^2) - (\mathbb{E}X)^2\right) = a^2\mathrm{Var}X$$

oraz

$$\begin{aligned}
\mathrm{Var}(X + a) &= \mathbb{E}(X + a)^2 - (\mathbb{E}(X + a))^2 = \mathbb{E}(X^2 + 2aX + a^2) - (\mathbb{E}X + a)^2 \\
&= \mathbb{E}(X^2) + 2a\mathbb{E}X + a^2 - (\mathbb{E}X)^2 - 2a\mathbb{E}X - a^2 = \mathrm{Var}X.
\end{aligned}$$

□

Definicja 1. *O zmiennych losowych $X_1, X_2, \ldots, X_n$ mówimy, że są* **parami niezależne**, *jeśli dla każdego $1 \leqslant i < j \leqslant n$ zmienne losowe X_i i X_j są niezależne.*

Przykład 1. Załóżmy, że zmienna losowa X przyjmuje wartości $0, 1, \ldots, n$ i może zostać przedstawiona w postaci

$$X = X_1 + X_2 + \ldots + X_n,$$

gdzie każda ze zmiennych losowych X_i, $i = 1, 2, \ldots, n$, ma rozkład dwupunktowy o zbiorze atomów $\{0, 1\}$ z prawdopodobieństwem $\mathbb{P}(X_i = 1) = p_i$. Takie zmienne losowe X_i nazywamy **zmiennymi losowymi indykatorowymi**. Ponadto załóżmy, że te zmienne losowe są parami niezależne, co z kolei implikuje $\mathrm{Cov}(X_i, X_j) = 0$. Wówczas możemy wyznaczyć wartość oczekiwaną oraz wariancję zmiennej losowej X w prosty sposób odwołując się do rozkładów zmiennych indykatorowych X_i. Mianowicie, dla każdego $i = 1, 2, \ldots, n$ mamy:

$$\mathbb{E}X_i = 0 \cdot \mathbb{P}(X_i = 0) + 1 \cdot \mathbb{P}(X_i = 1) = \mathbb{P}(X_i = 1) = p_i$$

oraz

$$\mathrm{Var}X_i = \mathbb{E}(X_i^2) - (\mathbb{E}X_i)^2 = 0^2 \cdot \mathbb{P}(X_i = 0) + 1^2 \cdot \mathbb{P}(X_i = 1) - p_i^2 = p_i - p_i^2.$$

A zatem otrzymujemy:

$$\mathbb{E}X = \mathbb{E}X_1 + \mathbb{E}X_2 + \ldots + \mathbb{E}X_n = p_1 + p_2 + \ldots + p_n$$

oraz

$$\mathrm{Var}X = \mathrm{Var}X_1 + \mathrm{Var}X_2 + \ldots + \mathrm{Var}X_n = (p_1 - p_1^2) + (p_2 - p_2^2) + \ldots + (p_n - p_n^2).$$

Przykład 2. Wyznaczmy wartość oczekiwaną oraz wariancję zmiennej losowej X o rozkładzie dwumianowym $Bin(n, p)$. Zauważmy, że możemy przedstawić tę zmienną losową w postaci sumy zmiennych losowych indykatorowych

$$X = X_1 + X_2 + \ldots + X_n,$$

gdzie dla $i = 1, 2, \ldots, n$ zmienna losowa X_i odpowiada wynikowi i-tej próby Bernoulliego, tzn. przyjmuje wartość 1 w przypadku sukcesu (dzieje się to z prawdopodobieństwem p) lub wartość 0 w przypadku porażki (z prawdopodobieństwem $1 - p$). Mamy zatem

$$\mathbb{E}X = \mathbb{E}X_1 + \mathbb{E}X_2 + \ldots + \mathbb{E}X_n = n \cdot \mathbb{E}X_1 = np.$$

Zauważmy ponadto, że ponieważ wyniki poszczególnych prób Bernoulliego nie wpływają na siebie, zmienne losowe $X_1, X_2, \ldots, X_n$ są niezależne. Dzięki tej obserwacji otrzymujemy

$$\mathrm{Var}X = \mathrm{Var}(X_1 + X_2 + \ldots + X_n) = \sum_{i=1}^{n} \mathrm{Var}X_i = n \cdot \mathrm{Var}X_1 = n \cdot \left(\mathbb{E}(X_1^2) - (\mathbb{E}X_1)^2\right) = n \cdot (p - p^2) = np(1 - p).$$

Przykład 3. Niech X będzie zmienną losową o rozkładzie geometrycznym $Ge(p)$ z parametrem p. Wiemy już, że wartość oczekiwana tej zmiennej losowej dana jest wzorem

$$\mathbb{E}X = \frac{1}{p}.$$

Chcemy teraz policzyć jej wariancję. Niestety w tym przypadku nie możemy wyrazić zmiennej losowej X w postaci sumy zmiennych indykatorowych, ponieważ nie wiemy a priori ile prób Bernoulliego zostanie oddanych do momentu uzyskania pierwszego sukcesu. Zatem wyznaczymy wariancję zmiennej losowej X wprost z definicji. W tym celu musimy wyznaczyć na początek wartość oczekiwaną zmiennej losowej X^2. Zanim jednak to zrobimy przytoczymy wzór, który pośrednio wyprowadziliśmy na jednym z poprzednich wykładów podczas wyznaczania wartości oczekiwanej zmiennej losowej o rozkładzie geometrycznym:

$$\sum_{k=1}^{\infty} k x^{k-1} = \frac{1}{(1 - x)^2} \quad \text{dla} \quad |x| < 1.$$

Dowodząc powyższą tożsamość skorzystaliśmy z własności sumowania szeregów, w tym z całkowania i różniczkowania. Używając podobnych narzędzi możemy również udowodnić poniższą tożsamość:

$$\sum_{k=1}^{\infty} k(k-1)x^{k-2} = \frac{2}{(1-x)^3} \quad \text{dla} \quad |x| < 1.$$

Przejdźmy teraz do wyznaczenia wartości oczekiwanej zmiennej losowej X^2. Mamy:

$$\begin{aligned}
\mathbb{E}X^2 &= \sum_{k=1}^{\infty} k^2 p(1-p)^{k-1} = p\sum_{k=1}^{\infty}\Big(k(k-1)(1-p)^{k-1} + k(1-p)^{k-1}\Big) \\
&= p\left((1-p)\sum_{k=1}^{\infty} k(k-1)(1-p)^{k-2} + \sum_{k=1}^{\infty} k(1-p)^{k-1}\right) \\
&= p\left(\frac{2(1-p)}{p^3} + \frac{1}{p^2}\right) = \frac{2-2p+p}{p^2} = \frac{2-p}{p^2}\,.
\end{aligned}$$

Zatem wariancja zmiennej losowej o rozkładzie geometrycznym $Ge(p)$ wynosi:

$$\mathrm{Var}X = \mathbb{E}(X^2) - (\mathbb{E}X)^2 = \frac{2-p}{p^2} - \frac{1}{p^2} = \frac{1-p}{p^2}\,.$$

Przykład 4. Rozważmy jeszcze raz eksperyment, w którym powtarzamy niezależne próby Bernoulliego, ale tym razem interesuje nas liczba prób potrzebnych do uzyskania łącznie r sukcesów. Zmienna losowa X zwracająca liczbę takich prób ma **rozkład Pascala** z parametrami p oraz r, gdzie p jak zawsze jest prawdopodobieństwem sukcesu w pojedynczej próbie. Zbiorem atomów zmiennej losowej X jest zbiór $\{r, r+1, r+2, \ldots\}$, natomiast funkcja masy prawdopodobieństwa dana jest wzorem:

$$\mathbb{P}(X = k) = \binom{k-1}{r-1} p^r (1-p)^{k-r} \quad \text{dla} \quad k = r, r+1, \ldots$$

Gdybyśmy chcieli wyznaczyć wartość oczekiwaną i wariancję zmiennej losowej X wprost z definicji, okazałoby się, że rachunki mocno się komplikują. Dlatego też zrobimy to w sposób sprytny. Mianowicie, jeżeli podzielimy nasz eksperyment na r etapów, gdzie każdy kolejny etap oznacza powtarzanie prób Bernoulliego do uzyskania kolejnego sukcesu, to możemy zmienną losową X wyrazić w postaci sumy

$$X = X_1 + X_2 + \ldots + X_r,$$

gdzie X_1 oznacza liczbę prób oddanych do uzyskania pierwszego sukcesu, a dla $i = 2, \ldots, r$, zmienna losowa X_i oznacza liczbę prób oddanych po sukcesie numer $i-1$ aż do uzyskania i-tego sukcesu. W szczególności każda ze zmiennych losowych X_i, $i = 1, 2, \ldots, r$, ma rozkład geometrycznym z parametrem p. Co więcej, zmienne losowe $X_1, X_2, \ldots, X_r$ są niezależne. Możemy zatem szybko wyznaczyć wartość oczekiwaną oraz wariancję zmiennej losowej X:

$$\mathbb{E}X = \mathbb{E}X_1 + \mathbb{E}X_2 + \ldots + \mathbb{E}X_r = r \cdot \mathbb{E}X_1 = \frac{r}{p}$$

oraz

$$\mathrm{Var}X = \mathrm{Var}X_1 + \mathrm{Var}X_2 + \ldots + \mathrm{Var}X_r = r \cdot \mathrm{Var}X_1 = \frac{r(1-p)}{p^2}\,.$$

Przykład 5. Rzucamy 100 razy symetryczną monetą. Niech X oznacza liczbę orłów wyrzuconych w pierwszych 80 rzutach, a Y – w całej serii. Czy zmienne losowe X i Y są niezależne? Ile wynosi $\rho(X, Y)$.

Spróbujmy najpierw odpowiedzieć na pytanie czy zmienne losowe X i Y są niezależne? Wydaje się, że odpowiedź powinna być negatywna, skoro na pewno zachodzi $X \leqslant Y$, czyli wartość zmiennej losowej X ma wpływ na wartość zmiennej losowej Y. Nie będziemy na razie dowodzić tego formalnie, tylko policzymy na początek $\rho(X, Y)$, bo jeśli okaże się, że $\rho(X, Y) \neq 0$, wówczas na pewno X i Y nie mogą być niezależne.

Tak jak w poprzednich przykładach możemy wprowadzić zmienne losowe indykatorowe $X_1, X_2, \ldots, X_{100}$ odpowiadające wynikom poszczególnych rzutów, tzn. dla $i = 1, 2, \ldots, 100$ mamy

$$X_i = \begin{cases} 1, & \text{jeśli w i-tym rzucie wypadł orzeł,} \\ 0, & \text{w przeciwnym przypadku.} \end{cases}$$

Wtedy:

$$X = X_1 + X_2 + \ldots + X_{80}, \quad Y = X_1 + X_2 + \ldots + X_{100}.$$

W szczególności wprowadzając nową zmienną losową

$$Z = X_{81} + X_{82} + \ldots + X_{100}$$

otrzymujemy

$$Y = X + Z.$$

Zmienne losowe X i Z mają rozkład dwumianowy

$$X \sim Bin\left(80, \frac{1}{2}\right), \quad Z \sim Bin\left(20, \frac{1}{2}\right).$$

Ponadto zmienne losowe X i Z są niezależne, ponieważ pierwsza z nich dotyczy początkowych 80 rzutów, a druga ostatnich 20. A zatem otrzymujemy:

$$\begin{aligned}
\mathrm{Cov}(X,Y) &= \mathrm{Cov}(X, X+Z) = \mathbb{E}(X(X+Z)) - \mathbb{E}X\mathbb{E}(X+Z) \\
&= \mathbb{E}(X^2) + \mathbb{E}(XZ) - (\mathbb{E}X)^2 - \mathbb{E}X\mathbb{E}Z \\
&= \mathbb{E}(X^2) - (\mathbb{E}X)^2 + \mathbb{E}(XZ) - \mathbb{E}(XZ) = \mathrm{Var}X.
\end{aligned}$$

Teraz wystarczy wyznaczyć wariancje zmiennych losowych X i Y:

$$\mathrm{Var}X = 80 \cdot \frac{1}{2} \cdot \frac{1}{2} = 20,$$

$$\mathrm{Var}Y = \mathrm{Var}(X+Z) = \mathrm{Var}X + \mathrm{Var}Z = 20 + 20 \cdot \frac{1}{2} \cdot \frac{1}{2} = 25.$$

Ostatecznie otrzymujemy:

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}X}\sqrt{\mathrm{Var}Y}} = \frac{20}{\sqrt{20} \cdot \sqrt{25}} = \frac{2\sqrt{5}}{5}.$$

W szczególności możemy stwierdzić, że zmienne losowe X i Y nie są niezależne.