Statystyczna analiza danych SAD-2020-2021

Wykład 5

Ciągłe zmienne losowe

<u>Definicja.</u> Zmienną losową ***X*** nazywamy **ciągłą** zmienną losową (zmienną losową typu ciągłego), jeśli istnieje nieujemna funkcja *f*, zwana **gęstością (gęstością prawdopodobieństwa)**, taka że dla dowolnych *a*, *b* , $-\infty \leq a \leq b \leq \infty$,

$$\boxed{P(a \leq X \leq b) = \int_a^b f(x)dx}.$$

Rozkład prawdopodobieństwa c.z.l.

$$P\big(X \in (a,b)\big) = P(X \in [a,b]) = \int_a^b f(x)dx = F(b) - F(a)$$

$$P(X \in (-\infty, x]) = F(x) = \int_{-\infty}^{x} f(t)dt \; =$$

dystrybuanta ciągłej zmiennej losowej

Własności dystrybuanty

$$F(x) = P(X \le x):$$

- $0 \le F(x) \le 1, \qquad x \in (-\infty, \infty)$
- funkcja rosnąca
- funkcja prawostronnie ciągła (ciągła dla przypadku ciągłej z.l.)
- $F(x) - F(x^{-}) = P(X = x)$
- $\lim_{x \to -\infty} F(x) = 0$
- $\lim_{x \to \infty} F(x) = 1$

Gęstości a histogramy unormowane

Niech **x_1, x_2,..., x_n** oznaczają obserwacje cechy **ciągłej** ***X***, otrzymywane niezależnie. Przy nieograniczenie rosnącej liczności próbki n, **łamane częstości histogramów unormowanych** (takich, że suma pól słupków = 1, gdy wysokość słupka = częstość/*h*, *h* = długość przedziału) **zbliżają się do krzywej ciągłej,** nazywanej **krzywą gęstości** lub **gęstością cechy *X***

Gęstości a histogramy unormowane

Gdy liczba przedziałów histogramu wzrasta, wysokości sąsiednich słupków są zbliżone, więc **łamana częstości** staje się coraz bardziej gładka, zbliża się nieograniczenie do pewnej idealnej krzywej ciągłej (**gęstości**). Zatem, dla dużej liczności próbki:

Pole pod krzywą gęstości = 1

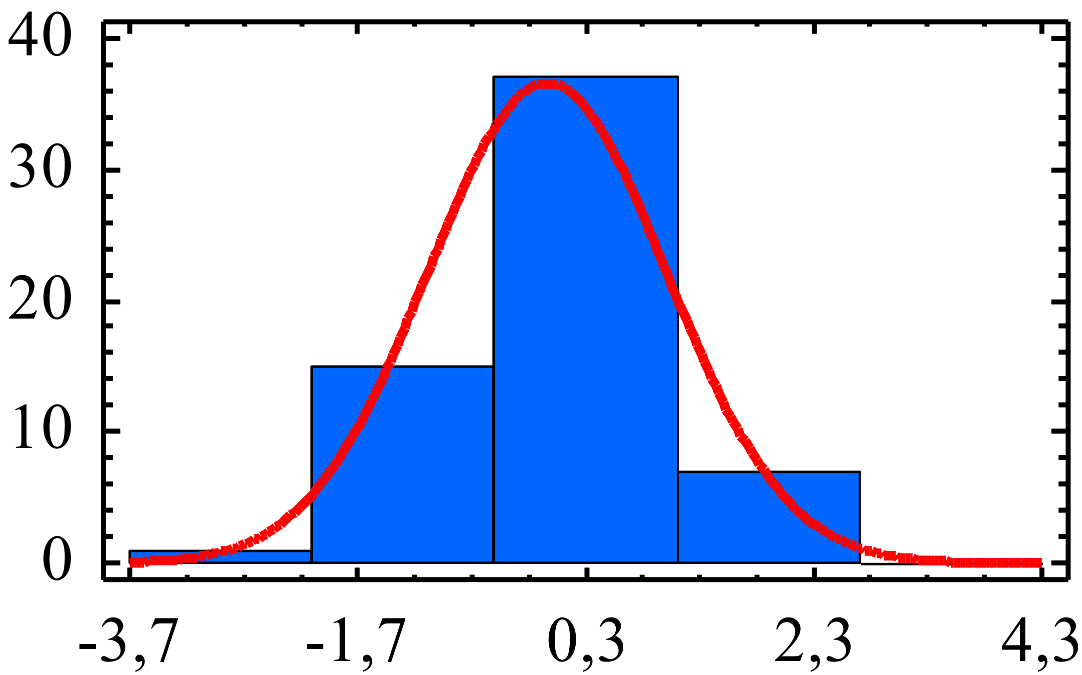
Gęstości a histogramy unormowane

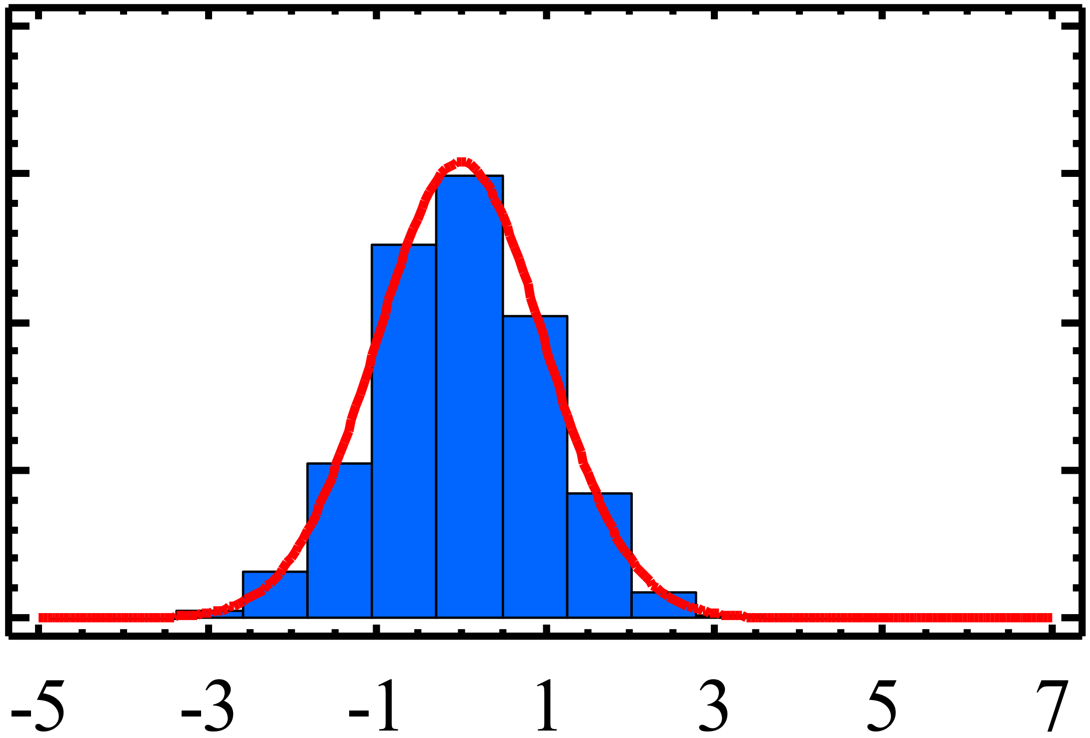
Gdy liczba przedziałów histogramu wzrasta, wysokości sąsiednich słupków są zbliżone, więc **łamana częstości** staje się coraz bardziej gładka, zbliża się nieograniczenie do pewnej idealnej krzywej ciągłej (**gęstości**). Zatem, dla dużej liczności próbki:

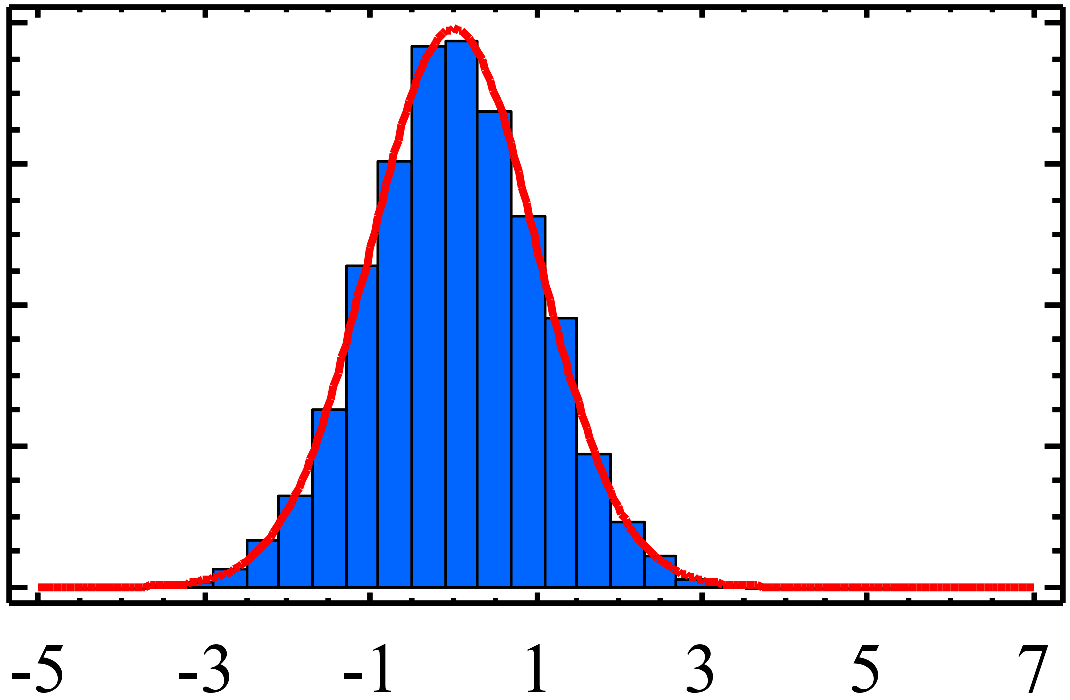
częstość obserwacji w przedziale =

wysokość słupka x *h* = w przybliżeniu

pole pod wykresem gęstości dla tego przedziału.

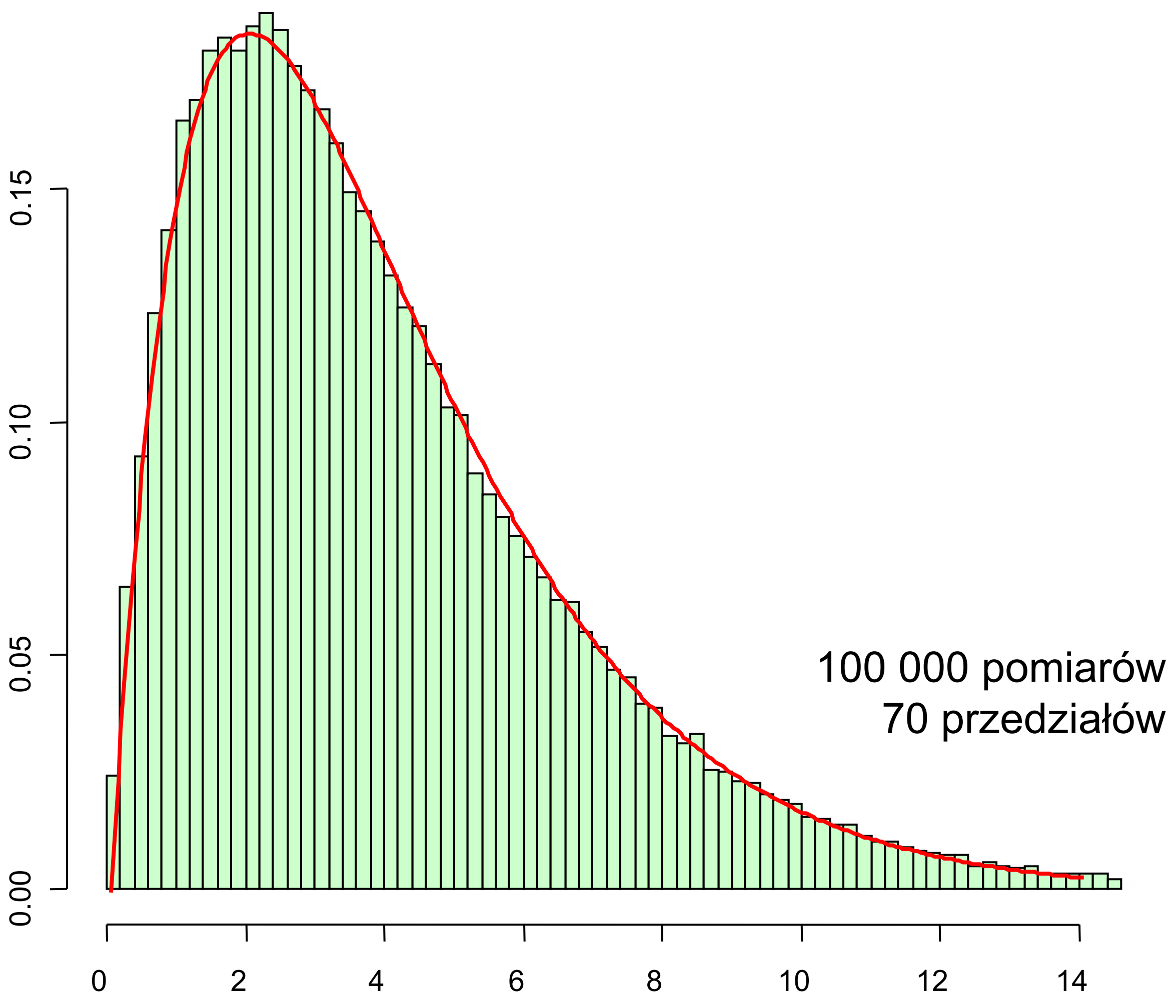

40
30
20
10
0
-3,7
-1,7
0,3
2,3
4,3


-5
-3
-1
1
3
5
7

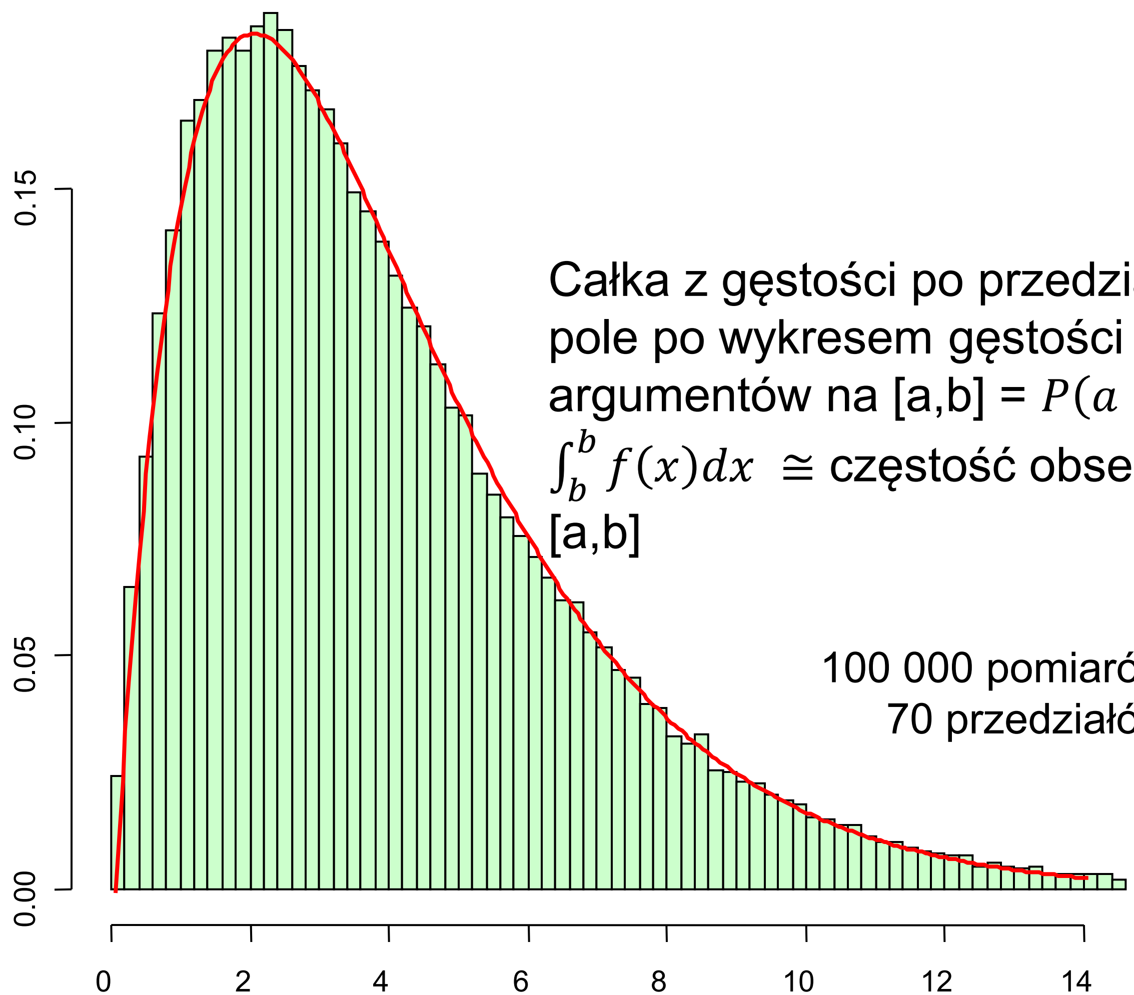


-5
-3
-1
1
3
5
7

Funkcja gęstości rozkładu a histogram unormowany



Funkcja gęstości rozkładu a histogram unormowany

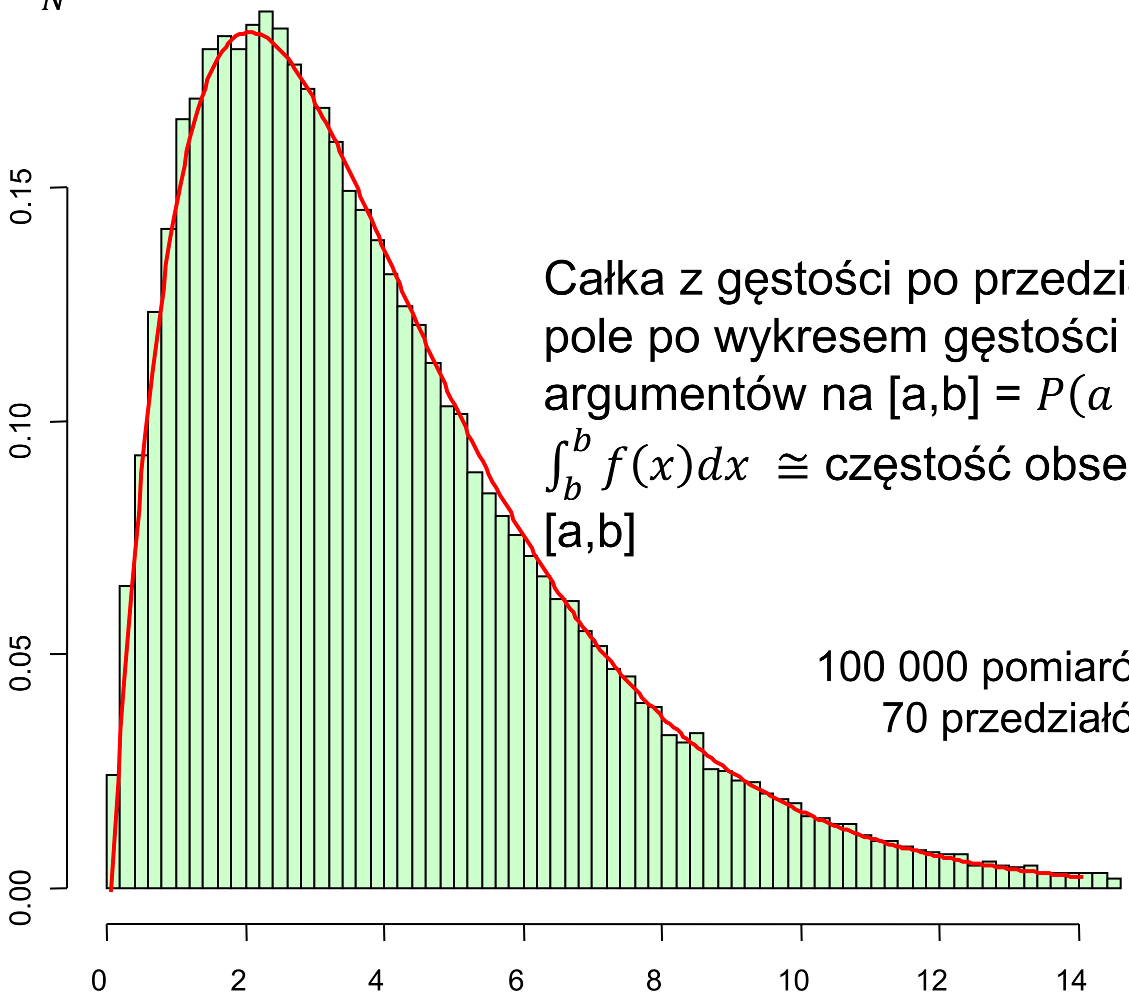


Całka z gęstości po przedziale [a,b] = pole po wykresem gęstości i nad osią argumentów na [a,b] = $P(a \leq X \leq b) = \int_b^b f(x)dx$ $\cong$ częstość obserwacji w [a,b]

100 000 pomiarów
70 przedziałów

Funkcja gęstości rozkładu

$\frac{N[a,b]}{N} \to P(a \le X \le b)$ gdy $N \to \infty$



Całka z gęstości po przedziale [a,b] = pole po wykresem gęstości i nad osią argumentów na [a,b] = $P(a \le X \le b) = \int_b^b f(x)dx$ ≅ częstość obserwacji w [a,b]

100 000 pomiarów
70 przedziałów

Przykład rozkładu c.z.l.

Przykład. Błąd przyrządu pomiarowego (w cm) jest zmienną losową *X* typu ciągłego, której gęstość określona jest wzorem

$$f(x) = \begin{cases} 0 & dla \quad x < -1 \;\; lub \;\; x \geq 1 \\ x + b & dla \quad -1 \leq x < 0 \\ -x + b & dla \quad 0 \leq x < 1 \end{cases}$$

Wyznaczyć

a) stałą *b*   (b) prawdopodobieństwo, że wartość bezwględna błędu nie przekroczy 0,5 cm. (c) Jaki procent niezależnych pomiarów ma błąd nie większy niż -0,5 cm.

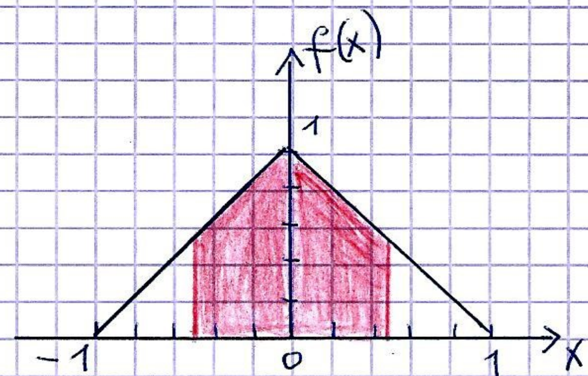
(d) dystrybuantę $F(0)$

(e) stałą c taką, że

1) $P(X \leq c) = 0{,}1$

2) $P(X \leq c) = 0{,}25$

3) $P(X \geq c) = 0{,}75$



$$f(x)=\begin{cases} 0, & x<-1 \text{ lub } x>1 \\ x+1, & -1\le x<0 \\ -x+1, & 0\le x\le 1 \end{cases}$$

(c) $P(X \le -0{,}5) = \frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2} = \frac{1}{8}$ = pole lewego białego trójkąta

(a) $b = 1$, bo pole pod wykresem gęstości = suma pól dwu trójkątów o polach ½.

Przykład rozkładu ciągłego

(b) $P(|X| \leq 0{,}5) = P(-0{,}5 \leq X \leq 0{,}5) = 2P(-0{,}5 \leq X \leq 0)$

$$= 2\int_{-0{,}5}^{0}(x+1)dx = 2\left(\int_{-0{,}5}^{0} xdx + \int_{-0{,}5}^{0} 1dx\right) =$$

$$= 2\left(\frac{x^2}{2}\bigg|_{-0{,}5}^{0} + x|_{-0{,}5}^{0}\right) = 2\left(\frac{0^2}{2} - \frac{(-0{,}5)^2}{2} + (0-(-0{,}5))\right)$$

$$= 0{,}25 + 0{,}5 = 0{,}75$$

Przykład rozkładu ciągłego

d) dystrybuantę $F(0)$

$$F(0) = P(X \leq 0) = 0{,}5$$

(e) stałą c taką, że

1) $P(X \leq c) = 0{,}1 \iff \int_{-1}^{c} (x+1)dx = 0{,}1, \quad (c < 0)$

$$\int_{-1}^{c} x dx + \int_{-1}^{c} dx = \frac{x^2}{2}\bigg|_{-1}^{c} + \big(c - (-1)\big) = \frac{c^2}{2} - \frac{1}{2} + c + 1 = 0{,}1$$

$(c+1)^2 = 0{,}2 \iff c = \sqrt{0{,}2} - 1.$

2) $P(X \leq c) = 0{,}25$ 3) $P(X \geq c) = 0{,}75$

Z definicji kwantyli: 1) $c = q_{0{,}1}$ 2) $c = q_{0{,}25}$ 3) $c = q_{0{,}25}$

Charakterystyki liczbowe zmiennych losowych

Wskaźniki położenia i rozproszenia dla ciągłych zmiennych losowych

<u>**Definicja.**</u> **Wartością średnią** (**oczekiwaną**) ciągłej zmiennej losowej ***X*** mającej gęstość *f* nazywamy liczbę

$$\boxed{\mu_X = \int_{-\infty}^{\infty} s f(s) ds}$$

Wartość średnia zmiennej losowej

Definicja. Niech X będzie ciągłą zmienną losową o gęstości f, a h funkcją określoną na zbiorze wartości ***X***. Wówczas **wartością oczekiwaną** (**średnią**) zmiennej losowej $Y = h(X)$ nazywamy liczbę

$$\boxed{\mu_Y = \int_{-\infty}^{\infty} h(s) f(s) ds}.$$

(jeśli całka istnieje).

<u>Definicja</u>. **Wariancją** ciągłej zmiennej losowej *X* o gęstości *f* nazywamy liczbę

$$\boxed{\sigma_X^2 = \int_{-\infty}^{\infty} (s-\mu_X)^2 f(s)\,ds}$$

Odchylenie standardowe:

$$\boxed{\sigma_X = \sqrt{\sigma_X^2}}.$$

<u>Uwaga.</u> Z definicji wariancji oraz wartości oczekiwanej funkcji zmiennej losowej

$$\boxed{\sigma_X^2 = E(X-\mu_X)^2}$$

Własności wartości średniej i wariancji

Twierdzenie. Jeśli ciągła zmienna losowa ma wariancję, to dla dowolnych liczb *a*, *b* zachodzą wzory

- $$\mu_{aX+b} = a\mu_X + b$$
- $$\sigma^2_{aX+b} = a^2\sigma^2_X$$
- $$\sigma^2_X = \mu_{X^2} - (\mu_X)^2.$$

Kwantyle zmiennej losowej

Dla próbki o dużej liczności i histogramu unormowanego:

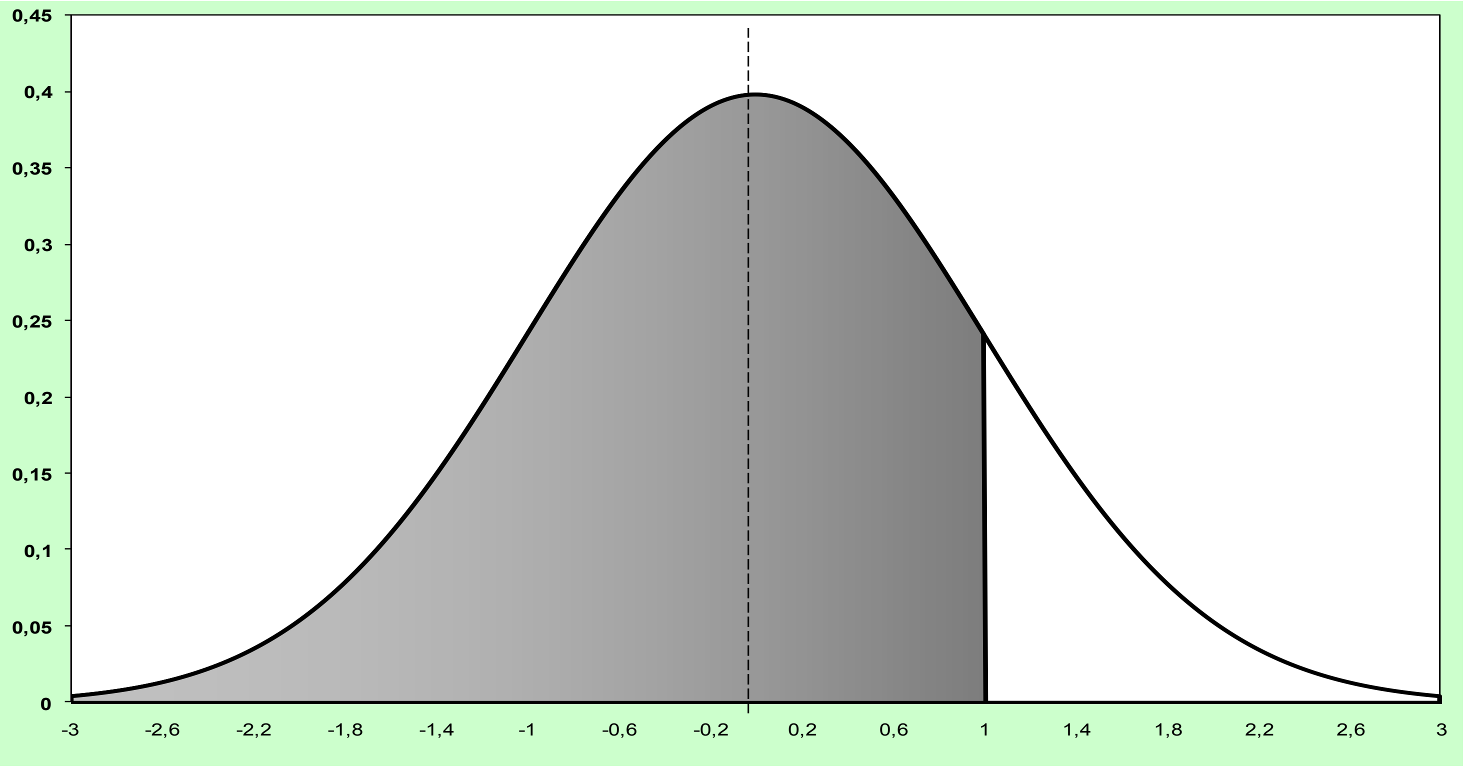
Częstość obserwacji $\leq$ *q* = prawy koniec j-tego przedziału $\approx$

$$\boxed{\sum_{i=1}^{j} \frac{n_i}{n} = \sum_{i=1}^{j} \frac{n_i}{nh} \times h}$$

$\approx$ pole pod wykresem gęstości *f(x)* dla $x \leq q$ =

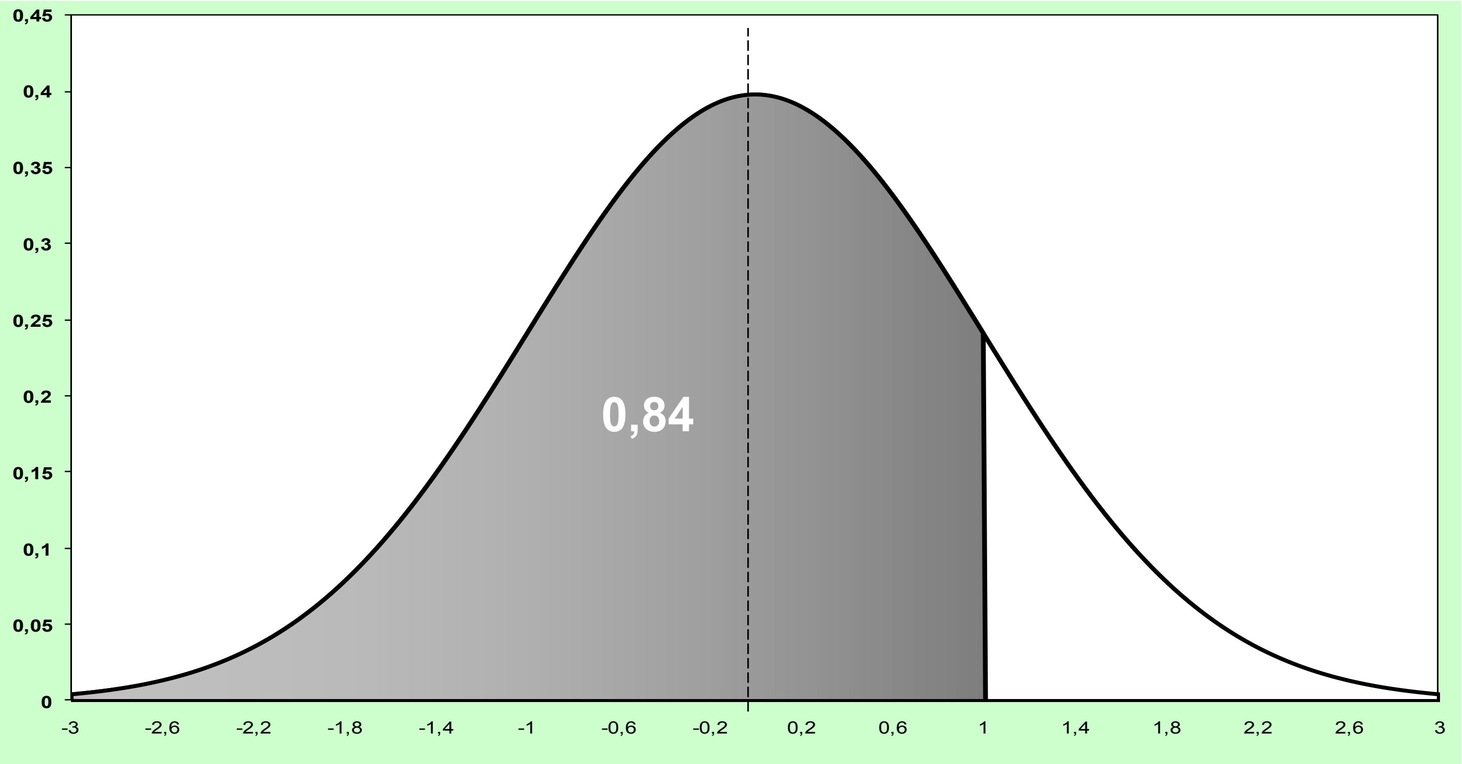
$$\boxed{\int_{-\infty}^{q} f(x)dx}$$

Kwantyle zmiennej losowej



Definicja. Niech $0 < p < 1$.

Kwantylem rzędu p nazywamy punkt q_p na osi poziomej, taki że pole pod gęstością na lewo od niego wynosi p



Pole zacieniowane = 0,84. Zatem kwantyl rzędu 0,84 = 1.

Kwantyle ciągłej zmiennej losowej

Przykład. Czas obsługi klienta w pewnej sieci masowej obsługi jest zmienną losową o rozkładzie wykładniczym. Średni czas obsługi wynosi 0,5 godziny.

(a) $q_{0,75} =?\quad F(q_{0,75}) = 0{,}75 \iff 1 - e^{-\lambda q_{0,75}} = 0{,}75$

gdzie $E(X) = \frac{1}{\lambda} = 0{,}5,$ stąd $\lambda = 2$

$$e^{-2q_{0,75}} = 0{,}25 \iff -2\, q_{0,75} = \ln 0{,}25 = -\ln 4,$$

$$q_{0,75} = \frac{1}{2}\ln 4 = \ln 2 \cong 0{,}6931$$

Interpretacja górnego kwartyla: 75% klientów jest obsługiwanych w czasie krótszym niż 0,6931 godz.

(b) Ile co najmniej czasu trwa obsługa 25% najdłużej obsługiwanych klientów: Czas obsługi tych klientów $\geq q_{0,75} = 0{,}6931$

Parametry gęstości (charakterystyki liczbowe rozkładu)

- **Mediana**: $q_{0,5}$
- **Pierwszy kwartyl:** $q_{0,25}$
- **Trzeci kwartyl:** $q_{0,75}$
- **Rozstęp międzykwartylowy:** $q_{0,75} - q_{0,25}$
- **Wartość średnia gęstości :** μ = środek ciężkości obszaru

płaskiego pomiędzy gęstością a osią poziomą:

$$\boxed{\mu = \int_{-\infty}^{\infty} x f(x) dx}$$

Mediana

Liczba $\boldsymbol{q_{0,5}}$, taka że pole pod wykresem gęstości na lewo od mediany wynosi 0,5. Zatem

$$\int_{-\infty}^{q_{0,5}} f(x)dx = \mathbf{0{,}5} = \int_{q_{0,5}}^{\infty} f(x)dx.$$

Charakterystyki liczbowe próbki i z.l.

Parametry próbki		**Parametry gęstości**	
♦ **Wartość średnia:**	$\bar{x}$	**Wartość średnia:**	μ
♦ **Odchylenie standardowe:**	s	**Odchylenie standardowe**	σ
♦ **Pierwszy kwartyl:**	Q_1	**Pierwszy kwartyl:**	$q_{0,25}$
♦ **Mediana:**	x_{med}	**Mediana:**	$q_{0,5}$
♦ **Trzeci kwartyl:**	Q_3	**Trzeci kwartyl:**	$q_{0,75}$

Ciągłe zmienne losowe

Przykłady ciągłych zmiennych losowych

- **Zmienna losowa o rozkładzie normalnym**

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

$$-\infty < x < \infty$$

Własności zmiennej losowej o rozkładzie normalnym

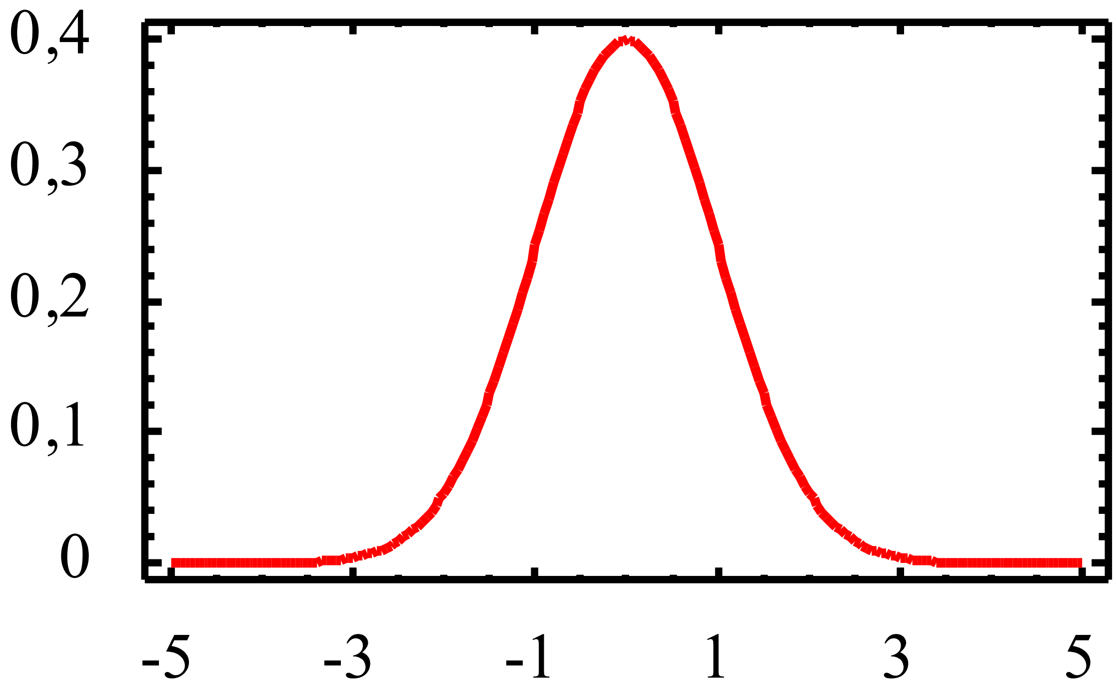
Twierdzenie. Niech

$$X \sim N(\mu, \sigma), \qquad Z = \frac{X - \mu}{\sigma}.$$

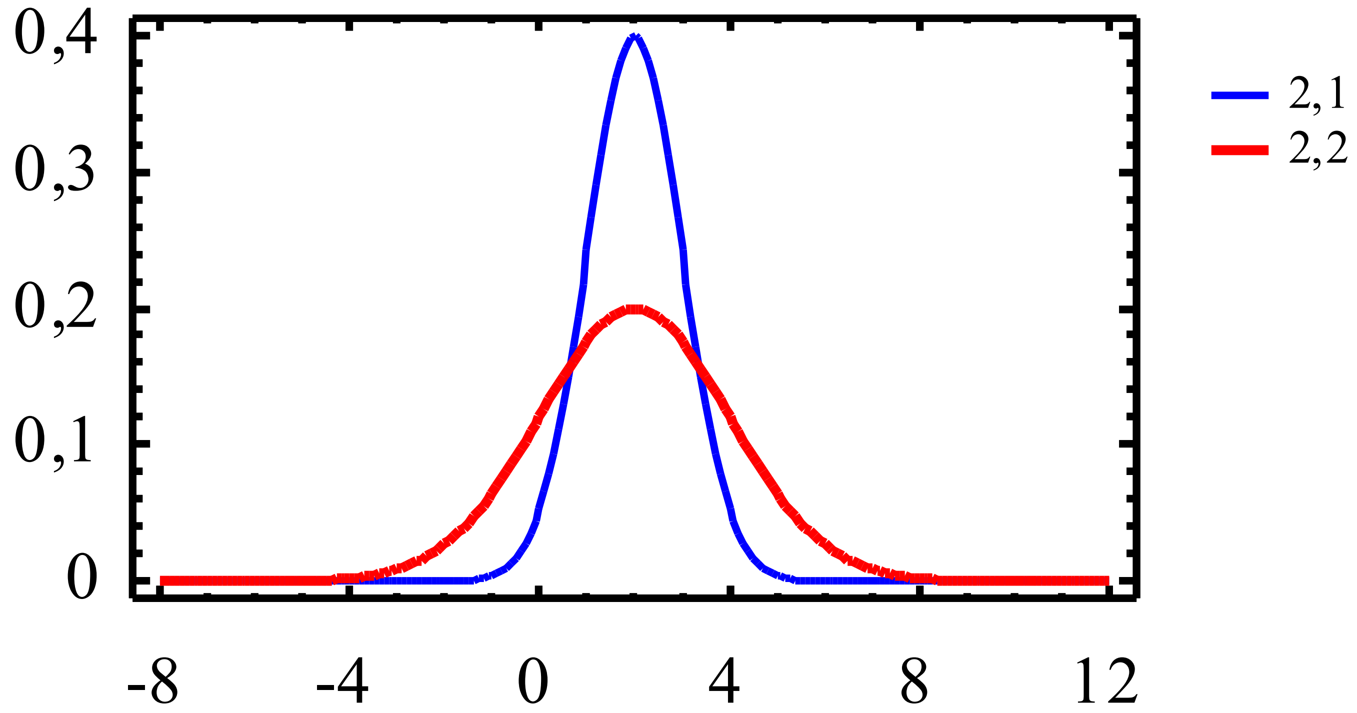
Wówczas

- $Z \sim N(0,1)$
- $\mu_X = \mu, \qquad \sigma_X^2 = \sigma^2$

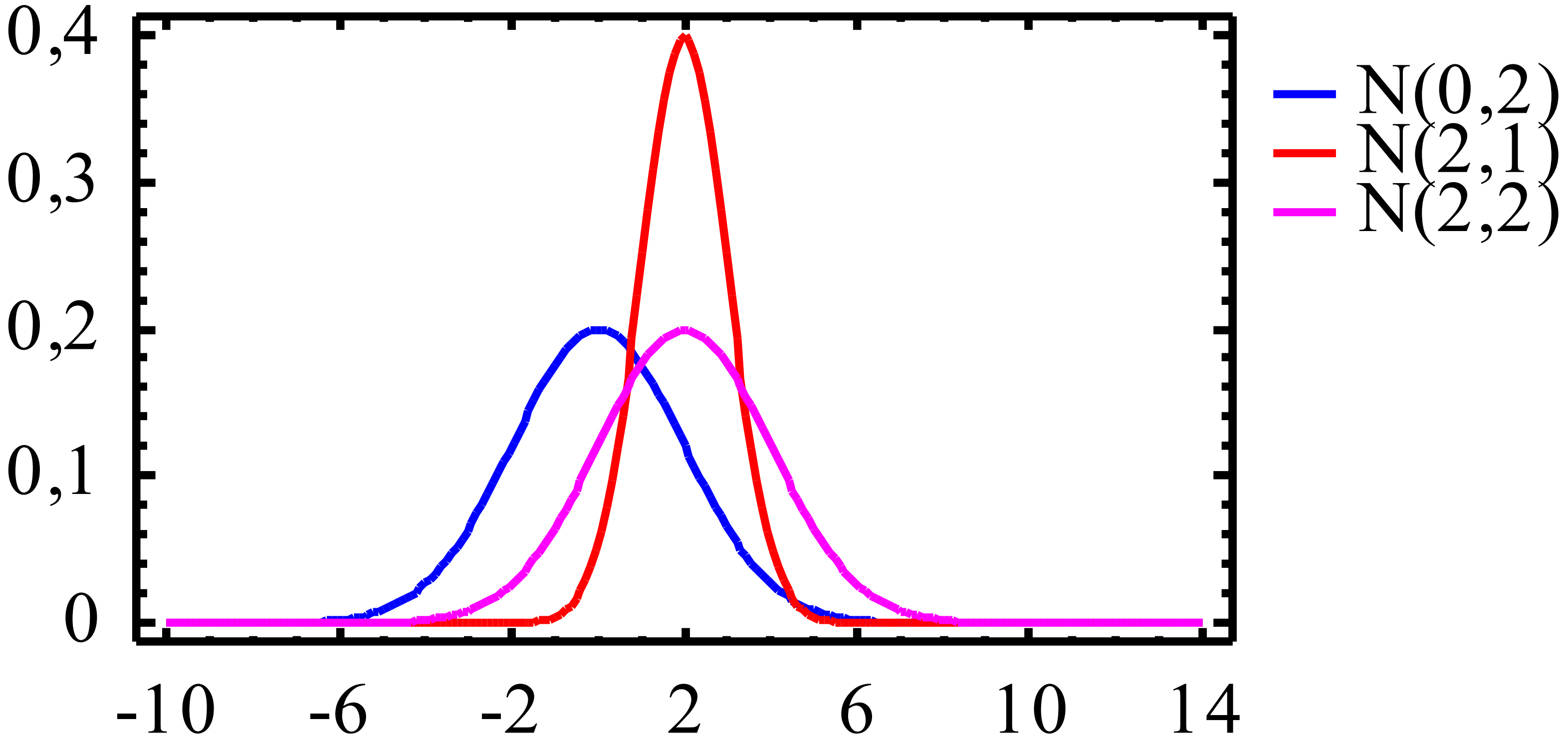
Standardowa gęstość normalna



Wykresy gęstości normalnych



Wykresy gęstości normalnych



Własności rozkładu normalnego

- **Reguła pięciu procent**:

Pole pod wykresem gęstości normalnej o parametrach μ, σ, dla przedziału $(\mu - 2\sigma, \mu + 2\sigma)$ jest równe 0,95. Stąd pole pod wykresem na zewnątrz tego odcinka wynosi 0,05.

Interpretacja. Przy dużej liczności próbki, jeśli cecha ma rozkład normalny, to częstość obserwacji w (μ - 2σ, μ + 2σ)

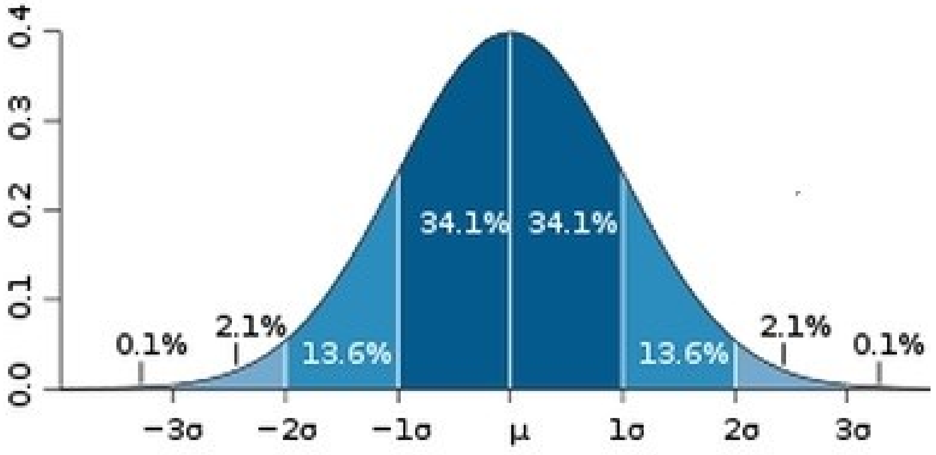
$\approx$ 0,95: **95% elementów w (μ - 2σ, μ + 2σ) .**

- **Prawo trzech sigm**:

częstość obserwacji w (μ - 3σ, μ + 3σ) $\approx$ 0,9972

Wykres gęstości $N(\mu, \sigma)$

https://www.naukowiec.org/wiedza/statystyka/rozklad-normalny



Własności normalnych zmiennych losowych

Twierdzenie

Jeśli $X \sim N(\mu, \sigma)$, $Y = aX + b$, to

$$Y \sim N(a\mu + b, \sqrt{a^2\sigma^2}).$$

Moda zmiennej losowej *X*

Definicja. **Modą** zmiennej losowej *X* nazywamy dowolne **maksimum lokalne**

- funkcji prawdopodobieństwa $p(\cdot)$ zmiennej *X*, gdy zmienna jest dyskretna,
- funkcji gęstości $f(\cdot)$ zmiennej *X*, gdy zmienna jest ciągła.

Uwaga. Moda może nie istnieć.

Moda rozkładu Poissona

Przykład. Liczba awarii sieci informatycznej w ciągu miesiąca jest zmienną losową X o rozkładzie Poissona $P(2)$. Znaleźć najbardziej prawdopodobną liczbę awarii w ciągu miesiąca.

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

$$\frac{p_{k+1}}{p_k} = \frac{e^{-\lambda}\lambda^{k+1}k!}{(k+1)!e^{-\lambda}\lambda^k} = \frac{\lambda}{k+1}, \quad k = 0,1,2,..$$

Moda rozkładu Poissona

Dla $\lambda = 2$:

$\frac{p_1}{p_0} = 2 > 1$, $\frac{p_2}{p_1} = 1$, $\frac{p_{k+1}}{p_k} < 1$ dla k = 2, 3, ...

$p_0 < p_1 = p_2 > p_3 > p_4 > .... > p_k$, stąd

1, 2 = najbardziej prawdopodobne liczby awarii.

Kwantyle zmiennej losowej

Definicja. **Kwantylem rzędu *p*,** (0 < *p* < 1), zmiennej losowej ***X*** o dystrybuancie *F* nazywamy liczbę q_p spełniającą warunki:

$$P(X < q_p) \leq p \leq P(X \leq q_p).$$

Kwantyl $q_{0,5}$ rzędu 0,5 nazywamy **medianą** zmiennej losowej *X*, a $q_{0,25}$ i $q_{0,75}$, odpowiednio, **kwartylem dolnym** i **kwartylem górnym.**

Wskaźniki rozproszenia zmiennej losowej *X*.

- **Wariancja** : $\mathrm{Var}(X) = \sigma_X^2 = E(X - \mu_X)^2$
- **Odchylenie standardowe**: $\sigma_X = \sqrt{Var(X)}$
- **Odchylenie przeciętne**: $d_X = E|X - \mu_X|$
- **Odstęp międzykwartylowy**: $q_{0,75} - q_{0,25}$
- **Współczynnik zmienności**: $v = \sigma_X / \mu_X$

Własności wartości oczekiwanej i wariancji.

- $E(c) = c$, *c* - dowolna stała.
- $E(aX) = aE(X)$, *a* – dowolna stała.
- $E(X + b) = E(X) + b$, *a*, *b* – dowolne stałe.
- $E(X + Y) = E(X) + E(Y)$.
- $X \geq Y \Rightarrow E(X) \geq E(Y)$.
- Var(*X*) nieujemna
- Var(*X*) = 0 wtedy i tylko wtedy, gdy *X* = c = stała

- Var(aX) = a^2 Var(X), a – stała
- Var($X+b$) = Var(X)
- Var(X) = $E(X^2)-(E(X))^2$

DWUWYMIAROWE ZMIENNE LOSOWE

Rozkład łączny pary zmiennych losowych (X, Y)

określonych na tej samej przestrzeni zdarzeń elementarnych:

$P((X, Y) \in A)$, *A* - dowolny podzbiór zbioru par wartości

zmiennych *X*, *Y*.

Definicja. Dystrybuantą zmiennej losowej (X, Y) nazywamy funkcję

$$F(x, y) = P(X \le x, Y \le y),$$

gdzie $-\infty < x < \infty, -\infty < y < \infty.$

Twierdzenie. Łączny rozkład prawdopodobieństwa zmiennej losowej (X,Y) określony jest jednoznacznie przez jej dystrybuantę.

Dyskretne zmienne losowe

Funkcja prawdopodobieństwa (łącznego) dwuwymiarowej zmiennej losowej dyskretnej:

$$f(x,y) = P(X = x, Y = y).$$

Własności:

- $f(x,y) \geq 0$, dla dowolnej pary wartości(x,y) ,
- $\sum_{x} \sum_{y} f(x,y) = 1$,
- $P((X,Y) \in A) = \sum_{(x,y)\in A} f(x,y)$,
- $F(x,y) = \sum_{s\leq x} \sum_{t\leq y} f(s,t)$.

Przykład. W każdym z dwóch etapów teleturnieju można otrzymać 0, 1, lub 2 punkty. Niech zmienne losowe *X*, *Y* oznaczają liczby punktów uzyskane w etapie I i II, odpowiednio, przez losowo wybranego uczestnika. Funkcję prawdopodobieństwa łącznego określa tabela:

Y \ X	0	1	2
0	0,5	0,05	0,01
1	0,2	0,1	0,06
2	0,02	0,03	*A*

Oblicz: stałą *A*, $f(2,2)$,

$P(Y=2)$, $F(1,1)$.

- $\sum_{x=0}^{2}\sum_{y=0}^{2} f(x,y)$ = **1.** Stąd **A = f(2,2) =**

1 – (0,5 + 0,05 + 0,01 + 0,2 + 0,1 + 0,06 + 0,02 + 0,03) =

1 – 0,97 = 0,03.

- $P(Y=2) = \sum_{x=0}^{2} P(X=x, Y=2)$ =

$f(0,2) + f(1,2) + f(2,2)$ = **0,01 + 0,06 + 0,03 = 0,1.**

- $F(1,1)$ = $P(X \le 1, Y \le 1)$ = $f(0,0) + f(0,1) + f(1,0) + f(1,1)$ =

= **0,5 + 0,05 + 0,2 + 0,1 = 0,85.**

Zmienne losowe ciągłe

Dwuwymiarowa zmienna losowa (X,Y) jest ciągłą zmienną losową, jeśli jej **łączny rozkład prawdopodobieństwa** określony jest przez **funkcję gęstości łącznej (łączną gęstość prawdopodobieństwa):**

$$P((X,Y)\in A)=\iint\limits_{A} f(x,y)\,dxdy$$

Dla $A=(-\infty,x]\times(-\infty,y]$:

$$\boxed{F(x,y) = P(X\le x, Y\le y)=\int_{-\infty}^{x}\int_{-\infty}^{y} f(s,t)\,dtds.}$$

Własności gęstości

- $f(x,y) \geq 0$

- $\int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{\infty} f(x,y)\,dx\,dy = 1$

- $f(x,y) = \dfrac{\partial^2}{\partial x \partial y} F(x,y), \quad -\infty < x < \infty, \; -\infty < y < \infty.$

Przykład. Zmienna losowa (X,Y) ma gęstość prawdopodobieństwa

$$f(x,y)=\begin{cases} x+y & \text{gdy} \quad 0\le x\le 1, 0\le y\le 1 \\ 0 & \quad przeciwnie \end{cases}.$$

Obliczyć

$$P(X\le 0{,}5, Y>0{,}25) = \int_0^{0{,}5}\int_{0{,}25}^{1}(x+y)\,dy\,dx =$$

$$\int_0^{0{,}5}(xy+y^2/2)\Big|_{0{,}25}^{1}dx = \int_0^{0{,}5}(x+0{,}5-0{,}25x-0{,}625/2)\,dx =$$

$$\left[0{,}75 \times \frac{x^2}{2} + 0{,}1875 \times x\right]_0^{0,5} = (0{,}1875 + 0{,}1875)/2 = 0{,}1875.$$

Rozkłady brzegowe

Niech (X,Y) będzie dwuwymiarową zmienną losową o rozkładzie prawdopodobieństwa określonym przez funkcję $f(x,y)$ (funkcja prawdopodobieństwa lub gęstość).

Rozkład brzegowy = rozkład prawdopodobieństwa

zmiennej losowej *X* lub zmiennej losowej *Y*.

- dla **dyskretnej** zmiennej (X, Y) , **brzegowe funkcje prawdopodobieństwa** są postaci

$$f_X(x) = P(X = x) = \sum_y f(x, y)$$

$$f_Y(y) = P(Y = y) = \sum_x f(x, y)$$

- dla **ciągłej** zmiennej (X, Y) , **brzegowe gęstości** są postaci

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)dy, \qquad f_Y(y) = \int_{-\infty}^{\infty} f(x, y)dx.$$

D.

- $$f_X(x) = P(X = x) = P(\bigcup_y \{X = x, Y = y\}) =$$

$$\sum_y P(X = x, Y = y) = \sum_y f(x, y).$$

- $$F_X(x) = P(X \le x) = P(X \le x, -\infty < Y < \infty) =$$

$$= \int_{-\infty}^{x} \int_{-\infty}^{\infty} f(s,t)\,dt\,ds.$$

Stąd

$$f_X(x) = \frac{d}{dx} F_X(x) = \int_{-\infty}^{\infty} f(x,t)\,dt$$