Wykład z analizy danych: powtórzenie zagadnień z rachunku prawdopodobieństwa

Marek Kubiak

Instytut Informatyki
Politechnika Poznańska

Plan wykładu

Cel stosowania rachunku prawdopodobieństwa w statystyce

Podstawowe pojęcia rachunku prawdopodobieństwa

Rozkład Bernoulliego

Rozkład normalny

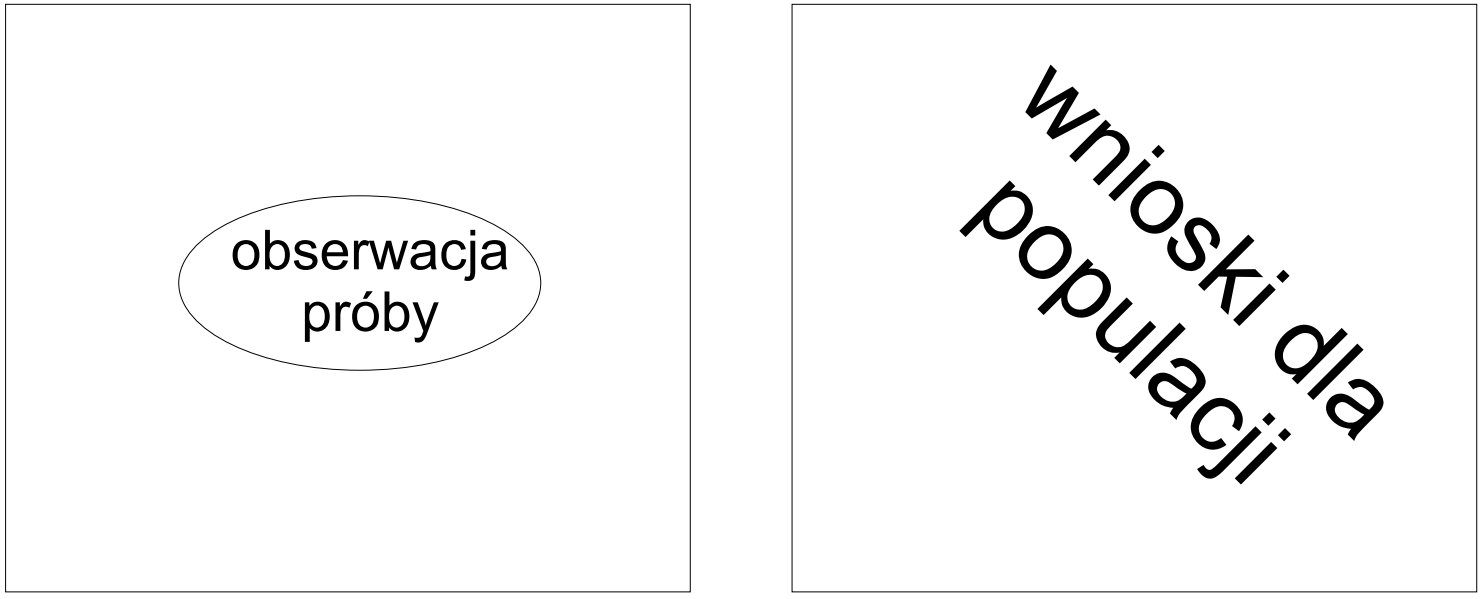
Centralne twierdzenie integralne Lindeberga–Levy'ego

Cel stosowania rachunku prawdopodobieństwa w statystyce

. . . czyli o co tutaj chodzi?

Cel stosowania rachunku prawdopodobieństwa w statystyce

Cel: wyciąganie sensownych i uzasadnionych wniosków o pewnej dużej zbiorowości (*populacji*) na podstawie obserwacji małego zbioru danych (*próby*).



Cel stosowania rachunku prawdopodobieństwa w statystyce

Przykłady:

- badanie przedwyborcze w grupie 1000 osób
  $\Rightarrow$ wnioski dotyczą milionów wyborców
- badanie jakości 200 sztuk amunicji
  $\Rightarrow$ wnioski dotyczą kilkuset tysięcy sztuk z dostawy dla armii
- badanie cech 1000 rozwiązań pewnego problemu optymalizacji
  $\Rightarrow$ wnioski dotyczą całej przestrzeni ($\approx 10^{100}$)

Cel stosowania rachunku prawdopodobieństwa w statystyce

Przykłady:

- badanie przedwyborcze w grupie 1000 osób
  $\Rightarrow$ wnioski dotyczą milionów wyborców
- badanie jakości 200 sztuk amunicji
  $\Rightarrow$ wnioski dotyczą kilkuset tysięcy sztuk z dostawy dla armii
- badanie cech 1000 rozwiązań pewnego problemu optymalizacji
  $\Rightarrow$ wnioski dotyczą całej przestrzeni ($\approx 10^{100}$)

To jest *wnioskowanie indukcyjne*!

Rachunek prawdopodobieństwa jest jedną z niewielu teorii pozwalających na *sensowne i uzasadnione wnioskowanie indukcyjne*.

Cel stosowania rachunku prawdopodobieństwa w statystyce

Statystyka *opisowa* vs. statystyka *matematyczna*.

Statystyka *opisowa*:
nie korzysta z rachunku prawdopodobieństwa
$\Rightarrow$ wszystkie wnioski dotyczą wyłącznie badanego zbioru danych.

Statystyka *matematyczna*:
korzysta z rachunku prawdopodobieństwa
$\Rightarrow$ przy odpowiednich założeniach wnioski dotyczą całej populacji.

Pojęcia podstawowe

Podstawowe obiekty probabilistyczne

- Przestrzeń probabilistyczna: $(\Omega, \mathcal{A}, P)$.
- Zmienna losowa: $X : \Omega \to \mathbf{R}$;
  intuicyjnie: wartość liczbowa zależna od przypadku.
- Rozkład prawdopodobieństwa zmiennej losowej:
  - funkcja prawdopodobieństwa $p_i = P(X = x_i)$ dla zmiennej skokowej (dyskretnej);
  - funkcja gęstości: $f(x)$ dla zmiennej ciągłej.
- Dystrybuanta zmiennej losowej: $F(x) = P(X < x)$,
  czyli kumulacja prawdopodobieństwa od strony $-\infty$:
  - $F(x_0) = \sum_{x_i < x_0} P(X = x_i)$ dla zmiennej dyskretnej;
  - $F(x_0) = \int_{-\infty}^{x_0} f(x)dx$ dla zmiennej ciągłej.

Pojęcia podstawowe

Podstawowe charakterystyki liczbowe zmiennej losowej

- Wartość oczekiwana (wartość średnia) zmiennej losowej dyskretnej X:
  $E(X) = \sum x \cdot P(X = x)$,
  miara położenia „środka" zmiennej.
- Wariancja zmiennej losowej dyskretnej X:
  $D^2(X) = \sum (x - E(X))^2 \cdot P(X = x)$,
  miara rozrzutu wokół „środka" zmiennej;
  sens kwadratu z wartości zmiennej X.
- Odchylenie standardowe (dewiacja) zmiennej losowej dyskretnej X:
  $D(X) = \sqrt{D^2(X)}$,
  miara rozrzutu wokół „środka" zmiennej;
  sens wartości zmiennej X.

Pojęcia podstawowe

Obliczenie charakterystyk prostej zmiennej dyskretnej

x	0	1	2	3	4	$\sum$
$P(X = x)$	0,2	0,2	0,2	0,2	0,2	1

Pojęcia podstawowe

Obliczenie charakterystyk prostej zmiennej dyskretnej

x	0	1	2	3	4	$\sum$
$P(X = x)$	0,2	0,2	0,2	0,2	0,2	1
$x \cdot P(X = x)$	0	0,2	0,4	0,6	0,8	$2 = E(X)$

Pojęcia podstawowe

Obliczenie charakterystyk prostej zmiennej dyskretnej

x	0	1	2	3	4	$\sum$
$P(X = x)$	0,2	0,2	0,2	0,2	0,2	1
$x \cdot P(X = x)$	0	0,2	0,4	0,6	0,8	$2 = E(X)$
$(x - E(X))^2 \cdot P(X = x)$	0,8	0,2	0	0,2	0,8	$2 = D^2(X)$

Pojęcia podstawowe

Niektóre własności wartości oczekiwanej i wariancji

- Wartość oczekiwana sumy zmiennych losowych:

$$E(X + Y) = E(X) + E(Y)$$

- Wartość oczekiwana przeskalowanej zmiennej:

$$E(k \cdot X) = k \cdot E(X)$$

- Wariancja sumy *niezależnych* zmiennych losowych:

$$D^2(X + Y) = D^2(X) + D^2(Y)$$

- Wariancja przeskalowanej zmiennej:

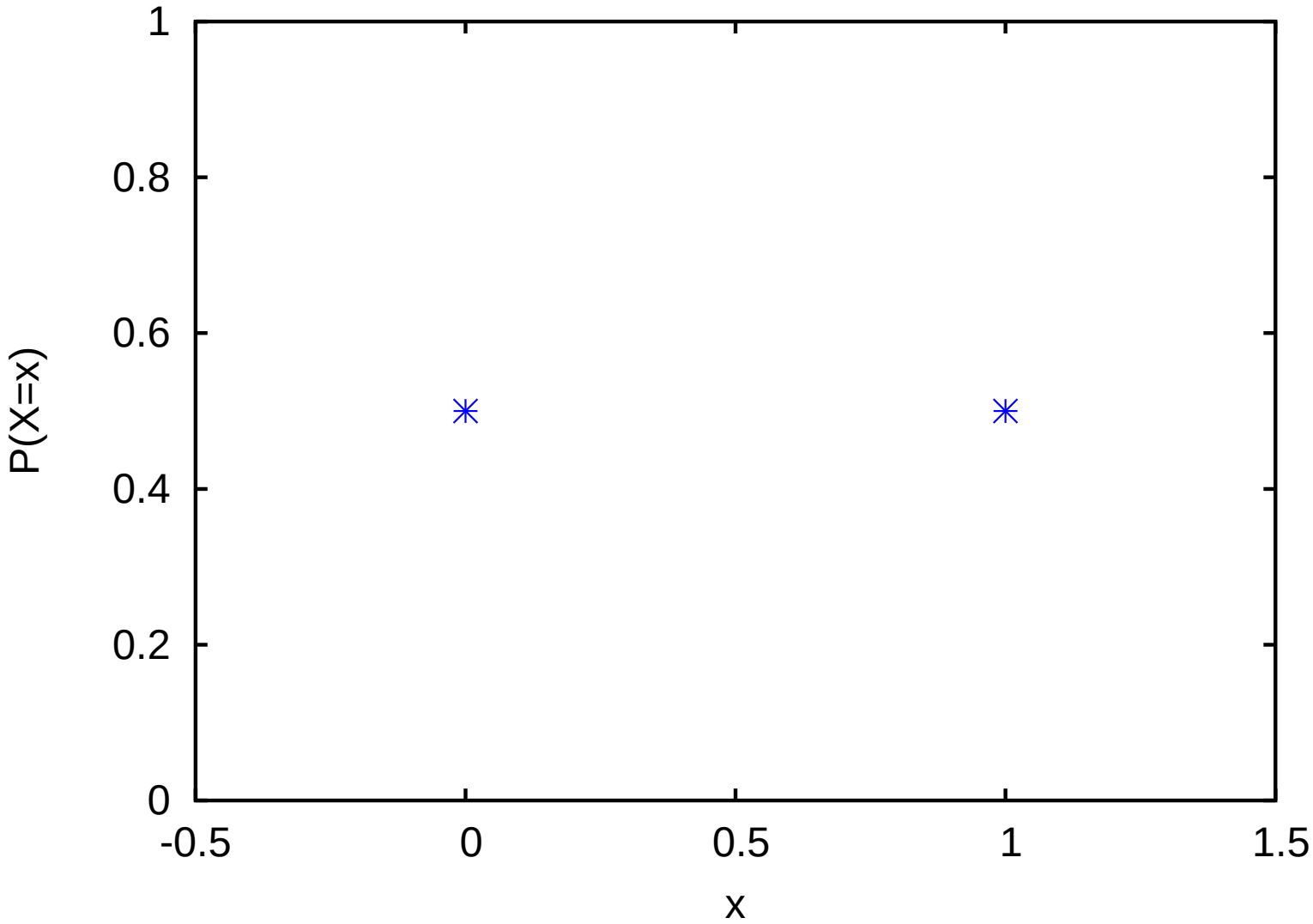
$$D^2(k \cdot X) = k^2 \cdot D^2(X)$$

Zmienna o rozkładzie dwupunktowym (zero–jedynkowym)

- oznaczenie: $X \sim B_1(p), \quad p \in (0,1)$.
- funkcja prawdopodobieństwa:
  $P(X = 1) = p,\ P(X = 0) = 1 - p$.
- wartość oczekiwana: $E(X) = p$.
- wariancja: $D^2(X) = p(1 - p)$.
- praktyczne występowanie: eksperymenty z dwoma możliwymi wynikami lub podział zbioru zdarzeń elementarnych na dwa rozłączne podzbiory.

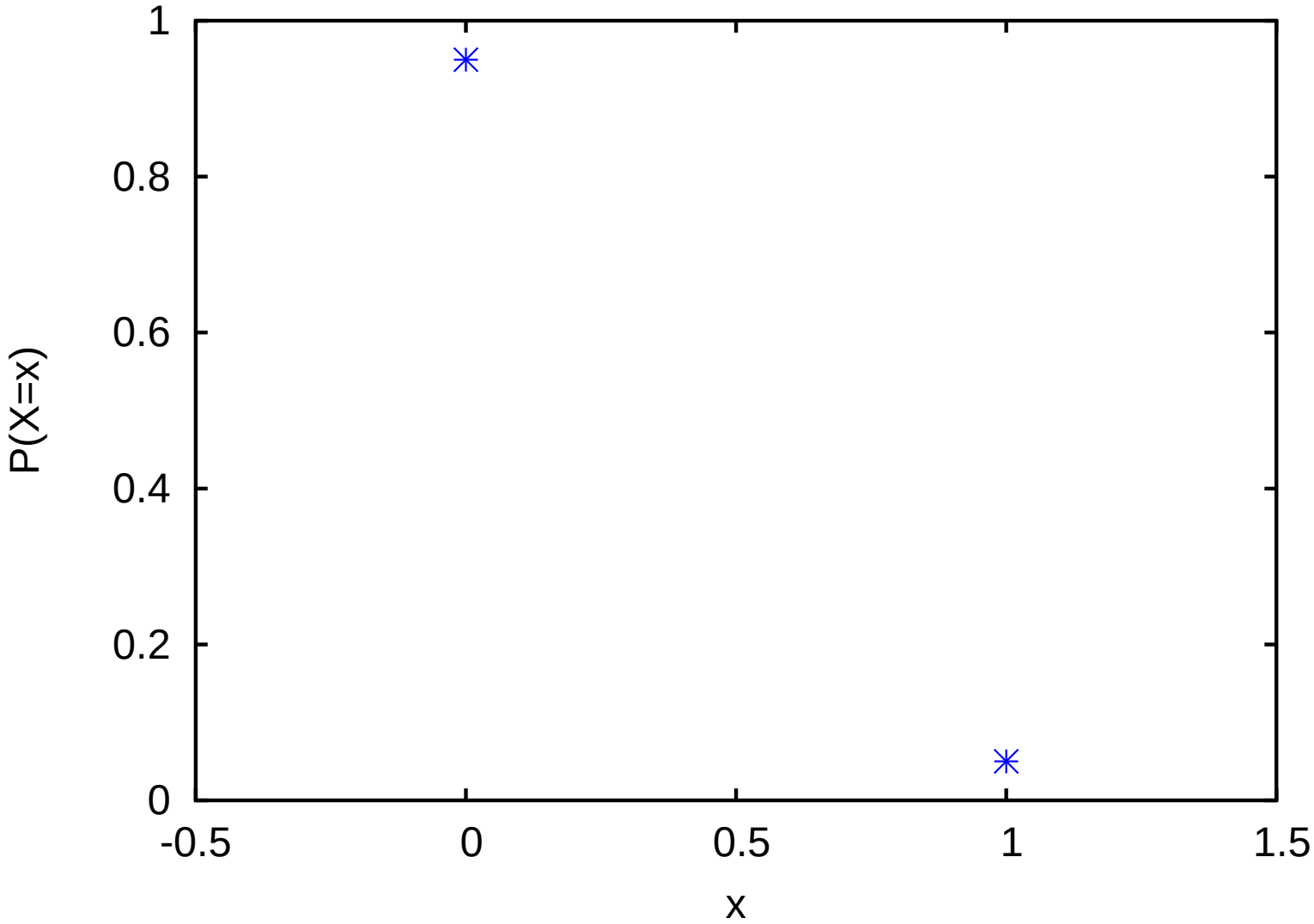
Zmienna o rozkładzie dwupunktowym (zero–jedynkowym)

Funkcja prawdopodobieństwa dla $p = 0,5$.



Zmienna o rozkładzie dwupunktowym (zero–jedynkowym)

Funkcja prawdopodobieństwa dla $p = 0,05$.



Zmienna o rozkładzie dwumianowym (Bernoulliego)

- oznaczenie: $X \sim B_n(p), \quad p \in (0,1), \ n \in \mathbf{N}$.
- funkcja prawdopodobieństwa:
  $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, \ldots, n.$
- wartość oczekiwana: $E(X) = np$.
- wariancja: $D^2(X) = np(1-p)$.
- praktyczne występowanie: ?

Zmienna o rozkładzie dwumianowym (Bernoulliego)

Twierdzenie Bernoulliego

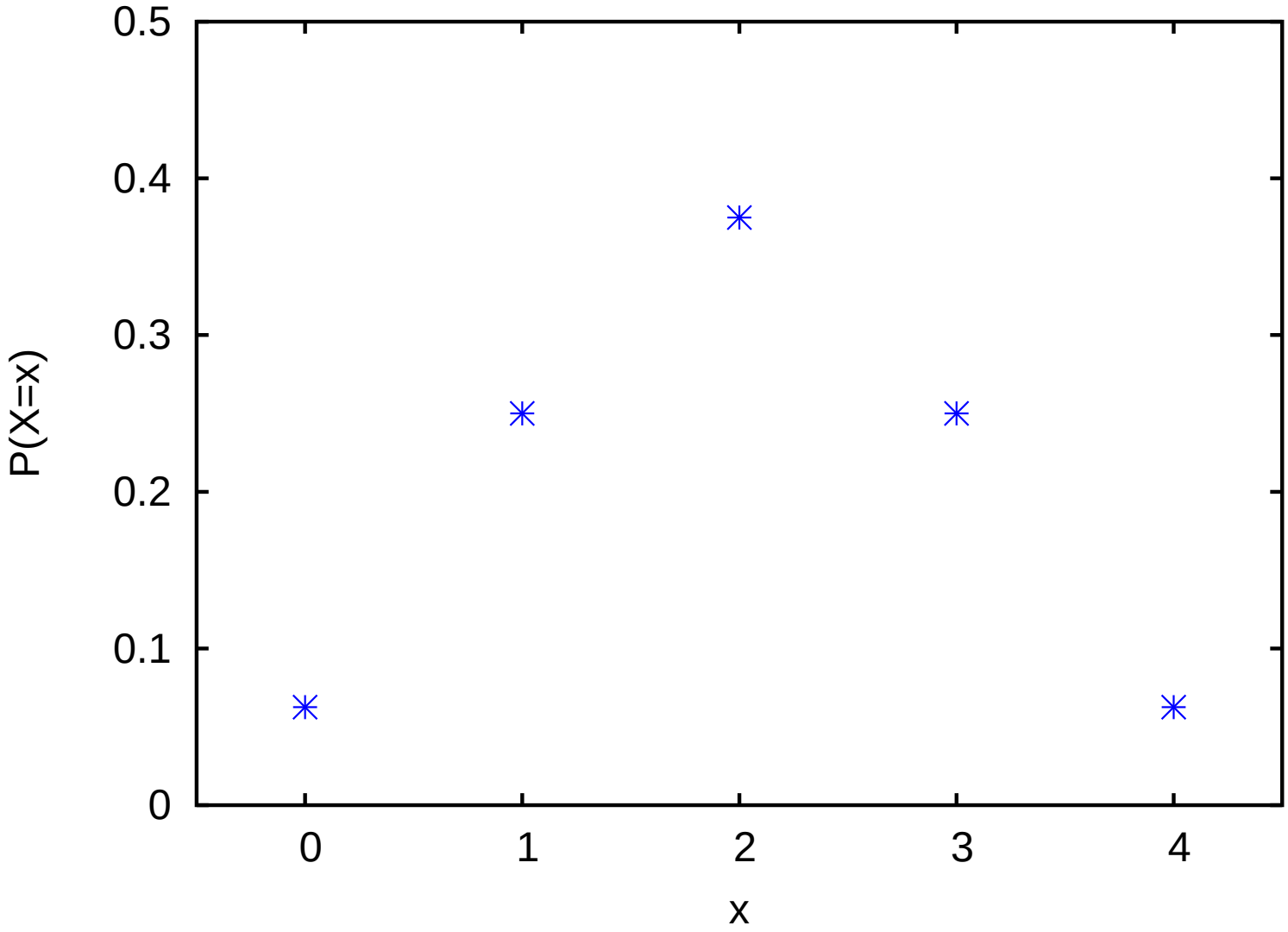
Niech $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ będzie wektorem n niezależnych zmiennych losowych o takim samym rozkładzie zero–jedynkowym $B_1(p)$, gdzie $p \in (0, 1)$.

Wtedy zmienna losowa $S_n = X_1 + X_2 + \ldots + X_n = \sum_{i=1}^{n} X_i$ ma rozkład dwumianowy $B_n(p)$.

To twierdzenie jest modelem praktycznego występowania rozkładu dwumianowego!
W statystyce będziemy korzystać z tego twierdzenia, gdy będziemy szacować nieznane prawdopodobieństwo p pewnego zdarzenia na podstawie wielu powtórzeń prostego doświadczenia i obserwacji, czy badane zdarzenie zaszło.

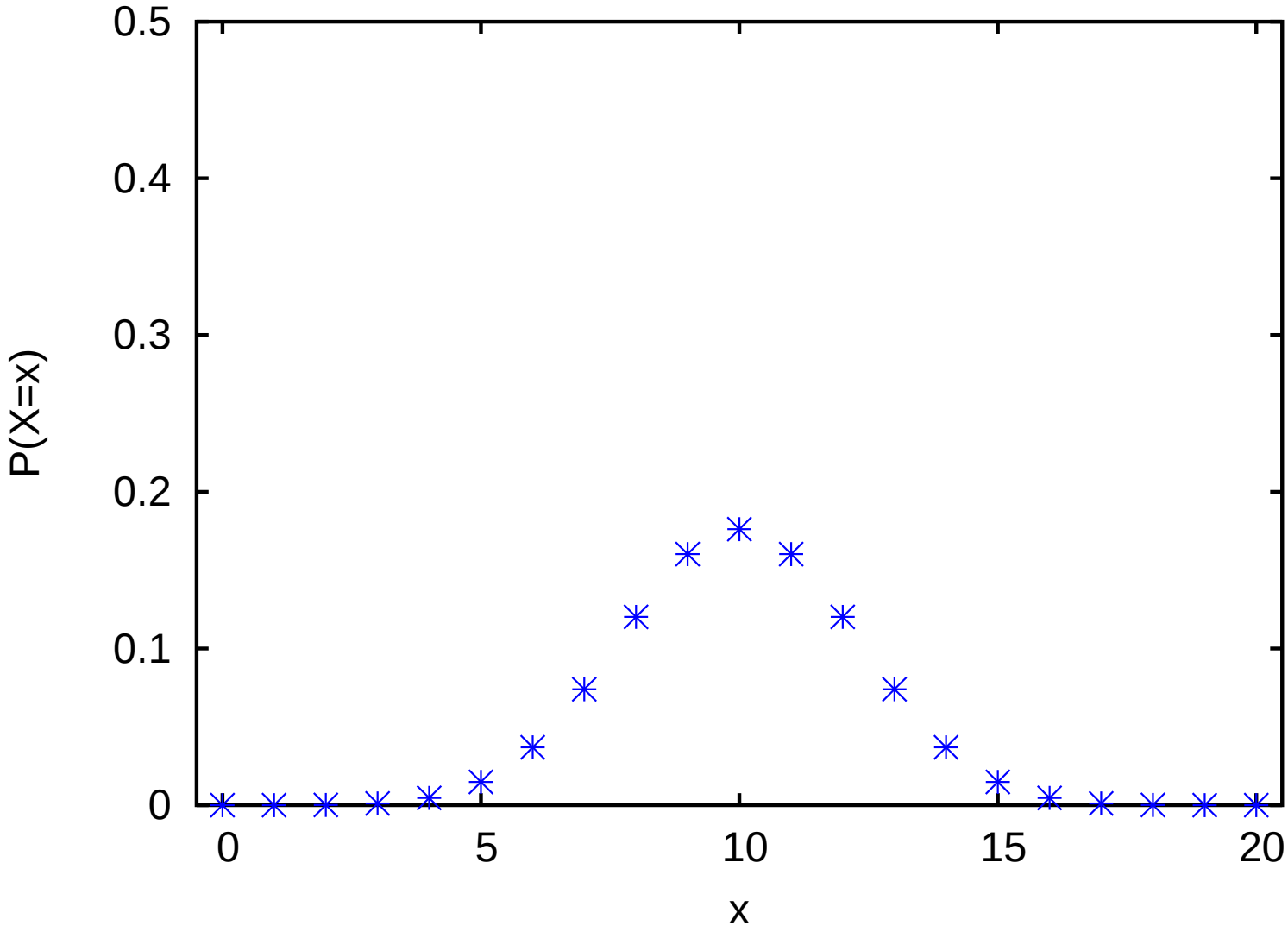
Zmienna o rozkładzie dwumianowym (Bernoulliego)

Funkcja prawdopodobieństwa dla $n = 4$ i $p = 0,5$.



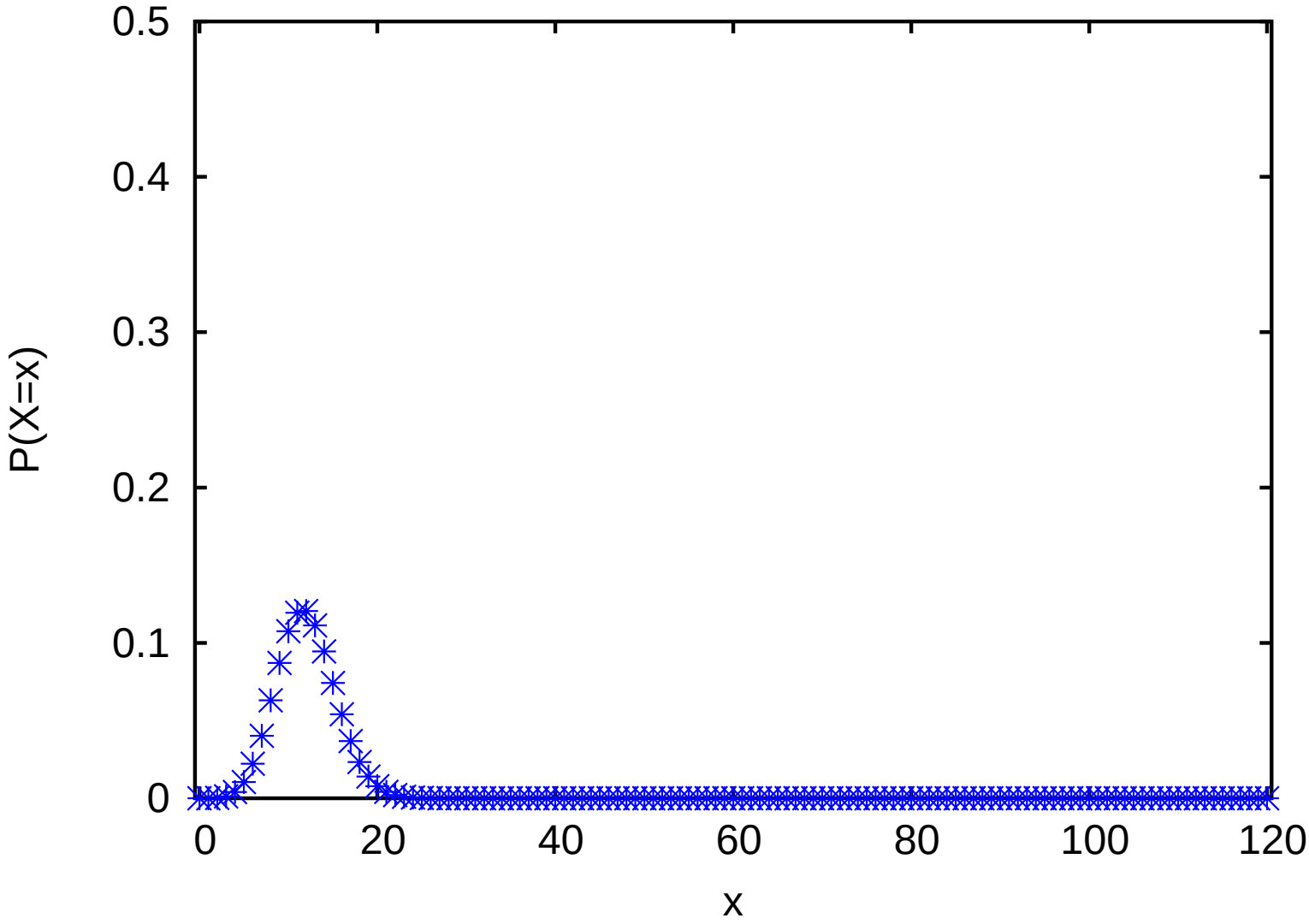
Zmienna o rozkładzie dwumianowym (Bernoulliego)

Funkcja prawdopodobieństwa dla $n = 20$ i $p = 0,5$.



Zmienna o rozkładzie dwumianowym (Bernoulliego)

Funkcja prawdopodobieństwa dla $n = 120$ i $p = 0,1$.



Zmienna o rozkładzie dwumianowym (Bernoulliego)

Przykłady obliczenia prawdopodobieństwa:

- $X \sim B_4(p = 0,5)$; obliczyć $P(X = 4)$.
- $X \sim B_{20}(p = 0,5)$;
  obliczyć $P(X \geq 18) = P(X = 18) + P(X = 19) + P(X = 20)$.
- $X \sim B_{120}(p = 0,1)$;
  obliczyć $P(X \geq 90) = P(X = 90) + \ldots + P(X = 120)$.

Przy obliczaniu prawdopodobieństw ostatniego typu występują praktyczne trudności:

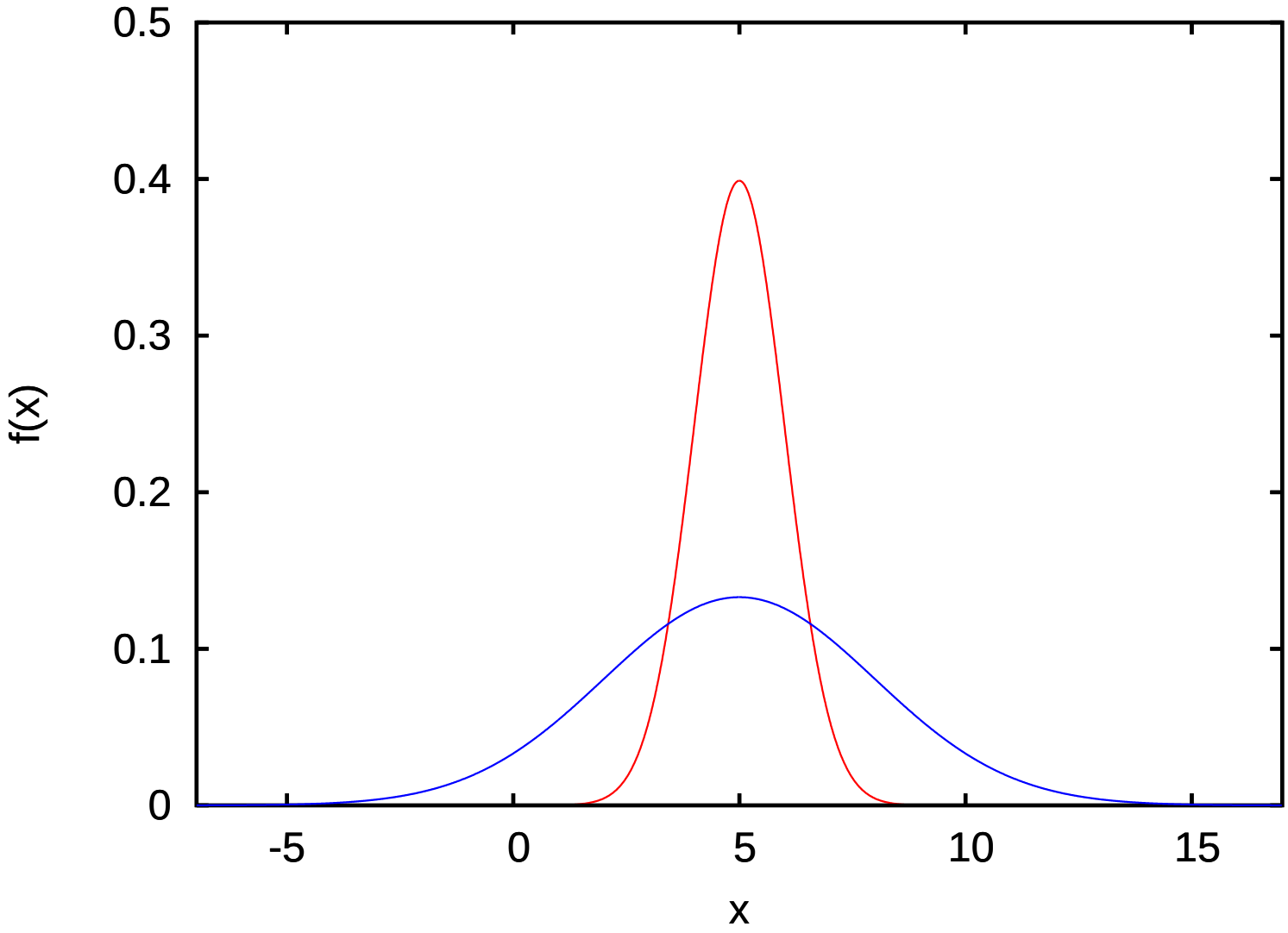
- sumowanie po wielu wartościach zmiennej (nieduży problem przy obliczeniach automatycznych),
- silnie o dużym argumencie, potęgi o wysokim wykładniku i małej podstawie (problemy numeryczne)

Zmienna o rozkładzie normalnym

- oznaczenie: $X \sim N(\mu, \sigma), \quad \mu \in \mathbf{R},\ \sigma \in \mathbf{R}_+$.
- funkcja gęstości: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\{-\frac{(x-\mu)^2}{2\sigma^2}\}, \quad x \in \mathbf{R}$.
- wartość oczekiwana: $E(X) = \mu$.
- wariancja: $D^2(X) = \sigma^2$.
- praktyczne występowanie: ?

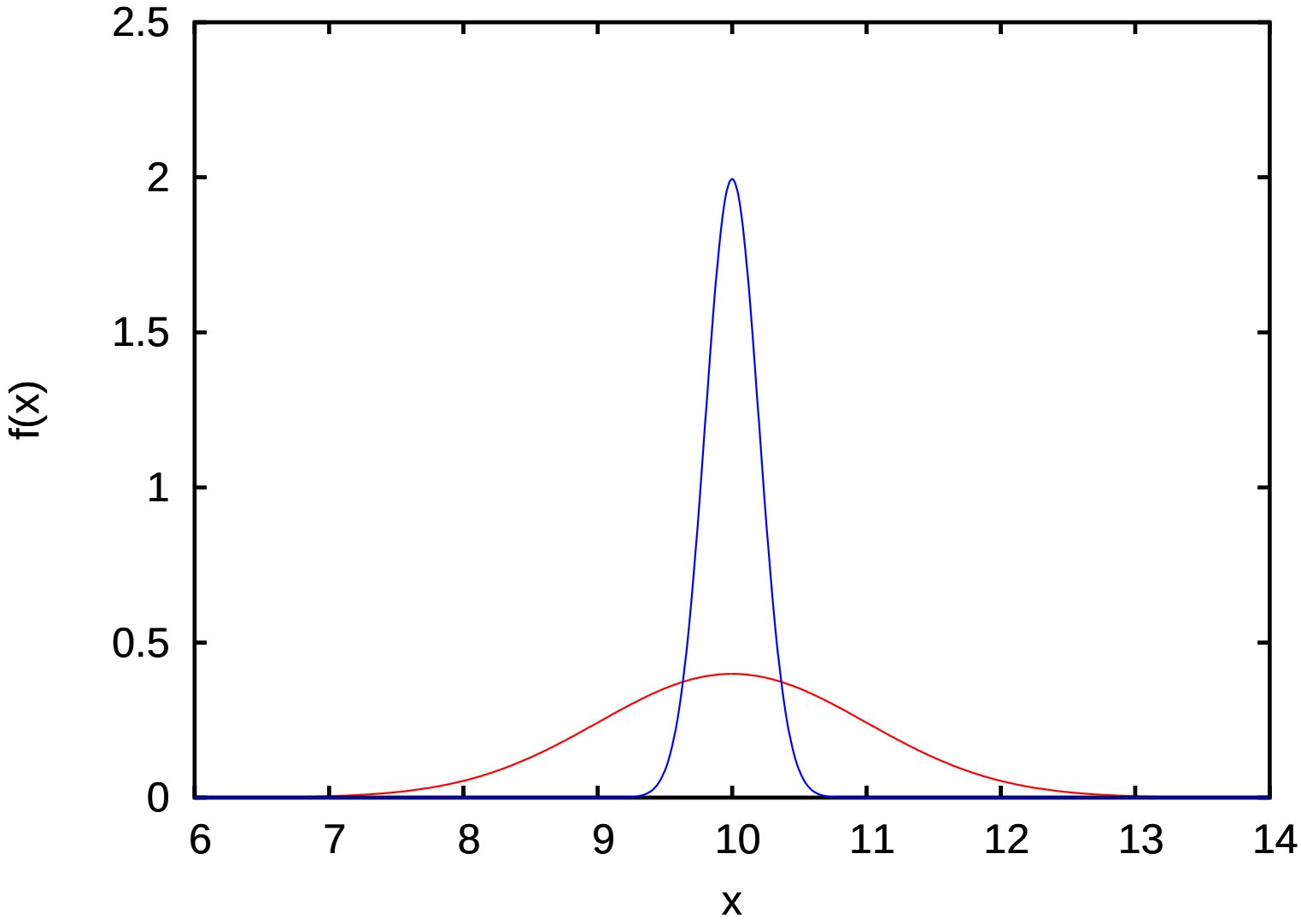
Zmienna o rozkładzie normalnym

Funkcje gęstości rozkładów $N(5,1)$ i $N(5,3)$.



Zmienna o rozkładzie normalnym

Funkcje gęstości rozkładów $N(10, 1)$ i $N(10; 0, 2)$.

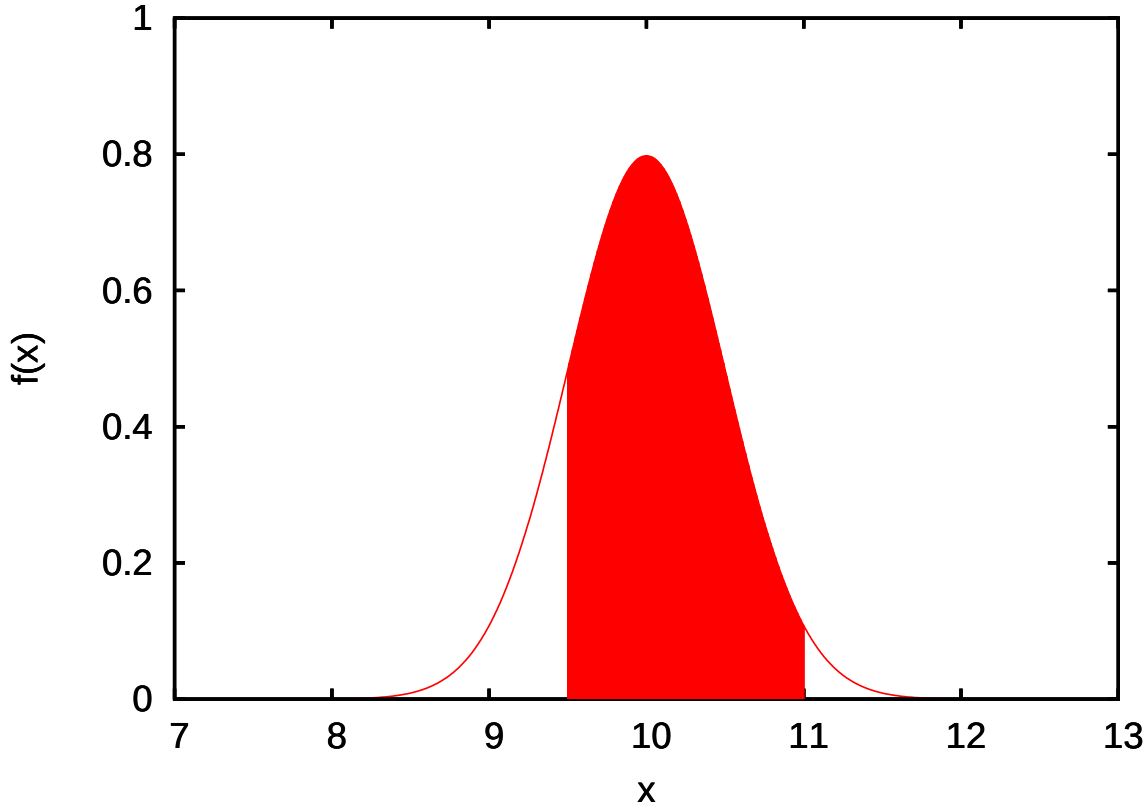


Zmienna o rozkładzie normalnym

Przykład obliczenia prawdopodobieństwa

Niech $X \sim N(10; 0,5)$; obliczyć $P(9,5 \leq X \leq 11)$.

Nie da się tego łatwo obliczyć analitycznie!



Standaryzacja zmiennej losowej

Standaryzacja zmiennej losowej X o niezerowej wariancji:

$$Y = \frac{X - E(X)}{D(X)}$$

Własności zmiennej losowej ustandaryzowanej Y:

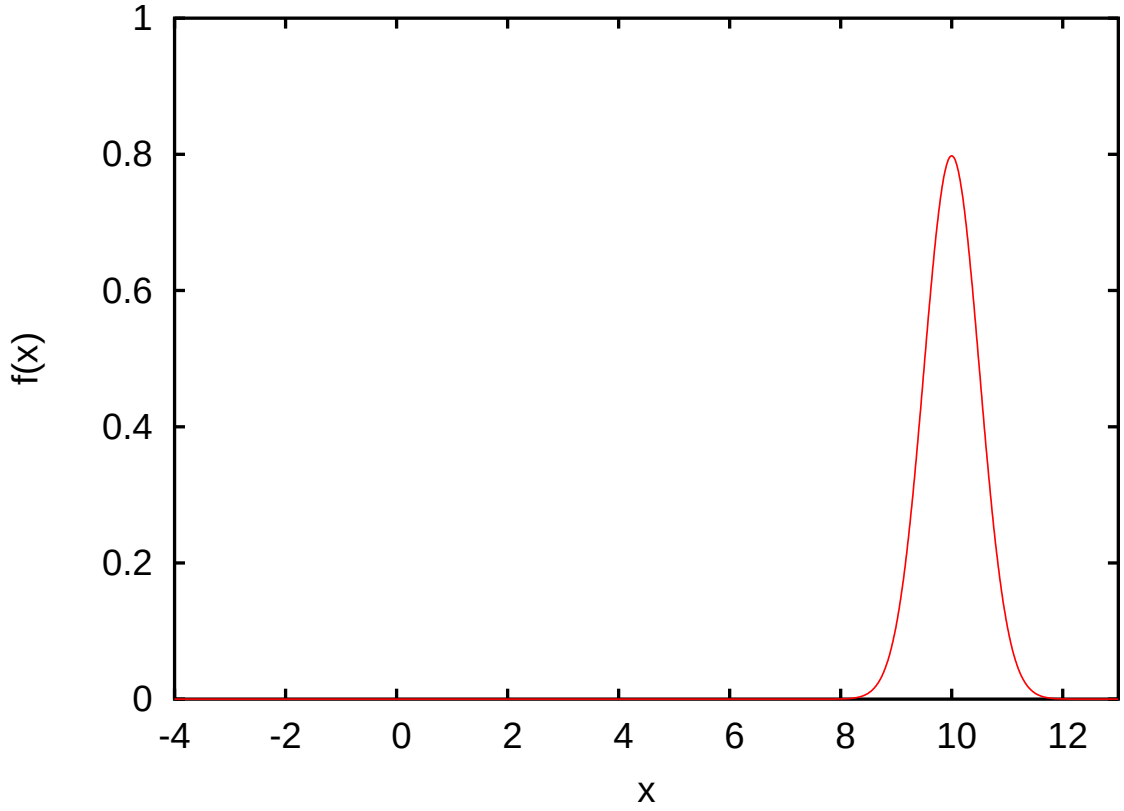
- $E(Y) = 0$,
- $D^2(Y) = 1$.

Standaryzacja: prosty zabieg techniczny, który pozwala traktować w ten sam sposób zmienne losowe o tym samym kształcie rozkładu, ale innych wartościach oczekiwanych i wariancjach.

Standaryzowany rozkład normalny $N(0,1)$

Przykład standaryzacji rozkładu normalnego

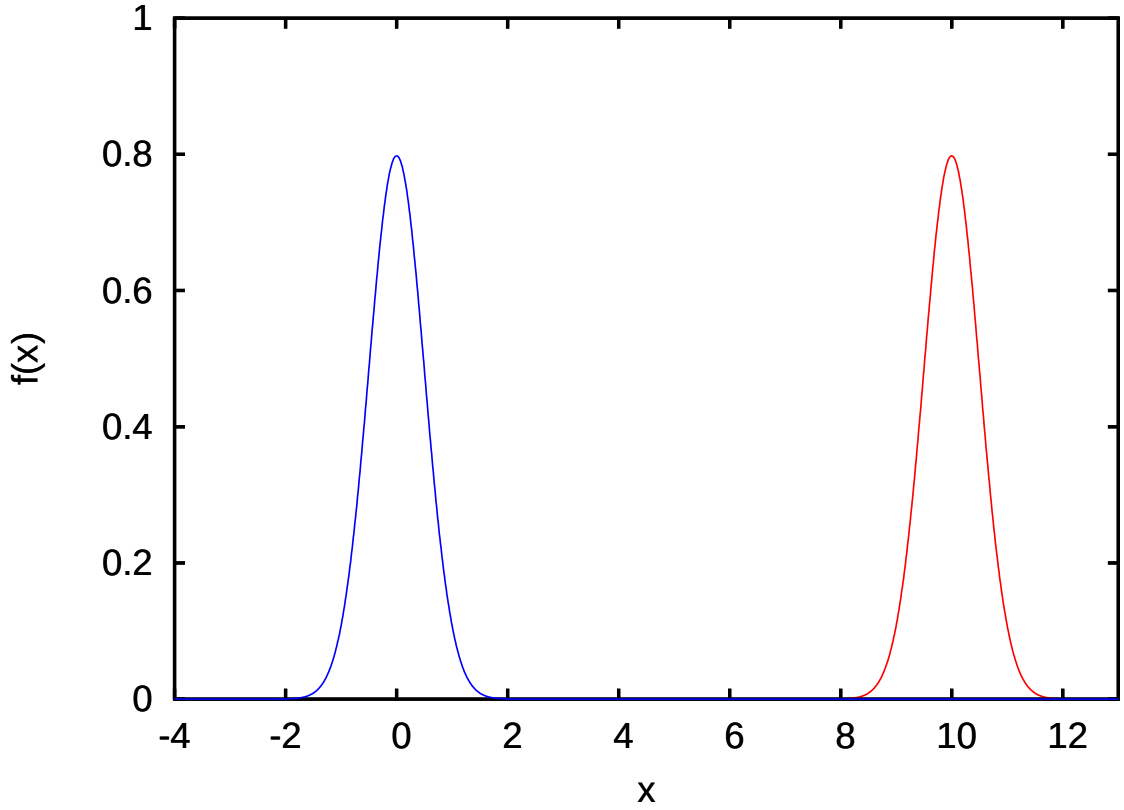
$X \sim N(10; 0,5)$



Standaryzowany rozkład normalny $N(0,1)$

Przykład standaryzacji rozkładu normalnego

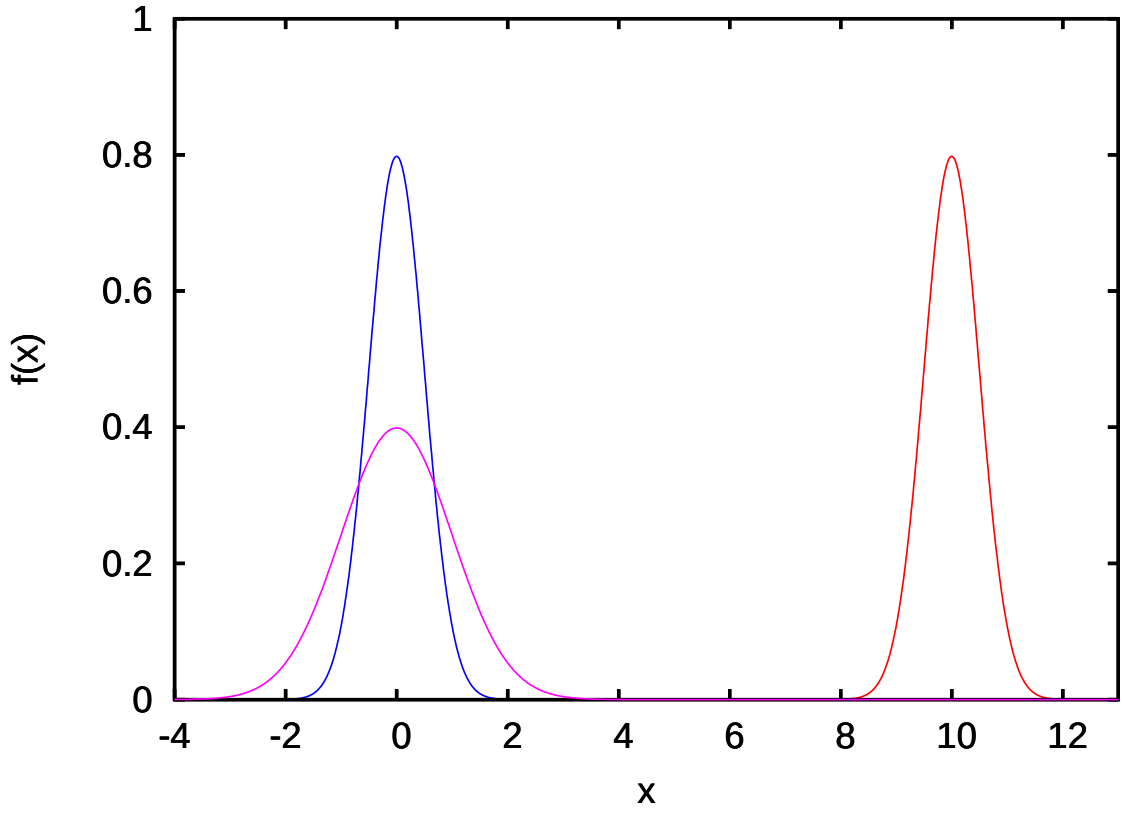
$$X' = X - E(X) = X - 10 \Rightarrow X' \sim N(0; 0,5)$$



Standaryzowany rozkład normalny $N(0,1)$

Przykład standaryzacji rozkładu normalnego

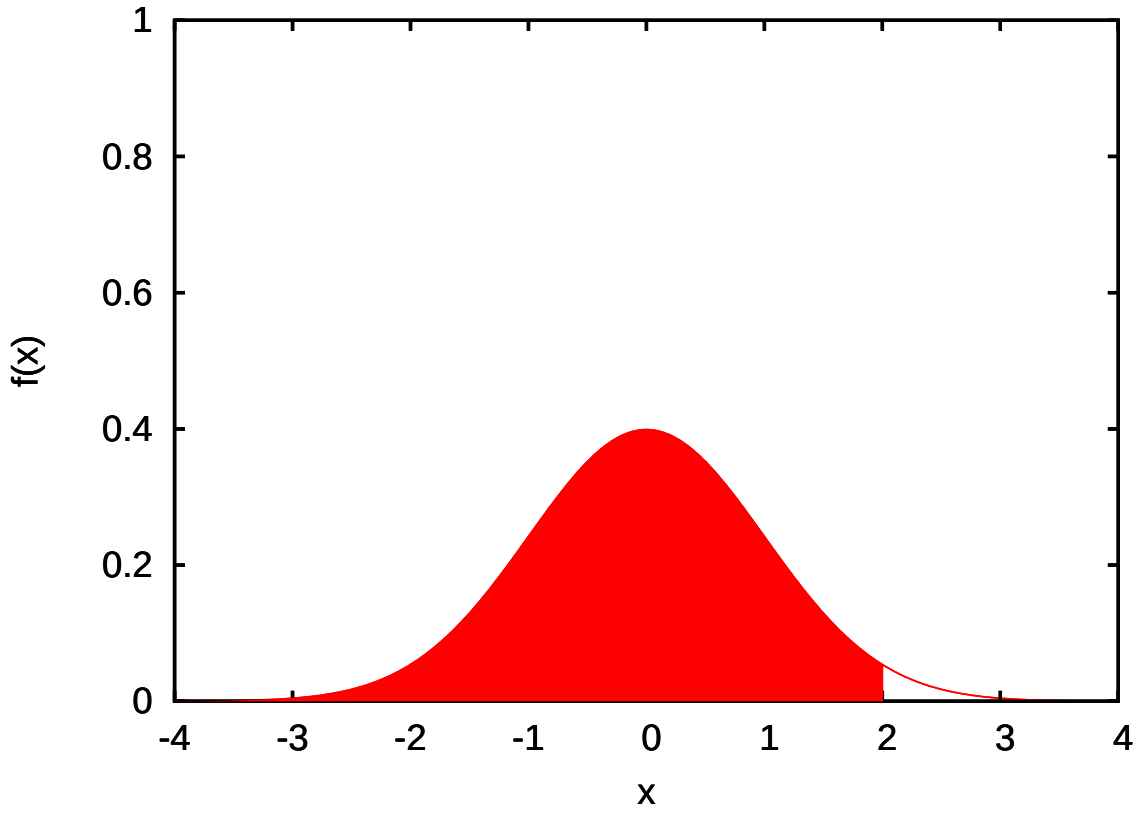
$$Y = \frac{X'}{D(X)} = \frac{X - E(X)}{D(X)} = \frac{X - 10}{0{,}5} \Rightarrow Y \sim N(0;1)$$



Dystrybuanta standaryzowanego rozkładu normalnego

Niech $X \sim N(0,1)$. Wartości dystrybuanty *tylko tej zmiennej*, czyli $\Phi(x) = P(X < x)$, są zawarte w tablicach, ale *tylko* dla wartości dodatnich x.

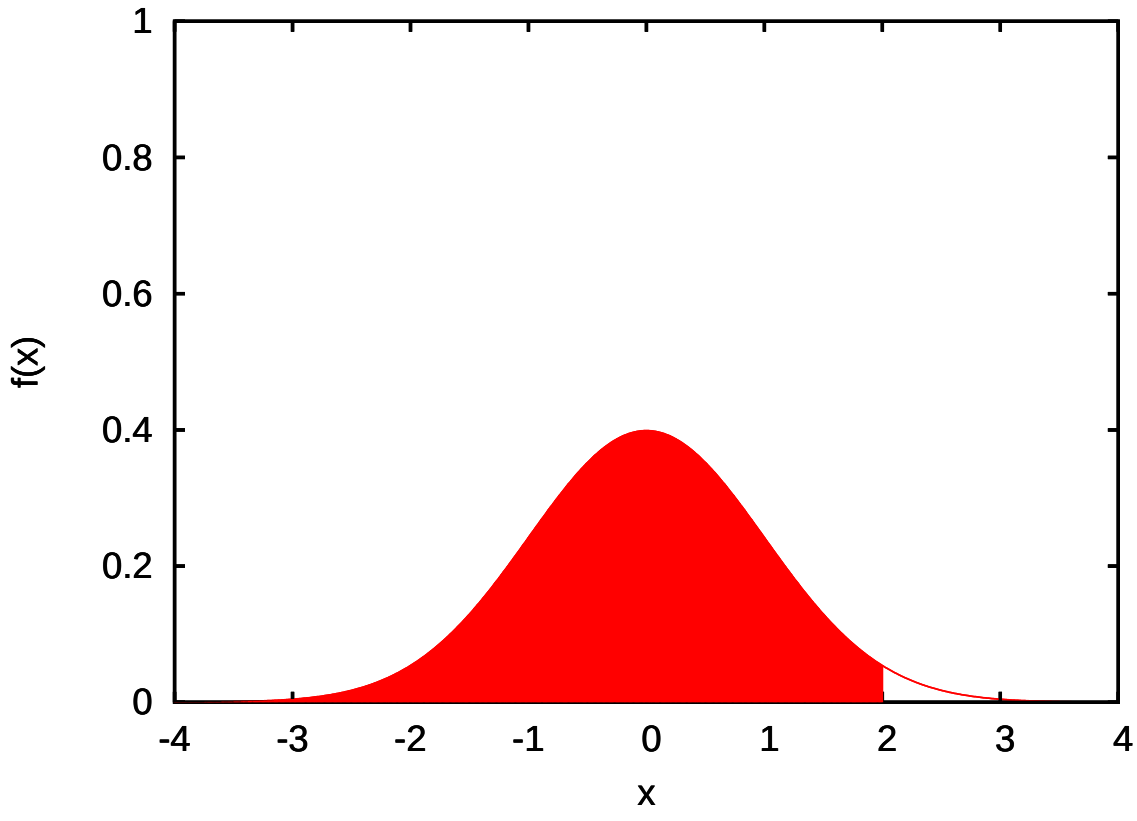
Np. $P(X < 2) = \Phi(2) =?$



Dystrybuanta standaryzowanego rozkładu normalnego

Niech $X \sim N(0,1)$. Wartości dystrybuanty *tylko tej zmiennej*, czyli $\Phi(x) = P(X < x)$, są zawarte w tablicach, ale *tylko* dla wartości dodatnich x.

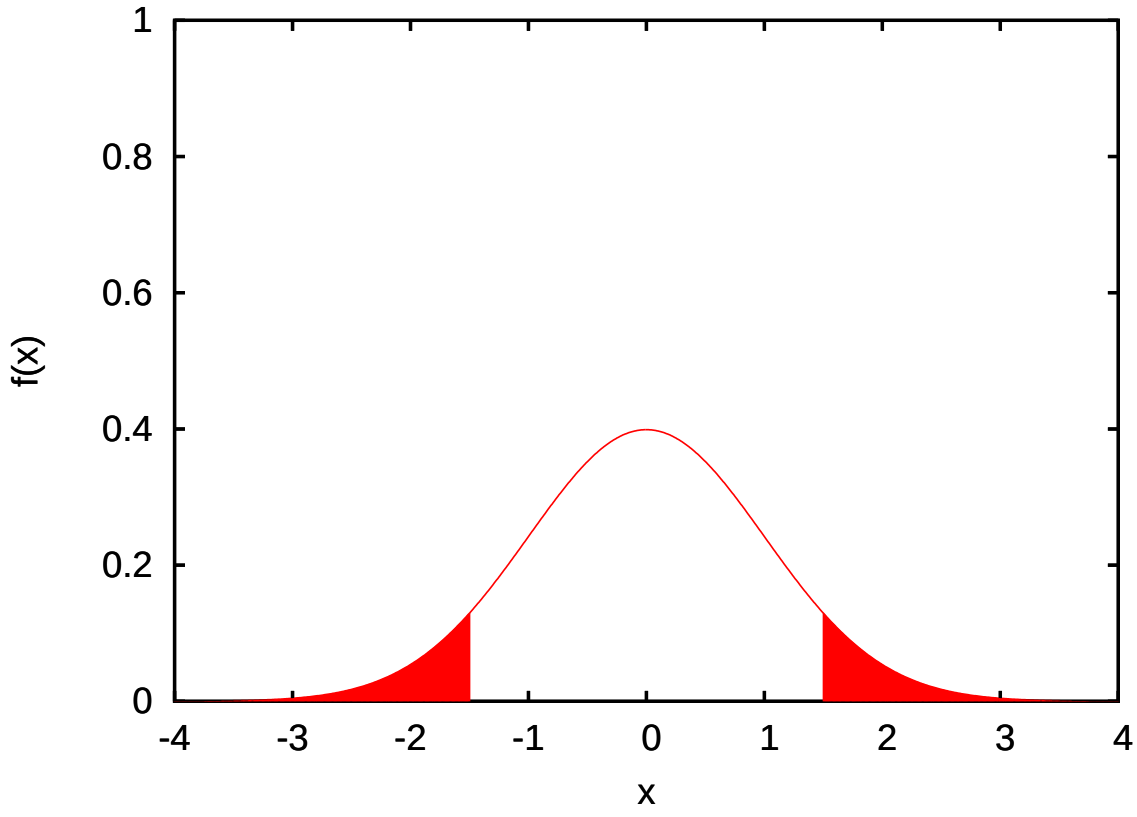
Np. $P(X < 2) = \Phi(2) = 0,9772$.



Dystrybuanta standaryzowanego rozkładu normalnego

Obliczenia dla wartości ujemnych x korzystają z symetrii standaryzowanego rozkładu normalnego względem $x = 0$.

Dla $x_0 \geq 0$ mamy $P(X > x_0) = P(X < -x_0)$



Dystrybuanta standaryzowanego rozkładu normalnego

Skoro mamy symetrię, to dla $x_0 \geq 0$ mamy:

$$\Phi(-x_0) = P(X < -x_0) = P(X > x_0)$$

Z własności zdarzeń przeciwnych i rozkładu ciągłego mamy:

$$P(X > x_0) = 1 - P(X < x_0) = 1 - \Phi(x_0)$$

W takim razie otrzymujemy:

$$\Phi(-x_0) = 1 - \Phi(x_0)$$

i ten wzór stosujemy w praktyce.

Np. $P(X < -2) = \Phi(-2) = 1 - \Phi(2) = 1 - 0,9772 = 0,0228$

Standaryzowany rozkład normalny $N(0,1)$

Przykład obliczenia prawdopodobieństwa

Niech $X \sim N(0;1)$; obliczyć $P(|X| > 3)$.

$$\begin{aligned} P(|X| > 3) &= P(X < -3) + P(X > 3) = \\ &= P(X < -3) + (1 - P(X < 3)) = \\ &= 1 - \Phi(3) + (1 - \Phi(3)) = \\ &= 2 \cdot (1 - \Phi(3)) = \\ &= 2 \cdot (1 - 0,99865) = \\ &= 0,0027. \end{aligned}$$

Zadanie domowe:

- $X \sim N(0,1)$; obliczyć: $P(X < -1,65)$,
- $X \sim N(0,1)$; obliczyć: $P(X \in [-1;1,5])$.

Znaczenie odchylenia standardowego w rozkładzie normalnym $N(0,1)$

Odchylenie od wartości oczekiwanej o wielokrotność odchylenia standardowego w rozkładzie:

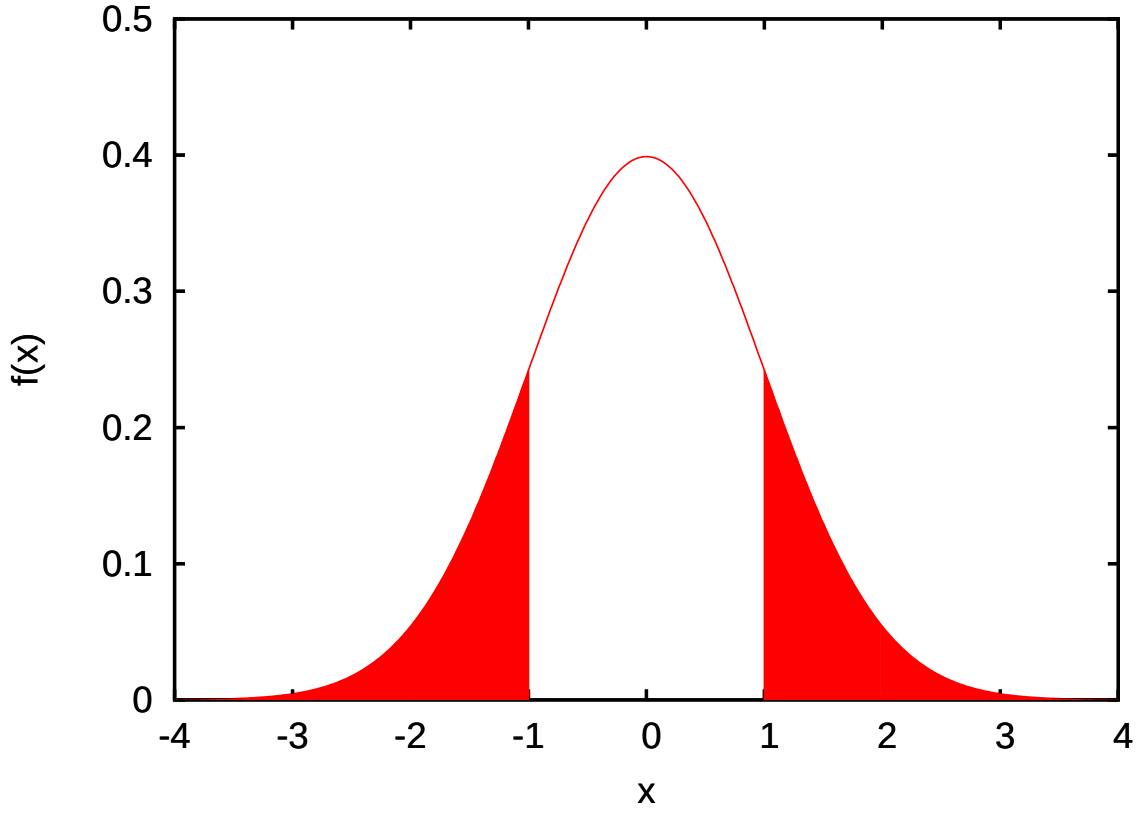
$$P(|X| > k) = ?$$

Uwaga! Standaryzacja rozkładu nie ma tutaj znaczenia; w rozkładach dowolnych, nieustandaryzowanych, rola wartości oczekiwanej i odchylenia standardowego jest taka sama, więc odpowiednie prawdopodobieństwa o formie:

$$P(|X - \mu| > k \cdot \sigma)$$

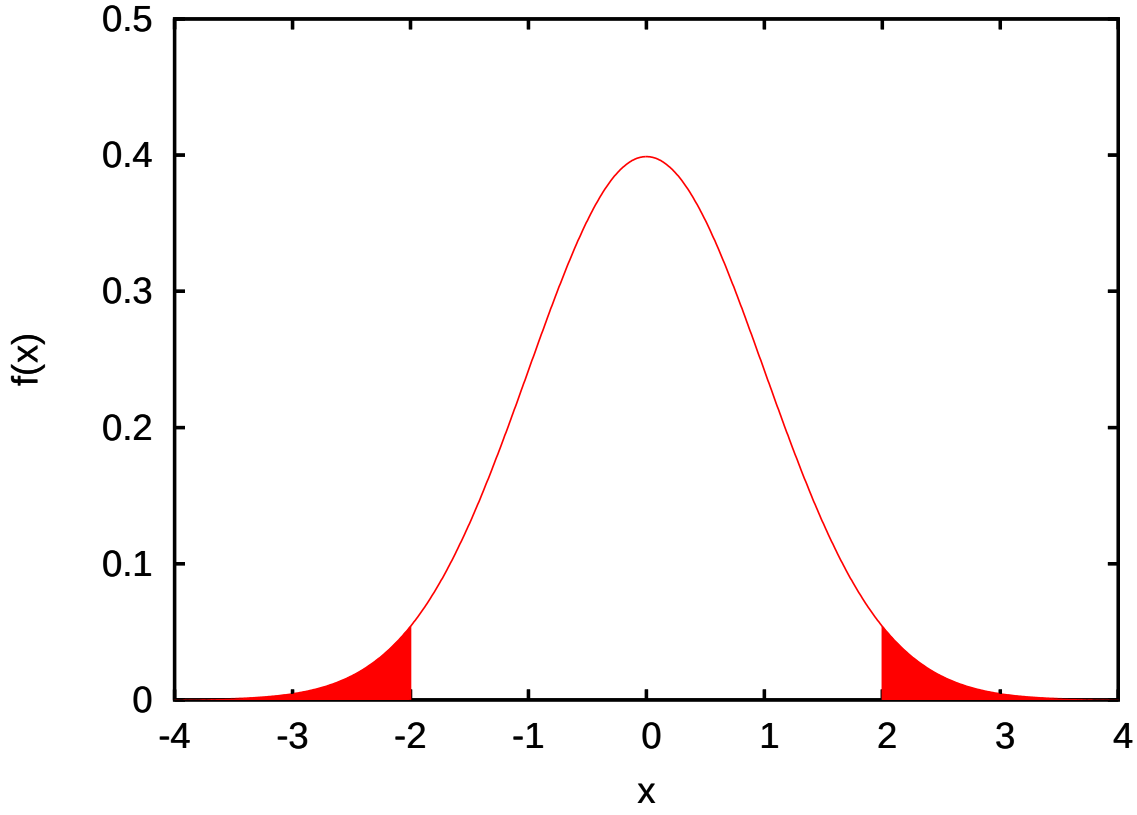
będą miały takie same wartości, jak te obliczone tutaj, dla tej samej wielokrotności k.

Znaczenie odchylenia standardowego w rozkładzie normalnym $N(0,1)$



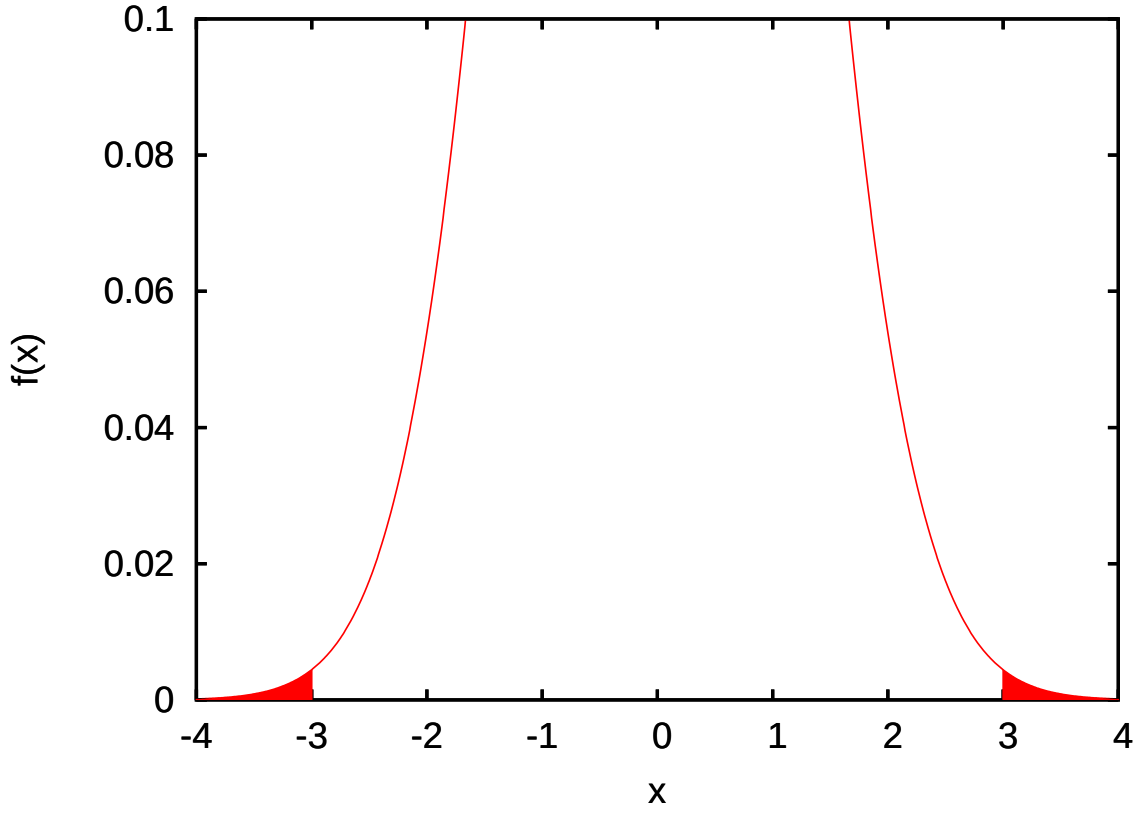
$$k = 1 \Rightarrow P(|X| > 1) = 0,3174.$$

Znaczenie odchylenia standardowego w rozkładzie normalnym $N(0,1)$



$$k = 2 \Rightarrow P(|X| > 2) = 0,0456.$$

Znaczenie odchylenia standardowego w rozkładzie normalnym $N(0,1)$



$$k = 3 \Rightarrow P(|X| > 3) = 0,0026.$$

Znaczenie odchylenia standardowego w rozkładzie normalnym $N(0,1)$

Reguła trzysigmowa: Prawdopodobieństwo odchylenia wartości zmiennej losowej o rozkładzie normalnym od wartości oczekiwanej o więcej niż 3 sigma wynosi ok. 0,0026 (czyli jest bardzo małe).

Dlaczego takie ważne są tego rozdzaju zdarzenia i ich prawdopodobieństwa?

Wszelkiego rodzaju testy hipotez (np. dla celów kontroli jakości produkcji) sprawdzają, czy zaobserwowane zdarzenie jest jeszcze prawdopodobnym odchyleniem od normy, czy już mało prawdopodobnym.

Zmienna o rozkładzie normalnym

Powrót do przykładu obliczenia prawdopodobieństwa

Niech $X \sim N(10; 0,5)$; obliczyć $P(9,5 \leq X \leq 11)$.

Co więcej wiemy?

- możemy wziąć zmienną Y przez ustandaryzowanie X,
- zmienna Y ma dystrybuantę podaną w tablicach.

Konieczne jest jednak „przetłumaczenie" zdarzenia dla zmiennej X w zdarzenie dla zmiennej Y.

Zmienna o rozkładzie normalnym

Tłumaczenie zdarzeń

Niech $X \sim N(10; 0,5)$.

$$Y = \frac{X - E(X)}{D(X)} = \frac{X - 10}{0,5} \Rightarrow Y \sim N(0,1)$$

$$9,5 \leq X \leq 11 \Leftrightarrow a_y \leq Y \leq b_y$$

Szukane: a_y, b_y (opis równoważnego przedziału dla Y).

$$9,5 \leq X \leq 11 \quad \text{odejmijmy } E(X) = 10$$

$$9,5 - 10 \leq X - 10 \leq 11 - 10 \quad \text{podzielmy przez } D(X) = 0,5$$

$$\frac{9,5 - 10}{0,5} \leq \frac{X - 10}{0,5} \leq \frac{11 - 10}{0,5} \quad \text{w środku otrzymaliśmy } Y$$

$$\text{Ostatecznie: } -1 \leq Y \leq 2$$

Zmienna o rozkładzie normalnym

Tłumaczenie zdarzeń

Ogólnie: niech $X \sim N(\mu; \sigma)$.

$$Y = \frac{X - E(X)}{D(X)} = \frac{X - \mu}{\sigma} \Rightarrow Y \sim N(0, 1)$$

$$a_x \leq X \leq b_x \Leftrightarrow a_y \leq Y \leq b_y$$

Szukane: a_y, b_y (opis równoważnego przedziału dla Y).

$$a_x \leq X \leq b_x \quad \text{odejmijmy } E(X) = \mu$$

$$a_x - \mu \leq X - \mu \leq b_x - \mu \quad \text{podzielmy przez } D(X) = \sigma$$

$$\frac{a_x - \mu}{\sigma} \leq \frac{X - \mu}{\sigma} \leq \frac{b_x - \mu}{\sigma} \quad \text{w środku otrzymaliśmy } Y$$

$$\text{Ostatecznie: } a_y = \frac{a_x - \mu}{\sigma} \quad b_y = \frac{b_x - \mu}{\sigma}$$

Zmienna o rozkładzie normalnym

Powrót do przykładu obliczenia prawdopodobieństwa

Niech $X \sim N(10; 0,5)$; obliczyć $P(9,5 \leq X \leq 11)$.

$$Y = \frac{X - E(X)}{D(X)} = \frac{X - 10}{0,5} \Rightarrow Y \sim N(0,1)$$

$$9,5 \leq X \leq 11 \Leftrightarrow -1 \leq Y \leq 2$$

$$P(9,5 \leq X \leq 11) = P(-1 \leq Y \leq 2) = \Phi(2) - \Phi(-1) =$$

$$= \Phi(2) - (1 - \Phi(1)) = \Phi(2) - 1 + \Phi(1) = (\text{z tablic})$$

$$0,9772 - 1 + 0,8413 = 0,8185.$$

Zmienna o rozkładzie normalnym

Zadanie domowe:

- $X \sim N(15, 3)$, obliczyć $P(X \leq 10)$.
- $X \sim N(10,7; 2,1)$, obliczyć $P(X \in [6,5; 14,9])$.

Centralne twierdzenie integralne Lindeberga–Levy'ego

Treść twierdzenia

Niech $X_1, X_2, \ldots, X_n, \ldots$ będzie ciągiem zmiennych losowych:

- niezależnych,
- o takim samym rozkładzie,
- takich, że $E(X_i) = \mu < \infty$,
- takich, że $0 < D^2(X_i) = \sigma^2 < \infty$.

Niech:

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{i} \quad U_n = \frac{\overline{X}_n - \mu}{\sigma} \cdot \sqrt{n}.$$

Wtedy:

$$\forall u \in \mathbf{R} \qquad \lim_{n\to\infty} P(U_n < u) = \Phi(u).$$

Centralne twierdzenie integralne Lindeberga–Levy'ego

Komentarz do założeń

Niezależność i taki sam rozkład zmiennych to model eksperymentu, w którym pojedyncze doświadczenia wykonywane są niezależnie od siebie i w dokładnie takich samych warunkach.

Założenie o skończonej wartości oczekiwanej jest wymaganiem formalnym: istnieją rozkłady, dla których nie można obliczyć wartości oczekiwaną, ale rzadko występują takie w praktyce.

Założenie o skończonej i niezerowej wariancji jest wymaganiem formalnym: istnieją rozkłady bez skończonej wariancji; dodatkowo sytuacja nie może być deterministyczna (musi być losowość).

Wniosek: założenia są bardzo praktyczne, gdyż pozwalają na rozsądne konstruowanie eksperymentów statystycznych!

Centralne twierdzenie integralne Lindeberga–Levy'ego

Komentarz do dodatkowych definicji

Ciąg zmiennych losowych $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ to nic innego jak ciąg kolejnych średnich arytmetycznych dla ciągu $X_1, X_2, \ldots, X_n, \ldots$. Zauważmy, że:

$$E(\overline{X}_n) = \mu \quad \text{i} \quad D^2(\overline{X}_n) = \frac{\sigma^2}{n}.$$

W takim razie:

$$U_n = \frac{\overline{X}_n - E(\overline{X}_n)}{D(\overline{X}_n)} = \frac{\overline{X}_n - \mu}{\sigma} \cdot \sqrt{n}$$

jest po prostu *ustandaryzowaną średnią arytmetyczną*.

Standaryzacja ta pozwala po prostu wyrazić twierdzenie dla rozkładów o różnych wartościach oczekiwanych i wariancjach.

Centralne twierdzenie integralne Lindeberga–Levy'ego

Komentarz do tezy

Teza stwierdza, że po standaryzacji średniej arytmetycznej $\overline{X}_n$ możemy do obliczania prawdopodobieństw dla tej zmiennej wykorzystać standaryzowany rozkład normalny.

Teza jest *bardzo silna*! Nawet gdy *nie znamy* rozkładu prawdopodobieństwa badanej zmiennej, to rozkład standaryzowanych średnich arytmetycznych $\overline{X}_n$ wielu takich zmiennych zbiega się do rozkładu $N(0,1)$ dla rozsądnych n.

Z tego twierdzenia wynika powszechność stosowania średnich arytmetycznych i pojawiania się rozkładów normalnych w eksperymentach statystycznych.

Centralne twierdzenie integralne Lindeberga–Levy'ego

Komentarz do tezy (c.d.)

Niektóre komentarze tej tezy mówią, że wyjaśnia ona także pojawianie się rozkładów normalnych w naturze (np. rozkład długości kości u zwierząt i człowieka, rozkład długości łodygi u roślin, itp.). Wg nich w naturze na wiele zjawisk i wielkości fizycznych wpływa wiele niezależnych, niekontrolowanych czynników, które łącznie mogą „uśredniać się" do rozkładu normalnego.

Do tego wniosku należy jednak podchodzić ostrożnie, gdyż jest to tylko rozumowanie przybliżone.

Centralne twierdzenie integralne Lindeberga–Levy'ego

Modyfikacja tezy

Twierdzenie jest prawdziwe (przy tych samych założeniach) także po wyrażeniu tezy dla standaryzowanych sum, zamiast dla standaryzowanych średnich arytmetycznych.

Niech:

$$S_n = \sum_{i=1}^{n} X_i \quad \text{i} \quad Z_n = \frac{S_n - n \cdot \mu}{\sigma \cdot \sqrt{n}}.$$

Wtedy:

$$\forall z \in \mathbf{R} \qquad \lim_{n \to \infty} P(Z_n < z) = \Phi(z).$$

Centralne twierdzenie integralne Lindeberga–Levy'ego

A co się dzieje, gdy niezależne zmienne X_i mają taki sam rozkład zero–jedynkowy $B_1(p)$?

Przy takich założeniach $S_n = \sum_{i=1}^{n} X_i$ ma rozkład $B_n(p)$ (por. twierdzenie Bernoulliego).

Jednak ciąg standaryzowanych zmiennych

$$Z_n = \frac{S_n - E(S_n)}{D(S_n)} = \frac{S_n - n \cdot p}{\sqrt{n \cdot p \cdot (1-p)}},$$

zgodnie z tw. Lindeberga–Levy'ego, ma dystrybuantę zbieżną do dystrybuanty $N(0,1)$.

Wniosek: rozkład Bernoulliego jest zbieżny do rozkładu normalnego. Przybliżenie rozkładem normalnym można bezpiecznie stosować dla $n \cdot p \geq 5$ i $n \cdot (1-p) \geq 5$.

Zastosowanie twierdzenia Lindeberga–Levy'ego

Rzucono 720 razy kostką sześcienną. Jakie jest P, że całkowita liczba wyrzuconych czwórek (czterech oczek) jest zawarta w granicach od 100 do 150?

Niech wynik pojedynczego rzutu kostką będzie zmienną losową X. Interesujące nas zdarzenie wylosowania czterech oczek w rzucie oznaczymy jako sukces ($X = 1$). Pozostałe wyniki losowania będziemy traktowali jako porażkę ($X = 0$). Jasne jest, że $X \sim B_1(p = 1/6)$.

W eksperymencie mamy 720 rzutów, a więc mamy do czynienia z ciągiem zmiennych losowych $X_1, \ldots, X_{720}$ o takim samym rozkładzie $B_1(p = 1/6)$. Rzuty są wykonywane niezależnie od siebie, więc i zmienne X_i są od siebie niezależne.

Zastosowanie twierdzenia Lindeberga–Levy'ego

Mamy obliczyć P dla całkowitej liczby wyrzuconych czwórek we wszystkich 720 rzutach, a więc będziemy się posługiwać zmienną:

$$S_{720} = \sum_{i=1}^{720} X_i.$$

Poszukiwane prawdopodobieństwo ma w takim razie postać $P(100 \leq S_{720} \leq 150)$.

Z tw. Bernoulliego wynika, że $S_{720} \sim B_{720}(p = 1/6)$, czyli:

$$E(S_{720}) = n \cdot p = 720 \cdot 1/6 = 120.$$

$$D(S_{720}) = \sqrt{D^2(S_{720})} = \sqrt{n \cdot p \cdot (1-p)} = \sqrt{720 \cdot 1/6 \cdot 5/6} = 10.$$

Jednak bezpośrednie korzystanie z rozkładu dwumianowego jest kłopotliwe, ze względu na dużą liczbę operacji i błędy numeryczne.

Zastosowanie twierdzenia Lindeberga–Levy'ego

Zauważmy, że ciąg zmiennych $X_i,\ i = 1, \ldots, 720$ spełnia założenia tw. Lindeberga–Levy'ego. Twierdzenie to, w wersji dla sum, mówi, że zmienna S_{720} po standaryzacji ma w przybliżeniu rozkład normalny $N(0, 1)$.

$$Z_{720} = \frac{S_{720} - E(S_{720})}{D(S_{720})} = \frac{S_{720} - 120}{10} \sim N(0, 1).$$

Przybliżenie to będzie wystarczająco dobre, gdyż $n \cdot p = 120 \geq 5$ i $n \cdot (1 - p) = 600 \geq 5$

Należy jednak najpierw przetłumaczyć interesujące nas zdarzenie dla S_{720} w zdarzenie dla Z_{720}:

$$100 \leq S_{720} \leq 150 \Leftrightarrow a_z \leq Z_{720} \leq b_z$$

$$a_z = \frac{100 - 120}{10} = -2 \quad b_z = \frac{150 - 120}{10} = 3.$$

Zastosowanie twierdzenia Lindeberga–Levy'ego

W takim razie poszukiwaną wartość można zapisać jako:

$$P(100 \leq S_{720} \leq 150) = P(-2 \leq Z_{720} \leq 3)$$

Teraz można skorzystać z tezy tw. Lindeberga–Levy'ego, czyli dla zmiennej Z_{720} wykorzystać dystrybuantę $N(0,1)$

$$P(100 \leq S_{720} \leq 150) = P(-2 \leq Z_{720} \leq 3) = \Phi(3) - \Phi(-2) =$$

$$= \Phi(3) - (1 - \Phi(2)) = \Phi(3) + \Phi(2) - 1 = (\text{z tablic})$$

$$= 0,9987 + 0,9772 - 1 = 0,9759.$$

Wniosek: prawdopodobieństwo, że w 720 rzutach sześcienną kostką uzyskamy całkowitą liczbę czwórek pomiędzy 100 a 150 wynosi 0,9757 (ponad 97,5%).

Zastosowanie twierdzenia Lindeberga–Levy'ego

Zadanie domowe

- $X \sim B_{120}(p = 0,1)$, obliczyć $P(X \geq 90)$.
- Partia towaru ma wadliwość 7%. Pobrano próbę 800–elementową tego towaru. Obliczyć prawdopodobieństwo, że liczba sztuk wadliwych w tej próbie jest zawarta w granicach 6%-9%. Skorzystać z tw. Lindeberga–Levy'ego.

Wymagana wiedza i umiejętności

- Rachunek prawdopodobieństwa a statystyka:
  - cel stosowania rachunku prawdopodobieństwa w statystyce;
  - statystyka opisowa vs. statystyka matematyczna.
- Pojęcia podstawowe:
  - pojęcie zmiennej losowej;
  - funkcja prawdopodobieństwa i gęstości: definicja i sens;
  - dystrybuanta: definicja i sens;
  - charakterystyki liczbowe zmiennej losowej $E(X)$, $D^2(X)$ i $D(X)$: definicje, sens, sposób obliczania.
- Rozkład Bernoulliego:
  - funkcja prawdopodobieństwa, wartość oczekiwana, wariancja;
  - sens twierdzenia Bernoulliego dla statystyki.

Wymagana wiedza i umiejętności

- Rozkład normalny:
  - wykres funkcji gęstości i jego zależność od parametrów, wartość oczekiwana, wariancja;
  - standaryzacja zmiennej normalnej;
  - korzystanie z tablic rozkładu $N(0,1)$;
  - obliczanie prawdopodobieństw w rozkładzie $N(0,1)$;
  - tłumaczenie zdarzeń dla dowolnej zmiennej normalnej w zdarzenia dla zmiennej $N(0,1)$;
  - obliczanie prawdopodobieństw w dowolnym rozkładzie normalnym.
- Twierdzenie Lindeberga–Levy'ego:
  - treść dla średnich arytmetycznych i sum;
  - sens założeń i ich znaczenie dla statystyki;
  - sens tezy i jej znaczenie dla statystyki;
  - korzystanie z twierdzenia przy obliczaniu prawdopodobieństw.

Literatura

- W. Starzyńska, *Statystyka praktyczna*, Wydawnictwo Naukowe PWN, 2000.
- J. Koronacki, J. Mielniczuk, *Statystyka dla studentów kierunków technicznych i przyrodniczych*, WNT, 2001.
- G. A. Ferguson, Y. Takane, *Analiza statystyczna w psychologii i pedagogice*, Wydawnictwo Naukowe PWN, Warszawa, 2003.
- L. T. Kubik, *Zastosowanie elementarnego rachunku prawdopodobieństwa do wnioskowania statystycznego. Wykład i uwagi krytyczne*, Wydawnictwo Naukowe PWN, Warszawa, 1998.
- W. Krysicki i inni, *Rachunek prawdopodobieństwa i statystyka matematyczna w zadaniach, Część I, Rachunek prawdopodobieństwa, Część II, Statystyka matematyczna*, Wydawnictwo Naukowe PWN, Warszawa, 1995.
- T. Gerstenkorn, T. Śródka, *Kombinatoryka i rachunek prawdopodobieństwa*, PWN, Warszawa, 1972.

Dziękuję za uwagę!