Zmienne losowe

dr Mariusz Grządziel

Katedra Matematyki, Uniwersytet Przyrodniczy we Wrocławiu

rok akademicki 2017/2018 — semestr letni

Pojęcie zmiennej losowej

Nieformalne określenie— wynik liczbowy doświadczenia losowego.
Przykładami zmiennej losowej są:
suma oczek otrzymanych po dwukrotnym rzucie kostką;
cena losowo wybranego mieszkania (z listy mieszkań oferowanych do sprzedaży);
temperatura człowieka, zmierzona w losowo wybranej chwili.

Definicja 1

Funkcję określoną na przestrzeni zdarzeń elementarnych będziemy nazywać zmienną losową.

Uwaga 1

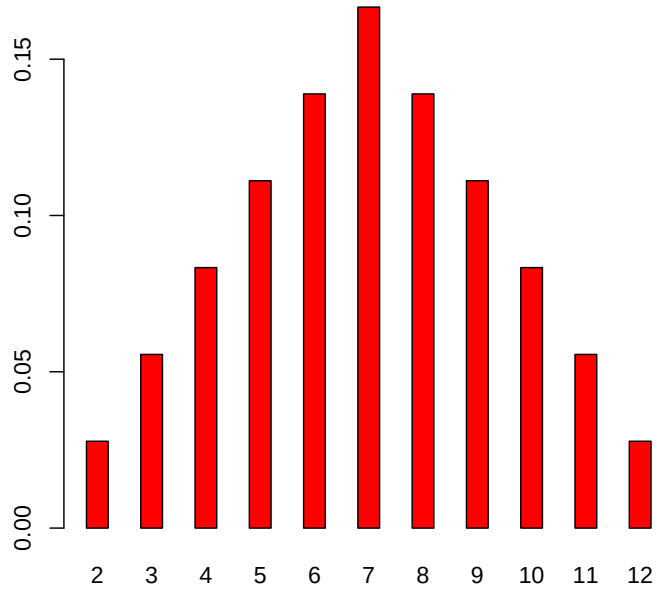
Gdy przestrzeń zdarzeń elementarnych nie jest przeliczalna (jej elementów nie można ustawić w ciąg), powyższa definicja jest uproszczoną wersją definicji zmiennej losowej przyjętej w monografiach poświęconych teorii prawdopodobieństwa.

Zmienna losowa— przykład

Rzucamy dwukrotnie kostką.
Niech X — suma oczek;
X— przykład zmiennej losowej.
X przyjmuje wartości $2, 3, \ldots, 11, 12$ z prawdopodobieństwami:

k	2	3	4	5	6	7	8	9	10	11	12
$P(X = k)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Funkcja przyporządkowująca $k \in \{2, 3, \ldots, 11, 12\}$
prawdopodobieństwo $P(X = k)$— rozkład zmiennej X.
Notacja: $X = k$—zbiór zdarzeń elementarnych ω takich, że $X(\omega) = k$.
Analogicznie: $X < k$—zbiór zdarzeń elementarnych ω takich, że $X(\omega) < k$.



Rysunek: Wykres słupkowy przedstawiający rozkład zmiennej losowej X, sumy oczek otrzymanych w dwukrotnym rzucie kostką

Definicja 2

Zmienną losową nazywamy dyskretną (skokową), jeśli zbiór jej wartości $x_1, x_2, \ldots,$ można ustawić w ciąg.

W pewnych podręcznikach można znaleźć bardziej ogólną definicję dysktretnej zmiennej losowej.

G. Cantor (1873): wszystkich liczb rzeczywistych nie da się ustawić w ciąg.

Rozkład dyskretnej zmiennej losowej

Zbiór wartości dyskretnej zmiennej losowej X— ciąg $x_1, x_2, \ldots,$ (skończony lub nieskończony).

Definicja 3

Rozkład („rozłożenie masy prawdopodobieństwa") zmiennej losowej dyskretnej X jest określony przez układ nieujemnych liczb $p_1, p_2, \ldots$ spełniających warunki:

$$\sum p_i = 1, \tag{1}$$

$$p_i = P(X = x_i). \tag{2}$$

Uwaga 2

Zbiór par uporządkowanych $(x_1, p_1), (x_2, p_2) \ldots$ będziemy nazywać funkcją prawdopodobieństwa. Jeżeli liczba zdarzeń elementarnych jest skończona, funkcję tę można przedstawić przy pomocy tabelki.

Uwaga 3

Rozkład prawdopodobieństwa zmiennej losowej X jest definiowany w literaturze jako funkcja określona na pewnej rodzinie podzbiorów prostej $\mathbb{R}$ („rodzinie borelowskich podzbiorów $\mathbb{R}$") o wartościach w odcinku $[0, 1]$; rodzina ta zawiera wszystkie odcinki, półproste i proste,sumy odcinków (także nieskończone — przeliczalne).

Dyskretne zmienne losowe— przykłady

Przykłady:

- Rozkład X, sumy oczek w dwukrotnym rzucie kostką;
- Rozkład dwumianowy (por. Wykład 14, sem. 1);
- Rozkład zmiennej losowej dyskretnej W o rozkładzie danej tabelką:

x_i	-1	2	5
p_i	0,5	0,3	0,2

- Rozkład Z, gdzie Z oznacza liczbę rzutów monetą, po której po raz pierwszy wypada orzeł (zdarzeniu polegającemu na tym, że orzeł wypadnie już w pierwszym rzucie, odpowiada wartość zmiennej Z równa 0).

z niezależności zdarzeń:

$$P(Z = k) = \left(\frac{1}{2}\right)^{k+1}, \quad k = 0, 1, 2, \ldots.$$

Zmienna losowa Z przykład dyskretnej zmiennej losowej, dla której zbiór wartości: $\{0, 1, 2, \ldots\}$ nie jest skończony.

Rzuty osobiste— przykład

Niech X- liczba trafień w wykonywanym przez koszykarza A rzucie osobistym.Niech: T odpowiada trafieniu do kosza, C odpowiada chybieniu.
Przestrzeń zdarzeń elementarnych: $\mathcal{S} = \{C, T\}$.
Niech X- liczba trafionych rzutów. Zmienna X jest funkcją określoną na $\mathcal{S}$;

$$X(C) = 0, \quad X(T) = 1.$$

Zakładamy, że prawdopodobieństwo trafienia wynosi 0,9.
Rozkład zmiennej losowej X można przedstawić przy pomocy tabelki:

k	0	1
$P(X = k)$	0,1	0,9

Liczba trafień Y w dwóch rzutach

Niech Y- liczba trafień w dwóch wykonywanych przez koszykarza A rzutach osobistych.
Przyjmujemy, że prawdopodobieństwo trafienia w jednym rzucie osobistym wynosi 0,9 i zdarzenie trafienia/chybienia w drugim rzucie jest niezależne od analogicznego zdarzenia w pierwszym rzucie.
Można pokazać, że:

$$P(Y=0) = \left(\frac{1}{10}\right)^2 = 0{,}01,$$

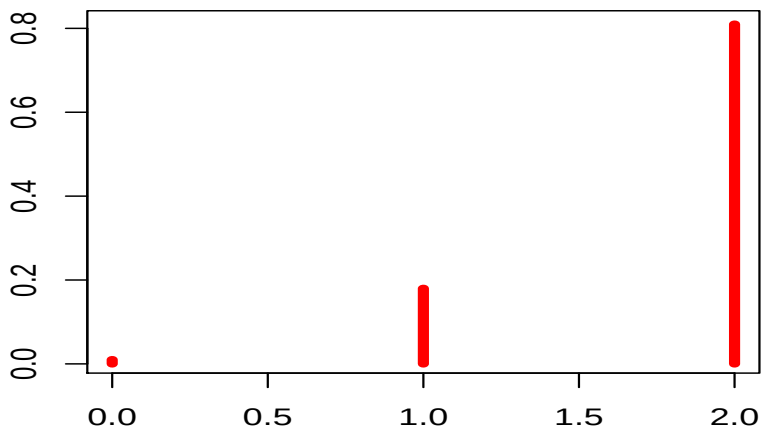
$$P(Y=1) = 2 \times \frac{1}{10} \times \frac{9}{10} = 0{,}18,$$

$$P(Y=2) = \left(\frac{9}{10}\right)^2 = 0{,}81.$$

Liczba trafień Y w dwóch rzutach— c.d.

Rozkład można przedstawić w postaci tabelki lub wykresu słupkowego:

k	0	1	2
$P(X = k)$	0,01	0,18	0,81



Rozkład dwumianowy

Symbol Newtona $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ jest równy liczbie podzbiorów k-elementowych zbioru n-elementowego ($0 \leqslant k \leqslant n$).

Definicja 4

Mówimy, że zmienna losowa X ma rozkład dwumianowy z parametrami $n \in \mathbb{N}$ i $0 < p < 1$, co w skrótowo zapisujemy $X \sim Bin(n, p)$ (lub $X \sim Bin(n; p)$), jeśli

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, 2, \ldots, n.$$

„Dziesięciokrotny rzut monetą”— przykład

Niech V oznacza liczbę orłów otrzymanych w dziesięciokrotnym rzucie monetą (zakładamy, że moneta jest „rzetelna”, tj. prawdopodobieństwo otrzymania orła jest równe $\frac{1}{2}$ oraz że wyniki kolejnych rzutów są od siebie niezależne). Chcemy obliczyć prawdopodobieństwo;

$$P(V \geqslant 9).$$

Rozwiązanie $V \sim Bin(10; 0{,}5)$,

$$\begin{aligned} P(V \geqslant 9) &= P(V = 9) + P(V = 10) = \\ &= \binom{10}{9} (0{,}5)^9 (0{,}5)^1 + \binom{10}{10} (0{,}5)^{10} (0{,}5)^0 = \\ &= \frac{10}{1024} + \frac{1}{1024} = \frac{11}{1024}. \end{aligned}$$

Pojęcie dystrybuanty rozkładu

W obliczeniach podobnych do tych z poprzedniego przykładu użyteczne może się okazać pojęcie dystrybuanty zmiennej losowej.

Definicja 5

Niech X będzie dowolną zmienną losową. Dystrybuantą zmiennej losowej X nazywamy funkcję F określoną jako:

$$F(x) = P(X \leqslant x).$$

Uwaga 4

Zmienne losowe X i Y mają taki sam rozkład, jeśli mają te same dystrybuanty tzn. $F_X(t) = F_Y(t)$ dla każdego $t \in \mathbb{R}$.

„Dziesięciokrotny rzut monetą”—c.d.

V-liczba wyrzuconych orłów w dziesięciokrotnym rzucie monetą;

$$P(V \geqslant 9) = P(V = 9) + P(V = 10) = F_V(10) - F_V(8)$$

gdzie F_V jest dystrybuantą zmiennej losowej V.
Obliczenia wykonane w R-rze:

```
> pbinom(10,10,0.5)- pbinom(8,10,0.5)
[1] 0.01074219
```

pbinom- pierwsza litera odpowiada „dystrybuancie”, binom odpowiada rodzajowi rozkładu (ang. *binomial*- dwumianowy). Korzystając z polecenia *pbinom* można obliczać wartości dystrybuanty rozkładu $Bin(n, p)$ dla dużych wartości n.

Pojęcie dystrybuanty zmiennej losowej

Twierdzenie 1

Dystrybuanta F_X zmiennej losowej X ma następujące własności:

1. *F_X jest niemalejąca;*
2. $\lim_{t\to\infty} F_X(t) = 1$; $\lim_{t\to-\infty} F_X(t) = 0$;
3. *F_X jest prawostronnie ciągła.*

Twierdzenie 2

Jeżeli funkcja $F : \mathbb{R} \to \mathbb{R}$ spełnia warunki 1-3, to jest dystrybuantą pewnej zmiennej losowej.

Przykład

Dystrybuanta F_W zmiennej losowej dyskretnej W o rozkładzie danej tabelką:

x_i	-1	2	5
p_i	0,5	0,3	0,2

ma postać

$$F_W(t) = \begin{cases} 0 & \text{dla } t < -1; \\ 0{,}5 & \text{dla } t \in [-1; 2); \\ 0{,}8 & \text{dla } t \in [2; 5); \\ 1 & \text{dla } t \geqslant 5. \end{cases}$$

Uwaga 5

Jeśli znana jest dystrybuanta zmiennej losowej dyskretnej X, można wyznaczyć jej funkcję prawdopodobieństwa.

Funkcja kwantylowa

Dla zmiennej losowej X określamy wzorem

$$Q_X(u) = \min\{t \in \mathbb{R} : F_X(t) \geqslant u\}, \quad u \in (0,1),$$

gdzie F_X oznacza dystrybuantę zmiennej losowej X.

Wartość funkcji kwantylowej zmiennej losowej X dla argumentu $u \in (0,1)$: kwantyl rzędu u rozkładu zmiennej losowej X.

Przykład 1

Dla zmiennej losowej W funkcja kwantylowa Q_W ma postać

$$Q_W(u) = \begin{cases} -1 & \text{dla } u \in (0;0{,}5]; \\ 2 & \text{dla } u \in (0{,}5;0{,}8]; \\ 5 & \text{dla } u \in (0{,}8;1). \end{cases}$$

Zmienna o rozkładzie Poissona

Definicja 6

Mówimy, że zmienna losowa X ma rozkład Poissona z parametrem $\lambda > 0$, jeśli przyjmuje ona wartości w zbiorze $\{0, 1, 2, \ldots, \}$ oraz

$$P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k = 0, 1, 2, \ldots.$$

Rozkład Poissona może być zastosowany z powodzeniem do opisu takich cech jak liczba nasion chwastów wśród nasion trawy, liczba klientów zgłaszających się dziennie do banku, liczba wypadków drogowych na placu Grunwaldzkim w danym dniu itd.

Zmienna o rozkładzie geometrycznym

Mówimy, że zmienna losowa X ma rozkład geometryczny z parametrem p, $0 < p < 1$, jeżeli przyjmuje ona wartości w zbiorze $\{1, 2, \ldots, \}$ oraz

$$P(X = k) = p(1 - p)^{k-1}.$$

Przykłady

- Rozważmy zmienną losową Z, liczba reszek poprzedzających pierwsze wypadnięcie orła. Zmienna $Z + 1$ ma rozkład geometryczny z parametrem $p = \frac{1}{2}$.
- Rzucamy kostką sześcienną tak długo, dopóki nie wypadnie szóstka. Liczba rzutów ma rozkład geometryczny z parametrem $p = \frac{1}{6}$.

Pole trapezu krzywoliniowego

Przypomnienie: figurę ograniczoną przez:

- wykres funkcji $y = f(x)$, gdzie f jest funkcją ciągłą;
- proste $x = a$, $x = b$, $a < b$,
- oś OX (tj. prostą $y = 0$)

będziemy nazywać *trapezem krzywoliniowym* (odpowiadającym funkcji f oraz odcinkowi $[a, b]$).

Definicja 7 (wstępne określenie całki oznaczonej)

Niech $y = f(x)$ będzie funkcją ciągłą i nieujemną na odcinku $[a, b]$. Pole trapezu krzywoliniowego odpowiadającego funkcji f i odcinkowi $[a, b]$ będziemy nazywać całką oznaczoną funkcji f na przedziale $[a, b]$ i oznaczać symbolem

$$\int_a^b f(x)dx.$$

Pole trapezu krzywoliniowego — przykład

Pole trapezu krzywoliniowego odpowiadającego funkcji $f(x) = \sqrt{1 - x^2}$ i odcinkowi $[1, 1]$ jest równe $\frac{\pi}{2}$ (połowie pola koła o promieniu 1). Przy użyciu notacji rachunku całkowego

$$\int_{-1}^{1} f(x)dx = \frac{\pi}{2}.$$

Funkcje przedziałami ciągłe I

Rozważmy odcinek $[a, b]$. Możemy go przedstawić jako sumę n podprzedziałów wybierając $n-1$ punktów $x_1, x_2, \ldots, x_{n-1}$ spełniających nierówności:

$$a < x_1 < x_2 < \ldots < x_{n-1} < b. \tag{3}$$

Będziemy przyjmować oznaczenia: $x_0 = a$, $x_n = b$.

Układ punktów spełniających warunek (3) będziemy nazywać podziałem P odcinka $[a, b]$. Będziemy używać symbolu

$$P = \{x_0, x_1 < x_2 < \ldots < x_{n-1}, x_n\}$$

na oznaczenie tego podziału. Odcinek $(x_{k-1} < x_k)$ będziemy nazywać k-tym otwartym podziałem odcinka P.

Funkcje przedziałami ciągłe II

Definicja 8

Powiemy,że funkcja s jest przedziałami stała na przedziale $[a, b]$*, jeżeli istnieje podział* $P = \{x_0, x_1 < x_2 < \ldots < x_{n-1}, x_n$ *odcinka* $[a, b]$ *taki, że s jest stała na każdym otwartym odcinku podziału P, tj. dla każdego* $k = 1, 2, \ldots, n$ *istnieje liczba* s_k *taka, że*

$$s(x) = s_k \quad dla \quad x \in (x_{k-1}, x_k)$$

Funkcje przedziałami ciągłe III

Definicja 9

Całkę funkcji przedziałami stałej na przedziale $[a, b]$*, która na* k*-ym otwartym odcinku podziału* $P = \{a = x_0, x_1 < x_2 < \ldots < x_{n-1}, x_n = b$ *przyjmuje wartość równą* c_k*,* $k = 1, 2, \ldots, n$*, będziemy definiować wzorem*

$$\sum_{k=1}^{n} c_k(x_k - x_{k-1}).$$

Definicję całki dla funkcji przedziałami ciągłej można uznać za punkt startowy do określenia definicji całki oznaczonej Riemanna dla szerokiej klasy funkcji (które mogą mieć bardzo wiele punktów nieciągłości i wielokrotnie zmieniać znak).

Całki niewłaściwe

Można uzasadnić, że funkcji f określonej wzorem

$$f(x) = \frac{1}{x^2}$$

prawdziwa jest równość

$$\int_1^T \frac{1}{x^2} = 1 - \frac{1}{T}.$$

Jeżeli T zbiega do nieskończoności, to całka ta będzie zbiegać do 1; granicę tę oznaczymy

$$\int_1^\infty f(x)dx.$$

W podobny sposób definiujemy całkę niewłaściwą na półprostej $[a, \infty)$ (lub półprostej $(-\infty, a]$) dla dowolnej funkcji f (spełniającej pewne założenia), gdzie $a \in \mathbb{R}$. Całkę tę oznaczamy symbolem $\int_a^\infty f(x)dx$ (lub $\int_{-\infty}^a f(x)dx$)

Całka niewłaściwa na prostej

Definicja 10

Dla funkcji f określonej na prostej $\mathbb{R}$ całkę na prostej $\mathbb{R}$, którą będziemy oznaczać

$$\int_{-\infty}^{\infty} f(x)dx,$$

definiujemy jako sumę:

$$\int_{-\infty}^{a} f(x)dx + \int_{a}^{\infty} f(x)dx, \tag{4}$$

gdzie a jest dowolną liczbą rzeczywistą; całka na prostej funkcji f istnieje, jeżeli obie całki w sumie (4) są określone.

Zmienne losowe typu ciągłego

Definicja 11

Mówimy, że zmienna losowa X jest typu ciągłego, jeśli istnieje nieujemna funkcja g taka, że dla każdych $-\infty \leqslant a < b \leqslant \infty$

$$P(a \leqslant X \leqslant b) = \int_a^b g(x)dx.$$

Funkcja g — to tzw. funkcja gęstości zmiennej losowej X (lub gęstość rozkładu zmiennej losowej X).

Rozkład jednostajny na odcinku $[0, 1]$

Przykładem zmiennej losowej typu ciągłego jest rozkład jednostajny na odcinku $[0, 1]$ (oznaczenie: $U(0, 1)$). Jego funkcja gęstości u dana jest wzorem

$$u(x) = \begin{cases} 1, & \text{jeśli } 0 \leqslant x \leqslant 1, \\ 0 & \text{jeśli } x < 0 \text{ lub } x > 1. \end{cases}$$

Rozkład ten może opisywać np. czas oczekiwania na autobus A, odjeżdżający do miejscowości B co godzinę, przez pasażera C; zakładamy, że C nie zna rozkładu jazdy dla tej linii i że przychodzi na przystanek w losowym momencie.

Prawdopodobieństwa odpowiadające nierównościom ostrym i słabym

Dla zmiennej losowej X o rozkładzie typu ciągłego mamy:

$$P(a < X < b) = P(a \leqslant X < b) = P(a < X \leqslant b) = P(a \leqslant X \leqslant b).$$

Równość ta wynika z własności całki oznaczonej.

Rozkład jednostajny na odcinku $[0, 1]$ — przykład obliczeń

Czas oczekiwania na autobus — zmienna losowa $Y \sim U(0, 1)$.
Prawdopodobieństwo $P(\frac{1}{3} < Y < \frac{1}{2})$ jest równe:

$$P\Big(\frac{1}{3} < Y < \frac{1}{2}\Big) = \int_{1/3}^{1/2} 1 dx = \frac{1}{6}.$$

Rozkład normalny

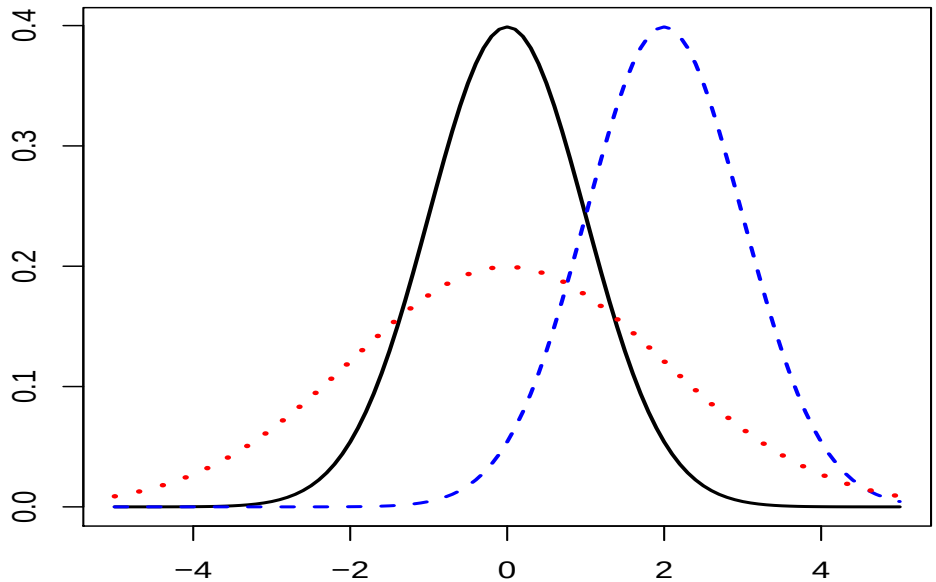
Szczególnie ważnym w zastosowaniach jest rozkład normalny.

Definicja 12

Mówimy, że zmienna losowa X ma rozkład normalny z parametrami μ i σ, gdzie $\mu \in \mathbb{R}$ i $\sigma > 0$, jeżeli gęstość jej rozkładu jest określona wzorem:

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

Skrótowy zapis: $X \sim N(\mu, \sigma)$. Dla $\mu = 0$ i $\sigma = 1$ będziemy pisać zamiast $\phi_{0,1}(x)$ krótko $\phi(x)$.



Rysunek: Wykresy gęstości rozkładów normalnych: $N(0,1)$ (linia ciągła), $N(0,2)$ (linia „kropkowana”), $N(2,1)$ (linia „kreskowana”).

Rozkład normalny— zastosowania

Wiele cech (zmiennych losowych) w życiu gospodarczym i w świecie przyrody ma rozkład zbliżony do normalnego. Wynika to z tzw. centralnego twierdzenia granicznego, z którego wynika, że średnia $\frac{1}{n}(X_1 + X_2 + \ldots + X_n)$, gdzie $X_1, X_2, \ldots, X_n$ są niezależnymi zmiennymi losowymi o tym samym rozkładzie, ma rozkład zbliżony do normalnego $N(\mu, \sigma)$ dla pewnych μ i σ (przy pewnych założeniach, które zazwyczaj są spełnione). Dokładniejsze sformułowanie tego twierdzenia wymaga określenia wartości oczekiwanej i wariancji zmiennej losowej, por. [KM01, str. 141].

Standardowy rozkład normalnym $N(0,1)$ — obliczanie prawdopodobieństw zdarzeń

Dla $a < b$ prawdopodobieństwo $P(a < X < b)$, gdzie $X \sim N(0,1)$ jest równe:

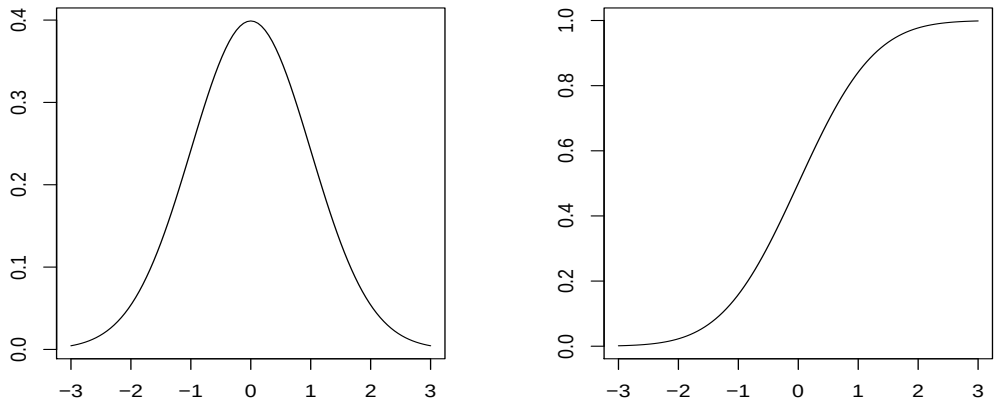
$$P(a < X < b) = \int_a^b \phi(x)dx = \Phi(b) - \Phi(a),$$

gdzie Φ jest określona przez:

$$\Phi(t) = \int_{-\infty}^t \phi(x)dx.$$

Funkcja Φ jest dystrybuantą rozkładu normalnego $N(0,1)$. Funkcji Φ nie da się wyrazić za pomocą skończonej liczby działań na podstawowych funkcjach elementarnych — stąd potrzeba sporządzania tablic statystycznych zawierających wartości funkcji Φ (można je znaleźć w prawie każdym podręczniku statystyki).

Własności funkcji Φ



Rysunek: Wykresy gęstości ϕ rozkładu normalnego (z lewej strony) $N(0,1)$ i dystrybuanty rozkładu normalnego Φ (z prawej strony)

Można pokazać, że $\Phi(0) = 0{,}5$ oraz $\Phi(t) = 1 - \Phi(-t)$ dla dowolnego t; stąd można się ograniczyć do tablicowania funkcji Φ dla $t \geqslant 0$.

Rozkład normalny — obliczanie prawdopodobieństw zdarzeń

Można pokazać, że jeśli $X \sim N(\mu, \sigma)$, to

$$\frac{X-\mu}{\sigma} \sim N(0,1).$$

Stąd dla $a < b$ prawdopodobieństwo $P(a < X < b)$, $X \sim N(\mu, \sigma)$, jest równe:

$$P(a < X < b) = P\Big(\frac{a-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{b-\mu}{\sigma}\Big) = \Phi\Big(\frac{b-\mu}{\sigma}\Big) - \Phi\Big(\frac{a-\mu}{\sigma}\Big).$$

Przykład rachunkowy

Niech X oznacza wzrost dorosłych mężczyzn w państwie A; zakładamy, że $X \sim N(177, 10)$.
Chcemy obliczyć: (a) $P(174 < X < 182)$, (b) $P(X > 182)$.
Obliczenia dla (a):

$$\begin{aligned} P(174 < X < 182) &= \Phi\Big(\frac{182-177}{10}\Big) - \Phi\Big(\frac{174-177}{10}\Big) = \\ &= \Phi(0{,}5) - \Phi(-0{,}3) = \Phi(0{,}5) - (1 - \Phi(0{,}3)) = \\ &= \Phi(0{,}5) + \Phi(0{,}3) - 1 \approx 0{,}6915 + 0{,}6179 - 1 = 0{,}3094. \end{aligned}$$

Obliczenia dla (b) można przeprowadzić w analogiczny sposób, korzystając z równości:

$$P(X > 182) = 1 - P(X < 182) = 1 - \Phi\Big(\frac{182-175}{10}\Big) = 1 - \Phi(0{,}5).$$

Gęstość zmiennej losowej typu ciągłego

Dowolna funkcja g spełniająca warunki:

- dziedziną funkcji g jest zbiór liczb rzeczywistych $\mathbb{R}$,
- $g(x) \geqslant 0$ dla $x \in \mathbb{R}$,
- $\int_{-\infty}^{\infty} g(x) = 1$,

jest funkcją gęstością pewnej zmiennej losowej;
Poza rozkładem normalnym i rozkładem jednostajnym $U(0,1)$ do opisu cech w życiu gospodarczym i naukach przyrodniczych stosuje się wiele innych rozkładów prawdopodobieństwa.

Gęstość emiryczna — przykład

Dane *normtemp*— zebrane w celu weryfikacji hipotezy mówiącej, że średnia wartość temperatury zdrowego człowieka jest równa 98,6 stopni w skali Fahrenheita (37,0 stopni w skali Celsjusza). Zbiór danych zawiera pomiary temperatury i tętna (temperatura wyrażona jest w stopniach Fahrenheita). Można go pobrać z odpowiedniego repozytorium a następnie zapisać do zbioru o nazwie np. **t** (tzw. „data frame"). Zbiór **t** składa się z trzech zmiennych: **temperature**, **gender** i **hr**. Aby uczynić naszą prezentację bardziej czytelną, zmieniamy nazwy zmiennych na odpowiednio: **temp**, **plec** i **tetno**. Odpowiednie polecenia systemu R są zapisane w pliku *t.R*. Wydruk tego pliku zamieszczamy poniżej (na następnym slajdzie):

Pobieranie zbioru danych z repozytorium systemu R

```
library(utils)
install.packages("UsingR",
repo="http://cran.r-project.org")
library(UsingR)
t<-normtemp
names(t)<-c("temp","plec","tetno")
```

Dane *normtemp* są również dostępne na stronie

`http://www.stat.ucla.edu/~rgould/m12s01/ttest.pdf`

System (pakiet) R można pobrać pod adresem
http://r.meteo.uni.wroc.pl/bin/windows/base/

Pobieranie zbioru danych z repozytorium systemu R

```
> names(t)
[1] "temp"  "plec"  "tetno"
> t[1:10,]
temp plec tetno
1   96.3    1    70
2   96.7    1    71
3   96.9    1    74
4   97.0    1    80
5   97.1    1    73
6   97.1    1    75
7   97.1    1    82
8   97.2    1    64
9   97.3    1    69
10 97.4    1    70
> sort(temp)
[1]  96.3  96.4  96.7  96.7  96.8  96.9  97.0  97.1
[9]  97.1  97.1  97.2  97.2  97.2  97.3  97.4  97.4
...
[129] 100.0 100.8
```

Szereg rozdzielczy

Dla zbioru danych liczbowych $\{y_1, y_2 \ldots, y_N\}$ niech:
$MIN1$ oznacza liczbę mniejszą od najmniejszej z liczb $\{y_1, y_2 \ldots, y_N\}$,
$MAX1$ oznacza liczbę większą lub równą od największej z liczb $\{y_1, y_2 \ldots, y_N\}$.
$MIN1$ i $MAX1$ mogą być odpowiednimi zaokrągleniami wartości, odpowiednio, minimalnej i maksymalnej rozważanego zbioru danych, $MIN1 < MAX1$.
Podzielmy odcinek $(MIN1, MAX1]$ na k przedziałów (zwanych klasami) o równej długości:

$$(x_0, x_1], (x_1, x_2], \ldots, (x_{k-1}, x_k],$$

gdzie $x_0 = MIN1$, $x_k = MAX1$. Funkcję przyporządkowującą poszczególnym przedziałom liczbę elementów rozważanego zbioru danych do nich należących będziemy nazywać szeregiem rozdzielczym.
Liczbę klas k w szeregu rozdzielczym wyznaczamy na przykład korzystając z wzorów:

$$k \approx log_2 n + 1 \quad \text{lub} \quad k \approx \frac{3\sqrt{n}}{4}.$$

Szereg rozdzielczy: dane NT

Przyjmujemy: $MIN1 = 96$, $MAX1 = 101$, $k = 10 \approx \frac{3\sqrt{130}}{4}$.
Zakładając, że dane dotyczące temperatury zdrowych ludzi znajdują się w zmiennej `t$temp` konstruujemy szereg rozdzielczy w środowisku R:

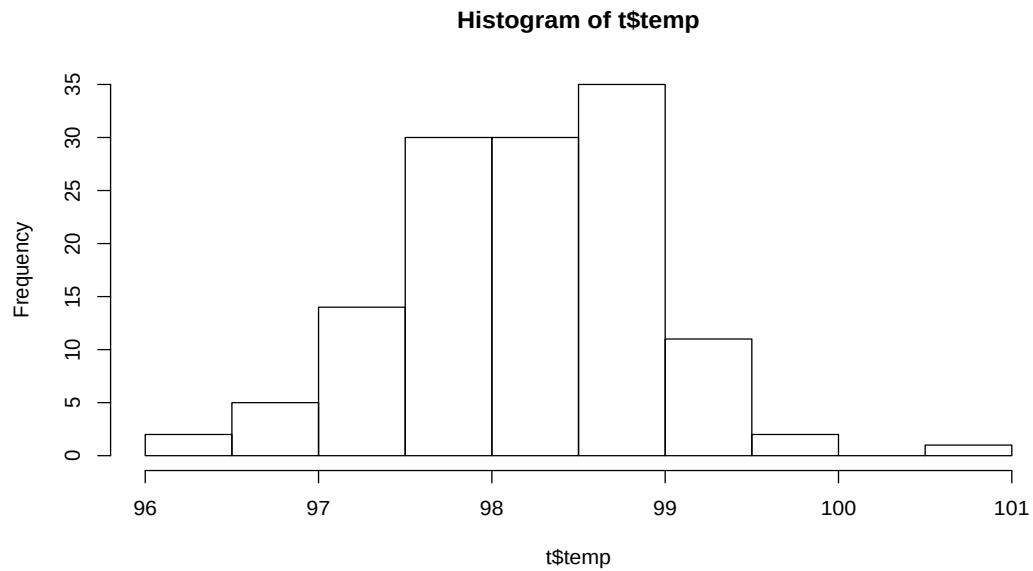
```
> table(cut(t$temp, breaks = c(96,96.5,97,97.5,98,
+ 98.5,99,99.5,100,100.5,101)))

(96,96.5]   (96.5,97]   (97,97.5]   (97.5,98]
2           5           14          30
(98,98.5]   (98.5,99]   (99,99.5]   (99.5,100]
30          35          11          2
(100,100.5] (100.5,101]
0           1
```

Szereg rozdzielczy — przedstawiony w formie tabeli

(96,96.5]	(96.5,97]	(97,97.5]	(97.5,98]	(98,98.5]	(98.5,99]	(99,99.5]	(99.5,100]	(100,100.5]	(100.5,101]
2	5	14	30	30	35	11	2	0	1

Histogram: dane NT

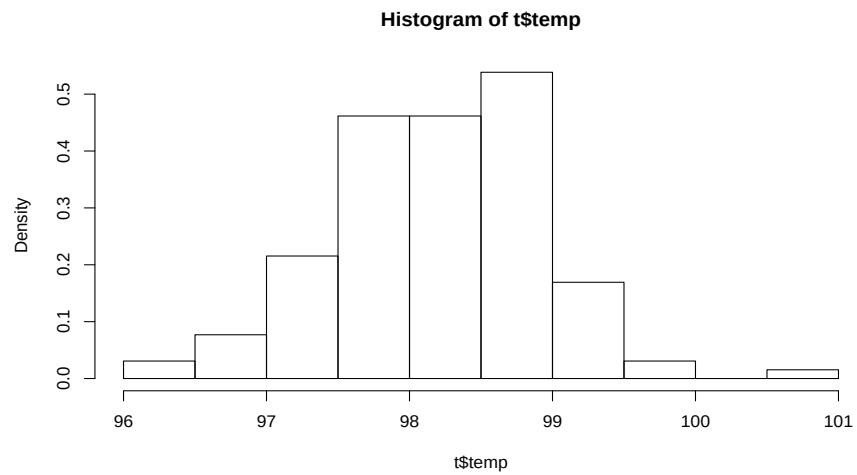
Wykres słupkowy odpowiadający szeregowi rozdzielczemu (podstawami prostokątów — słupków są kolejne klasy, ich wysokości są równe liczebnościom odpowiednich klas): histogram liczebności



Rysunek: Histogram liczebności dla danych NT odpowiadający szeregowi rozdzielczemu z poprzedniego slajdu

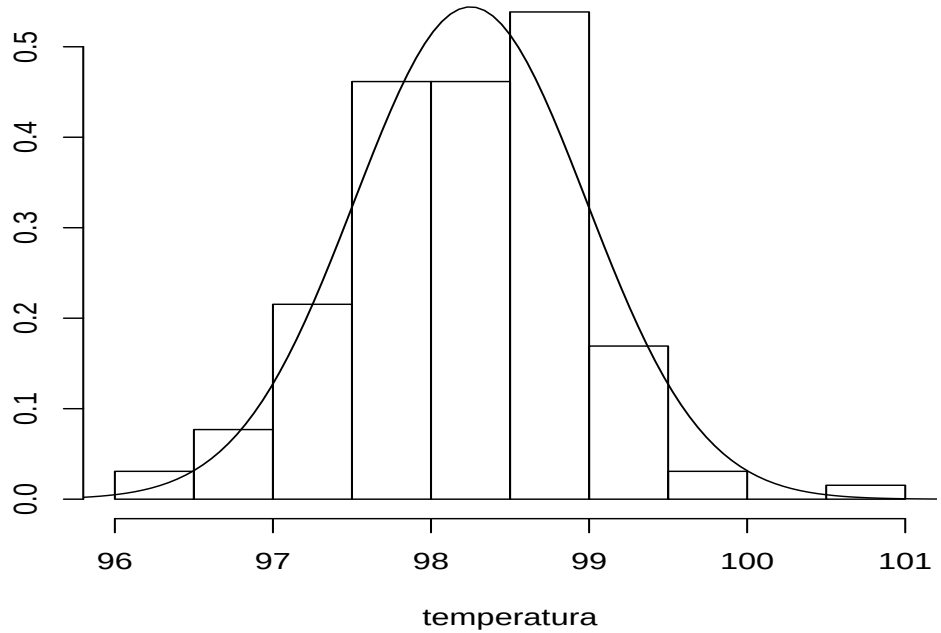
Histogram probabilistyczny i gęstość empiryczna - dane NT

Histogram probabilistyczny: histogram tak wyskalowany, aby "pole pod nim było równe 1":
wyskości słupków: $\frac{2}{130\times 0{,}5}, \frac{5}{130\times 0{,}5}, \ldots;$
histogram probabilistyczny — przez niektórych definiowany jako funkcja przedziałami ciągła (stała), której wykres "pokrywa się" ze zdefiniowanym wyżej wykresem słupkowym; inna nazwa tak określonej funkcji: gęstość empiryczna.



Rysunek: Gęstość empiryczna dla danych NT

Gęstość empiryczna + krzywa normalna



Rysunek: Gęstość empiryczna dla danych NT z dołączonym wykresem gęstości normalnej z parametrami $\mu = 98{,}25$ i $\sigma = 0{,}733$.

Gęstość empiryczna —funkcja przedziałami ciągła

$$h(x) = \begin{cases} \frac{2}{130\times 0{,}5}, & \text{dla } x \in (96; 96{,}5] \\ \frac{5}{130\times 0{,}5}, & \text{dla } x \in (96{,}5, 97] \\ \vdots & \\ \frac{1}{130\times 0{,}5}, & \text{dla } x \in (100{,}5; 101] \\ 0, & \text{dla } x \notin (96; 101] \end{cases}$$

Frakcja obserwacji należących do $(96; 97]$;

$$\int_{96}^{97} h(x)dx = \frac{1}{2}\left(\frac{2}{130 \times 0{,}5} + \frac{5}{130 \times 0{,}5}\right) = \frac{7}{130} \approx 0{,}054.$$

Konstrukcja histogramu probabilistycznego (gęstości empirycznej) — przypadek ogólny

W ogólnym przypadku wysokość k-tego słupka histogramu probabilistycznego (wartość funkcji dla argumentów należących do k-tej klasy) jest równa $\frac{n_k}{nd}$, gdzie

- n_k jest liczebnością k-tej klasy,
- d jest długością klasy,
- n jest liczbą obserwacji.

Funkcja h — „histogram probabilistyczny” (gęstość empiryczna) — zdefiniowana wzorem:

$$h(x) = \begin{cases} \frac{n_k}{nd}, & \text{x należy do } k\text{-tej klasy}, \\ 0, & \text{x nie należy do żadnej z klas.} \end{cases}$$

Dziedziną funkcji h jest zbiór liczb rzeczywistych $\mathbb{R}$.

Funkcja wykładnicza

Definicja 13

Dla liczby dodatniej a różnej od 1 *funkcję*

$$f(x) = a^x, \tag{5}$$

gdzie $x \in \mathbb{R}$*, będziemy nazywać funkcją wykładniczą.*

Fakt 1

Wzór (5) można również zapisać następująco:

$$f(x) = e^{bx}, \quad x \in \mathbb{R}, \tag{6}$$

gdzie $e = 2{,}718...$ *jest granicą ciągu*

$$a_1 = (1+1)^1 = 2, a_2 = (1+1/2)^2 = 2\frac{1}{4}, \ldots,$$

$$a_{100} = (1+\frac{1}{100})^{100} = 2{,}705, \ldots, a_{1000} = (1+\frac{1}{1000})^{1000} = 2{,}717, \ldots$$

a b jest liczbą rzeczywistą różną od 0*.*

Fakt 2

Niech U będzie funkcją wykładniczą. Dla dowolnych $x, y \in \mathbb{R}$

$$U(x+y) = U(x)U(y). \tag{7}$$

Twierdzenie 3 (Feller)

Niech $U(t)$ będzie funkcją określoną dla $t > 0$ i ograniczoną w każdym skończonym przedziale. Jeżeli $U(t)$ spełnia warunek

$$U(t_1 + t_2) = U(t_1)U(t_2), \quad \textit{dla} \quad t_1, t_2 \geqslant 0,$$

wówczas bądź $U(t) = 0$ dla wszystkich t, bądź $U(t) = e^{bt}$ dla pewnego $b \in \mathbb{R}$.

Dowód tego twierdzenia można znaleźć w [Fel80, rodz. XVII, &6].
Rozwiązaniem równania (7) jest albo funkcja wykładnicza albo funkcja tożsamościowo równa 1.

Zmienna losowa o rozkładzie wykładniczym

Mówimy, że zmienna losowa typu ciągłego X ma rozkład wykładniczy z parametrem $\lambda > 0$ jeżeli jej funkcja gęstości g ma postać

$$g(t) = \begin{cases} 0, & t < 0, \\ \lambda e^{-\lambda t}, & t \geqslant 0. \end{cases} \tag{8}$$

Można sprawdzić, że dystrybuanta F_X zmiennej losowej X ma postać

$$F_X(t) = \begin{cases} 1 - e^{-\lambda t}, & t \geqslant 0; \\ 0, & t < 0. \end{cases}$$

Zmienna losowa o rozkładzie wykładniczym: zastosowania

- czas bezawaryjnej pracy pewnych przedmiotów (urządzeń), takich jak np. żarówka;
- czas oczekiwania na rozpad cząsteczki radioaktywnej;
- wartości maksymalne dziennego opadu w danym roku.

Rozkład wykładniczy: „brak pamięci”

Niech X będzie zmienną losową o rozkładzie wykładniczym z parametrem $\lambda > 0$. Dla dowolnych a i b dodatnich prawdziwa jest równość:

$$P(X > a + b | X \geqslant a) = P(X > b).$$

Istotnie

$$P(X > a + b | X > a) = \frac{P(X > a + b) \wedge P(X > a)}{P(X > a)} =$$

$$= \frac{P(X > a + b)}{P(X > a)} = \frac{\exp\left[-\lambda(a + b)\right]}{\exp\left[-\lambda(a)\right]} = \exp\left[-\lambda b\right] = P(X > b).$$

Interpretacja: X jest czasem bezawaryjnej pracy pewnego urządzenia, to niezależnie od dotychczasowego czasu pracy tego urządzenia, dalszy czas pracy ma taki sam rozkład jak „całkowity czas pracy” (czyli rozkład wykładniczy z parametrem λ).

Można uzasadnić (korzystając z Twierdzenia 3, że jeżeli zmienna losowa X spełnia warunek $P(X \geqslant 0) = 1$ i ma rozkład typu ciągłego który ma własność braku pamięci, to ma rozkład wykładniczy.

Zmienna losowa o rozkładzie wykładniczym: przykład zastosowań

Czas życia żarówki X jest zmienną losową o rozkładzie wykładniczym z wartością oczekiwaną równą 10 (jednostką pomiaru jest miesiąc). Chcemy obliczyć:

- prawdopodobieństwo, że żarówka będzie „bezawaryjnie funkcjonować" przez 15 miesięcy;
- kwantyl rzędu 0,95 czasu życia żarówki.

Rozwiązanie Zmienna losowa X ma rozkład wykładniczy z parametrem $\lambda = 1/10$.

$$P(X \geqslant 15) = 1 - \int_0^{15} \frac{1}{10} e^{-x/10} dx = 1 - [-e^{x/10}]_0^{15} = e^{-3/2} \approx 0{,}2231302.$$

Prawdopodobieństwo to można obliczyć wykorzystując pakiet R w następujący sposób:

```
> 1-pexp(3/2)
[1] 0.2231302
```

lub

```
> 1-pexp(15,0.1)
[1] 0.2231302
```

Rozkład logarytmicznie normalny

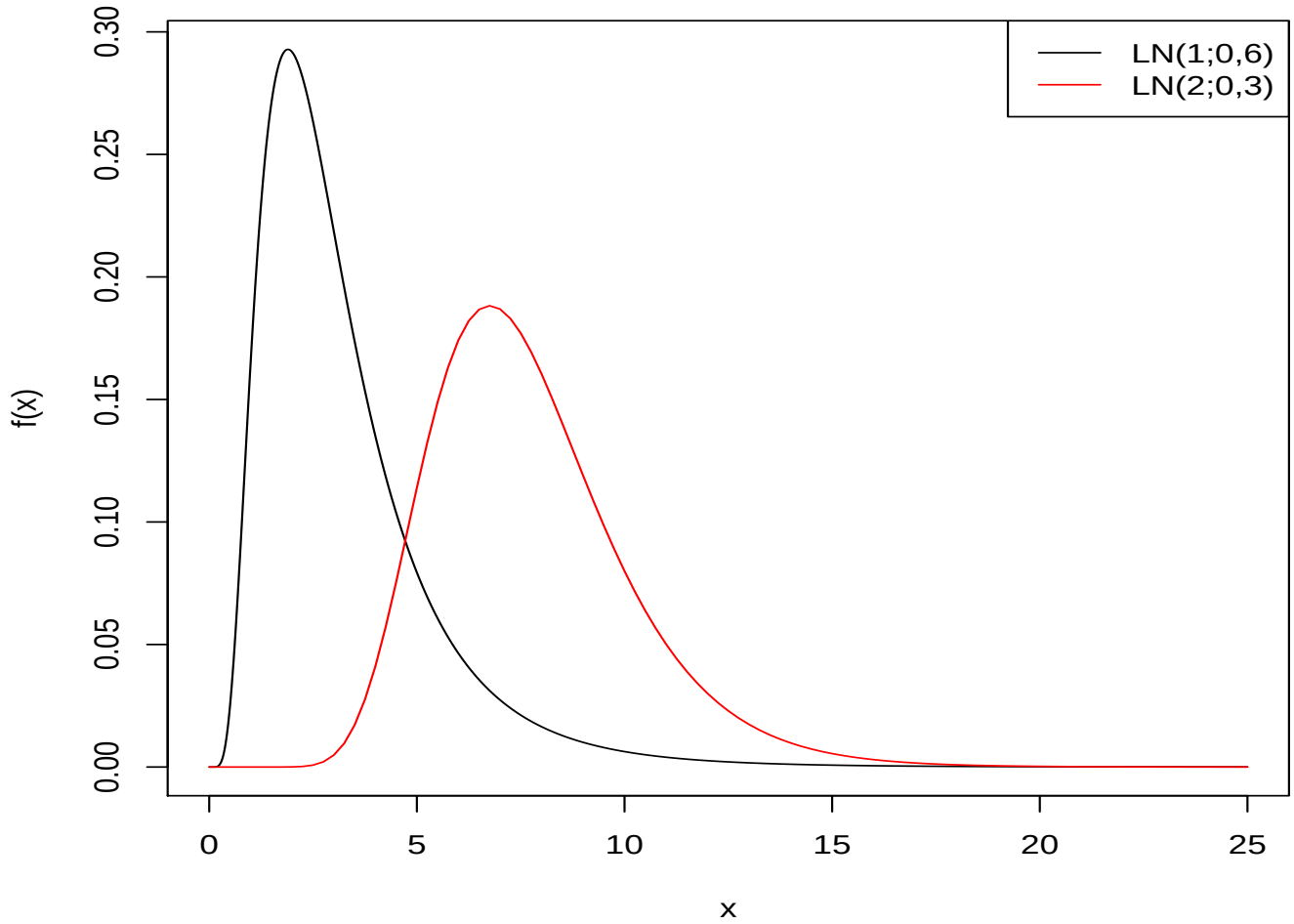
Mówimy, że zmienna losowa X ma rozkład logarytmicznie normalny, jeżeli zmienna losowa $Y = \ln X$ ma rozkład normalny $N(\mu, \sigma)$ dla pewnych $\mu \in \mathbb{R}$ i $\sigma > 0$; fakt ten zapiszemy: $Y \sim LN(\mu, \sigma)$; $\ln a$ oznacza $\log_e a$, gdzie $e = 2{,}718281....$ Wyznaczając gęstość $g_{\mu,\sigma}$ rozkładu zmiennej losowej Y wykorzystujemy fakt:

$$X = e^Y.$$

Można pokazać, że

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

Dużo zastosowań w naukach medycznych i inżynierii środowiska.



Rysunek: Wykresy gęstości rozkładów logarytmicznie normalnych $LN(1;0,6)$ i $LN(2;0,3)$.

Rozkład logarytmicznie normalny — wyjaśnienie intuicyjne

Whereas the normal distribution is the sum/difference of lots of things, the lognormal (because it is the log transform) is the product/quotient of lots of things. So if you are multiplying a bunch of variables together, the resultant distribution approaches lognormal as the number of variables gets large.

Rob Moss, `h`ttps://www.quora.com/What-is-intuition-explanation-of-log-normal-distribution

Rozkład logaytmicznie normalny — przykład zastosowań

R. Makuch, D. Freeman Jr., M. F. Johnson, Justification for the lognormal distribution as a model for blood pressure, Journal of Chronic Diseases, Vol. 32, Issue 3, 1979, Pages 245-250.

Ze streszczenia:

By viewing the cumulative effects on an individual's blood pressure as multifactorial and progressive in nature, this paper provides biological and statistical justification for the apparent lognormal distribution of blood pressure observed in population samples drawn in this country and elsewhere.

Centralne twierdzenie graniczne (wersja dla zmiennych losowych o niekoniecznie identycznych rozkładach) — wyjaśnienie intuicyjne

Roughly speaking, this theorem asserts that if one takes a statistic that is a combination of many independent and randomly fluctuating components, with no one component having a decisive influence on the whole, then that statistic will be approximately distributed according to a law called the normal distribution (or Gaussian distribution).
Terrence Tao, A second draft of a non-technical article on universality, artykuł dostępny pod adresem:
`https://terrytao.wordpress.com/2010/09/14/a-second-draft-of-a-non-technical-article-on-universality/`

Funkcja gamma

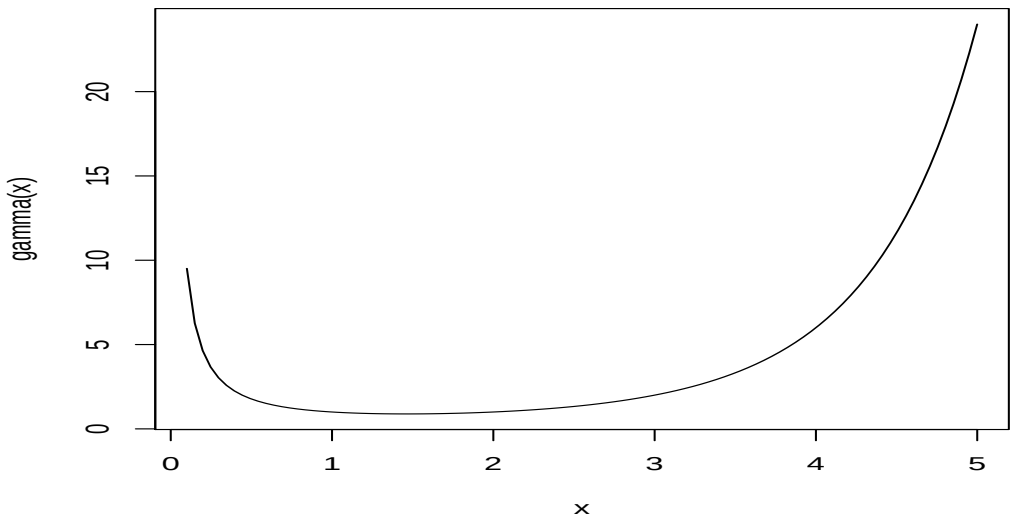
Funkcja Γ jest określona wzorem

$$\Gamma(t) = \int_0^\infty x^{t-1} e^{-x} dx, \quad t > 0;$$

Można uzasadnić, że:

1. $\Gamma(t) > 0$ dla $t > 0$;
2. $\Gamma(1) = 1$;
3. $\Gamma(t+1) = t\Gamma(t)$ dla $t > 0$;
4. $\Gamma(1/2) = \sqrt{\pi}$;
5. $\lim_{t \to 0_+} \Gamma(t) = \infty$;
6. $\lim_{t \to \infty} \Gamma(t) = \infty$;

Funkcja gamma



Rysunek: Wykres funkcji Γ na odcinku $[0{,}1; 5]$.

Rozkład gamma

Definicja 14

Powiemy, że zmienna losowa X typu ciągłego ma rozkład gamma z parametrami α i λ, (co zapisujemy $X \sim \Gamma(\alpha, \lambda)$) gdzie $\lambda, \alpha > 0$, jeżeli jej gęstość ma postać

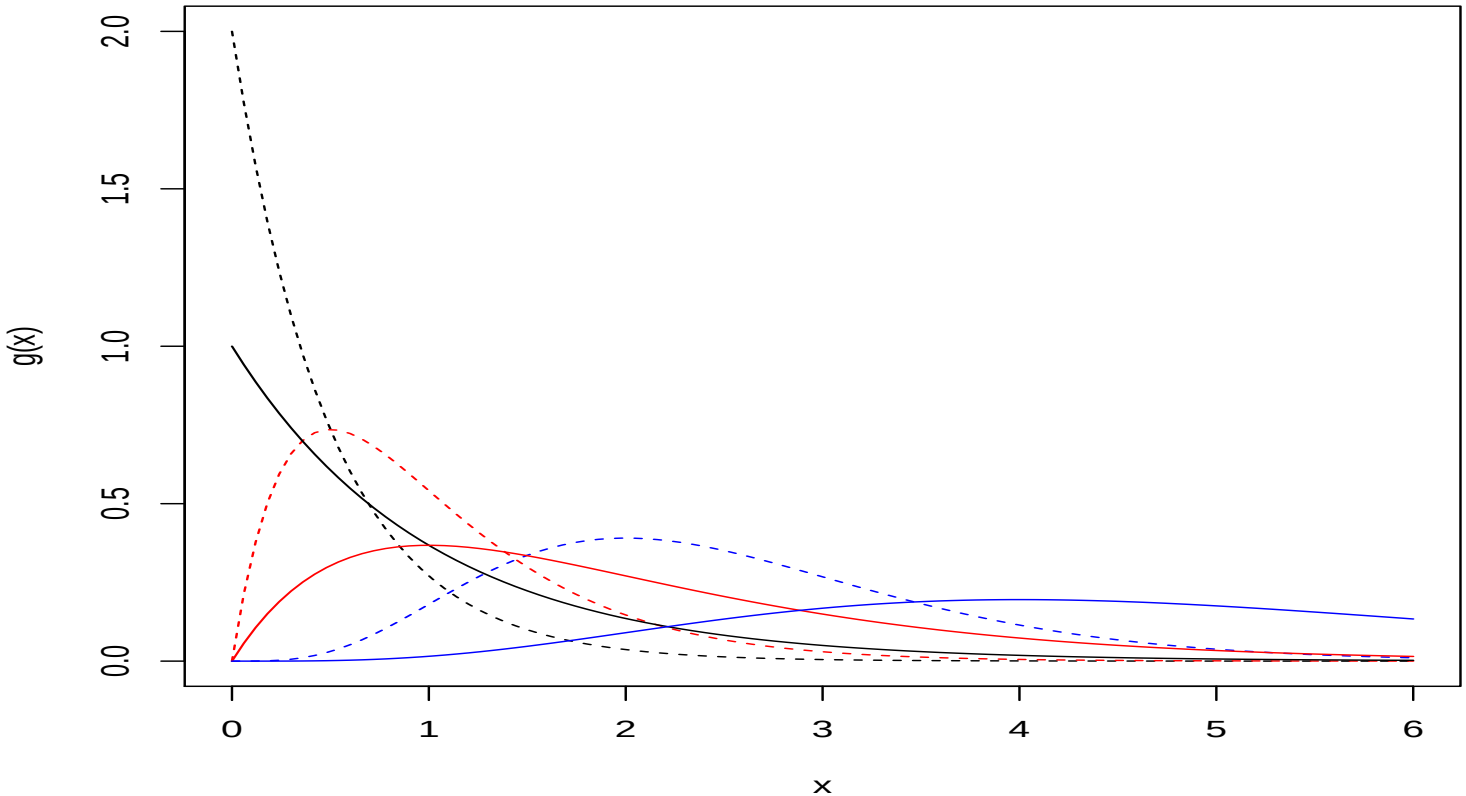
$$g(x; \alpha, \lambda) = \begin{cases} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, & dla \quad x \geqslant 0, \\ 0, \quad dla & x < 0. \end{cases}$$

Uwaga 6

Jeżeli:

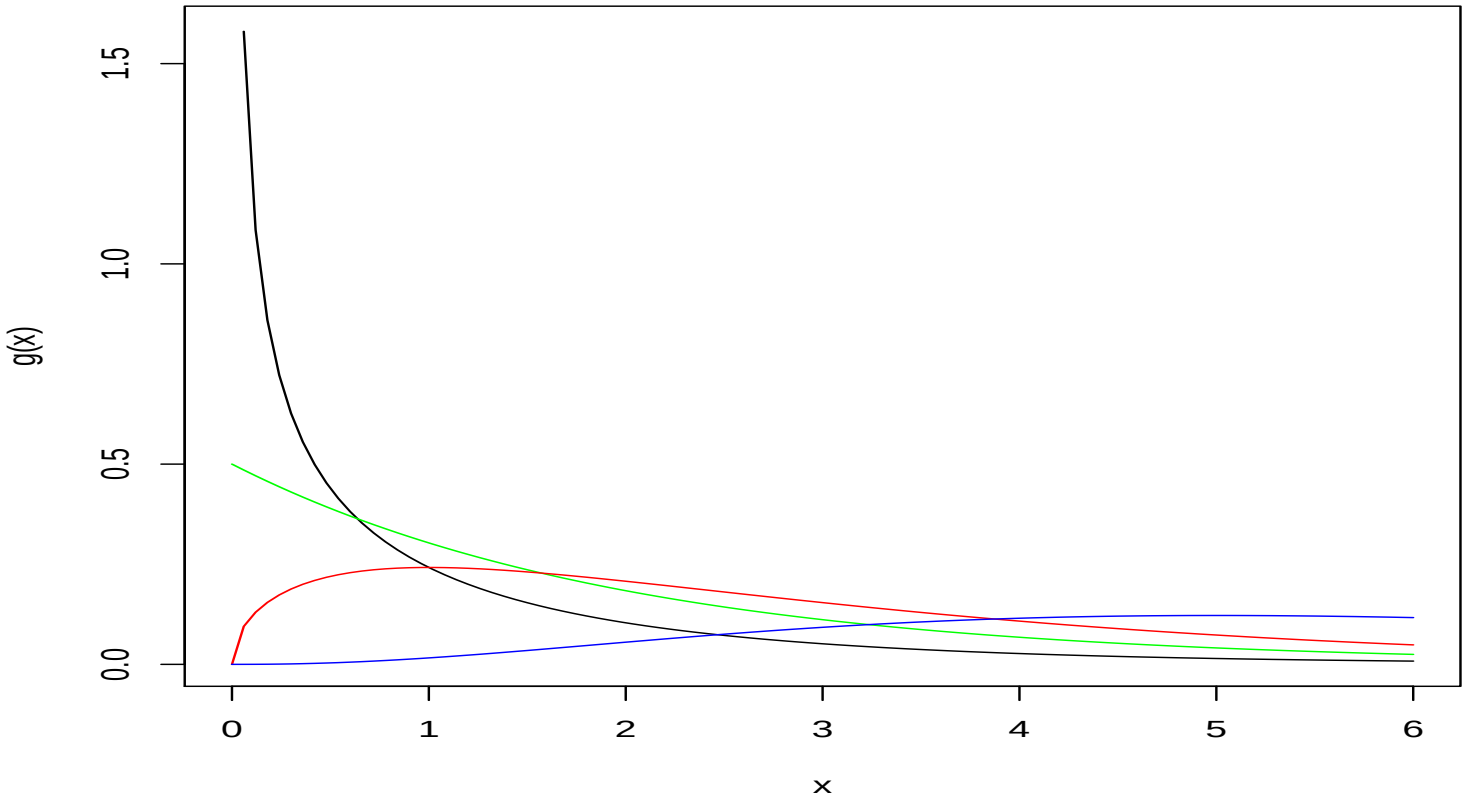
- $\alpha = 1$ *zmienna losowa X ma rozkład wykładniczy z parametrem* λ*;*
- $\alpha = \frac{n}{2}$ *i* $\lambda = \frac{1}{2}$*, gdzie n jest liczbą naturalną, powiemy, że zmienna losowa X ma rozkład* χ^2 *(chi kwadrat) z n stopniami swobody.*

Rozkład gamma — c.d.



Rysunek: Gęstości rozkładów gamma: $\Gamma(1,1)$ (linia czarna), $\Gamma(2,1)$ (linia czerwona ciągła), $\Gamma(5,1)$ (linia niebieska ciągła, $\Gamma(1,2)$ (linia czarna przerywana), $\Gamma(2,2)$ (linia czerwona przerywana), $\Gamma(5,2)$ (linia niebieska przerywana).

Rozkład χ^2



Rysunek: Gęstości rozkładów χ^2: z jednym stopniem swobody (linia czarna), z dwoma stopniami swobody (linia zielona), z trzema stopniami swobody (linia czerwona), z siedmioma stopniami swobody (linia niebieska).

Rozkład wariancji próbkowej

Niech $X_1, X_2, \ldots, X_n$ będą niezależnymi zmiennymi losowymi o (jednakowym) rozkładzie normalnym $N(\mu, \sigma)$ (losową próbą prostą). Zmienna losowa

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

jest nieobciążonym („efektywnym”) estymatorem wariancji σ^2. Zmienna losowa

$$\frac{nS_n^2}{\sigma^2}$$

ma rozkład χ^2 z $(n-1)$ stopniami swobody.

Funkcja kwantylowa — przypadek zmiennych losowych typu ciągłego

Jeśli zmienna losowa typu ciągłego X ma dystrybuantę ciągłą i rosnącą na $\mathbb{R}$ (zbiorem wartości F_X jest $(0,1)$), to funkcja kwantylowa Q_X jest funkcją odwrotną do dystrybuanty F_X.

Jeśli zmienna losowa typu ciągłego X ma dystrybuantę ciągłą i rosnącą na przedziale $\mathcal{I} \subset \mathbb{R}$, zbiorem wartości $F_X|_{\mathcal{I}}$ (zawężenia F_X do $\mathcal{I}$) jest odcinek $(0,1)$, to funkcja kwantylowa Q_X jest funkcją odwrotną do $F_X|_{\mathcal{I}}$;

czyli funkcja Q_X przyporządkowuje argumentowi $t \in (0,1)$ jedyne rozwiązanie równania

$$F_X(u) = t.$$

Pojęcie kwantyla

Dla ustalonej zmiennej losowej (ustalonego rozkładu prawdopodobieństwa) wartość funkcji kwantylowej dla $c \in (0,1)$ nazywamy kwantylem rzędu c. Kwantyl rzędu 0,5 nazywamy medianą, kwantyle rzędu 0,25 i 0,75 nazywamy, odpowiednio, kwartylem dolnym i kwartylem górnym (danego rozkładu).

Uwaga 7

Niektórzy autorzy definiują pojęcie kwantyla rozkładu zmiennej losowej w sposób dopuszczający niejednoznaczność w sytuacji, gdy dystrybuanta tej zmiennej losowej nie jest ciągła lub nie jest ściśle monotoniczna.

Uwaga 8

Istnieje wiele sposobów określania kwantyli próbkowych, których wartości miałyby oceniać wartości kwantyli ustalonych rzędów (w pakiecie R co najmniej 9 sposobów).

Przykład

Dla zmiennej losowej Y o rozkładzie jednostajnym $U(0,1)$, której gęstość g dana jest wzorem

$$g(x) = \begin{cases} 0 & x < 0 \\ 1 & x \in [0,1] \\ 0 & x \in (1,\infty) \end{cases}$$

funkcja kwantylowa Q_Y jest dana wzorem:

$$Q_Y(t) = t, \quad t \in (0,1).$$

Przykład

Dla zmiennej losowej W typu ciągłego, której gęstość h dana jest wzorem

$$h(x) = \begin{cases} 0 & x < 0 \\ x/2 & x \in [0,2] \\ 0 & x \in (2,\infty) \end{cases}$$

funkcja kwantylowa Q_W jest dana wzorem:

$$Q_W(t) = 2\sqrt{t}, \quad t \in (0,1).$$

Zmienna losowa o rozkładzie wykładniczym — znajdowanie wartości kwantyla

Oznaczmy kwantyl rzędu 0,95 dla rozkładu zmiennej X mającej rozkład wykładniczy z parametrem $\lambda = 0{,}1$ przez c. Stała c spełnia jest rozwiązaniem równania:

$$\int_0^c \frac{1}{10} e^{-x/10} dx = 0{,}95.$$

Stąd:

$$\begin{aligned} 1 - e^{-c/10} &= 0{,}95; \\ e^{-c/10} &= 0{,}05; \\ -c/10 &= \ln 0{,}05 = -\log 20; \\ c &= 10 \ln 20 \approx 29{,}95732. \end{aligned}$$

Kwantyl ten można obliczyć przy użyciu R:

```
> qexp(0.95,0.1)
[1] 29.95732
```

Funkcja kwantylowa Q_X ma postać

$$Q_X(t) = -10 \ln(1-t), \quad t \in (0,1).$$

Lektura uzupełniająca

[Bed04] Bednarski, T., Elementy matematyki w naukach ekonomicznych. Oficyna ekonomiczna. Kraków 2004, str. 228–234.

[Fel80] Feller, W., Wstęp do rachunku prawdopodobieństwa, t. 1, PWN 1980.

[JS06] Jakubowski, J., Sztencel, R., Rachunek prawdopodobieństwa dla (prawie) każdego. Wydawnictwo Script. Warszawa 2006.

[KM01] Koronacki, J., Mielniczuk, J. Statystyka dla studentów kierunków technicznych i przyrodniczych. WNT. Warszawa 2001, podrozdział 2.2.1, str. 94–110.

[Kry99] Krysicki, W., Bartos, J., Dyczka, W., Królikowska, K. i Wasilewski, M., Rachunek prawdopodobieństwa i statystyka matematyczna w zadaniach, Cz. 1, Rachunek prawdopodobieństwa, PWN 1999.