Matematyka dla biologów — Zajęcia nr 12.

Rachunek prawdopodobieństwa

Zmienne losowe

Przebieg różnych zjawisk losowych wygodnie jest opisywać za pomocą specjalnie wybranych funkcji, określonych na przestrzeni probabilistycznej, które zawierają najważniejsze informacje o przebiegu danego zjawiska. Jako przykład może służyć funkcja, która po ustaleniu stawek, opisuje wartość wygranej przy grze polegającej na rzutach monetą. Wartości tej funkcji niosą najważniejszą informacje dla gracza o rezultacie gry. Tego typu funkcje nazywa się **zmiennymi losowymi**.

Definicja zmiennej losowej

Definicja

Zmienną losową nazywamy funkcję przyjmującą wartości w zbiorze liczb rzeczywistych określoną na zbiorze zdarzeń elementarnych.

Niech $(\Omega\,, P)$ będzie przestrzenią probabilistyczną. Jeśli zmienna losowa $X : \Omega \to \mathbb{R}$ przyjmuje wartości dyskretne tzn. jej zbiór wartości jest skończony $x_1\,, x_2\,, \ldots x_n$ to wtedy **rozkładem zmiennej losowej** X nazywamy zbiór R_X par , z których każda określa z jakim prawdopodobieństwem zmienna losowa przyjmuje daną wartość

$$R_X = \{(x_1\,, p_1)\,, (x_2\,, p_2) \ldots (x_n\,, p_n)\}$$

gdzie

$$p_i = P(\{\omega : X(\omega) = x_i\}) \text{ lub w skróconym zapisie } p_i = P(X = x_i).$$

Zbiór R_X ma tyle elementów ile różnych wartości przyjmuje zmienna X. Ze względu na to, że zbiór wartości zmiennej X jest skończony taki rozkład nazywa się **rozkładem dyskretnym zmiennej losowej**, a samą zmienną losową nazywamy się wtedy zmienną losową dyskretną.

Trzeba podkreślić, że sam rozkład prawdopodobieństwa nie niesie pełnej informacji o zmiennej losowej jako o funkcji, określa jedynie z jakimi prawdopodobieństwami dana zmienna losowa przyjmuje swoje wartości.

Przykład

Wykorzystując rzut kością określimy grę losową: jeśli wypadnie ”6” gracz otrzymuje 90zł, a jeśli wypadnie nieparzysta liczba oczek otrzymuje 10 zł i nic nie traci ani nic nie otrzymuje w pozostałych przypadkach. Wtedy :

$$\Omega = \{1,2,3,4,5,6\}\,, P(\{i\}) = \frac{1}{6} \overset{ozn.}{=} q_i \;\; i = 1,2\ldots,6\,.$$

Zmienną losową, która opisuje wartości wygranych, oznaczymy przez Y. Przyjmuje ona tylko trzy wartości: $0\,, 10\,, 90$, a więc $P(Y=0) = q_2 + q_4 = \frac{1}{3}\,, P(Y=10) = q_1 + q_3 + q_5 = \frac{1}{2}\,, P(Y=90) = q_6 = \frac{1}{6}$ i jej rozkład jest następujący

$$\left\{\left(0\,,\frac{1}{3}\right)\,,\,\left(10\,,\frac{1}{2}\right)\,,\,\left(90\,,\frac{1}{6}\right)\right\}\,.$$

Duże znaczenie w rachunku prawdopodobieństwa mają charakterystyki liczbowe zmiennych losowych - wartość oczekiwana EX oraz wariancja, oznaczana jako $VarX$ lub D^2X. Można je wyrazić znając jedynie ich rozkłady. Wartość oczekiwaną zdefiniował w 1658 roku Huygens (Christiaan Huygens (1629-1695)) w pracy poświęconej teorii gry w kości.

Wartość oczekiwana zmiennej losowej

Definicja

Wartością oczekiwaną dyskretnej zmiennej losowej X o rozkładzie R_X nazywamy liczbę

$$EX = \sum_{i=1}^{n} x_i p_i \,.$$

Dla zmiennej Y z przykładu ; $EY = 10 \cdot \frac{3}{6} + 0 + 90\frac{1}{6} = 20\,.$

Rozkład jednostajny

Rozpatrzmy rozkład R_J zmiennej losowej przyjmującej n wartości z tym samym prawdopodobieństwem $\frac{1}{n}$

$$R_J = \{(x_1 , \frac{1}{n}) , (x_2 , \frac{1}{n}) \dots (x_i , \frac{1}{n}) , (x_n , \frac{1}{n})\} .$$

Przykładem zmiennej losowej o takim rozkładzie jest zmienna losowa J określona na przestrzeni probabilistycznej

$$\Omega = \{\omega_i : i = 1 , 2 , \dots , n\}$$

takiej, że zdarzenia są jednakowo prawdopodobne tzn. $P(\{\omega_i\}) = \frac{1}{n}$ dla każdego i. Wartością oczekiwaną zmiennej losowej określonej jako $J(\omega_i) = x_i$ jest **wartość średnia** ze wszystkich wartości tej zmiennej

$$EJ = \sum_{i=1}^{n} x_i \frac{1}{n} = \frac{\sum_{i=1}^{n} x_i}{n} .$$

Prawdopodobieństwa jako wagi

Z tego punktu widzenia w przypadku ogólnym ujętym w definicji można powiedzieć, że wartość oczekiwana jest średnią ważoną, której wagi czyli prawdopodobieństwa p_i określają jak wielki jest wkład poszczególnych wartości x_i do całkowitej ”średniej”.

Podstawowe własności wartości oczekiwanej

1. Jeśli G jest zmienną losową będącą złożeniem zmiennej $X : \Omega \to \mathbb{R}$ i funkcji $g : \mathbb{R} \to \mathbb{R}$ tzn.

$$G(\omega_i) = g(X(\omega_i)), \quad \omega_i \in \Omega,$$

to

$$EG = E(g(X)) = \sum_{i=1}^{n} g(x_i)p_i.$$

2. Nietrudno udowodnić wprost z definicji, że jeśli X i Y są dwiema zmiennymi losowymi określonymi na tej samej przestrzeni probabilistycznej a a i b są liczbami to

$$E(aX + bY) = aEX + bEY.$$

3. Jeśli X_c jest zmienną losową przyjmującą tylko jedną wartość c to

$$EX_c = c\sum_{i=1}^{n} p_i = c.$$

Wariancja

Miarą tego jak bardzo wartości zmiennej losowej X odbiegają od wartości oczekiwanej jest **wariancja** oznaczana jako $VarX$ lub D^2X i **dyspersja** $DX = \sqrt{D^2X}$. Dyspersja określa innymi słowy średni rozrzut zmiennej losowej. Dyspersja bywa też nazywana **odchyleniem standardowym**.

Definicja

Oznaczając przez m_x wartość oczekiwaną zmiennej X jej **wariancją** *w przypadku dyskretnej zmiennej losowej nazywamy liczbę*

$$D^2X = VarX := E((X - m_x)^2) = \sum_{i=1}^{n}(x_i - m_x)^2 p_i$$

Skoro definiowana wielkość ma być miarą średniego rozrzutu wartości zmiennej losowej, to uzasadnione jest pytanie dlaczego nie określić jej jako wartości oczekiwanej odległości pomiędzy wartością zmiennej od średniej tzn.

$$E|X - m_x| = \sum_{i=1}^{n} |x_i - m_x| p_i \,.$$

To nie jest zły pomysł, ale niepraktyczny z rachunkowego punktu wiedzenia, gdyż moduł nie ma tak dobrych własności arytmetycznych. Oto przykład obliczenia wariancji wprost z definicji

$$VarX := E((X - m_x)^2) = E(X^2) - 2(m_x)EX + (m_x)^2 = E(X^2) - (m_x)^2 \,. \quad (1)$$

Aby obliczyć wariancję wystarczy zatem obliczyć EX i $E(X^2)$.

Łatwo sprawdzić, inną ważną własność:

jeśli X jest zmienną losową, a a i b pewnymi liczbami to

$$Var(aX + b) = a^2 VarX \,. \quad (2)$$

Niezależność zmiennych losowych

Skoro zmienne losowe opisują rezultaty różnych zjawisk losowych to naturalne jest pytanie o wzajemne związki pomiędzy różnymi zmiennymi losowymi (określonymi na tej samej przestrzeni zdarzeń elementarnych). Tego typu związki mogą wyrażać istnienie związków przyczynowo-skutkowych pomiędzy tymi zjawiskami. Równie ważne bywa określenie braku tego typu związku. Dobrym przykładem jest pobieranie próbek z jakiejś populacji. Posłużymy się przykładem zaczerpniętym z książki Łomnickiego ”Statystyka dla Biologów”.

Płeć osobnika wybranego z jakiejś populacji możemy uznać za realizację zmiennej losowej dwuwartościowej. Niezależność prób oznacza tu, że odłowienie osobnika jednej płci nie ma wpływu na następny odłów. W przypadku ptaków, które występują w czasie rozrodu parami (jest tak u synogarlic tureckich) warunek ten może nie być spełniony bo odławiając samicę zwiększamy prawdopodobieństwo schwytania w następnym odłowie jej partnera. Jeśli zaś ptaki trzymają się w grupach jednopłciowych, odłowienie samicy zwiększa prawdopodobieństwo schwytania w drugim odłowie następnej samicy, kolejne próby nie są zatem niezależne.

Niezależnośc zmiennych losowych

Definicja

Dwie zmienne losowe $X : \Omega \mapsto \mathbb{R}$ *i* $Y : \Omega \mapsto \mathbb{R}$ *o rozkładach dyskretnych*

$$R_X = \{(x_1 , p_1) , \ldots (x_i , p_i) , \ldots (x_n , p_n)\} \tag{3}$$

oraz

$$R_Y = \{(y_1 , q_1) , \ldots (y_j , q_i) \ldots , (y_m , q_m)\} \tag{4}$$

są niezależne *jeśli dla dowolnych wartości* x_i *oraz* y_j *które przyjmują, zachodzi*

$$P(X = x_i , Y = y_j) = p_i q_j$$

gdzie $P(X = x_i , Y = y_i)$ *oznacza prawdopodobieństwo zdarzenia, że X przyjęła wartość* x_i *i zmienna losowa Y przyjęła wartość* y_j *. Podobnie definiuje się niezależność dowolnej liczby zmiennych losowych większej od dwóch.*

Niezależne zmienne losowe mają kilka bardzo ważnych własności.

Stwierdzenie

Jeżeli zmienne losowe X i Y o rozkładach (3)-(4)są niezależne to

$$E\,XY = EX \cdot EY$$

Dowód. Ustalmy iloczyn $x_i y_j$. Zauważmy, że wśród wartości zmiennych X i Y może być więcej par liczb, które po wymnożeniu dają $x_i y_j$ np. $x_i = 5$, $y_j = 10$ oraz $x_k = 2$ i $y_l = 25$. Aby znaleźć rozkład zmiennej XY, dla każdego takiego iloczynu $x_i y_j$ trzeba zatem posumować wszystkie prawdopodobieństwa zdarzeń postaci $P(X = x_k\ , Y = y_l)$ o tej własności, że $x_i y_j = x_k y_l$. Ponieważ zmienne losowe są niezależne to

$$P(X = x_k\ , Y = y_l) = p_k q_l$$

Stąd i z definicji wartości oczekiwanej można wydedukować, że

$$E\,XY = \sum x_i y_j p_i q_j\,, \tag{5}$$

gdzie suma brana jest po wszystkich i i j takich, że $1 \leqslant i \leqslant n\,, 1 \leqslant j \leqslant m\,.$ Z drugiej strony łatwo sprawdzić, że

$$(EX)(EY) = (\sum_{i=1}^{n} x_i p_i)(\sum_{j=1}^{m} y_j q_j)$$

równe jest właśnie (5). Pamiętamy, że wartość oczekiwana sumy zmiennych losowych jest sumą ich wartości oczekiwanych. Odpowiemy teraz na pytanie postawione wcześniej.

Bardzo często w rachunku prawdopodobieństwa i statystyce rozważa się sumy

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

które reprezentować mogą na przykład kolejne wyniki pomiarów jakieś wielkości i chcemy policzyć np. średnią tych wyników. Powstaje naturalne pytanie, **jaki rozkład ma suma zmiennych losowych jeśli znamy rozkłady każdej ze zmiennych składowych?**

Wariancja sumy

Czy wariancja sumy zmiennych losowych jest sumą wariancji?

Okazuje się, że w ogólności tak być nie musi, ale jest to prawdą jeśli zmienne losowe są niezależne.

Kowariancja zmiennych losowych

Oznaczmy: $EX = m_x$, $EY = m_y$.

Definicja

Kowariancją *zmiennych losowych X i Y nazywa się liczbę*

$$Cov(X, Y) = E((X - m_x)(Y - m_y))$$

Łatwo sprawdzić, że $Cov(X, Y) = E(XY) - m_x m_y$ Stąd wynika ważny wniosek.

Jeśli zmienne losowe X i Y są niezależne to $Cov(X, Y) = 0$.

Stwierdzenie

Przyjmijmy założenia takie jak w poprzednim twierdzeniu i oznaczenia jak wyżej. Wtedy

$$Var(X + Y) = VarX + VarY + 2Cov(X, Y).$$

Korelacje

Współczynnikiem korelacji zmiennych losowych X i Y nazywamy liczbę z przedziału $[-1\,,1]$ równą

$$\varrho(X,Y) = \frac{Cov(X,Y)}{\sqrt{VarX}\cdot\sqrt{VarY}}\,.$$

- Jeśli $\varrho(X, Y) = 0$ to zmienne losowe nazywa się nieskorelowanymi. Stąd wynika, że **jeśli dwie zmienne losowe są niezależne to są nieskorelowane.**
- Jeśli $\varrho(X, Y) = 1(-1)$ to zmienne nazywa się dodatnio (ujemnie) skorelowanymi.
- Łatwo sprawdzić, korzystając z (1) i (2), że jeśli $Y = aX$, gdzie a to pewna liczba, to

$$\varrho(X, Y) = \begin{cases} 1 & \text{gdy} \quad a > 0\,. \\ -1 & \text{gdy} \quad a < 0\,, \end{cases}$$

  Zmienne X i aX są zatem dodatnio lub ujemnie skorelowane w zależności od znaku a.

Stwierdzenie, że dwa zjawiska (losowe) np. cechy osobnicze w badanej populacji są dodatnio skorelowane oznacza w praktyce, że zjawiska te współwystępują i może, ale nie musi występować pomiędzy nimi związek przyczynowo skutkowy.

Ciąg prób Bernoulliego

Szczególną rolę w rachunku prawdopodobieństwa i statystyce pełni opis rezultatów serii powtórzeń jakiegoś doświadczenia w przypadku gdy kolejne doświadczenia są wzajemnie niezależne. Typowym przykładem jest seria rzutów monetą lub kością do gry jeśli zapewni się przy każdym rzucie idealnie takie same warunki dotyczące stanu przedmiotu którym rzucamy.

Rozważmy ciąg n niezależnych doświadczeń taki, że w wyniku każdego doświadczenia może zajść zdarzenie $A \subset \Omega$ lub przeciwne do niego zdarzenie $\bar{A}$. Często jedno ze zdarzeń A lub $\bar{A}$ nazywa się umownie sukcesem. Przyjmijmy, że zdarzenie A (sukces) zachodzi z prawdopodobieństwem p a zdarzenie $\bar{A}$ z prawdopodobieństwem $q\,,\ p+q=1$. Niech $X_i : \Omega_n \to \{1\,,0\}$ będzie zmienną losową równą 1 gdy w i-tym doświadczeniu zaszło zdarzenie A i równą 0 w przeciwnym przypadku.

Definicja

Ciąg niezależnych zmiennych losowych $X_1\,, X_2\,, \ldots X_n$ z których każda ma ten sam rozkład

$$\{(1\,,p)\,,(0\,,q)\}$$

nazywa się **ciągiem prób Bernoulliego**.

Ciąg prób Bernoulliego dla $p = q = \frac{1}{2}$ jest oczywiście modelem probabilistycznym opisującym wyniki serii rzutów monetą symetryczną. Ponieważ zmienne losowe w ciągu prób Bernoulliego są niezależne, prawdopodobieństwo wystąpienia serii określonych wyników jest iloczynem prawdopodobieństw otrzymania każdego wyniku w poszczególnych próbach. Na przykład w przypadku czterech prób

$$\begin{aligned} &P(X_1 = 0\,, X_2 = 1\,, X_3 = 1\,, X_4 = 1) \\ &= P(X_1 = 0)P(X_2 = 1)P(X_3 = 1)P(X_4 = 1) = p^3 q\,. \end{aligned}$$

Rozkład dwumianowy

Rozkład ten ma kluczowe znaczenie w zastosowaniach rachunku prawdopodobieństwa gdyż opisuje liczbę sukcesów w ciągu prób Bernoulliego. Szukamy zatem rozkładu zmiennej losowej

$$S_n = \sum_{i=1}^{n} X_i \,,$$

która przyjmuje wartość k jeśli dokładnie k razy w ciągu prób Bernoulliego zaszło zdarzenie A. Aby znaleźć rozkład tej zmiennej obliczamy prawdopodobieństwo zdarzenia, że dokładnie k składników w powyższej sumie przyjmuje wartość 1 a pozostałe wartość 0. Prawdopodobieństwo zdarzenia, że wybrane k zmiennych przyjęło wartość 1 a pozostałe wartość 0 wynosi $p^k(1-p)^{n-k}$. Zwróćmy uwagę, że k składników w n-składnikowej sumie można wybrać na tyle sposobów ile jest kombinacji k-elementowych ze zbioru n-elementowego, czyli

$$\binom{n}{k} = \frac{n!}{(n-k)!k!}$$

Sumując prawdopodobieństwa rozłącznych zdarzeń odpowiadającym seriom zawierającym dokładnie k sukcesów osiągniętych w różnych próbach otrzymujemy

$$P(S_n = k) \stackrel{ozn.}{=} b_{n,p}(k) = \binom{n}{k} p^k (1-p)^{n-k} .$$

Taki rozkład zmiennej losowej nazywamy **rozkładem dwumianowym (ang. binomial distribution) o parametrach n i p**. Łatwo sprawdzić stosując Wniosek(1.8), że

$$ES_n = np \quad VarS_n = np(1-p) .$$

Przykład

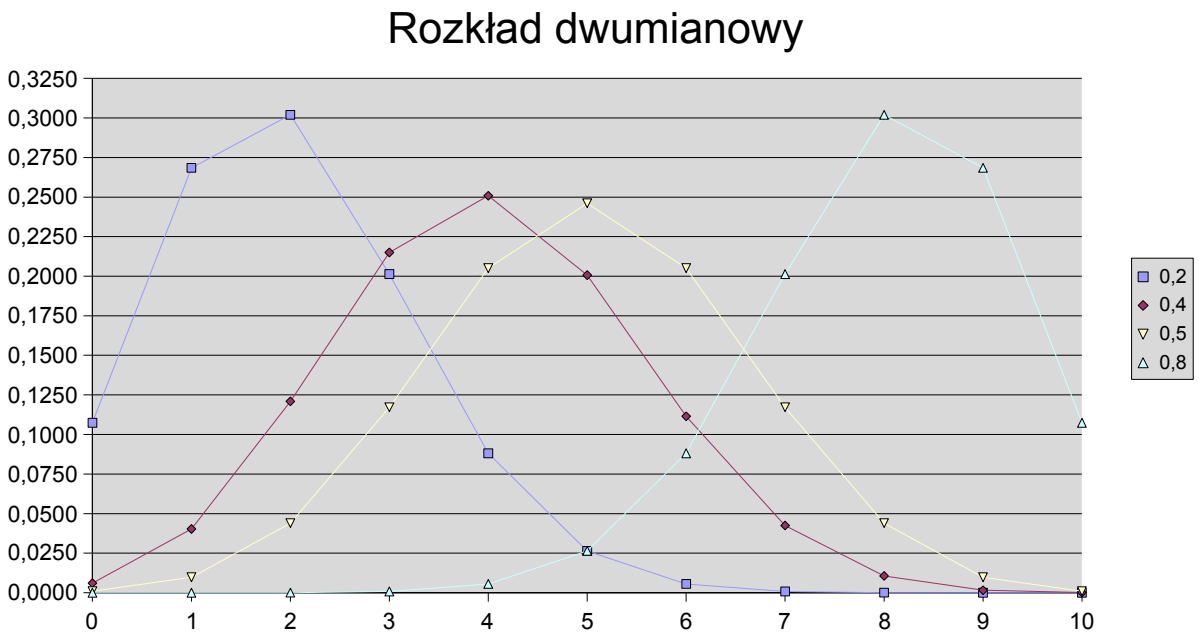
Typowe zastosowanie rozkładu dwumianowego znajdujemy w klasycznej (mendlowskiej) **genetyce populacyjnej.** Rozpatrzmy mianowicie dużą populację organizmów diploidalnych i zajmijmy się jednym locus, któremu odpowiadają dwie alle A i a. Przyjmijmy, że allel A występuje w populacji z częstością p a allel a z częstością q , $p + q = 1$, oraz, że nie występują mutacje. Wiemy, że zamiast częstości możemy tu mówić o prawdopodobieństwie występowania danego allelu. Przyjmując **całkowicie losowe kojarzenie w pary** i brak sprzężeń pomiędzy genami możemy przyjąć, że allel, który trafia na odpowiednie miejsce na jednym z chromosomów homologicznych jest losowany z puli genów, niezależnie dla każdego z dwóch homologicznych chromosomów.

To tak jakbyśmy dla obsadzenia obu miejsc na chromosomach homologicznych dwukrotnie rzucali monetą. W przypadku allelu A prawdopodobieństwo jego wylosowania wynosi p a dla allelu a odpowiednio q. Jeśli wylosowanie A nazwiemy sukcesem to prawdopodobieństwo występowania w kolejnej generacji ”genotypu”

- AA odpowiada prawdopodobieństwu wystąpienia dwóch sukcesów w dwóch próbach Bernoulliego czyli $\binom{2}{2} p^2 q^0 = p^2 \overset{ozn.}{=} u$,
- Aa odpowiada prawdopodobieństwu wystąpienia jednego sukcesu w dwóch próbach Bernoulliego czyli $\binom{2}{1} p^1 q^1 = 2pq \overset{ozn.}{=} v$,
- aa odpowiada prawdopodobieństwu nie wystąpienia ani jednego sukcesu w dwóch próbach Bernoulliego czyli $\binom{2}{0} p^0 q^2 = q^2 \overset{ozn.}{=} w$.

Jest to rozkład dwumianowy.

Rozkład dwumianowy



Rozkład dwumianowy: $n = 10, p = 0,2\ ,\ 0,4\ ,\ 0,5\ ,\ 0,8\ ..$

Zadania

1. W pewnym lesie występuje N zwierząt danego gatunku, w tym M zaobrączkowanych. Znaleźć wzór, który określa prawdopodobieństwo, że wśród n losowo schwytanych zwierząt k jest niezaobrączkowanych. (**wskaz.** Zastosować schemat Bernoulliego.)

2. Wykonujemy niezależne rzuty monetą symetryczną. W czterech kolejnych rzutach otrzymaliśmy orły. Czy prawdopodobieństwo wyrzucenia reszki w piątym rzucie jest większe od $\frac{1}{2}$?

3. Pewna gra polega na rzucie kostką i monetą. Wygrana występuje przy łącznym wyrzuceniu piątki i orła. Jakie jest prawdopodobieństwo, że w trzech grach wygrana wystąpi dokładnie raz?

wskaz. Zastosować schemat Bernoulliego do ciągu doświadczeń, w którym sukcesem jest jednoczesne wyrzucenie orła i piątki. Rzuty kością i monetą traktujemy jako zdarzenia niezależne. Odp