Statystyczna analiza danych SAD-2022/23

Wykład 4

Ciągłe zmienne losowe

Definicja. Zmienną losową ***X*** nazywamy **ciągłą** zmienną losową, jeśli istnieje nieujemna funkcja *f*, zwana **gęstością**, taka że dla dowolnych a, b , $-\infty \le a \le b \le \infty$,

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

Dystrybuanta

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

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

dystrybuanta ciągłej zmiennej losowej

Dystrybuanta

<u>Definicja.</u> Funkcję $\boxed{F(x) = \int_{-\infty}^{x} f(s)ds}$, $x \in (-\infty, \infty)$,

nazywamy **dystrybuantą** zmiennej losowej ***X*.**

- $F(x) = P(X \leq x)$, dla każdego *x*
- $P(-\infty \leq X \leq \infty) = P(-\infty < X < \infty) = \int_{-\infty}^{\infty} f(s)ds = 1$
- $f(x) \geq 0$, dla każdego *x*
- $P(X = c) = 0$, dla każdej stałej *c*

Własności dystrybuanty

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

- $0 \le F(x) \le 1, \quad x \in (-\infty, \infty)$
- 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

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***

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:

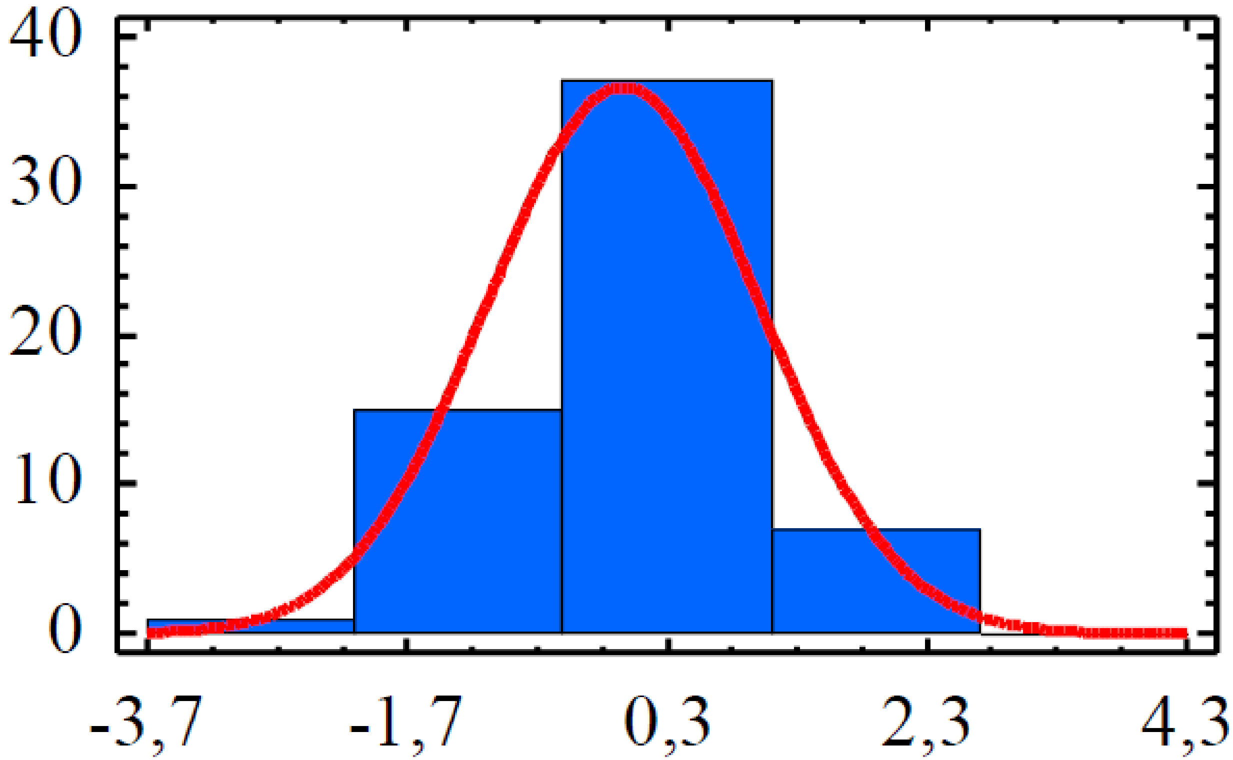
Pole pod krzywą gęstości = 1

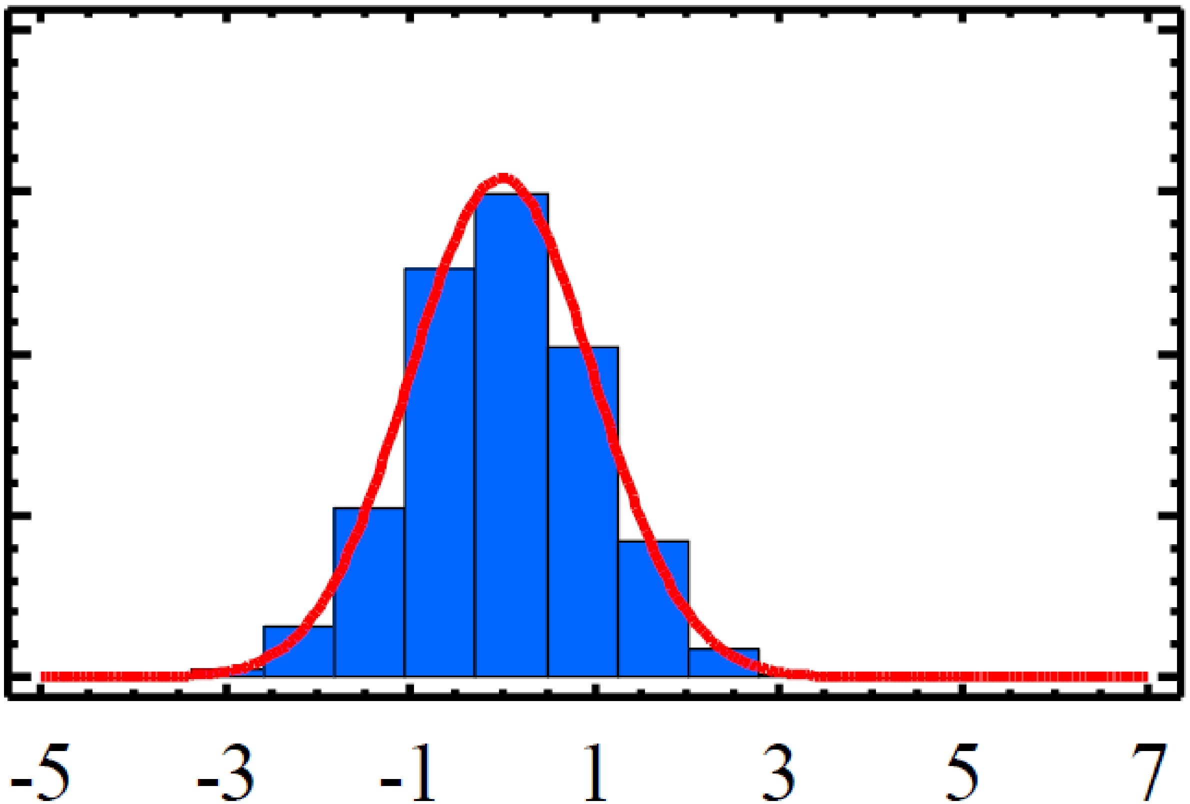
Gdy liczba przedziałów histogramu wzrasta, wysokości sąsiednich słupków są zbliżone, więc **łamana częstości** staje się coraz bardziej gładka, zbliża się nieograniczenie do pewnej idealnej krzywej ciągłej (**gęstości**). Zatem, dla dużej liczności próbki:

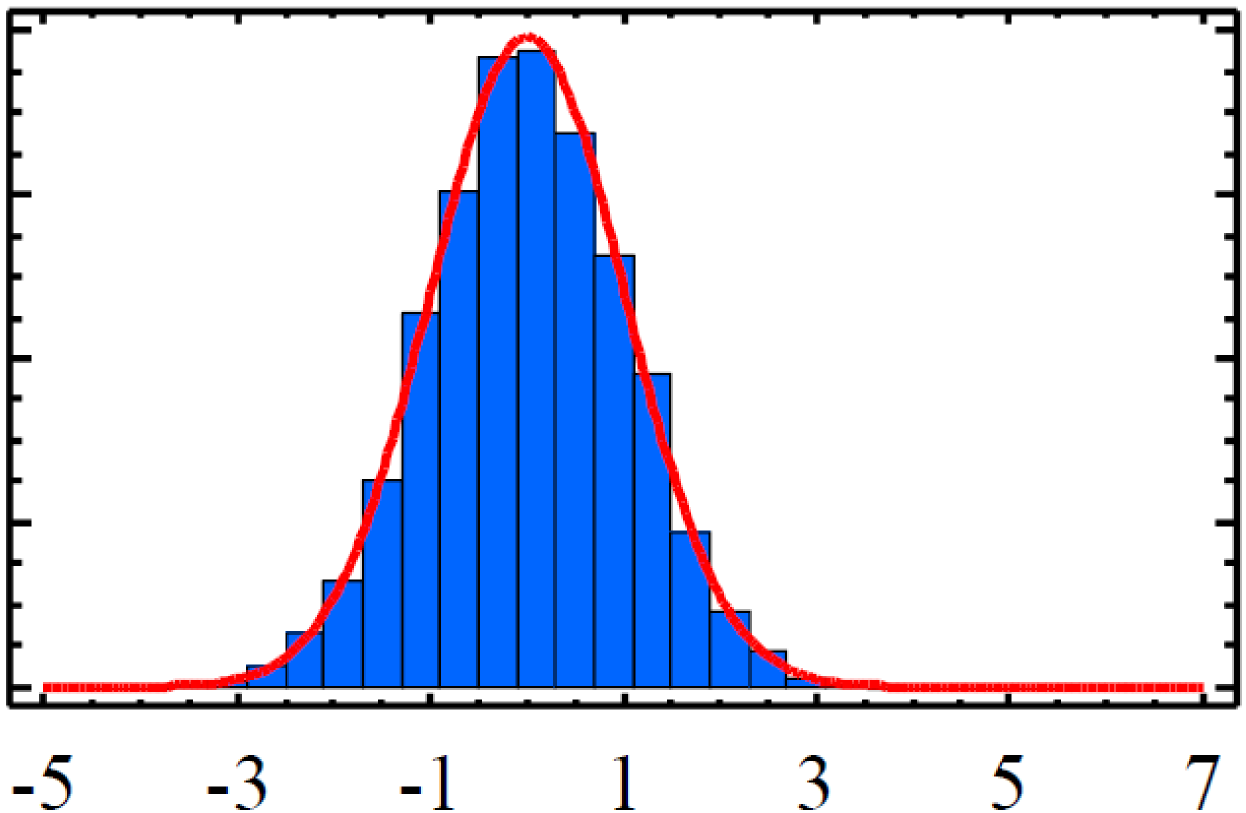
częstość obserwacji w przedziale =

wysokość słupka x *h* = w przybliżeniu

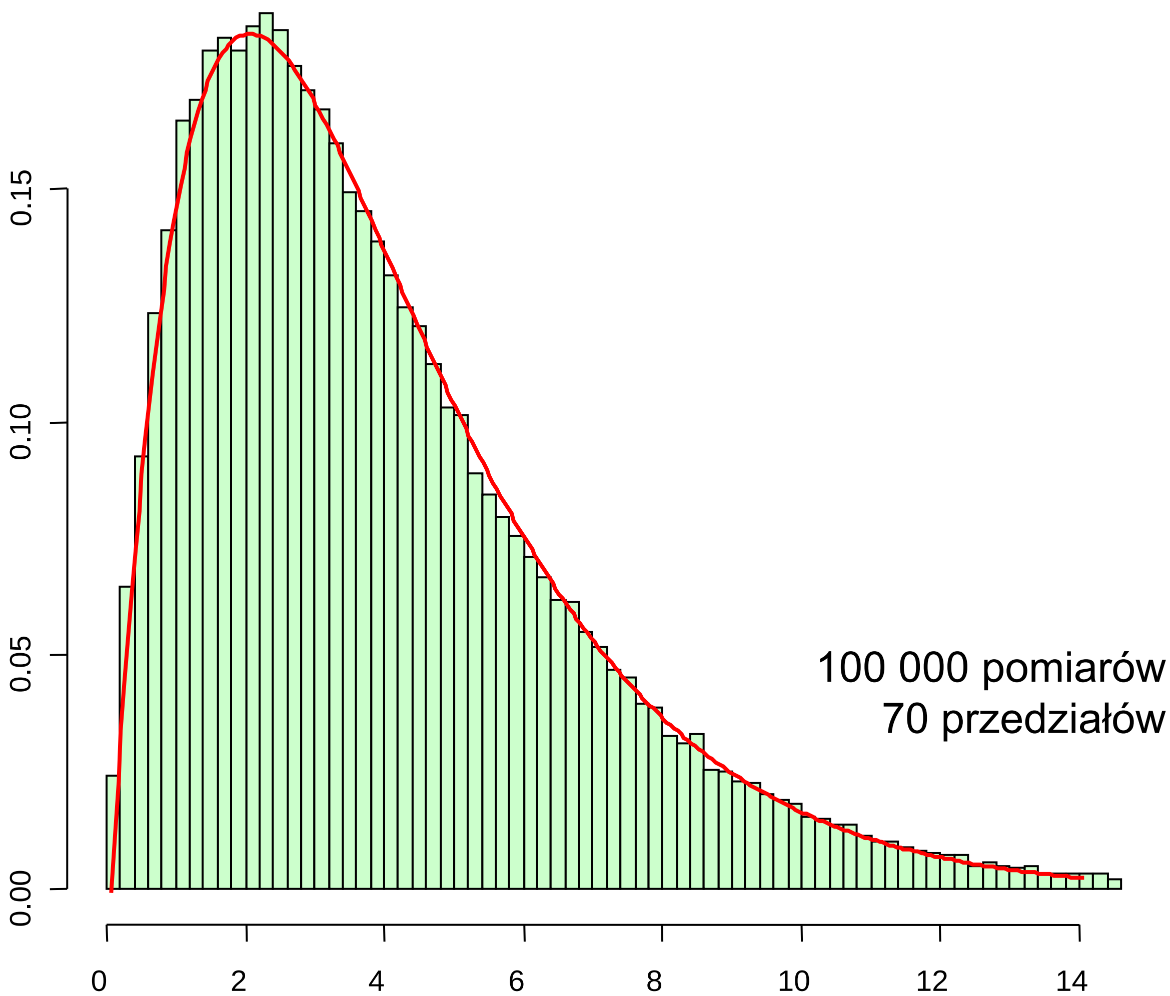
pole pod wykresem gęstości dla tego przedziału.


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

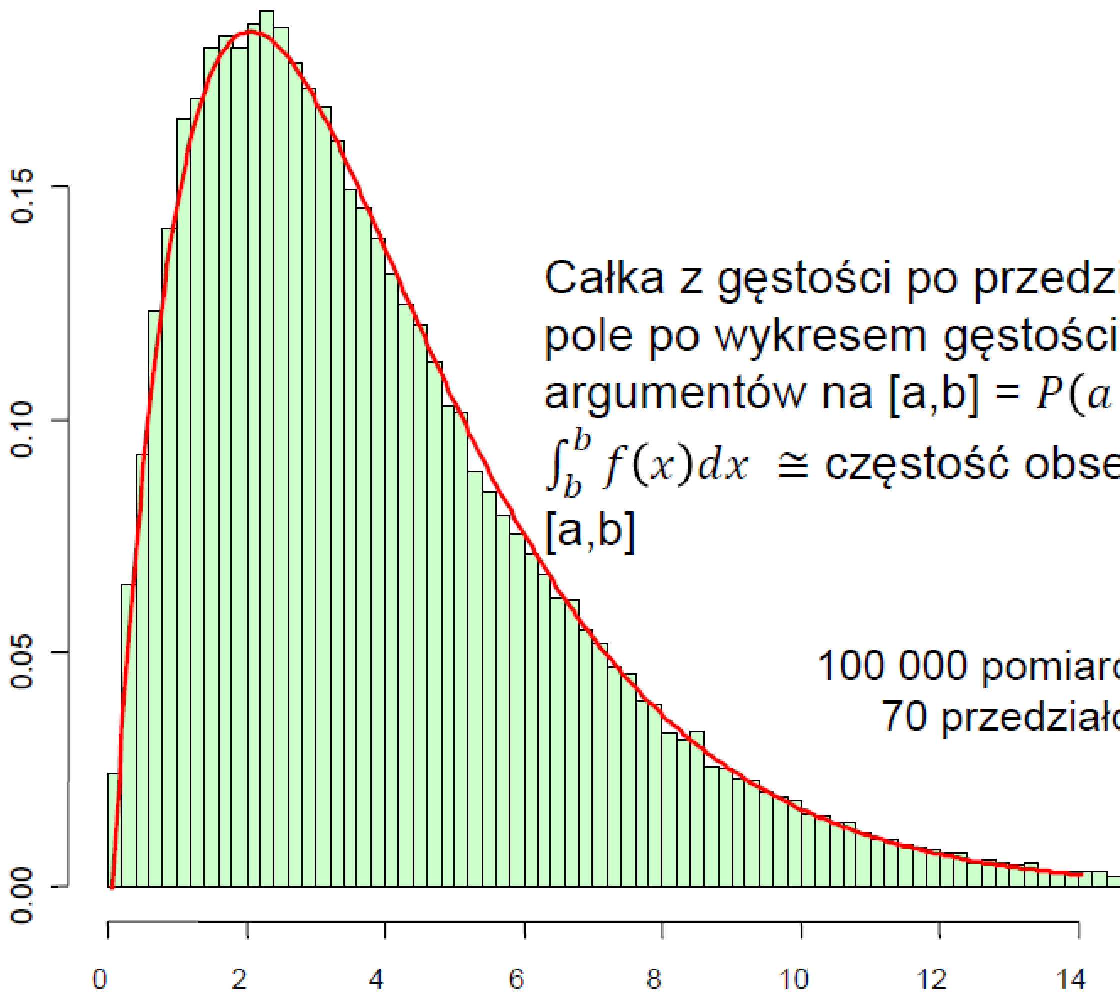

-5
-3
-1
1
3
5
7


-5
-3
-1
1
3
5
7

Funkcja gęstości rozkładu a histogram unormowany



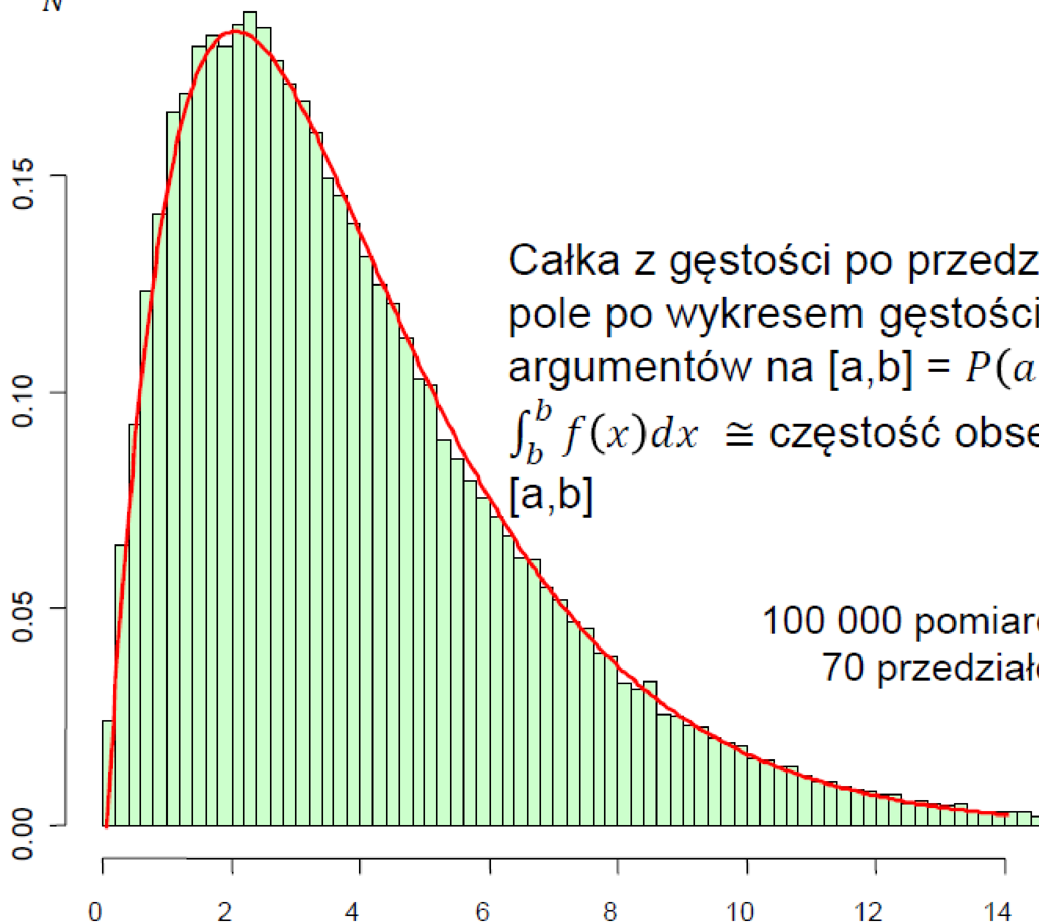
Funkcja gęstości rozkładu a histogram unormowany



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

Funkcja gęstości rozkładu

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



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

Przykład

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 $F(0)$

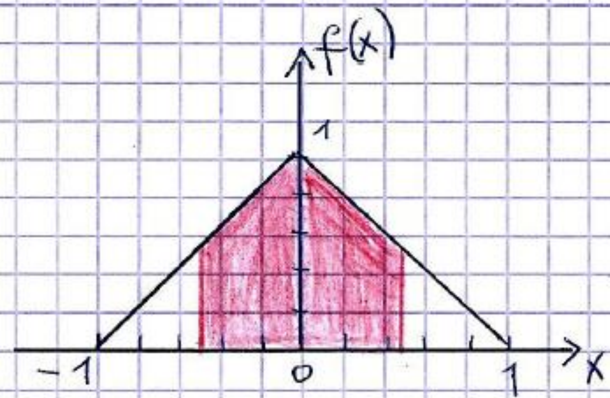
(e) stałą c taką, że

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

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

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

Przykład



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

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

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

Przykład

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

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

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

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

Przykład

d) dystrybuantę $F(0)$

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

(e) stałą c taką, że

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

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

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

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) $c = q_{0,25}$

Wskaźniki położenia i rozproszenia dla ciągłych zmiennych losowych

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

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

Wartość średnia zmiennej losowej

<u>Definicja</u>. 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

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

(jeśli całka istnieje).

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

$$\mathrm{Var}(X) = \mathrm{V}(X) = \sigma_X^2 = \int_{-\infty}^{\infty} (s-\mu_X)^2 f(s)\,ds$$

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:

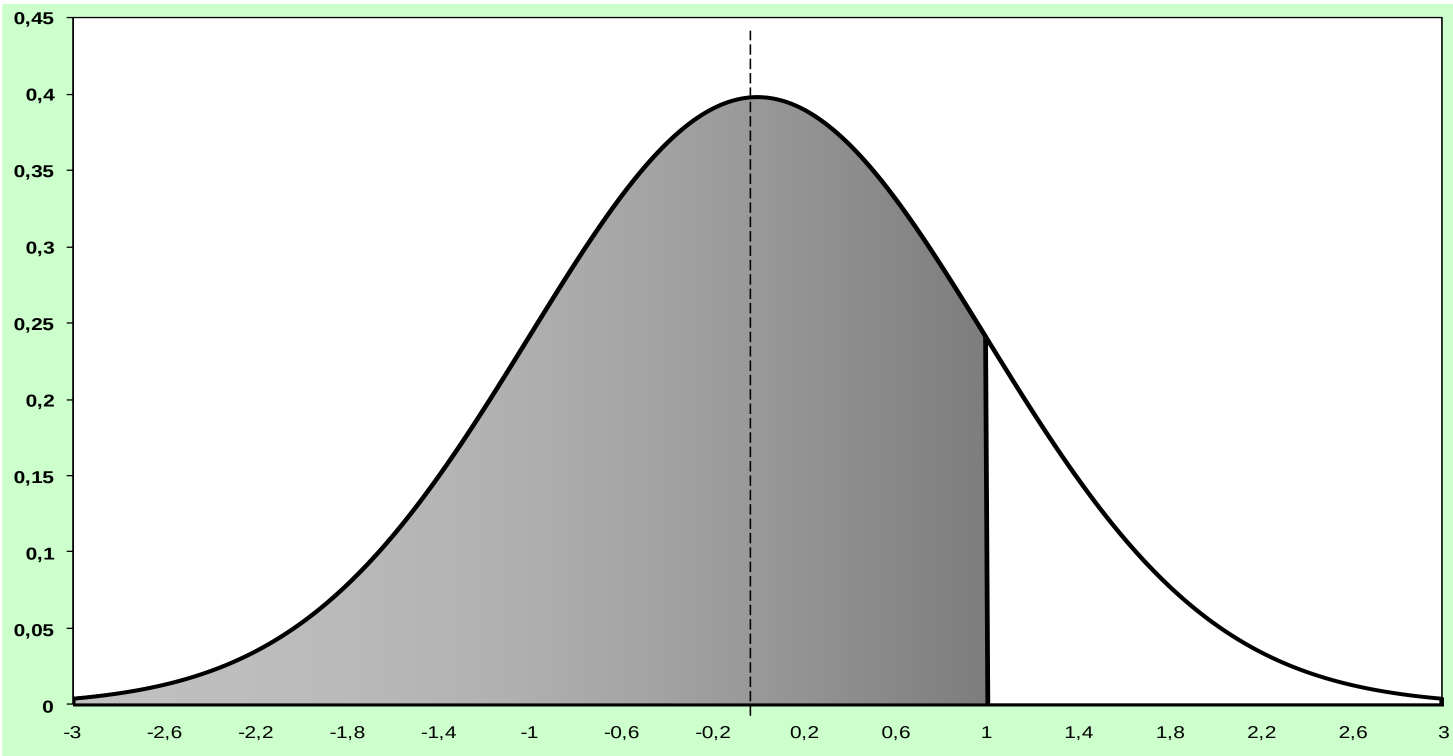
Częstość obserwacji $\boldsymbol{\leq q \approx}$

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

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

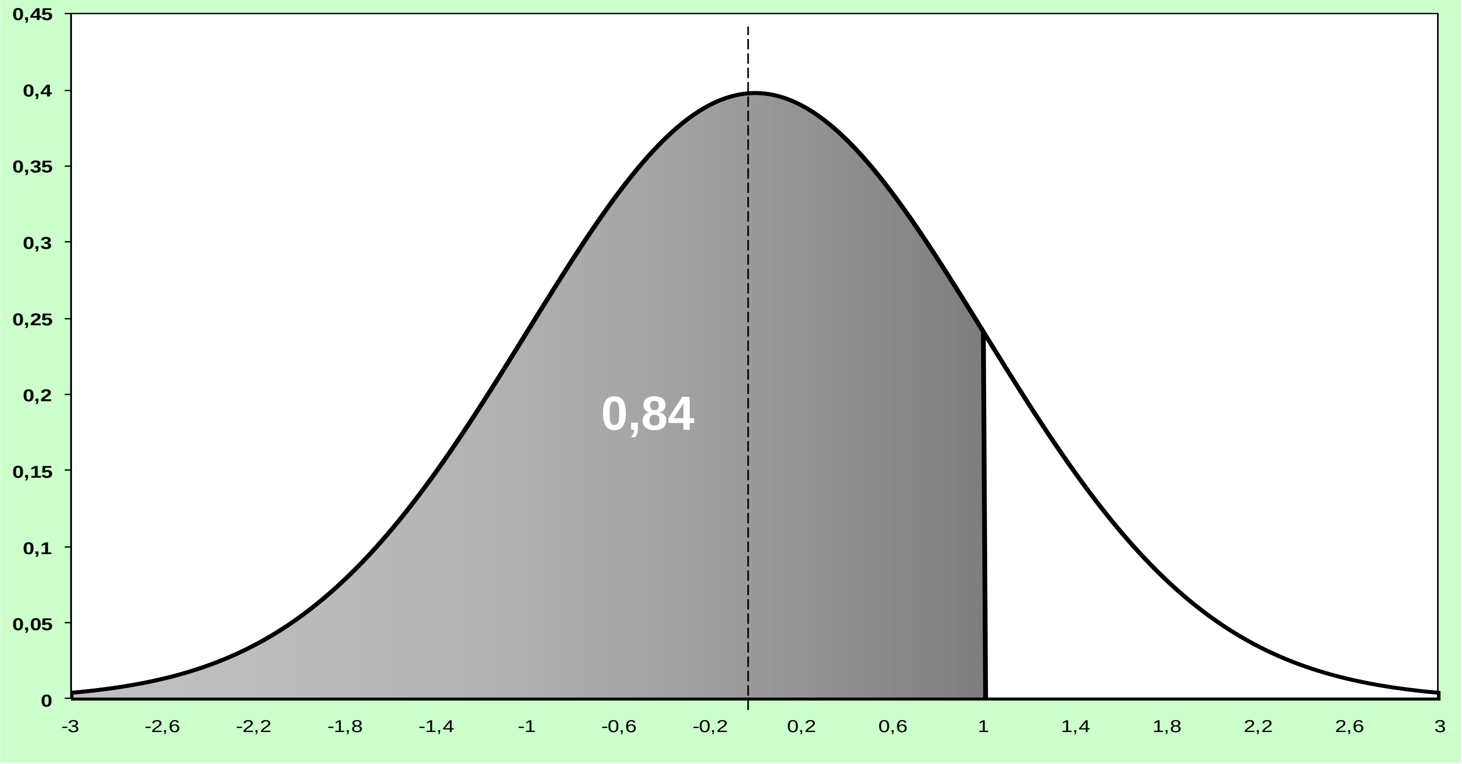
$$\boxed{\int_{-\infty}^{q} f(x)dx}$$

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

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

- **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

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

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:**	$q_{0,75}$

Standaryzacja

<u>Stwierdzenie.</u> (standaryzacja)

Jeśli zmienna losowa ***X*** ma wartość średnią $\boxed{\mu_X}$ oraz wariancję $\boxed{\sigma_X^2}$, to standaryzowana zmienna losowa

$$Z = X^* = \frac{X - \mu_X}{\sigma_X}$$

ma wartość średnią **0** i wariancję **1.**

- Dowód:

$$\mu_Z = E\left(\frac{X-\mu_X}{\sigma_X}\right) = E\left(\frac{1}{\sigma_X}\times(X-\mu_X)\right)$$

$$\frac{1}{\sigma_X}\times E(X-\mu_X) = \frac{1}{\sigma_X}\times(E(X)-E(X)) = 0.$$

$$\sigma_Z^2 = E\left(\frac{X-\mu_X}{\sigma_X}\right)^2 = \left(\frac{1}{\sigma_X}\right)^2 \times E(X-\mu_X)^2 = 1.$$

Nierówność Czebyszewa

<u>**Twierdzenie.**</u> Niech zmienna losowa ***X*** ma wartość średnią μ oraz wariancję σ^2. Wówczas

$$\boxed{P(|X-\mu| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}},$$

dla dowolnego $\varepsilon > 0$.