WIKIPEDIA

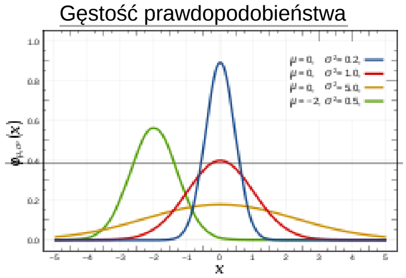
Rozkład normalny

Rozkład normalny, **rozkład Gaussa** (w literaturze francuskiej zwany rozkładem Laplace’a-Gaussa) – jeden z najważniejszych rozkładów prawdopodobieństwa odgrywający ważną rolę w statystyce. Wykres funkcji prawdopodobieństwa tego rozkładu jest krzywą w kształcie dzwonu (tak zwaną krzywą dzwonową).

Przyczyną jego znaczenia jest częstość występowania w naturze. Jeśli jakaś wielkość jest sumą lub średnią bardzo wielu drobnych losowych czynników, to niezależnie od rozkładu każdego z tych czynników jej rozkład będzie zbliżony do normalnego (centralne twierdzenie graniczne) – dlatego można go bardzo często zaobserwować w danych[a]. Ponadto rozkład normalny ma interesujące właściwości matematyczne, dzięki którym oparte na nim metody statystyczne są proste obliczeniowo[b].

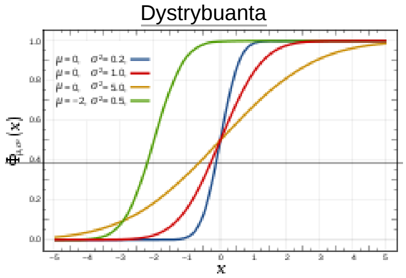
Rozkład normalny

Gęstość prawdopodobieństwa



Czerwona linia odpowiada standardowemu rozkładowi normalnemu

Dystrybuanta



Kolory odpowiadają wykresowi powyżej

Parametry	μ położenie (liczba rzeczywista) $\sigma^2 > 0$ podniesiona do kwadratu skala (liczba rzeczywista)
Nośnik	$x \in \mathbb{R}$
Gęstość prawdopodobieństwa	$\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$
Dystrybuanta	$\frac{1}{2}\left(1+\operatorname{erf}\frac{x-\mu}{\sigma\sqrt{2}}\right)$
Wartość oczekiwana (średnia)	μ
Mediana	μ
Moda	μ
Wariancja	σ^2
Współczynnik skośności	0
Kurtoza	0
Entropia	$\ln\left(\sigma\sqrt{2\,\pi\,e}\right)$
Funkcja tworząca momenty	$M_X(t) = \exp\left(\mu\,t + \frac{\sigma^2 t^2}{2}\right)$
Funkcja charakterystyczna	$\chi_X(t) = \exp\left(\mu\,i\,t - \frac{\sigma^2 t^2}{2}\right)$
Odkrywca	Abraham de Moivre (1733)[1]

Spis treści

Definicja rozkładu normalnego

Istnieje wiele równoważnych sposobów zdefiniowania rozkładu normalnego. Należą do nich: funkcja gęstości, dystrybuanta, momenty, kumulanty, funkcja charakterystyczna, funkcja tworząca momenty i funkcja tworząca kumulanty. Wszystkie kumulanty rozkładu normalnego wynoszą 0 oprócz pierwszych dwóch.

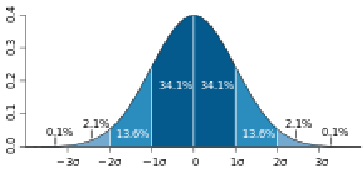
Funkcja gęstości

Funkcja gęstości prawdopodobieństwa rozkładu normalnego ze średnią μ i odchyleniem standardowym σ (równoważnie: wariancją σ^2) jest przykładem funkcji Gaussa. Dana jest ona wzorem:

$$f_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$$

Fakt, iż zmienna losowa X ma rozkład normalny z wartością oczekiwaną μ i wariancją σ^2 zapisuje się często $X \sim \mathcal{N}(\mu, \sigma^2)$.

Uwaga: W wielu źródłach rozkład normalny jest oznaczany przez $\mathcal{N}(\mu, \sigma)$.



Ilustracja reguły trzech sigm

Jeśli $\mu = 0$ i $\sigma = 1$, to rozkład ten nazywa się *standardowym rozkładem normalnym*, jego funkcja gęstości opisana jest wzorem:

$$\phi_{0,1}(x) = \phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$

We wszystkich rozkładach normalnych funkcja gęstości jest symetryczna względem wartości średniej rozkładu.
Około 68,3% pola pod wykresem krzywej znajduje się w odległości jednego odchylenia standardowego od średniej, około 95,5% w odległości dwóch odchyleń standardowych i około 99,7% w odległości trzech (reguła trzech sigm). Punkt przegięcia krzywej znajduje się w odległości jednego odchylenia standardowego od średniej.

Dystrybuanta

Dystrybuanta jest definiowana jako prawdopodobieństwo tego, że zmienna *X* ma wartości mniejsze bądź równe *x* i w kategoriach funkcji gęstości wyrażana jest (dla rozkładu normalnego) wzorem:

$$P(X \leq x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}\, dx.$$

Całki powyższej nie da się obliczyć dokładnie metodą analityczną. W konkretnych zagadnieniach do obliczenia wartości dystrybuanty stosuje się zatem tablice statystyczne (bądź też odpowiednie kalkulatory czy oprogramowanie komputerów). Tablice zawierają dane dla dystrybuanty *standardowego* rozkładu normalnego, tradycyjnie oznaczanej jako Φ i zdefiniowanej jako rozkład o parametrach μ = 0 i σ = 1:

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}}\, e^{-\frac{z^2}{2}}\, dz.$$

Związek dystrybuanty Φ i dystrybuanty rozkładu normalnego *X* o dowolnie zadanych parametrach μ i σ otrzymuje się za pomocą *standaryzowania rozkładu* (zob. też poniżej).

$$P(X \leq x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

Dystrybuanta standardowego rozkładu normalnego może być wyrażona poprzez funkcję specjalną (nieelementarną, przestępną), tzw. funkcję błędu jako:

$$\Phi(z) = \frac{1}{2}\left(1 + \operatorname{erf} \frac{z}{\sqrt{2}}\right)$$

Funkcje tworzące

Funkcja charakterystyczna

Funkcją charakterystyczną rozkładu normalnego jest

$$\varphi(t) = \exp\left(i\mu t - \frac{\sigma^2 t^2}{2}\right).$$

W przypadku standardowego rozkładu normalnego ma ona postać:

$$\varphi(t) = \exp\left(-\frac{t^2}{2}\right).$$

Własności

1. Jeśli $X \sim N(\mu, \sigma^2)$ oraz a, b są liczbami rzeczywistymi, to $aX + b \sim N(a\mu + b, (a\sigma)^2)$.
2. Jeśli $X_1 \sim N(\mu_1, \sigma_1^2)$ i $X_2 \sim N(\mu_2, \sigma_2^2)$ oraz zmienne X_1, X_2 są niezależne, to $X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.
3. Jeśli $X_1, \ldots, X_n$ są niezależnymi zmiennymi losowymi o standardowym rozkładzie normalnym, to zmienna $X_1^2 + \cdots + X_n^2$ ma rozkład chi-kwadrat z n stopniami swobody

Parametry rozkładu

- wartość oczekiwana: μ
- mediana: μ
- wariancja: σ^2
- odchylenie standardowe σ
- skośność: 0
- kurtoza: 0 (lub 3, przyjmując dawniej używaną definicję).

Dowód, że μ jest wartością oczekiwaną

Zgodnie z definicją, wartość oczekiwaną rozkładu normalnego można wyznaczyć obliczając wartość całki

$$E(X) = \int_{-\infty}^{+\infty} x \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} x e^{-\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)^2} dx$$

w celu uproszczenia można wprowadzić nową zmienną y

$$y = \frac{x-\mu}{\sigma\sqrt{2}} \Rightarrow x = y\sqrt{2}\sigma + \mu \text{ (a więc granice całkowania nie zmieniają się)}$$
$$dy = \frac{1}{\sqrt{2}\sigma} dx \Rightarrow dx = \sqrt{2}\sigma dy$$

przy czym granice całkowania pozostają bez zmian. Zatem

$$E(X) = \frac{1}{\sqrt{2}\sqrt{\pi}\sigma} \int_{-\infty}^{+\infty} (y\sqrt{2}\sigma + \mu) e^{-y^2} \sqrt{2}\sigma dy = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{+\infty} (y\sqrt{2}\sigma + \mu) e^{-y^2} dy =$$
$$= \frac{1}{\sqrt{\pi}} \left(\sqrt{2}\sigma \int_{-\infty}^{+\infty} y e^{-y^2} dy + \mu \int_{-\infty}^{+\infty} e^{-y^2} dy \right) = \frac{\sqrt{2}\sigma}{\sqrt{\pi}} \underbrace{\left(\int_{-\infty}^{+\infty} y e^{-y^2} dy \right)^{(1)}}_{=0} + \frac{\mu}{\sqrt{\pi}} \underbrace{\left(\int_{-\infty}^{+\infty} e^{-y^2} dy \right)^{(2)}}_{=\sqrt{\pi}} =$$
$$= \frac{\sqrt{2}\sigma}{\sqrt{\pi}} \cdot 0 + \frac{\mu}{\sqrt{\pi}} \cdot \sqrt{\pi} = \mu$$

cbdo. Przy wyprowadzeniu skorzystano z:

$$(1) = \int_{-\infty}^{+\infty} y e^{-y^2} dy = \frac{1}{2} \int_{+\infty}^{+\infty} e^{-z} dz = 0$$
$$z = y^2 \text{ (granice całkowania się zmieniają)}$$
$$dz = 2y dy \Rightarrow dy = \frac{dz}{2y}$$

$$(2) = \int_{-\infty}^{+\infty} e^{-y^2} dy \text{ – całka Poissona}$$

Standaryzowanie zmiennych losowych o rozkładzie normalnym

Konsekwencją własności 1 jest możliwość przekształcenia wszystkich zmiennych losowych o rozkładzie normalnym do standardowego rozkładu normalnego.

Jeśli *X* ma rozkład normalny ze średnią μ i wariancją σ², wtedy:

$$Z = \frac{X - \mu}{\sigma}$$

Z jest zmienną losową o standardowym rozkładzie normalnym N(0, 1). Ważną konsekwencją jest postać dystrybuanty:

$$P(X \leq x) = \Phi\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{2}\left(1 + \operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right)$$

Odwrotnie, jeśli *Z* jest zmienną losową o standardowym rozkładzie normalnym, to:

$$X = \sigma Z + \mu$$

jest zmienną o rozkładzie normalnym ze średnią μ i wariancją σ².

Standardowy rozkład normalny został stablicowany i inne rozkłady normalne są prostymi transformacjami rozkładu standardowego. W ten sposób możemy używać tablic dystrybuanty standardowego rozkładu normalnego do wyznaczenia wartości dystrybuanty rozkładu normalnego o dowolnych parametrach.

Generowanie wartości losowych o rozkładzie normalnym

W symulacjach komputerowych zdarza się, że potrzebujemy wygenerować wartości zmiennej losowej o rozkładzie normalnym. Istnieje kilka metod, najprostszą z nich jest odwrócenie dystrybuanty standardowego rozkładu normalnego. Są jednak metody bardziej wydajne, jedną z nich jest transformacja Boxa-Mullera, w której dwie zmienne losowe o rozkładzie jednostajnym (prostym do wygenerowania – patrz generator liczb losowych) są transformowane na zmienne o rozkładzie normalnym.

Transformacja Boxa-Mullera jest konsekwencją własności 3 i faktu, że rozkład chi-kwadrat z dwoma stopniami swobody jest rozkładem wykładniczym (łatwym do wygenerowania).

Centralne twierdzenie graniczne

Jedną z najważniejszych własności rozkładu normalnego jest fakt, że (przy pewnych założeniach) rozkład sumy dużej liczby zmiennych losowych jest w przybliżeniu normalny. Jest to tak zwane centralne twierdzenie graniczne.

W praktyce twierdzenie to ma zastosowanie, jeśli chcemy użyć rozkładu normalnego jako przybliżenia dla innych rozkładów

- Rozkład dwumianowy z parametrami (n, p) jest w przybliżeniu normalny dla dużych n i p nie leżących zbyt blisko 1 lub 0. Przybliżony rozkład ma średnią równą $\mu = np$ i odchylenie standardowe $\sigma = \sqrt{np(1-p)}$.
- Rozkład Poissona z parametrem λ jest w przybliżeniu normalny dla dużych wartości λ. Przybliżony rozkład normalny ma średnią $\mu = \lambda$ i odchylenie standardowe $\sigma = \sqrt{\lambda}$.

Dokładność przybliżenia tych rozkładów zależy od celu użycia przybliżenia i tempa zbieżności do rozkładu normalnego. Zazwyczaj takie przybliżenia są mniej dokładne w ogonach rozkładów

Nieskończona podzielność

Rozkład normalny należy do rozkładów mających własność nieskończonej podzielności

Występowanie

Rozkład normalny (lub wielowymiarowy rozkład normalny) jest często stosowanym założeniem, w praktyce jednak nigdy nie jest ściśle realizowany. Rozkład normalny ma bowiem niezerową gęstość prawdopodobieństwa dla dowolnej wartości zmiennej losowej, podczas gdy w rzeczywistości zmienne są zawsze ograniczone, a często nieujemne.

Mimo to rzeczywisty rozkład jest często bardzo zbliżony do normalnego, stąd zwykle zakłada się, że zmienna ma rozkład normalny. Nie należy jednak robić tego bez sprawdzenia jak wielkie są rozbieżności. Rozkłady dalekie od normalnego (np. z elementami odstającymi) mogą sprawić, że wyniki metod statystycznych będą mylnie interpretowane.

Przykładem są tu metody regresji liniowej oraz korelacji Pearsona, które, choć zdefiniowane dla dowolnych rozkładów, mają sensowną interpretację tylko dla wielowymiarowego rozkładu normalnego wektora próbki. Jeśli w próbce występują elementy odstające, co jest szczególnym przypadkiem rozkładu dalekiego od normalnego, korelacja może przyjąć dowolną wartość między −1 a +1, bez względu na rzeczywistą zależność między zmiennymi losowymi. Także regresja będzie dawała błędne rezultaty

Inteligencja

Inteligencja mierzona testami inteligencji uważana jest za zmienną o rozkładzie normalnym. W praktyce testy dają wyniki skwantowane, a nie ciągłe. W dodatku ich wyniki są ograniczone do pewnego przedziału. Przybliżenie jest jednak wystarczające.

Kontrowersyjny pogląd o nieadekwatności tego modelu został zaprezentowany w książce The Bell Curve.

Wzrost

Podobnie wzrost człowieka może być uznany w przybliżeniu za zmienną o rozkładzie normalnym. Musimy wtedy oczywiście założyć, że wartość oczekiwana rozkładu wynosi na przykład 170 cm, a odchylenie standardowe jest wystarczająco małe, aby przypadek ludzi o ujemnym wzroście miał znikomo małe prawdopodobieństwo.

Natężenie źródła światła

Natężenie światła z pojedynczego źródła zmienia się w czasie i zazwyczaj zakłada się, że ma rozkład normalny. Zgodnie z mechaniką kwantową światło jest strumieniem fotonów. Zwykłe źródło światła, świecące dzięki termicznej emisji, powinno świecić w krótkich przedziałach czasu zgodnie z rozkładem Poissona. W dłuższym przedziale czasowym (dłuższym niż czas koherencji) dodawanie się do siebie niezależnych zmiennych prowadzi w przybliżeniu do rozkładu normalnego.

Błędy pomiaru

Wielokrotne powtarzanie tego samego pomiaru daje wyniki rozrzucone wokół określonej wartości. Jeśli wyeliminujemy wszystkie większe przyczyny błędów, zakłada się, że pozostałe mniejsze błędy muszą być rezultatem dodawania się do siebie dużej liczby niezależnych czynników, co daje w efekcie rozkład normalny. Odchylenia od rozkładu normalnego rozumiane są jako wskazówka, że zostały pominięte błędy systematyczne. To stwierdzenie jest centralnym założeniem teorii błędów.

Uwagi

a. Ściślej: można zaobserwować rozkłady bardzo zbliżone do rozkładu normalnego. Rozkład normalny zakłada niezerowe prawdopodobieństwo dla każdej możliwej liczby rzeczywistej. W rzeczywistości wszelkie zmienne są ograniczone, na przykład nie ma ludzi o ujemnym wzroście ani o wzroście kilometra, jednak rozkłady spotykane w praktyce są tak bardzo zbliżone do rozkładu normalnego, że różnica ta nie ma znaczenia.
b. Te właściwości to na przykład: Suma i różnica dwóch zmiennych o rozkładach normalnych ma rozkład normalny. Logarytm z gęstości rozkładu normalnego to funkcja kwadratowa, dzięki czemu metoda najmniejszych kwadratów stosowana w regresji liniowej dla rozkładu normalnego błędów jest metodą największej wiarygodności.

Przypisy

1. Abraham de Moivre, „Approximatio ad Summam Terminorum Binomii $(a + b)^n$ in Seriem expansi” (wydrukowany 12 listopada 1733 w Londynie).

Bibliografia

- J. Wawrzynek: *Metody opisu i wnioskowania statystycznego*. Wrocław: Wydawnictwo Akademii Ekonomicznej im. Oskara Langego we Wrocławiu, 2007, s. 62. ISBN 978-83-7011-859-4.