Rozkłady prawdopodobieństwa

Statystyka i analiza danych 2019

Jurek Błaszczyński

17 marca 2019

Rozkład normalny

Zmienna losowa $X \sim N(\mu, \sigma^2)$ ma następującą gęstość prawdopodobieństwa (PDF):

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

można pokazać, że:

$$E[X] = \mu, D^2[X] = \sigma^2$$

Rodzina rozkładów normalnych ma przydatną własność

Dowolna kombinacja liniowa zmiennych losowych o rozkładzie normalnym ma rozkład normalny

Przykład całkowania metodą Monte Carlo

Metodą Monte Carlo można obliczyć pole figury zdefiniowanej nierównością:

$$x^2 + y^2 \leq r^2,$$

czyli koła o promieniu r i środku w punkcie $(0, 0)$.

1. Losuje się n punktów z opisanego na tym kole kwadratu – dla koła o $r = 1$ współrzędne wierzchołków $(-1, -1)$, $(-1, 1)$, $(1, 1)$, $(1, -1)$.
2. Po wylosowaniu każdego z tych punktów trzeba sprawdzić czy jego współrzędne spełniają powyższą nierówność (tj. czy punkt należy do koła).

Wynikiem losowania jest informacja, że z n wszystkich prób k było trafionych, zatem (dla $r = 1$) pole koła wynosi:

$$P_k = P \times \frac{k}{n},$$

gdzie P jest polem kwadratu opisanego na tym kole ($P = 4$).

Jak wylosować liczbę z rozkładu normalnego mając do dyspozycji tylko generator liczb z zakresu $[0, 1]$ (jednostajnie)?

$$\mathit{Unif}\,[0,1] \rightarrow 0 \ldots 1,$$

$$X \sim \mathit{Unif}\,[0,1],$$
$$Z = \Phi^{-1}(X),$$

gdzie Φ^{-1} jest kwantylem rozkładu $N(0,1)$. To znaczy:

$$Z \sim N(0,1), \Phi(z) = P(Z < z) = p, \Phi^{-1}(p) = z.$$

- **Dwupunktowy** $B_1(p)$, dla zmiennej losowej $X \in \{0, 1\}$ o funkcji prawdopodobieństwa, wartości oczekiwanej i wariancji:

$$P(X = 1) = 1 - P(X = 0) = p, \quad E[X] = p, \quad D^2[X] = p(1 - p).$$

- **Dwumianowy** $B_n(p)$:

$$X = \sum_{k=1}^{n} X_k, \qquad \text{gdzie} \;\; X_k \sim B_1(p).$$

Suma n niezależnych zmiennych dwupunktowych. Zachodzi:

$$E[X] = np, \qquad D^2[X] = np(1 - p).$$

Użyteczne własności

Niech $X_1, \ldots, X_n$ – **i.i.d.**[1] oraz $E[X_k] = \mu$, $D^2[X_k] = \sigma^2$.

- Dla sumy $X = \sum_{k=1}^{n} X_k$:

$$E[X] = E[X_1] + \ldots + E[X_n] = n\mu,$$
$$D^2[X] = D^2[X_1] + \ldots + D^2[X_n] = n\sigma^2.$$

- Dla średniej $\overline{X} = \frac{1}{n}\sum_{k=1}^{n} X_k = \frac{1}{n}X$:

$$E[\overline{X}] = \frac{1}{n}E[X] = \mu,$$
$$D^2[\overline{X}] = \frac{1}{n^2}D^2[X] = \frac{\sigma^2}{n}, \qquad \text{czyli} \quad D[\overline{X}] = \frac{\sigma}{\sqrt{n}}.$$

[1] i.i.d.: *independent and identically distributed* – niezależne zmienne losowe o tym samym rozkładzie

Prawo wielkich liczb

$X_1, \ldots, X_n$ – **i.i.d.**[2], $E[X_k] = \mu$, $D^2[X_k] = \sigma^2$.

$$E[\overline{X}] = \mu, \qquad D[\overline{X}] = \frac{\sigma}{\sqrt{n}}.$$

Rozkład średniej **koncentruje się** wokół wartości oczekiwanej μ.

[2] i.i.d.: *independent and identically distributed* – niezależne zmienne losowe o tym samym rozkładzie

Centralne Twierdzenie Graniczne

Znane też jest pod nazwą **Twierdzenie Lindeberga-Lévy'ego**. Niech $X_1, \dots, X_n$ **– i.i.d.**, $E[X_k] = \mu$, $D^2[X_k] = \sigma^2$.

Zdefiniujmy **ustandaryzowaną** średnią $Z = \frac{\overline{X}-\mu}{\sigma}\sqrt{n}$.

$$E[Z] = 0, \qquad D^2[Z] = 1.$$

Zachodzi:

$$Z \xrightarrow{D} N(0,1).$$

Zgodnie z Centralnym Twierdzeniem Granicznym rozkład normalny występuje powszechnie.