

# Bootcamp Data Science

Statystyka

Przemysław Spurek

O co chodzi z tą zmienną losową?

# Prawdopodobieństwo klasyczne (z liceum)

Teoria prawdopodobieństwa zajmuje się zdarzeniami pojawiającymi się przy wykonywaniu doświadczeń losowych, czyli takich, których wyniku nie da się z góry przewidzieć, a jednocześnie dających się powtarzać w tych samych warunkach.

# Prawdopodobieństwo klasyczne (z liceum)

Teoria prawdopodobieństwa zajmuje się zdarzeniami pojawiającymi się przy wykonywaniu doświadczeń losowych, czyli takich, których wyniku nie da się z góry przewidzieć, a jednocześnie dających się powtarzać w tych samych warunkach.

Impuls do rozwoju teorii prawdopodobieństwa dała analiza gier hazardowych (XVII wiek), a także, w późniejszych czasach, analiza zjawisk masowych.

# Prawdopodobieństwo klasyczne (z liceum)

Teoria prawdopodobieństwa zajmuje się zdarzeniami pojawiającymi się przy wykonywaniu doświadczeń losowych, czyli takich, których wyniku nie da się z góry przewidzieć, a jednocześnie dających się powtarzać w tych samych warunkach.

Impuls do rozwoju teorii prawdopodobieństwa dała analiza gier hazardowych (XVII wiek), a także, w późniejszych czasach, analiza zjawisk masowych.

## Przykład

Pojedynczy rzut monetą. Możliwe wyniki to orzeł lub reszka. Doświadczenie można powtarzać wielokrotnie w tych samych warunkach. Czego można oczekiwać w wyniku wielokrotnego powtórzenia tego doświadczenia?

## Zadanie

Rzucamy trzema identycznymi monetami. Oblicz prawdopodobieństwo zdarzenia, polegającego na wyrzuceniu co najmniej dwóch orłów.

## Zadanie

Rzucamy trzema identycznymi monetami. Oblicz prawdopodobieństwo zdarzenia, polegającego na wyrzuceniu co najmniej dwóch orłów.

$$\Omega = \{(O, O, O); (O, O, R); (O, R, O); (R, O, O); (R, R, R); (R, R, O); (R, O, R); (O, R, R); \}$$

## Zadanie

Rzucamy trzema identycznymi monetami. Oblicz prawdopodobieństwo zdarzenia, polegającego na wyrzuceniu co najmniej dwóch orłów.

$$\Omega = \{(O, O, O); (O, O, R); (O, R, O); (R, O, O); \\ (R, R, R); (R, R, O); (R, O, R); (O, R, R); \}$$

$$A = \{(O, O, O); (O, O, R); (O, R, O); (R, O, O)\}$$



## Zadanie

Rzucamy trzema identycznymi monetami. Oblicz prawdopodobieństwo zdarzenia, polegającego na wyrzuceniu co najmniej dwóch orłów.

$$\Omega = \{(O, O, O); (O, O, R); (O, R, O); (R, O, O); \\ (R, R, R); (R, R, O); (R, O, R); (O, R, R); \}$$

$$A = \{(O, O, O); (O, O, R); (O, R, O); (R, O, O)\}$$

Więc moc obu zbiorów wynosi:

$$|\Omega| = 8, \quad |A| = 4$$

## Zadanie

Rzucamy trzema identycznymi monetami. Oblicz prawdopodobieństwo zdarzenia, polegającego na wyrzuceniu co najmniej dwóch orłów.

$$\Omega = \{(O, O, O); (O, O, R); (O, R, O); (R, O, O); \\ (R, R, R); (R, R, O); (R, O, R); (O, R, R); \}$$

$$A = \{(O, O, O); (O, O, R); (O, R, O); (R, O, O)\}$$

Więc moc obu zbiorów wynosi:

$$|\Omega| = 8, \quad |A| = 4$$

Obliczamy prawdopodobieństwo:

$$P(A) = \frac{|A|}{|\Omega|} = \frac{4}{8} = \frac{1}{2}.$$

## Definicja

Niech  $\Omega$  będzie dowolnym zbiorem, zwanym przestrzenią zdarzeń elementarnych. Elementy  $\omega$  tej przestrzeni nazywamy *zdarzeniami elementarnymi*.

## Definicja

Niech  $\Omega$  będzie dowolnym zbiorem, zwanym przestrzenią zdarzeń elementarnych. Elementy  $\omega$  tej przestrzeni nazywamy *zdarzeniami elementarnymi*.

**$\sigma$ -ciało podzbiorów  $\Omega$ .** Zdarzeniami nazywać będziemy wyłącznie podzbiory należące do  $\mathcal{F}$ . Przełożmy powyższe wymagania na formalne własności matematyczne zbioru  $\mathcal{F}$ :

- 1  $\emptyset \in \mathcal{F}$ ,
- 2 jeżeli  $A \in \mathcal{F}$  to  $A' = \Omega \setminus A \in \mathcal{F}$ ,
- 3 jeżeli  $A_i \in \mathcal{F}$ , dla  $i = 1, 2, \dots$  to  $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$ .

## Definicja

Niech  $\Omega$  będzie dowolnym zbiorem, zwanym przestrzenią zdarzeń elementarnych. Elementy  $\omega$  tej przestrzeni nazywamy *zdarzeniami elementarnymi*.

**$\sigma$ -ciało podzbiorów  $\Omega$ .** Zdarzeniami nazywać będziemy wyłącznie podzbiory należące do  $\mathcal{F}$ . Przełożmy powyższe wymagania na formalne własności matematyczne zbioru  $\mathcal{F}$ :

- 1  $\emptyset \in \mathcal{F}$ ,
- 2 jeżeli  $A \in \mathcal{F}$  to  $A' = \Omega \setminus A \in \mathcal{F}$ ,
- 3 jeżeli  $A_i \in \mathcal{F}$ , dla  $i = 1, 2, \dots$  to  $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$ .

## Definicja

Rodzinę zdarzeń  $\mathcal{F}$  spełniającą warunki (S1-S3) nazywamy  $\sigma$ -ciałem (podzbiorów zbioru  $\Omega$ ).

# Prawdopodobieństwo klasyczne (z liceum)

Możemy teraz zdefiniować miarę probabilistyczną na wprowadzonym  $\sigma$ -ciele.

## Definicja

Prawdopodobieństwem nazywamy dowolną funkcję  $P: \mathcal{F} \rightarrow \mathbb{R}$  o wartościach rzeczywistych, określoną na  $\sigma$ -ciele zdarzeń  $\mathcal{F} \subset 2^\Omega$ , spełniającą warunki:

- 1  $P(A) \geq 0$  dla każdego  $A \in \mathcal{F}$ ,
- 2  $P(\Omega) = 1$ .
- 3 Jeżeli  $A_i \in \mathcal{F}$ ,  $i = 1, 2, \dots$  oraz  $A_i \cap A_j = \emptyset$  dla  $i \neq j$ , to

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

Matematyczny model doświadczenia losowego to trójka

$$(\Omega, \mathcal{F}, P),$$

gdzie  $P$  jest prawdopodobieństwem, określonym na pewnym  $\sigma$ -ciele podzbiorów zbioru zdarzeń elementarnych  $\Omega$ .

Taką trojkę nazywamy **przestrzenią probabilistyczną**.

## Definicja

*Prawdopodobieństwo warunkowe* zdarzenia  $A$  pod warunkiem zdarzenia  $B$  określone jest wzorem:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

przy założeniu, że  $P(B) > 0$ .

## Własność

$$P(A \cap B) = P(A|B) \cdot P(B)$$

dla  $P(B) > 0$ .



## Zadanie

Ktoś rzucił 3 razy monetą i poinformował nas, że wypadła nieparzysta liczba orłów (zdarzenie B). Jaka jest szansa, że wypadły 3 orły (zdarzenie A).

## Zadanie

Ktoś rzucił 3 razy monetą i poinformował nas, że wypadła nieparzysta liczba orłów (zdarzenie B). Jaka jest szansa, że wypadły 3 orły (zdarzenie A).

$\Omega =$

$\{(o, o, o), (o, o, r), (o, r, o), (o, r, r), (r, o, o), (r, o, r), (r, r, o), (r, r, r)\}$

## Zadanie

Ktoś rzucił 3 razy monetą i poinformował nas, że wypadła nieparzysta liczba orłów (zdarzenie B). Jaka jest szansa, że wypadły 3 orły (zdarzenie A).

$\Omega =$

$\{(o, o, o), (o, o, r), (o, r, o), (o, r, r), (r, o, o), (r, o, r), (r, r, o), (r, r, r)\}$

Prawdopodobieństwo wylosowania trzech orłów pod rząd:  $P(A) = \frac{1}{8}$

Prawdopodobieństwo wylosowania nieparzystej ilości orłów:  $P(B) = \frac{4}{8}$

## Zadanie

Ktoś rzucił 3 razy monetą i poinformował nas, że wypadła nieparzysta liczba orłów (zdarzenie B). Jaka jest szansa, że wypadły 3 orły (zdarzenie A).

$\Omega =$

$\{(o, o, o), (o, o, r), (o, r, o), (o, r, r), (r, o, o), (r, o, r), (r, r, o), (r, r, r)\}$

Prawdopodobieństwo wylosowania trzech orłów pod rząd:  $P(A) = \frac{1}{8}$

Prawdopodobieństwo wylosowania nieparzystej ilości orłów:  $P(B) = \frac{4}{8}$

Prawdopodobieństwo warunkowe zdarzenia A pod warunkiem zdarzenia B

$$\begin{aligned} P(A|B) &= \frac{P(A \cap B)}{P(B)} = \frac{P(\{(r, r, r)\} \cap \{(o, o, r), (o, r, o), (r, o, o), (r, r, r)\})}{P(\{(o, o, r), (o, r, o), (r, o, o), (r, r, r)\})} = \\ &= \frac{\frac{1}{8}}{\frac{4}{8}} = \frac{1}{8} \cdot \frac{8}{4} = \frac{1}{4} \end{aligned}$$

## Wzór na prawdopodobieństwo całkowite

Jeżeli  $\{H_1, H_2, \dots, H_n\}$  jest **rozbiciem**  $\Omega$  na zdarzenia o dodatnich prawdopodobieństwach, to dla dowolnego zdarzenia  $A$

$$P(A) = \sum_{i=1}^n P(A|H_i)P(H_i).$$

## Wzór na prawdopodobieństwo całkowite

Jeżeli  $\{H_1, H_2, \dots, H_n\}$  jest **rozbiciem**  $\Omega$  na zdarzenia o dodatnich prawdopodobieństwach, to dla dowolnego zdarzenia  $A$

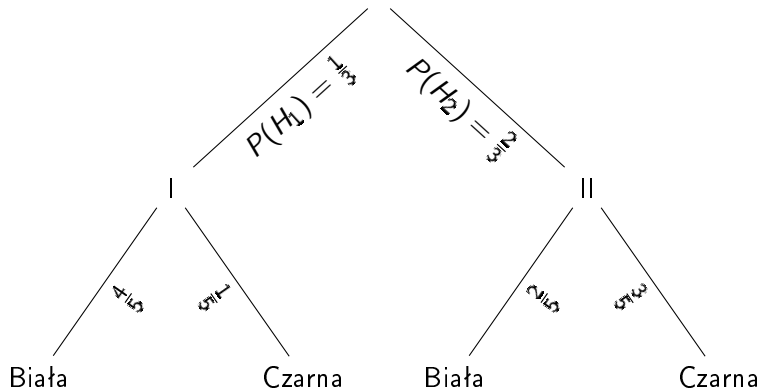
$$P(A) = \sum_{i=1}^n P(A|H_i)P(H_i).$$

## Zadanie

Pierwsza urna zawiera 4 białe i jedną czarną kulę, druga – 2 białe i 3 czarne. Losujemy urnę tak, by szansa wybrania pierwszej urny była dwukrotnie mniejsza niż drugiej. Następnie z wybranej urny losujemy kulę. Jakie jest prawdopodobieństwo wylosowania kuli białej.

# Prawdopodobieństwo całkowite

Możemy narysować drzewo:



# Prawdopodobieństwo całkowite

Oznaczmy:

$H_1$  - wyboru pierwszej urny ( $P(H_1) = \frac{1}{3}$ )

$H_2$  - wyboru drugiej urny ( $P(H_2) = \frac{2}{3}$ )

$A$  - wylosowano kulę białą

Z treści zadania wiemy, że  $P(H_2) = 2 \cdot P(H_1)$ . Wiadomo również, że  $P(H_1 \cup H_2) = 1$  oraz  $P(H_1 \cap H_2) = \emptyset$ . Czyli  $P(H_1) = \frac{1}{3}$  oraz  $P(H_2) = \frac{2}{3}$ . Podstawiając do wzoru na prawdopodobieństwo całkowite otrzymujemy:

$$P(A) = P(A|H_1) \cdot P(H_1) + P(A|H_2) \cdot P(H_2) = \frac{4}{5} \cdot \frac{1}{3} + \frac{2}{5} \cdot \frac{2}{3} = \frac{8}{15}$$

gdzie:

$P(A|H_1)$  – prawdopodobieństwo wylosowania białej kuli z pierwszej urny

$P(A|H_2)$  – prawdopodobieństwo wylosowania białej kuli z drugiej urny



## Wzór Bayesa

Jeżeli  $\{H_i\}_{i \in I}$  jest przeliczalnym **rozbiem**  $\Omega$  na zdarzenia o dodatnich prawdopodobieństwach oraz  $P(A) > 0$ , to dla dowolnego  $j \in I$  mamy

$$P(H_j|A) = \frac{P(A|H_j)P(H_j)}{\sum_{i \in I} P(A|H_i)P(H_i)}.$$

## Wzór Bayesa

Jeżeli  $\{H_i\}_{i \in I}$  jest przeliczalnym **rozbiciem**  $\Omega$  na zdarzenia o dodatnich prawdopodobieństwach oraz  $P(A) > 0$ , to dla dowolnego  $j \in I$  mamy

$$P(H_j|A) = \frac{P(A|H_j)P(H_j)}{\sum_{i \in I} P(A|H_i)P(H_i)}.$$

## Uwaga

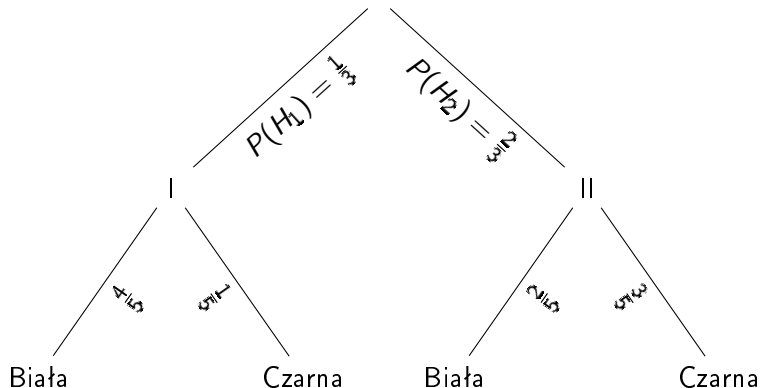
Prawdopodobieństwo hipotetyczne  $P(H_i)$  nazywamy prawdopodobieństwem **a priori** (przed doświadczeniem),  $P(H_i|A)$  prawdopodobieństwem **a posteriori** (po doświadczeniu).

## Zadanie

W sytuacji z poprzedniego zadania oblicz prawdopodobieństwo, że losowano z drugiej urny gdy wynikiem losowania jest kula biała.

# Prawdopodobieństwo całkowite

Możemy narysować drzewo:



Oznaczmy:

$H_1$  - wyboru pierwszej urny ( $P(H_1) = \frac{1}{3}$ )

$H_2$  - wyboru drugiej urny ( $P(H_2) = \frac{2}{3}$ )

$A$  - wylosowano kulę białą Z Rozwiązania poprzedniego zadania wiemy, że

$$P(A) = \frac{8}{15}.$$

W takiej sytuacji:  $P(H_2|A)$  - oznacza prawdopodobieństwo, że losowano z drugiej urny gdy wynikiem losowania jest kula biała. Z tw. Bayesa mamy:

$$P(H_2|A) = \frac{P(A|H_2)P(H_2)}{P(A|H_1) \cdot P(H_1) + P(A|H_2) \cdot P(H_2)} = \frac{\frac{2}{5} \cdot \frac{2}{3}}{\frac{8}{15}} = \frac{\frac{4}{15}}{\frac{8}{15}} = \frac{1}{2}.$$

## Definicja

Zdarzenia  $A$  i  $B$  nazywamy niezależnymi, gdy

$$P(A \cap B) = P(A) \cdot P(B).$$

## Zadanie

Z 52 kart ciągniemy jedną. Czy zdarzenia w następujących parach są niezależne:

- 1  $A$  – Wyciągnięcie damy,  $B$  – wyciągnięcie karo
- 2  $A$  – Wyciągnięcie czerwonej figury,  $B$  – wyciągnięcie kiera

## Zadanie

Z 52 kart ciągniemy jedną. Czy zdarzenia w następujących parach są niezależne:

- ❶  $A$  – Wyciągnięcie damy,  $B$  – wyciągnięcie karo
- ❷  $A$  – Wyciągnięcie czerwonej figury,  $B$  – wyciągnięcie kiera

1.  $A$  – Wyciągnięcie damy,  $B$  – wyciągnięcie karo

$$P(A) = \frac{4}{52}, \quad P(B) = \frac{13}{52}$$

$$P(A \cap B) = \frac{1}{52}$$

$$P(A \cap B) = \frac{1}{52} = \frac{4}{52} \frac{13}{52} = P(A) \cdot P(B).$$



## Zadanie

Z 52 kart ciągniemy jedną. Czy zdarzenia w następujących parach są niezależne:

- ❶  $A$  – Wyciągnięcie damy,  $B$  – wyciągnięcie karo
- ❷  $A$  – Wyciągnięcie czerwonej figury,  $B$  – wyciągnięcie kiera

1.  $A$  – Wyciągnięcie damy,  $B$  – wyciągnięcie karo

$$P(A) = \frac{4}{52}, \quad P(B) = \frac{13}{52}$$

$$P(A \cap B) = \frac{1}{52}$$

$$P(A \cap B) = \frac{1}{52} = \frac{4}{52} \cdot \frac{13}{52} = P(A) \cdot P(B).$$

Zdarzenia są niezależne.

2.  $A$  – Wyciągnięcie czerwonej figury,  $B$  – wyciągnięcie kiera

$$P(A) = \frac{8}{52}, \quad P(B) = \frac{13}{52}$$

$$P(A) \cdot P(B) = \frac{8}{52} \cdot \frac{13}{52} = \frac{2}{52}, \quad P(A \cap B) = \frac{4}{52}$$

$$P(A \cap B) \neq P(A) \cdot P(B).$$

Zdarzenia są zależne.

# Zmienne losowe.

## Oznaczenie

$\mathcal{B}(\mathbb{R})$  – rodzina zbiorów Borelowskich.

## Oznaczenie

$\mathcal{B}(\mathbb{R})$  – rodzina zbiorów Borelowskich.

## Definicja

Niech będzie dana przestrzeń probabilistyczna  $(\Omega, \mathcal{F}, P)$ . Funkcję  $X: \Omega \rightarrow \mathbb{R}$  określoną na przestrzeni zdarzeń elementarnych nazywamy zmienną losową o wartościach w  $\mathbb{R}$  jeżeli dla każdego  $a \in \mathbb{R}$  zbiór  $X^{-1}((-\infty, a))$  jest zdarzeniem elementarnym, czyli  $X^{-1}((-\infty, a)) \in \mathcal{F}$ .

Zdefiniuj zmienną losową dla poniższych “zdarzeń”:

- Rzut symetryczną monetą:

Zdefiniuj zmienną losową dla poniższych “zdarzeń”:

- Rzut symetryczną monetą:

$X: \Omega \rightarrow \{0, 1\}$ , gdzie

$$\begin{aligned} X(O) &= 0, & X(R) &= 1, \\ P(X = 0) &= \frac{1}{2}, & P(X = 1) &= \frac{1}{2}. \end{aligned}$$

Zdefiniuj zmienną losową dla poniższych “zdarzeń”:

- Rzut symetryczną monetą:

$X: \Omega \rightarrow \{0, 1\}$ , gdzie

$$X(O) = 0, \quad X(R) = 1,$$

$$P(X = 0) = \frac{1}{2}, \quad P(X = 1) = \frac{1}{2}.$$

- Wybór jednej karty z tali:

Zdefiniuj zmienną losową dla poniższych “zdarzeń”:

- Rzut symetryczną monetą:

$X: \Omega \rightarrow \{0, 1\}$ , gdzie

$$X(O) = 0, \quad X(R) = 1,$$

$$P(X = 0) = \frac{1}{2}, \quad P(X = 1) = \frac{1}{2}.$$

- Wybór jednej karty z tali:

$X: \Omega \rightarrow \{1, 2, \dots, 52\}$ , gdzie

$$X(\text{As kier}) = 1, \dots, X(2 \text{ pik}) = 52,$$

$$P(X = i) = \frac{1}{52}, \text{ dla } i = 1, \dots, 52.$$



# Zmienna losowa

Zdefiniuj zmienną losową dla poniższych “zdarzeń”:

# Zmienna losowa

Zdefiniuj zmienną losową dla poniższych “zdarzeń”:

- Rzut kostką:

Zdefiniuj zmienną losową dla poniższych “zdarzeń”:

- Rzut kostką:

$X: \Omega \rightarrow \{1, 2, 3, 4, 5, 6\}$ , gdzie

$$X(\text{wypada } 1) = 1, \dots, X(\text{wypada } 6) = 6,$$

$$P(X = i) = \frac{1}{6}, \text{ dla } i = 1, \dots, 6.$$

Zdefiniuj zmienną losową dla poniższych “zdarzeń”:

- Rzut kostką:

$X: \Omega \rightarrow \{1, 2, 3, 4, 5, 6\}$ , gdzie

$$X(\text{wypada } 1) = 1, \dots, X(\text{wypada } 6) = 6,$$

$$P(X = i) = \frac{1}{6}, \text{ dla } i = 1, \dots, 6.$$

- Odbiór partii produktów, z których 98% jest dobra, a pozostała wybrakowana:

# Zmienna losowa

Zdefiniuj zmienną losową dla poniższych “zdarzeń”:

- Rzut kostką:

$X: \Omega \rightarrow \{1, 2, 3, 4, 5, 6\}$ , gdzie

$$X(\text{wypada } 1) = 1, \dots, X(\text{wypada } 6) = 6,$$

$$P(X = i) = \frac{1}{6}, \text{ dla } i = 1, \dots, 6.$$

- Odbiór partii produktów, z których 98% jest dobra, a pozostała wybrakowana:

$X: \Omega \rightarrow \{0, 1\}$ , gdzie

$$X(\text{produkt dobry}) = 0, \quad X(\text{produkt wybrakowany}) = 1,$$

$$P(X = 0) = \frac{98}{100}, \quad P(X = 1) = \frac{2}{100}.$$

## Definicja

Niech będzie dana przestrzeń probabilistyczna  $(\Omega, \mathcal{F}, P)$ . Rozkładem prawdopodobieństwa zmiennej losowej  $X: (\Omega, \mathcal{F}) \rightarrow (\mathbb{R}, \mathcal{B}(\mathbb{R}))$  nazywamy prawdopodobieństwo *miara* $_X$ , określone na  $\mathcal{B}(\mathbb{R})$  zależnością:

$$\text{miara}_X(B) = P(X^{-1}(B)), \quad B \in \mathcal{B}(\mathbb{R}).$$

## Definicja

Niech będzie dana przestrzeń probabilistyczna  $(\Omega, \mathcal{F}, P)$ . Rozkładem prawdopodobieństwa zmiennej losowej  $X: (\Omega, \mathcal{F}) \rightarrow (\mathbb{R}, \mathcal{B}(\mathbb{R}))$  nazywamy prawdopodobieństwo *miara* $_X$ , określone na  $\mathcal{B}(\mathbb{R})$  zależnością:

$$\text{miara}_X(B) = P(X^{-1}(B)), \quad B \in \mathcal{B}(\mathbb{R}).$$

## Oznaczenie

$P(X^{-1}(B))$  można również zapisywać:

$$P(X^{-1}(B)) = P(\{\omega \in \Omega: X(\omega) \in B\}) = P(X \in B).$$

Ostatniej, skrótowej wersji będziemy używać najczęściej.

Po co nam te zmienne losowe?



# Zmienna losowa

Zdefiniuj zmienną losową dla poniższych “zdarzeń”:

- Rzut monetą (nie necessarily symetryczną):

$$X: \Omega \rightarrow \{0, 1\}, \text{ gdzie } X(O) = 0, \quad X(R) = 1,$$

$$P(X = 0) = p, \quad P(X = 1) = (1 - p).$$

# Zmienna losowa

Zdefiniuj zmienną losową dla poniższych “zdarzeń”:

- Rzut monetą (nie koniecznie symetryczną):

$$X: \Omega \rightarrow \{0, 1\}, \text{ gdzie } X(O) = 0, \quad X(R) = 1,$$

$$P(X = 0) = p, \quad P(X = 1) = (1 - p).$$

- Płeć noworodka:

$$X: \Omega \rightarrow \{0, 1\}, \text{ gdzie } X(Ch) = 0, \quad X(Dzi) = 1,$$

$$P(X = 0) = p, \quad P(X = 1) = (1 - p).$$

Zdefiniuj zmienną losową dla poniższych “zdarzeń”:

- Rzut monetą (nie necessarily symetryczną):

$$X: \Omega \rightarrow \{0, 1\}, \text{ gdzie } X(O) = 0, \quad X(R) = 1,$$

$$P(X = 0) = p, \quad P(X = 1) = (1 - p).$$

- Płeć noworodka:

$$X: \Omega \rightarrow \{0, 1\}, \text{ gdzie } X(Ch) = 0, \quad X(Dzi) = 1,$$

$$P(X = 0) = p, \quad P(X = 1) = (1 - p).$$

- Wygrana w totolotka:

$$X: \Omega \rightarrow \{0, 1\}, \text{ gdzie } X(Wyg) = 0, \quad X(Prze) = 1,$$

$$P(X = 0) = p, \quad P(X = 1) = (1 - p).$$

# Zmienne losowe o rozkładzie dyskretnym

# Zmienne losowe o rozkładzie dyskretnym

## Definicja

Zmienna losowa  $X$  ma rozkład dyskretny, jeśli istnieje taki zbiór przeliczalny  $S \subset \mathbb{R}$ , taki że  $miara_X(S) = 1$ .

# Zmienne losowe o rozkładzie dyskretnym

## Definicja

Zmienna losowa  $X$  ma rozkład dyskretny, jeśli istnieje taki zbiór przeliczalny  $S \subset \mathbb{R}$ , taki że  $miara_X(S) = 1$ .

# Rozkład zero-jedynkowy (próba Bernoulliego)

[https://github.com/przem85/statistic\\_4/blob/master/D02\\_Z01\\_bernoulli\\_distribution.ipynb](https://github.com/przem85/statistic_4/blob/master/D02_Z01_bernoulli_distribution.ipynb)

Próba Bernoulliego (rozkład zero-jedynkowy) – dyskretny rozkład prawdopodobieństwa, dla którego zmienna losowa przyjmuje tylko wartości: 0 lub 1:

$$P(X = k) = \begin{cases} p & \text{gdy } k = 0 \\ 1 - p & \text{gdy } k \neq 0 \end{cases},$$

gdzie  $0 < p < 1$ , in  $\{0, 1\}$ .

Powyższą funkcję opisującą prawdopodobieństwo wystąpienia każdego z elementów nazywamy funkcją gęstości (**probability mass function (PMF)**).

# Rozkład zero-jedynkowy (próba Bernoulliego)

- Definiujemy zmienną losową

```
from scipy import stats  
p = 0.5  
bernoulliDist = stats.bernoulli(p)
```

- Możemy wypisać parametry

```
p_tails = bernoulliDist.pmf(0)  
p_heads = bernoulliDist.pmf(1)  
print(p_tails)  
print(p_heads)
```

- Możemy wylosować próbkę oraz narysować histogram

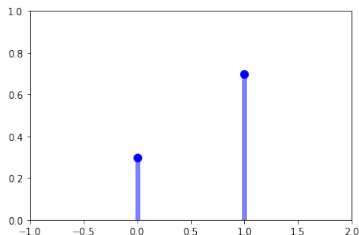
```
trials = bernoulliDist.rvs(100)  
trials  
plt.hist(trials)  
plt.show()
```



# Rozkład zero-jedynkowy (próba Bernoulliego)

- Rysujemy gęstość

```
p = 0.7
fig, ax = plt.subplots(1, 1)
x = np.arange(0, 2)
ax.plot(x, stats.bernoulli.pmf(x, p),
        'bo', ms=8, label='bernoulli pmf')
ax.vlines(x, 0, stats.bernoulli.pmf(x, p),
          colors='b', lw=5, alpha=0.5)
plt.show()
```

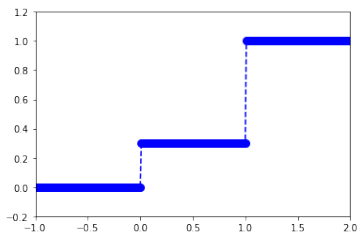


# Rozkład zero-jedynkowy (próba Bernoulliego)

- Rysujemy dystrybuantę

```
p = 0.7
fig, ax = plt.subplots(1, 1)

x = np.arange(-5, 5, 0.01)
ax.plot(x, stats.bernoulli.cdf(x, p),
        'bo--', ms=8, label='bernoulli cdf')
rv = stats.bernoulli(p)
plt.show()
```



# Rozkład dwumianowy

`https://github.com/przem85/statistic\_4/blob/master/D02\_Z02\_binomial\_distribution.ipynb`

- Jeśli wielokrotnie rzucimy monety i pytamy "Jak często pojawiłaby się reszka?" to dostajemy rozkład dwumianowy.

# Rozkład dwumianowy

[https://github.com/przem85/statistic\\_4/blob/master/D02\\_Z02\\_binomial\\_distribution.ipynb](https://github.com/przem85/statistic_4/blob/master/D02_Z02_binomial_distribution.ipynb)

- Jeśli wielokrotnie rzucimy monety i pytamy "Jak często pojawiłaby się reszka?" to dostajemy rozkład dwumianowy.
- Ogólnie rzecz biorąc, rozkład dwumianowy jest związany z pytaniem "Z danej (stałej) liczby prób, ile zakończyło się sukcesem?"

# Rozkład dwumianowy

[https://github.com/przem85/statistic\\_4/blob/master/D02\\_Z02\\_binomial\\_distribution.ipynb](https://github.com/przem85/statistic_4/blob/master/D02_Z02_binomial_distribution.ipynb)

- Jeśli wielokrotnie rzucimy monety i pytamy "Jak często pojawiłaby się reszka?" to dostajemy rozkład dwumianowy.
- Ogólnie rzecz biorąc, rozkład dwumianowy jest związany z pytaniem "Z danej (stałej) liczby prób, ile zakończyło się sukcesem?"
- Przykłady:

# Rozkład dwumianowy

[https://github.com/przem85/statistic\\_4/blob/master/D02\\_Z02\\_binomial\\_distribution.ipynb](https://github.com/przem85/statistic_4/blob/master/D02_Z02_binomial_distribution.ipynb)

- Jeśli wielokrotnie rzucimy monety i pytamy "Jak często pojawiłaby się reszka?" to dostajemy rozkład dwumianowy.
- Ogólnie rzecz biorąc, rozkład dwumianowy jest związany z pytaniem "Z danej (stałej) liczby prób, ile zakończyło się sukcesem?"
- Przykłady:
  - Dla dzieci urodzonych w danym szpitalu, w danym dniu, ile z nich będzie dziewczynkami?

[https://github.com/przem85/statistic\\_4/blob/master/D02\\_Z02\\_binomial\\_distribution.ipynb](https://github.com/przem85/statistic_4/blob/master/D02_Z02_binomial_distribution.ipynb)

- Jeśli wielokrotnie rzucimy monety i pytamy "Jak często pojawiłaby się reszka?" to dostajemy rozkład dwumianowy.
- Ogólnie rzecz biorąc, rozkład dwumianowy jest związany z pytaniem "Z danej (stałej) liczby prób, ile zakończyło się sukcesem?"
- Przykłady:
  - Dla dzieci urodzonych w danym szpitalu, w danym dniu, ile z nich będzie dziewczynkami?
  - Ilu uczniów w danej klasie ma zielone oczy?

# Rozkład dwumianowy

[https://github.com/przem85/statistic\\_4/blob/master/D02\\_Z02\\_binomial\\_distribution.ipynb](https://github.com/przem85/statistic_4/blob/master/D02_Z02_binomial_distribution.ipynb)

- Jeśli wielokrotnie rzucimy monety i pytamy "Jak często pojawiłaby się reszka?" to dostajemy rozkład dwumianowy.
- Ogólnie rzecz biorąc, rozkład dwumianowy jest związany z pytaniem "Z danej (stałej) liczby prób, ile zakończyło się sukcesem?"
- Przykłady:
  - Dla dzieci urodzonych w danym szpitalu, w danym dniu, ile z nich będzie dziewczynkami?
  - Ilu uczniów w danej klasie ma zielone oczy?
  - Ile komarów z roju umrze po zastosowaniu oprysku środkiem owadobójczym?



# Rozkład dwumianowy

Gdy zmienna losowa  $X$  ma rozkład dwumianowy z parametrami  $p$  i  $n$ , zapisujemy go jako  $X \sim B(n, p)$ , a gęstość wyrażona jest wzorem:

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k},$$

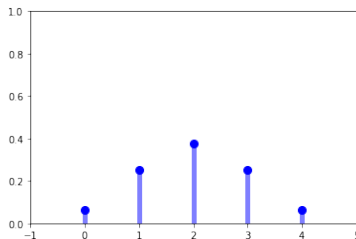
gdzie  $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ .

# Rozkład dwumianowy

- Rysujemy gęstość

```
fig, ax = plt.subplots(1, 1)
x = np.arange(0, 5)

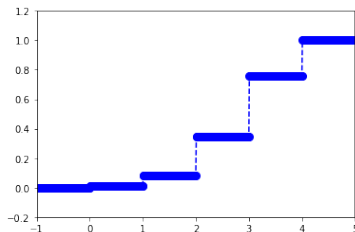
ax.plot(x, stats.binom.pmf(x, num, p),
        'bo', ms=8, label='bernoulli pmf')
ax.vlines(x, 0, stats.binom.pmf(x, num, p),
          colors='b', lw=5, alpha=0.5)
plt.show()
```



# Rozkład dwumianowy

- Rysujemy dystrybuantę

```
fig, ax = plt.subplots(1, 1)
x = np.arange(-5, 5, 0.01)
(p, num) = (0.5, 4)
ax.plot(x, stats.binom.cdf(x, num, p),
        'bo--', ms=8, label='bernoulli cdf')
plt.show()
```



[https://github.com/przem85/statistic\\_4/blob/master/D02\\_Z03\\_Poisson\\_distribution.ipynb](https://github.com/przem85/statistic_4/blob/master/D02_Z03_Poisson_distribution.ipynb)

- Rozkład Poissona jest bardzo podobny do rozkładu dwumianowego. Różnica jest subtelna.

[https://github.com/przem85/statistic\\_4/blob/master/D02\\_Z03\\_Poisson\\_distribution.ipynb](https://github.com/przem85/statistic_4/blob/master/D02_Z03_Poisson_distribution.ipynb)

- Rozkład Poissona jest bardzo podobny do rozkładu dwumianowego. Różnica jest subtelna.
- Rozkład dwumianowy sprawdza, ile razy rejestruje się sukces w stosunku do stałej liczby prób, a rozkład Poissona określa, ile razy występuje dyskretne zdarzenie (najczęściej w jakimś ustalonym czasie).

[https://github.com/przem85/statistic\\_4/blob/master/D02\\_Z03\\_Poisson\\_distribution.ipynb](https://github.com/przem85/statistic_4/blob/master/D02_Z03_Poisson_distribution.ipynb)

- Rozkład Poissona jest bardzo podobny do rozkładu dwumianowego. Różnica jest subtelna.
- Rozkład dwumianowy sprawdza, ile razy rejestruje się sukces w stosunku do stałej liczby prób, a rozkład Poissona określa, ile razy występuje dyskretne zdarzenie (najczęściej w jakimś ustalonym czasie).
- Nie ma “ustalonej” ilości możliwych sukcesów (parametru  $n$ ). Rozkład Poissona jest określony przez pojedynczy parametr  $\lambda$ .

[https://github.com/przem85/statistic\\_4/blob/master/D02\\_Z03\\_Poisson\\_distribution.ipynb](https://github.com/przem85/statistic_4/blob/master/D02_Z03_Poisson_distribution.ipynb)

- Rozkład Poissona jest bardzo podobny do rozkładu dwumianowego. Różnica jest subtelna.
- Rozkład dwumianowy sprawdza, ile razy rejestruje się sukces w stosunku do stałej liczby prób, a rozkład Poissona określa, ile razy występuje dyskretne zdarzenie (najczęściej w jakimś ustalonym czasie).
- Nie ma “ustalonej” ilości możliwych sukcesów (parametru  $n$ ). Rozkład Poissona jest określony przez pojedynczy parametr  $\lambda$ .
- Przykłady:

[https://github.com/przem85/statistic\\_4/blob/master/D02\\_Z03\\_Poisson\\_distribution.ipynb](https://github.com/przem85/statistic_4/blob/master/D02_Z03_Poisson_distribution.ipynb)

- Rozkład Poissona jest bardzo podobny do rozkładu dwumianowego. Różnica jest subtelna.
- Rozkład dwumianowy sprawdza, ile razy rejestruje się sukces w stosunku do stałej liczby prób, a rozkład Poissona określa, ile razy występuje dyskretne zdarzenie (najczęściej w jakimś ustalonym czasie).
- Nie ma “ustalonej” ilości możliwych sukcesów (parametru  $n$ ). Rozkład Poissona jest określony przez pojedynczy parametr  $\lambda$ .
- Przykłady:
  - Ile groszy znajdę podczas mojego spaceru do domu?



[https://github.com/przem85/statistic\\_4/blob/master/D02\\_Z03\\_Poisson\\_distribution.ipynb](https://github.com/przem85/statistic_4/blob/master/D02_Z03_Poisson_distribution.ipynb)

- Rozkład Poissona jest bardzo podobny do rozkładu dwumianowego. Różnica jest subtelna.
- Rozkład dwumianowy sprawdza, ile razy rejestruje się sukces w stosunku do stałej liczby prób, a rozkład Poissona określa, ile razy występuje dyskretne zdarzenie (najczęściej w jakimś ustalonym czasie).
- Nie ma “ustalonej” ilości możliwych sukcesów (parametru  $n$ ). Rozkład Poissona jest określony przez pojedynczy parametr  $\lambda$ .
- Przykłady:
  - Ile groszy znajdę podczas mojego spaceru do domu?
  - Ilu dzieci urodzi się dzisiaj w szpitalu?

[https://github.com/przem85/statistic\\_4/blob/master/D02\\_Z03\\_Poisson\\_distribution.ipynb](https://github.com/przem85/statistic_4/blob/master/D02_Z03_Poisson_distribution.ipynb)

- Rozkład Poissona jest bardzo podobny do rozkładu dwumianowego. Różnica jest subtelna.
- Rozkład dwumianowy sprawdza, ile razy rejestruje się sukces w stosunku do stałej liczby prób, a rozkład Poissona określa, ile razy występuje dyskretne zdarzenie (najczęściej w jakimś ustalonym czasie).
- Nie ma “ustalonej” ilości możliwych sukcesów (parametru  $n$ ). Rozkład Poissona jest określony przez pojedynczy parametr  $\lambda$ .
- Przykłady:
  - Ile groszy znajdę podczas mojego spaceru do domu?
  - Ilu dzieci urodzi się dzisiaj w szpitalu?
  - Ile jest dziur na 100 metrowym odcinku drogi?

Zamiast parametru  $p$ , który reprezentuje prawdopodobieństwo sukcesu w jednej próbie Bernoulliego (jak w rozkładzie dwumianowym), tym razem mamy parametr  $\lambda$ , który oznacza “średnią lub przewidywaną” liczbę zdarzeń, które mają wystąpić w naszym eksperymencie.

Rozkład prawdopodobieństwa zmiennej losowej  $X$  o rozkładzie Poissona z parametrem  $\lambda > 0$  wyraża się wzorem:

$$P(X = k) = e^{-\lambda} \cdot \frac{\lambda^k}{k!}.$$

[https://github.com/przem85/bootcamp/blob/master/statistics/D02\\_Z03.ipynb](https://github.com/przem85/bootcamp/blob/master/statistics/D02_Z03.ipynb)

## Zadanie

Proszę napisać skrypt w Pythonie, w którym:

- zdefiniujesz zmienną losową o rozkładzie Poissona  $\lambda = 2$ ,
- narysujesz dla niej gęstość i dystrybuantę,
- wylosujesz próbkę i narysujesz histogram,
- narysujesz kilka gęstości rozkładu Poissona  $\lambda = 1, 4, 10$ ,
- (dla chętnych) policzysz skośność i kurtozę dla gęstości Poissona  $\lambda = 1, 4, 10$ .

# Zmienne losowe o rozkładzie ciągłym

# Zmienne losowe o rozkładzie ciągłym

## Definicja

Zmienna losowa  $X$  ma rozkład ciągły, jeśli istnieje taka funkcja  $f: \mathbb{R} \rightarrow \mathbb{R}$ , że

$$\text{miara}_X(A) = \int_A f(x)dx, \quad A \in \mathcal{B}(\mathbb{R}).$$

Wtedy  $f$  nazywamy gęstością rozkładu  $\text{miara}_X$ .

## Przykład

Wiele pomiarów ma wyniki, który nie są ograniczone do wartości całkowitych/dyskretnych, np. waga osoby może być dowolną liczbą dodatnią.

W tym przypadku krzywa opisująca prawdopodobieństwo dla każdej wartości, to znaczy rozkład prawdopodobieństwa, jest funkcją i nazywamy ją funkcją gęstości prawdopodobieństwa (PDF).

Podobnie jak w przypadku dyskretnym mamy:



$$0 \leq f(x), \quad \forall x \in \mathbb{R}$$



$$\int_{-\infty}^{\infty} f(x) dx = 1.$$



Gęstość zmienna losowa  $X$  o rozkładzie jednostajnym na odcinku  $[a, b]$  ( $a < b$  oraz  $a, b \in \mathbb{R}$ ) jest dana przez:

$$\chi_{[a,b]}(x) = \begin{cases} \frac{1}{b-a} & \text{gdy } x \in [a, b] \\ 0 & \text{gdy } x \notin [a, b] \end{cases}$$

Przykład w Jupyter

[https://github.com/przem85/statistic\\_4/blob/master/D02\\_Z04\\_uniform\\_distribution.ipynb](https://github.com/przem85/statistic_4/blob/master/D02_Z04_uniform_distribution.ipynb)

Rozkład normalny to najważniejszy rozkład prawdopodobieństwa. Wynika to z faktu, że średnie wartości wszystkich rozkładów przybliża rozkład normalny.

Gęstość zmienna losowa  $X$  o rozkładzie normalnym z parametrami  $\mu$  i  $\sigma$  jest dana przez:

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

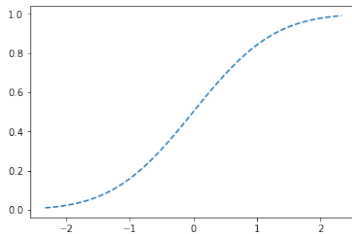
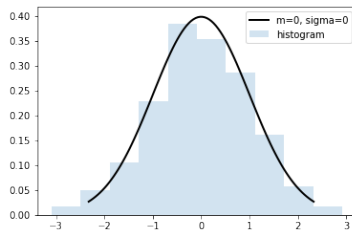
[https://github.com/przem85/statistic\\_4/blob/master/D02\\_Z05\\_normal\\_distribution.ipynb](https://github.com/przem85/statistic_4/blob/master/D02_Z05_normal_distribution.ipynb)

## Zadanie

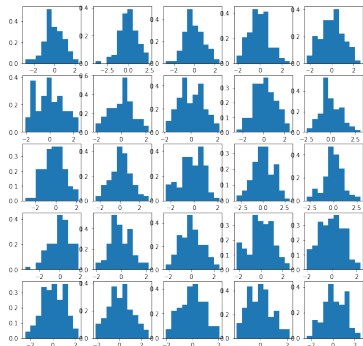
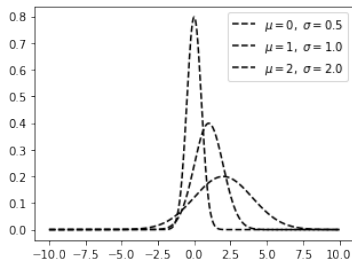
Proszę napisać skrypt w Pythonie, w którym:

- zdefiniujesz zmienną losową o rozkładzie normalnym  $\mu = 0$ ,  $\sigma = 1$ ,
- narysujesz dla niej gęstość i dystrybuantę,
- wylosujesz próbkę i narysujesz histogram (na jednym rysunku),
- narysujesz kilka gęstości rozkładu normalnego z różnymi parametrami,
- wylosujesz kilka próbek dla zmiennej losowej o rozkładzie normalnym  $\mu = 0$ ,  $\sigma = 1$ . (Czemu się od siebie różnią?),
- policzysz skośność i kurtozę dla gęstości Poissona  $\mu = 0$ ,  $\sigma = 1$ .

# Rozkład normalny, Gaussa



# Rozkład normalny, Gaussa



## Definicja

Niech będzie dana przestrzeń probabilistyczna  $(\Omega, \mathcal{F}, P)$ . Dystrybuantą zmiennej losowej  $X: \Omega \rightarrow \mathbb{R}$  nazywamy funkcję  $F_X: \mathbb{R} \rightarrow \mathbb{R}$ , określoną zależnością :

$$F_X(t) = P(X \leq t).$$

# Dystrybuanta zmiennej losowej

## Definicja

Niech będzie dana przestrzeń probabilistyczna  $(\Omega, \mathcal{F}, P)$ . Dystrybuantą zmiennej losowej  $X: \Omega \rightarrow \mathbb{R}$  nazywamy funkcję  $F_X: \mathbb{R} \rightarrow \mathbb{R}$ , określoną zależnością :

$$F_X(t) = P(X \leq t).$$

## Uwaga

Dystrybuanta  $F_X$  zmiennej losowej  $X$  ma następujące własności:

# Dystrybuanta zmiennej losowej

## Definicja

Niech będzie dana przestrzeń probabilistyczna  $(\Omega, \mathcal{F}, P)$ . Dystrybuantą zmiennej losowej  $X: \Omega \rightarrow \mathbb{R}$  nazywamy funkcję  $F_X: \mathbb{R} \rightarrow \mathbb{R}$ , określoną zależnością :

$$F_X(t) = P(X \leq t).$$

## Uwaga

Dystrybuanta  $F_X$  zmiennej losowej  $X$  ma następujące własności:

- a)  $F_X$  jest nie malejąca,



# Dystrybuanta zmiennej losowej

## Definicja

Niech będzie dana przestrzeń probabilistyczna  $(\Omega, \mathcal{F}, P)$ . Dystrybuantą zmiennej losowej  $X: \Omega \rightarrow \mathbb{R}$  nazywamy funkcję  $F_X: \mathbb{R} \rightarrow \mathbb{R}$ , określoną zależnością :

$$F_X(t) = P(X \leq t).$$

## Uwaga

Dystrybuanta  $F_X$  zmiennej losowej  $X$  ma następujące własności:

- a)  $F_X$  jest nie malejąca,
- b)  $\lim_{t \rightarrow \infty} F_X(t) = 1,$

## Definicja

Niech będzie dana przestrzeń probabilistyczna  $(\Omega, \mathcal{F}, P)$ . Dystrybuantą zmiennej losowej  $X: \Omega \rightarrow \mathbb{R}$  nazywamy funkcję  $F_X: \mathbb{R} \rightarrow \mathbb{R}$ , określoną zależnością :

$$F_X(t) = P(X \leq t).$$

## Uwaga

Dystrybuanta  $F_X$  zmiennej losowej  $X$  ma następujące własności:

- a)  $F_X$  jest nie malejąca,
- b)  $\lim_{t \rightarrow \infty} F_X(t) = 1$ ,
- c)  $\lim_{t \rightarrow -\infty} F_X(t) = 0$ ,

## Definicja

Niech będzie dana przestrzeń probabilistyczna  $(\Omega, \mathcal{F}, P)$ . Dystrybuantą zmiennej losowej  $X: \Omega \rightarrow \mathbb{R}$  nazywamy funkcję  $F_X: \mathbb{R} \rightarrow \mathbb{R}$ , określoną zależnością :

$$F_X(t) = P(X \leq t).$$

## Uwaga

Dystrybuanta  $F_X$  zmiennej losowej  $X$  ma następujące własności:

- a)  $F_X$  jest nie malejąca,
- b)  $\lim_{t \rightarrow \infty} F_X(t) = 1$ ,
- c)  $\lim_{t \rightarrow -\infty} F_X(t) = 0$ ,
- d)  $F_X$  jest prawostronnie ciągła

## Uwaga

Jeżeli  $f_X: \mathbb{R} \rightarrow [0, +\infty]$  jest gęstością ciągłej zmiennej losowej  $X$  to:

$$\int_{-\infty}^x f_X(t) dt = P((-\infty, x]) = F_X(x),$$

gdzie  $F_X$  jest dystrybuantą zmiennej losowej  $X$ .

# Dystrybuanta zmiennej losowej

## Uwaga

Jeżeli  $f_X: \mathbb{R} \rightarrow [0, +\infty]$  jest gęstością ciągłej zmiennej losowej  $X$  to:

$$\int_{-\infty}^x f_X(t) dt = P((-\infty, x]) = F_X(x),$$

gdzie  $F_X$  jest dystrybuantą zmiennej losowej  $X$ .

## Uwaga

Jeśli  $F_X$  jest dystrybuantą to jest ona prawie wszędzie różniczkowalna oraz jeśli  $F'_X$  (określona prawie wszędzie) jest prawie wszędzie różna od zera, to jest ona gęstością:

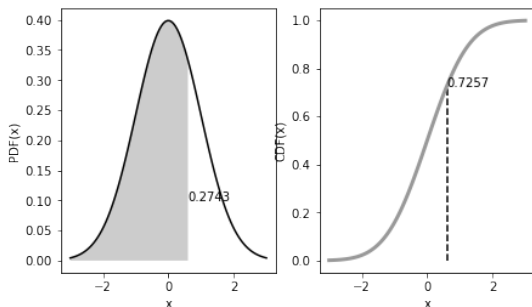
$$F'_X(x) = f(x).$$

# Dystrybuanta zmiennej losowej

[https://github.com/przem85/statistic\\_4/blob/master/D02\\_Z06\\_pdf\\_cdf.ipynb](https://github.com/przem85/statistic_4/blob/master/D02_Z06_pdf_cdf.ipynb)

Zgodnie z naszymi wzorami:

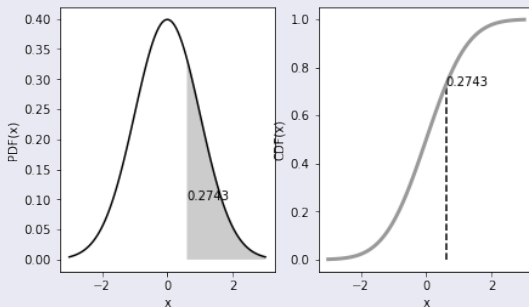
$$P(X \in [-\infty, x_0]) = \int_{-\infty}^{x_0} f_X(x) dx = P(X \leq x_0) = F_X(x_0)$$



## Zadanie 1

Napisz skrypt, który będzie liczył prawdopodobieństwo:

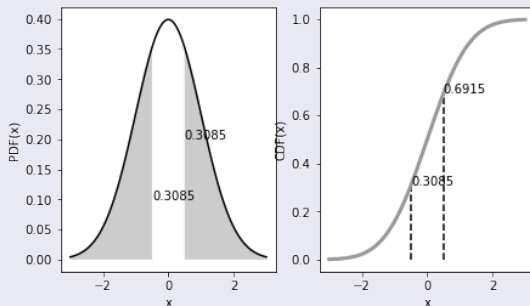
$$P(X \in [x_0, \infty]) = \int_{x_1}^{\infty} f_X(x) dx = P(X \geq x_0) = 1 - F_X(x_1)$$



## Zadanie 2

Napisz skrypt, który będzie liczył prawdopodobieństwo:

$$\begin{aligned} P(X \in [-\infty, x_1] \cup [x_2, \infty]) &= \int_{-\infty}^{x_1} f_X(x) dx + \int_{x_2}^{\infty} f_X(x) dx \\ &= P(X \leq x_2 \text{ or } X \geq x_2) = F_X(x_2) + 1 - F_X(x_1) \end{aligned}$$





[https://github.com/przem85/statistic\\_4/blob/master/D02\\_Z07\\_cdf.ipynb](https://github.com/przem85/statistic_4/blob/master/D02_Z07_cdf.ipynb)

## Zadanie 3

Niech  $X$  będzie zmienną losową o rozkładzie  $N(\mu = 0, \sigma^2 = 1)$ .

Obliczyć:

- $P(X \leq -0.4)$ ,
- $P(X \in (-0.2, 0.6))$ ,
- $P(X \geq -0.2)$ ,
- $P(|X| \leq 1)$ .

Rozwiązanie

- $P(X \leq -0.4) = CDF(-0.4) = 1 - CDF(0.4) = 1 - 0.6554 = 0.3446.$

Rozwiązanie

- $P(X \leq -0.4) = CDF(-0.4) = 1 - CDF(0.4) = 1 - 0.6554 = 0.3446$ .
- $P(-0.2 < X < 0.6) = CDF(0.6) - CDF(-0.2) =$

## Rozwiązanie

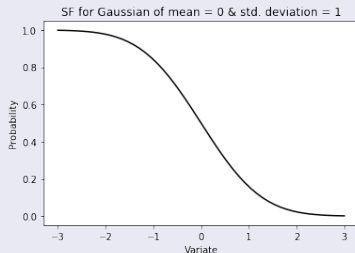
- $P(X \leq -0.4) = CDF(-0.4) = 1 - CDF(0.4) = 1 - 0.6554 = 0.3446$ .
- $P(-0.2 < X < 0.6) = CDF(0.6) - CDF(-0.2) =$   
 $= CDF(0.6) - (1 - CDF(0.2)) = CDF(0.6) - 1 + CDF(0.2) =$   
 $0.7257 - 1 + 0.5793 = 0.305$

# Survival Function

## Survival Function

Jak widzimy czasami w obliczeniach przydaje się funkcja  $1 - CDF(x)$ , którą nazywa się Survival Function

$$SF(x) = 1 - CDF(x)$$



## Rozwiązanie

- $P(X \leq -0.4) = CDF(-0.4) = 1 - CDF(0.4) = 1 - 0.6554 = 0.3446$ .
- $P(-0.2 < X < 0.6) = CDF(0.6) - CDF(-0.2) =$   
 $= CDF(0.6) - (1 - CDF(0.2)) = CDF(0.6) - 1 + CDF(0.2) =$   
 $0.7257 - 1 + 0.5793 = 0.305$

## Rozwiązanie

- $P(X \leq -0.4) = CDF(-0.4) = 1 - CDF(0.4) = 1 - 0.6554 = 0.3446.$
- $P(-0.2 < X < 0.6) = CDF(0.6) - CDF(-0.2) =$   
 $= CDF(0.6) - (1 - CDF(0.2)) = CDF(0.6) - 1 + CDF(0.2) =$   
 $0.7257 - 1 + 0.5793 = 0.305$
- $P(X \geq -0.2) = 1 - P(X \leq -0.2) = 1 - (1 - CDF(0.2)) =$   
 $1 - 1 + CDF(0.2) = CDF(0.2) = 0.5793.$

## Rozwiązanie

- $P(X \leq -0.4) = CDF(-0.4) = 1 - CDF(0.4) = 1 - 0.6554 = 0.3446.$
- $P(-0.2 < X < 0.6) = CDF(0.6) - CDF(-0.2) =$   
 $= CDF(0.6) - (1 - CDF(0.2)) = CDF(0.6) - 1 + CDF(0.2) =$   
 $0.7257 - 1 + 0.5793 = 0.305$
- $P(X \geq -0.2) = 1 - P(X \leq -0.2) = 1 - (1 - CDF(0.2)) =$   
 $1 - 1 + CDF(0.2) = CDF(0.2) = 0.5793.$
- $P(-1 \leq X \leq 1) = CDF(1) - CDF(-1) = CDF(1) - (1 - CDF(1)) =$   
 $CDF(1) - 1 + CDF(1) = 0.8413 - 1 + 0.8413 = 0.6826.$



## Definicja Nadzieja matematyczna (wartością oczekiwaną)

Niech  $(\Omega, \Sigma, P)$  będzie przestrzenią probabilistyczną, zaś  $X: \Omega \rightarrow \mathbb{R}$  - zmienną losową o rozkładzie dyskretnym:

$$P(X = x_i) = p_i, \quad i = 1, \dots, N, \quad N \leq \infty.$$

Nadzieję matematyczną nazywamy liczbę:

$$m = \mathbb{E}(X) = \mathbb{E}X = \sum_{i=1}^N x_i p_i.$$

Niech  $(\Omega, \Sigma, P)$  będzie przestrzenią probabilistyczną, zaś  $X: \Omega \rightarrow \mathbb{R}$  - zmienną losową o rozkładzie ciągłym z gęstością  $f$ , wtedy:

$$m = \mathbb{E}(X) = \mathbb{E}X = \int_{-\infty}^{\infty} x f(x) dx.$$

## Definicja Wariancja i odchylenie standardowe

Niech  $(\Omega, \Sigma, P)$  będzie przestrzenią probabilistyczną, zaś  $X: \Omega \rightarrow \mathbb{R}$  - zmienną losową, posiadającą skończoną wartość oczekiwaną  $m = \mathbb{E}(X)$ . Wariancją zmiennej losowej  $X$  nazywamy liczbę:

$$\sigma^2 = \mathbb{D}^2(X) = \mathbb{D}^2 X = \mathbb{E}((X - m)^2),$$

natomiast liczbę:

$$\sigma = \sqrt{\mathbb{D}^2(X)} = \sqrt{\mathbb{D}^2 X}$$

nazywamy odchyleniem standardowym zmiennej  $X$ .

## Uwaga

W przypadku zmiennej losowej o rozkładzie dyskretnym wariancję obliczamy ze wzoru:

$$\mathbb{D}^2(X) = \sum_{i=1}^N (x_i - m)^2 p_i.$$

## Uwaga

W przypadku zmiennej losowej o rozkładzie ciągłym wariancję obliczamy ze wzoru:

$$\mathbb{D}^2(X) = \int_{-\infty}^{\infty} (x - m)^2 f(x) dx.$$

Obliczymy wartość oczekiwaną zmiennej losowej o rozkładzie jednostajnym na przedziale o końcach  $a$  i  $b$ .

Otrzymujemy:

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} xf(x) dx = \int_a^b x \frac{1}{b-a} dx = \frac{1}{b-a} \int_a^b x dx = \frac{a+b}{2}.$$

Obliczymy wariancję zmiennej losowej o rozkładzie jednostajnym na przedziale o końcach  $a$  i  $b$ .

Wiemy już, że  $m = \mathbb{E}(X) = \frac{a+b}{2}$ . Mamy więc:

$$\mathbb{D}^2(X) = \int_{-\infty}^{\infty} (x - m)^2 f(x) dx = \frac{1}{b-a} \int_a^b \left(x - \frac{a+b}{2}\right)^2 dx = \frac{(b-a)^2}{12}.$$

[https://github.com/przem85/statistic\\_4/blob/master/D02\\_Z08\\_ppf.ipynb](https://github.com/przem85/statistic_4/blob/master/D02_Z08_ppf.ipynb)

Dla każdej dystrybuanty  $F$ , a więc też dla każdej zmiennej losowej, określa się tak zwany kwantyl rzędu  $p$ , gdzie  $0 < p < 1$ . Jest to liczba:

$$q_p = \min\{x : F(x) \geq p\}.$$

W przypadku gdy dystrybuanta jest funkcją odwracalną, określenie kwantyla znacznie się upraszcza:

$$q_p = F^{-1}(p).$$

W przypadku gdy dystrybuanta jest funkcją odwracalną, określenie kwantyla znacznie się upraszcza:

$$q_p = F^{-1}(p).$$

Wówczas kwantyl ma prostą interpretację w języku zmiennych losowych. Mianowicie:

$$P(X < q_p) = P(X \leq q_p) = F(q_p) = p,$$

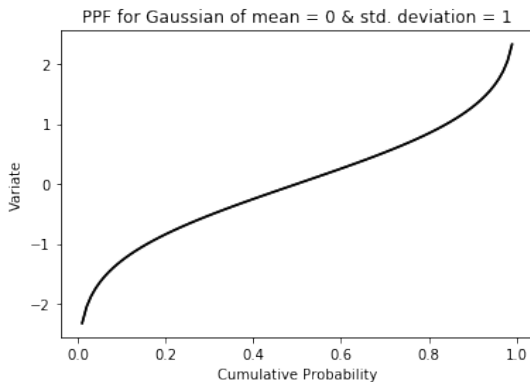
$$P(X > q_p) = 1 - P(X \leq q_p) = 1 - F(q_p) = 1 - p.$$



# Odwrotna do dystrybucyj

Jak widzimy czasami w obliczeniach przydaje się funkcja odwrotna do dystrybucyj  $CDF^{-1}(x)$ , którą nazywa się **Percentile Point Function** (PPF):

$$PPF(x) = CDF^{-1}(x)$$

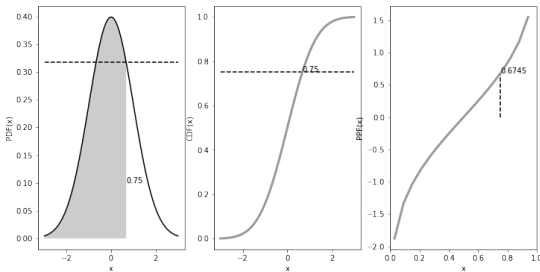


# Zadanie 1

Narysuj na oddzielnych wykresach:

- gęstość rozkładu normalnego,
- dystrybuantę rozkładu normalnego,
- funkcję odwrotną do dystrybuanty.

i zaznacz na nich odpowiednie wartości tak, by móc odtworzyć poniższy rysunek.



# Inverse Survival Function (ISF):

Pamiętamy, że Survival Function miała postać:

$$SF(x) = 1 - CDF(x).$$

Pamiętamy również, że Percentile Point Function (PPF) odwrotna do dystrybuanty miała postać:

$$PPF(x) = CDF^{-1}(x).$$

# Inverse Survival Function (ISF):

Pamiętamy, że Survival Function miała postać:

$$SF(x) = 1 - CDF(x).$$

Pamiętamy również, że Percentile Point Function (PPF) odwrotna do dystrybuanty miała postać:

$$PPF(x) = CDF^{-1}(x).$$

Funkcja ISF to funkcja odwrotna do  $SF$ :

$$ISF(x) = SF^{-1}(x).$$

# Inverse Survival Function (ISF):

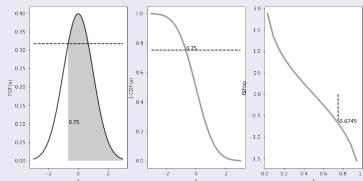
[https://github.com/przem85/statistic\\_4/blob/master/D02\\_Z09\\_isf.ipynb](https://github.com/przem85/statistic_4/blob/master/D02_Z09_isf.ipynb)

## Zadanie (dla chętnych)

Narysuj na oddzielnych wykresach:

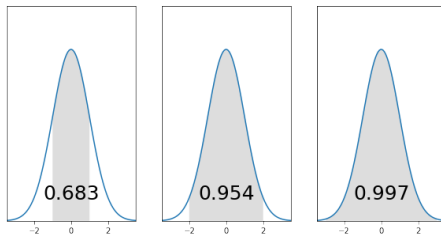
- gęstość rozkładu normalnego,
- dystrybuantę rozkładu normalnego,
- funkcję odwrotną do dystrybuanty.

i zaznacz na nich odpowiednie wartości tak, by móc odtworzyć poniższy rysunek.



# Rozkład normalny, Gaussa

Reguła Trzech Sigm dla danego rozkładu normalnego  $N(\mu, \sigma)$  oznacza, że w przedziale  $[\mu - 3\sigma, \mu + 3\sigma]$  znajduje się 99.7% wszystkich obserwacji.



[https://github.com/przem85/statistic\\_4/blob/master/D02\\_Z10\\_three\\_standard\\_deviations.ipynb](https://github.com/przem85/statistic_4/blob/master/D02_Z10_three_standard_deviations.ipynb)

## Zadanie

Napisz program sprawdzający regułę trzech sigm.

# Suma zmiennych losowych o rozkładzie normalnym

## Własność

Jeżeli zmienne losowe  $X_1, X_2, \dots, X_n$  są niezależne i zmienna  $X_i$  ma rozkład  $N(m_i, \sigma_i^2)$  (dla  $i = 1, 2, \dots, n$ ), to zmienna losowa:

$$Y = X_1 + X_2 + \dots + X_n,$$

ma rozkład normalny

$$N\left(\sum_{i=1}^n m_i, \sum_{i=1}^n \sigma_i^2\right).$$

## Zadanie

Niech  $X_1$  ma rozkład  $N(\mu = 10, \sigma^2 = 25)$  oraz  $X_2$  ma rozkład  $N(\mu = 1, \sigma^2 = 9)$  oraz  $X_3$  ma rozkład  $N(\mu = -4, \sigma^2 = 16)$ . Jaki rozkład ma

- a)  $X_1 + X_2 + X_3$ ,
- b)  $2X_1$  (do domu),
- c)  $2X_1 + 3X_2$  (do domu).



## Zadanie

Niech  $X_1$  ma rozkład  $N(\mu = 10, \sigma^2 = 25)$  oraz  $X_2$  ma rozkład  $N(\mu = 1, \sigma^2 = 9)$  oraz  $X_3$  ma rozkład  $N(\mu = -4, \sigma^2 = 16)$ . Jaki rozkład ma

- a)  $X_1 + X_2 + X_3$ ,
- b)  $2X_1$  (do domu),
- c)  $2X_1 + 3X_2$  (do domu).

Rozwiązanie

$$X_1 + X_2 + X_3 \sim N(10 + 1 - 4, 25 + 9 + 16) = N(7, 44).$$

## Centralne Twierdzenie Graniczne (Lindeberga–Lévy'ego)

Niech  $X_1, X_2, \dots, X_n$  będą niezależnymi zmiennymi losowymi o tym samym rozkładzie, wartości średniej  $m = EX$  i wariancji  $0 < \sigma^2 = D^2X < \infty$ .

Wtedy

$$\lim_{n \rightarrow \infty} P\left(\frac{X_1 + X_2 + \dots + X_n - nm}{\sigma\sqrt{n}} < x\right) = \Phi(x),$$

gdzie  $\Phi$  oznacza dystrybuantę standardowego rozkładu normalnego ( $N(0, 1)$ ).

## Zadanie

Zmienne losowe  $X_1, X_2, \dots, X_{60}$  są niezależne o rozkładzie jednostajnym na odcinku  $[1, 3]$ . Niech

$$X = \sum_{k=1}^{60} X_k.$$

Obliczyć przybliżoną wartość wyrażenia  $P(118 < X < 123)$ .

# Centralne twierdzenie graniczne

Korzystamy z CTG (Lindeberga–Lévy'ego).

# Centralne twierdzenie graniczne

Korzystamy z CTG (Lindeberga–Lévy'ego).

Mamy  $n = 60$  oraz ze wzorów na wartość oczekiwaną i wariancję zmiennej losowej oraz rozkładzie jednostajnym na odcinku

$$m = E(X) = \frac{a+b}{2} = 2, \quad \sigma = \sqrt{D^2(X)} = \sqrt{\frac{(b-a)^2}{12}} = \sqrt{\frac{1}{3}}.$$

# Centralne twierdzenie graniczne

Korzystamy z CTG (Lindeberga–Lévy'ego).

Mamy  $n = 60$  oraz ze wzorów na wartość oczekiwaną i wariancję zmiennej losowej oraz rozkładzie jednostajnym na odcinku

$$m = E(X) = \frac{a+b}{2} = 2, \quad \sigma = \sqrt{D^2(X)} = \sqrt{\frac{(b-a)^2}{12}} = \sqrt{\frac{1}{3}}.$$

Mamy

$$\begin{aligned} P(118 < \sum_{i=1}^{100} X_i < 123) &= P\left(\frac{118-120}{\sqrt{60 \cdot \frac{1}{3}}} < \frac{\sum_{i=1}^{100} X_i - nm}{\sigma \sqrt{n}} < \frac{123-120}{\sqrt{60 \cdot \frac{1}{3}}}\right) = \\ &= P\left(\frac{-2}{\sqrt{20}} < Z < \frac{3}{\sqrt{20}}\right) = P(-0.4472 < Z < 0.6708) = \\ &= \Phi(0.6708) - (1 - \Phi(0.4472)) = 0.7488 - 1 + 0.6726 = 0.4214. \end{aligned}$$

## Centralne Twierdzenie Graniczne (Moivre'a–Laplace'a)

Niech  $X_1, X_2, \dots, X_n$  będą niezależnymi zmiennymi losowymi o tym samym rozkładzie, takimi że:

$$S_n = X_1 + X_2 + \dots + X_n \sim \text{Bin}(n, p)$$

czyli rozkład dwumianowy z parametrami  $n$ ,  $p$ ,  $1 - p$ . Wtedy

$$\lim_{n \rightarrow \infty} P \left( \frac{X_1 + X_2 + \dots + X_n - np}{\sqrt{np(1-p)}} < x \right) = \Phi(x),$$

gdzie  $\Phi$  oznacza dystrybuantę standardowego rozkładu normalnego ( $N(0, 1)$ ).

## Zadanie

Prawdopodobieństwo uzyskania wygranej w pewnej grze losowej wynosi 0.1. Obliczyć prawdopodobieństwo, że spośród 500 grających osób wygra więcej, niż 60 osób.



# Centralne twierdzenie graniczne

Korzystamy z CTG (Moivre'a–Laplace'a).

# Centralne twierdzenie graniczne

Korzystamy z CTG (Moivre'a–Laplace'a).  
Mamy:  $n = 500$ ,  $p = 0.1$ ,  $(1 - p) = 0.9$ .

# Centralne twierdzenie graniczne

Korzystamy z CTG (Moivre'a–Laplace'a).

Mamy:  $n = 500$ ,  $p = 0.1$ ,  $(1 - p) = 0.9$ .

Musimy obliczyć:

$$\begin{aligned} P\left(\sum_{i=1}^{500} X_i > 60\right) &= P\left(\frac{\sum_{i=1}^{500} X_i - 50}{\sqrt{50 \cdot 0.9}} > \frac{60 - 50}{\sqrt{50 \cdot 0.9}}\right) = P\left(z > \frac{10}{\sqrt{45}}\right) = P(Z > 1.492) \\ &= 1 - P(Z < 1.492) = 1 - \Phi(1.492) = 1 - 0.93189 = 0.06811. \end{aligned}$$

# Centralne twierdzenie graniczne

[https://github.com/przem85/statistic\\_4/blob/master/D02\\_Z11\\_central\\_limit\\_theorem.ipynb](https://github.com/przem85/statistic_4/blob/master/D02_Z11_central_limit_theorem.ipynb)

Centralne twierdzenie graniczne mówi, że średnią wystarczająco dużej liczby zmiennych losowych o tym samym rozkładzie można przybliżyć za pomocą rozkładu normalnego (trzeba pamiętać o założeniach twierdzenia).

## Zadanie

Wygeneruj próbkę z rozkładu dwumianowego. Następnie podziel dane na zbiory po 2 i po 10 elementów. Policz średnie w zbiorach i stwórz z nich nową próbkę. Narysuj histogram dla próbki wszystkich trzech próbek.



## Definicja

Prostą próbą losową (lub krócej próbą losową) o liczności  $n$  nazywamy **ciąg niezależnych zmiennych losowych**  $X_1, X_2, \dots, X_n$  określonych na przestrzeni zdarzeń elementarnych  $\Omega$  i takich, że każda ze zmiennych ma taki sam rozkład.

## Definicja

Prostą próbą losową (lub krócej próbą losową) o liczności  $n$  nazywamy **ciąg niezależnych zmiennych losowych**  $X_1, X_2, \dots, X_n$  określonych na przestrzeni zdarzeń elementarnych  $\Omega$  i takich, że każda ze zmiennych ma taki sam rozkład.

## Uwaga

Konkretny ciąg wartości  $x_1, x_2, \dots, x_n$  (prostej) próby losowej  $X_1, X_2, \dots, X_n$  nazywamy realizacją (prostej) próby losowej lub próbką.

## Definicja

Prostą próbą losową (lub krócej próbą losową) o licznosci  $n$  nazywamy **ciąg niezależnych zmiennych losowych**  $X_1, X_2, \dots, X_n$  określonych na przestrzeni zdarzeń elementarnych  $\Omega$  i takich, że każda ze zmiennych ma taki sam rozkład.

## Uwaga

Konkretny ciąg wartości  $x_1, x_2, \dots, x_n$  (prostej) próby losowej  $X_1, X_2, \dots, X_n$  nazywamy realizacją (prostej) próby losowej lub próbką.

## Uwaga

Statystyką nazywamy każdą zmienną losową będącą ustaloną funkcją próby losowej  $X_1, X_2, \dots, X_n$ .

O co chodzi z tym ciągiem zmiennych losowych?