Analiza Danych - Podstawy Statystyczne

Statystyka opisowa a wnioskowanie statystyczne

Marek Rupniewski

16 marca 2019

Organizacja zajęć

Organizacja zajęć

- 4 bloki wykładów (2 x 1.5h każdy),
- 4 bloki ćwiczeń laboratoryjnych (2 x 1.5h każde).

Ocena:

- laboratoria: 4 x 5 punktów,
- test końcowy (0.5h na ostatnim wykładzie): 10 punktów.

Ocena końcowa wystawiana wg skali:

15p	18p	21p	24p	27p
2	3	3.5	4	4.5	5

Czym jest statystyka

Co to jest statystyka

Pierwotnie — czynności zbierania i opracowywania danych na potrzeby podejmowania decyzji w sferze spraw publicznych (spisy ludności, rejestry zgonów, itp.).

Obecnie — nauka zajmująca się opisem i analizą zjawisk „masowych” przy użyciu teorii i metod rachunku prawdopodobieństwa.

Czym się statystyka zajmuje

- Statystyka opisowa — metodami opracowywania (gromadzenia, prezentacji) danych („statystycznych”) bez odwoływania się do teorii rachunku prawdopodobieństwa
- Wnioskowanie statystyczne (Statystyka matematyczna) — metodami wyboru prób losowych i reguł decyzyjnych pozwalających na uogólnianie wniosków dotyczących prób na „populację”, z której te próby są wybierane.

Gdzie znajdują zastosowanie metody statystyczne

- w genetyce i bioinformatyce,
- w fizyce (np. kinetyczna teoria gazów),
- w projektowaniu systemów komputerowych (np. teoria kolejek),
- w projektowaniu i analizie systemów telekomunikacyjnych (np. analiza zakłóceń),
- w prognozowaniu pogody,
- w badaniach operacyjnych,
- w naukach aktuarialnych,
- w ekonomii i finansach,
- w naukach społecznych,
- w medycynie,
- ...

Elementy statystyki opisowej

Histogram (wykres słupkowy)

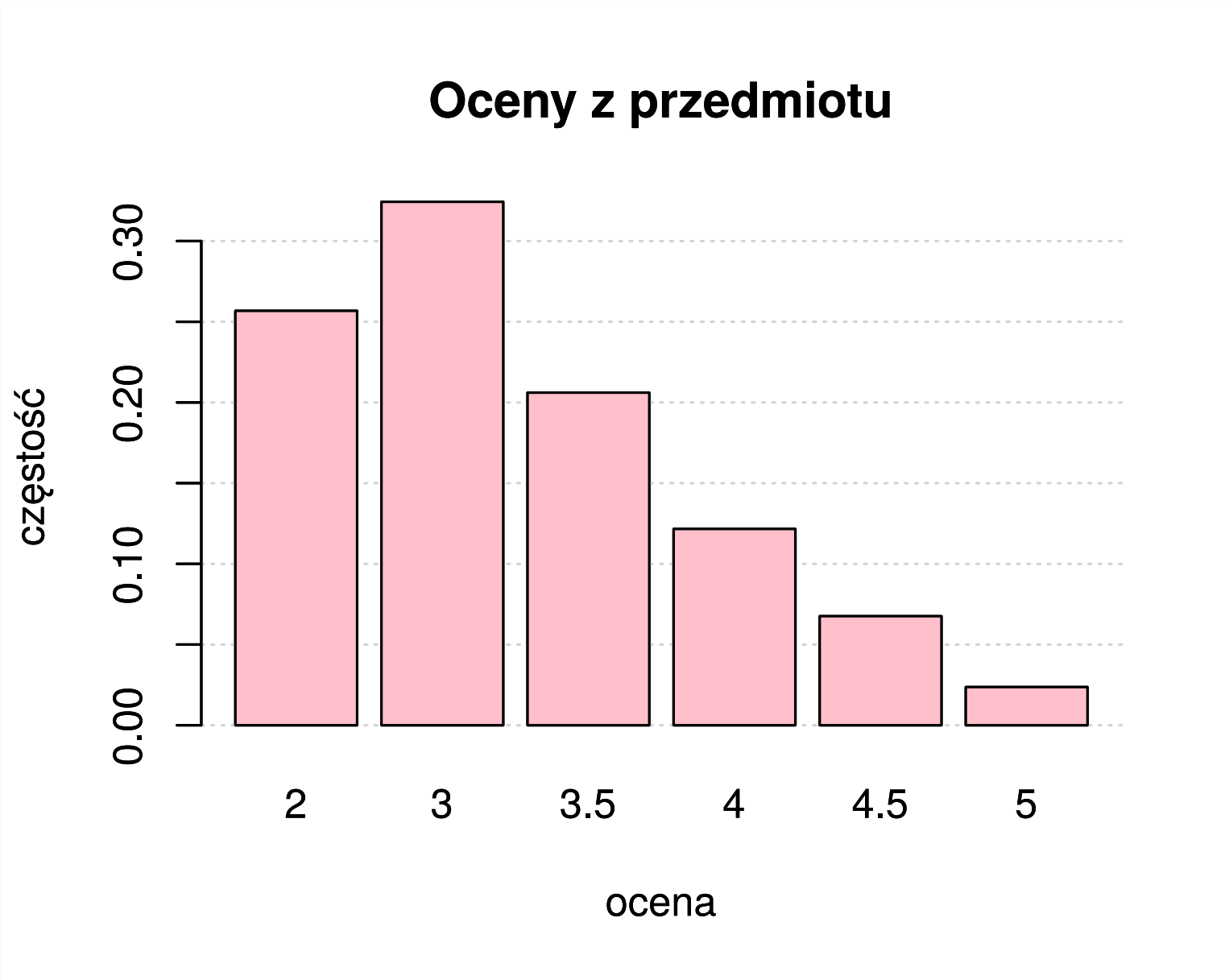


Diagram tortowy — przerost formy nad treścią

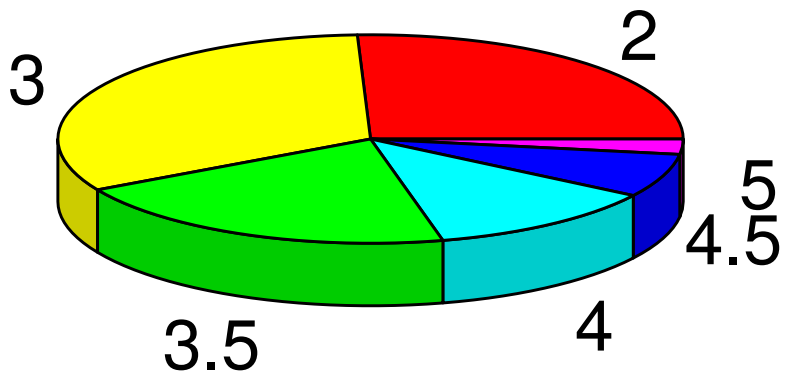
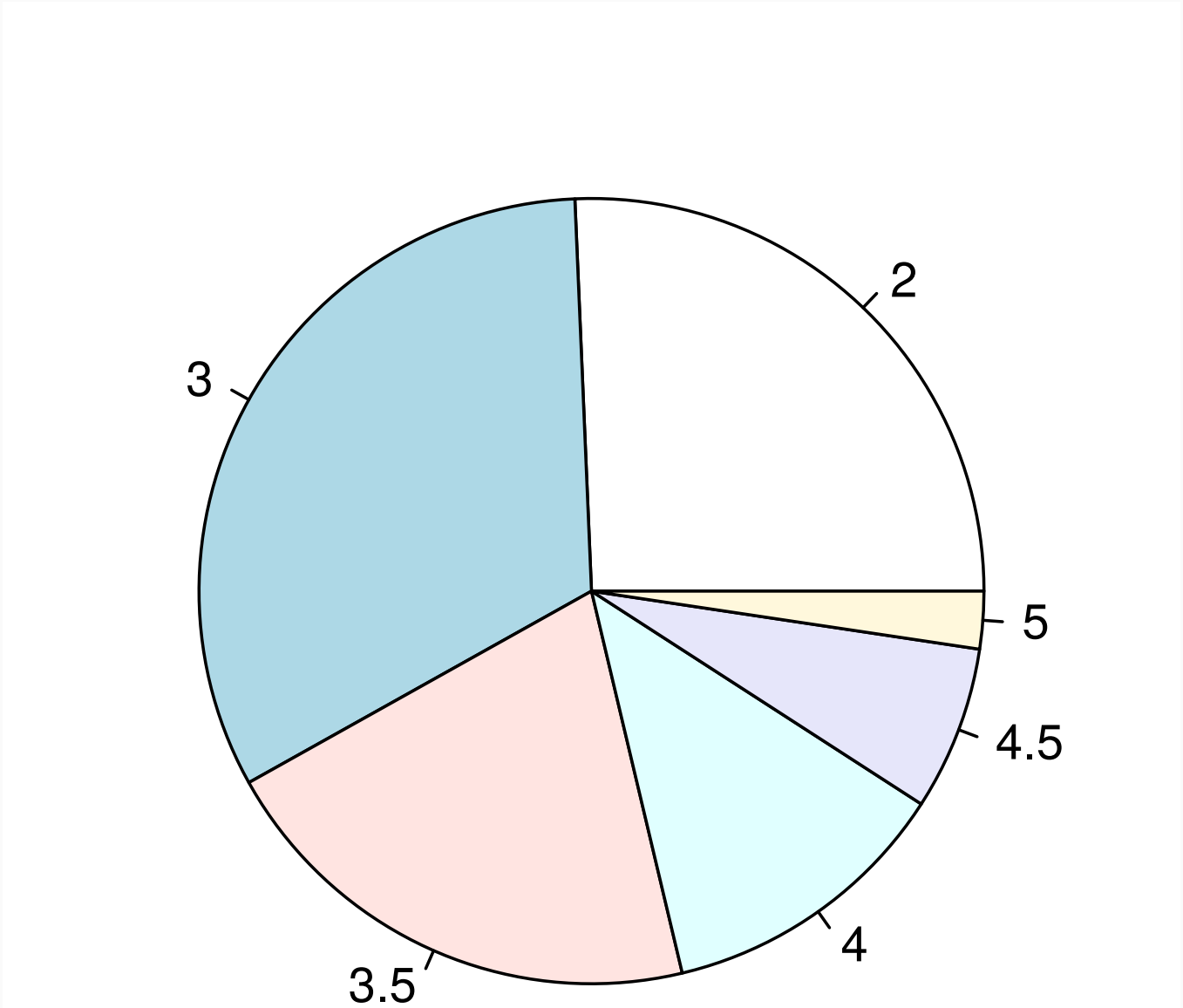
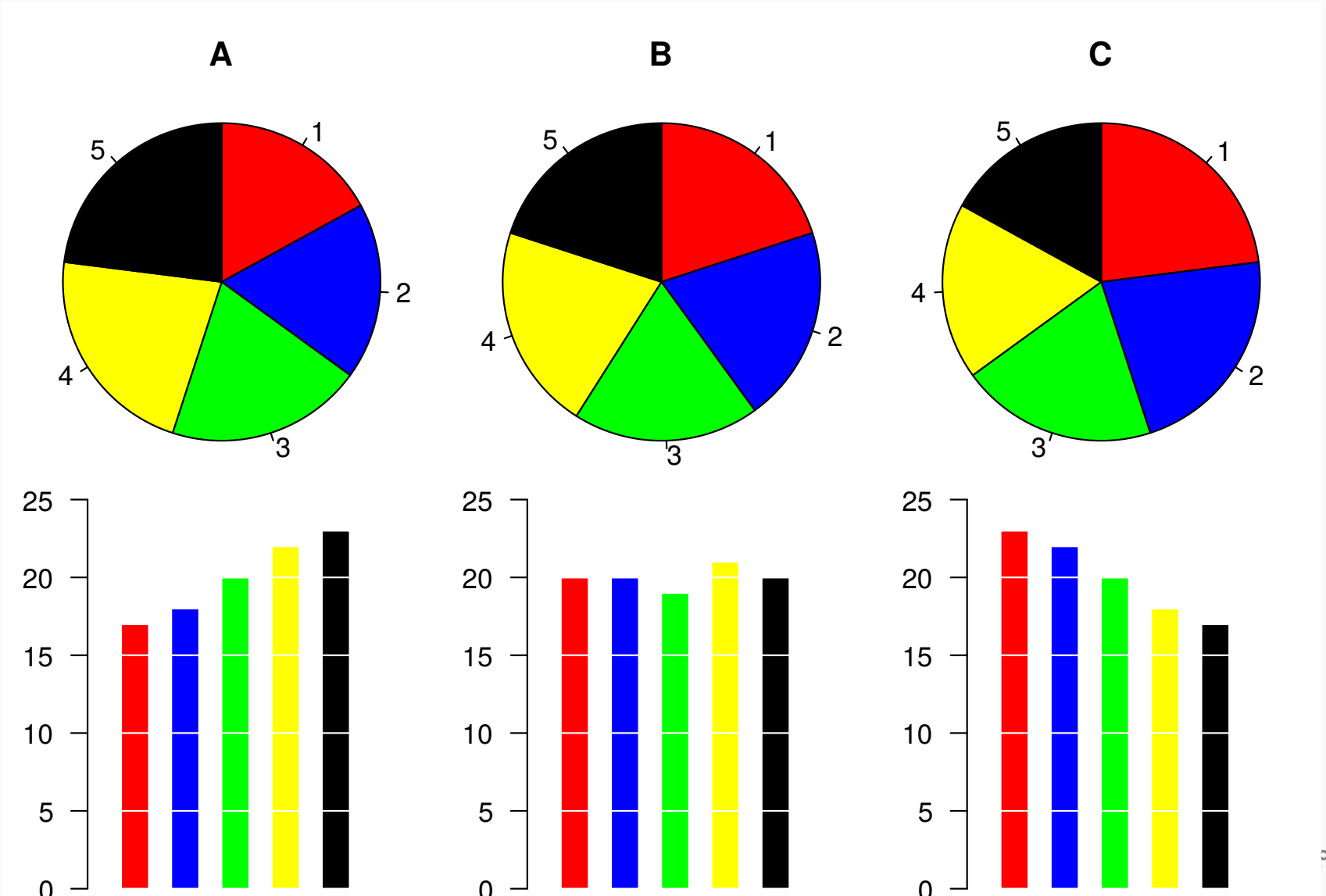


Diagram kołowy

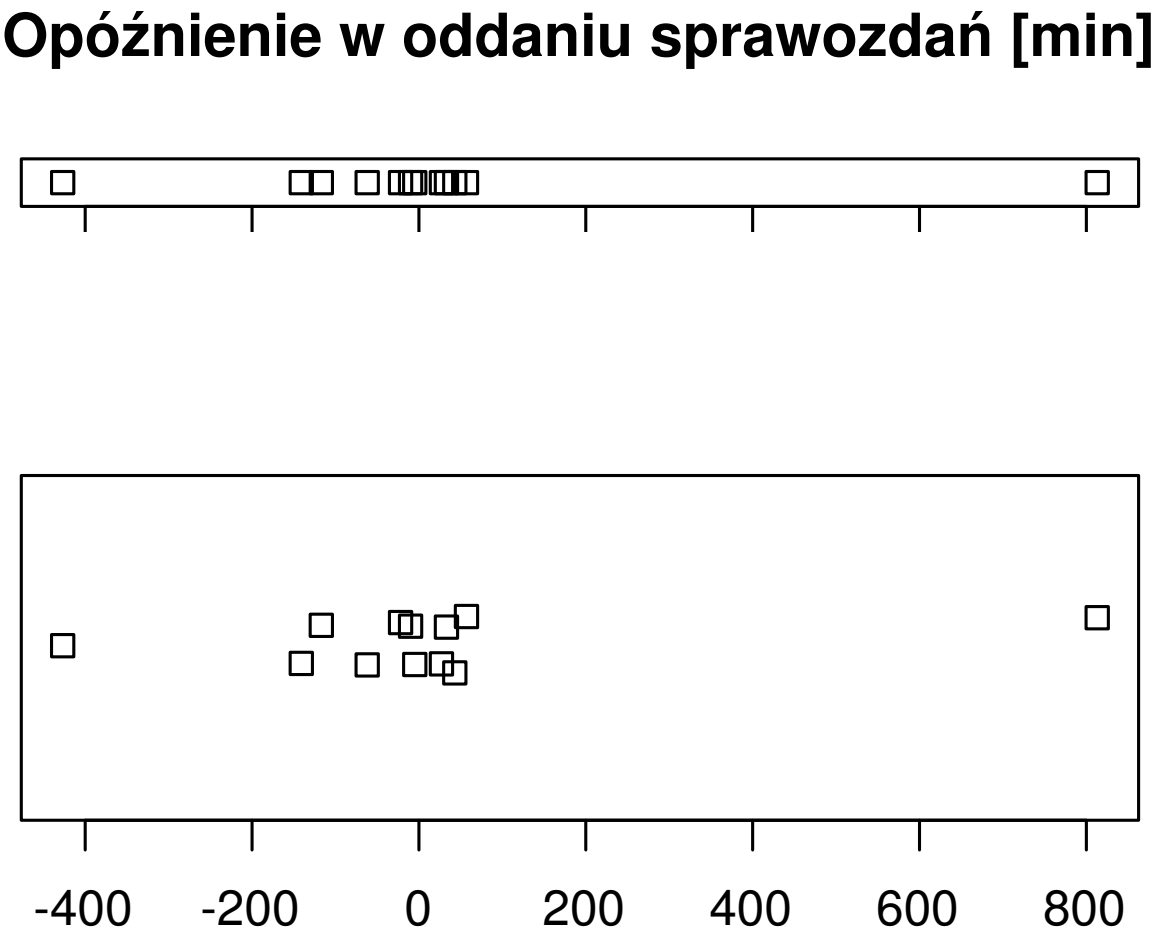


Dlaczego unikać diagramów kołowych

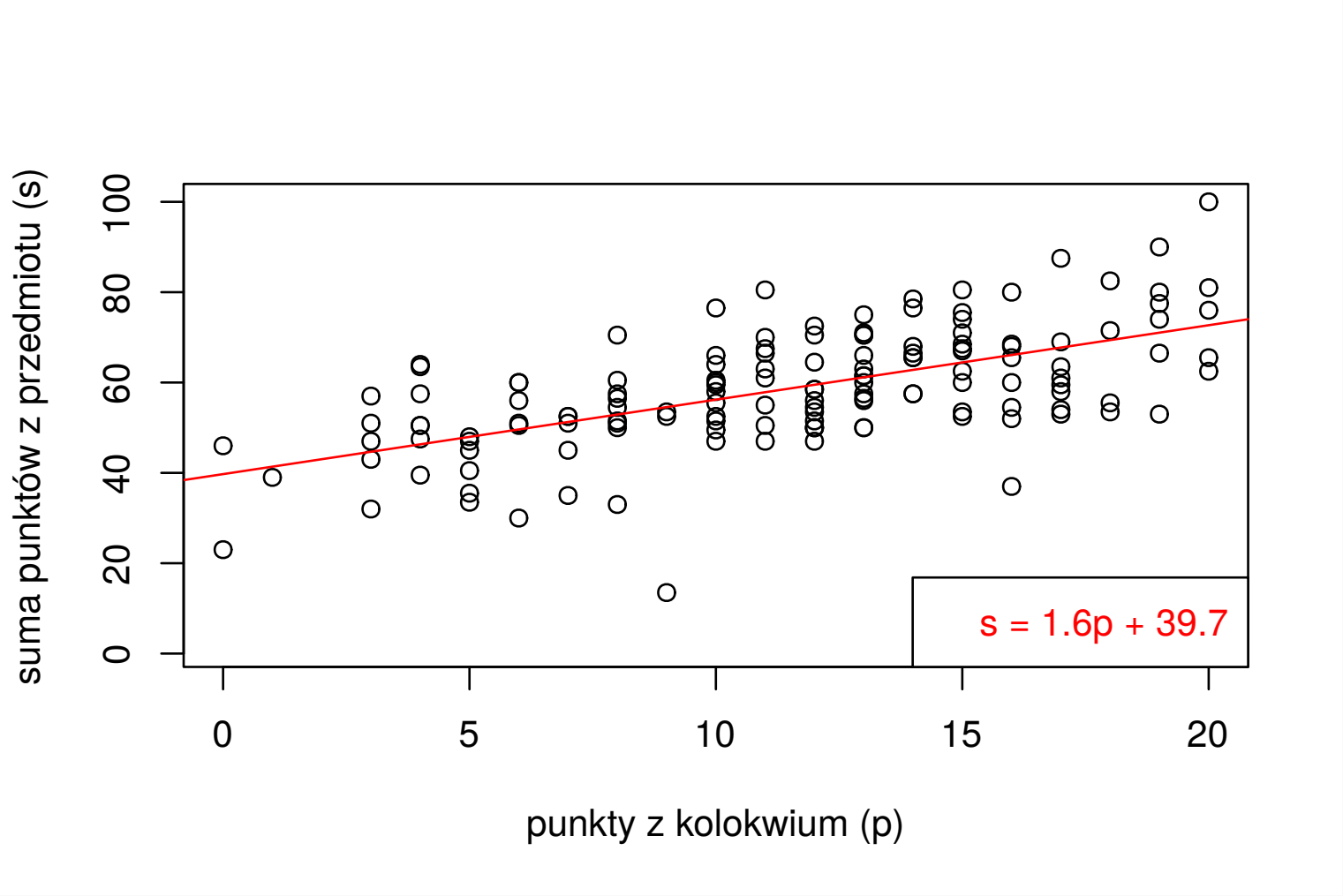
W każdym przypadku uszereguj kolorowe pola według wielkości.



Wykres paskowy



Wykres punktowy



Polecam stronę Fundacji Naukowej SmarterPoland.pl

http://smarterpoland.pl

Paradoks Simpsona

Przyjęci na studia (Univ. of California, Berkeley)

	Zgłoszeń	Przyjętych
Mężczyźni	2691	(1198) 45%
Kobiety	1835	(614) 33%

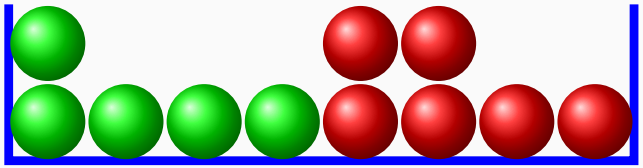
Przyjęci z podziałem na kierunki

	Mężczyźni		Kobiety	
K.	Zgłoszeń	Przyjętych	Zgłoszeń	Przyjętych
A	825	(512) 62%	108	(89) 82%
B	560	(353) 63%	25	(17) 68%
C	325	(120) 37%	593	(219) 37%
D	417	(138) 33%	375	(131) 35%
E	191	(53) 28%	393	(134) 34%
F	373	(22) 6%	341	(24) 7%

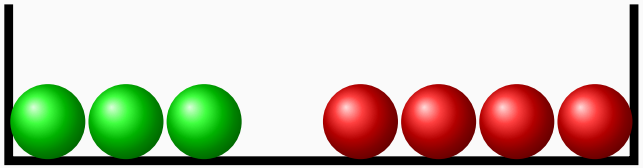
Paradoks Simpsona

Wybieramy koszyk, z koszyka losujemy kulę, wygrywa zielona.

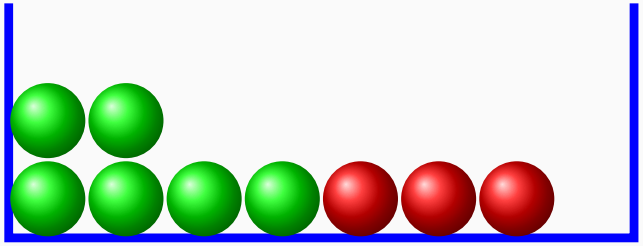
Który koszyk wybrać?



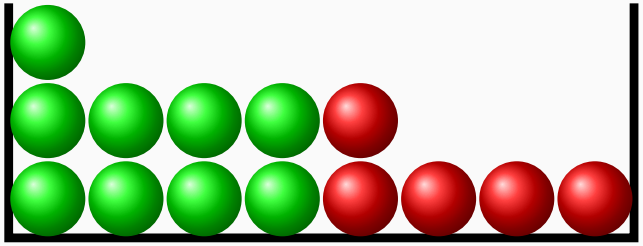
$\frac{5}{11} > \frac{3}{7}$



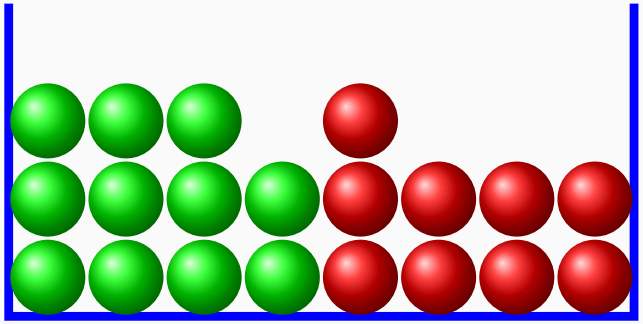
A teraz?

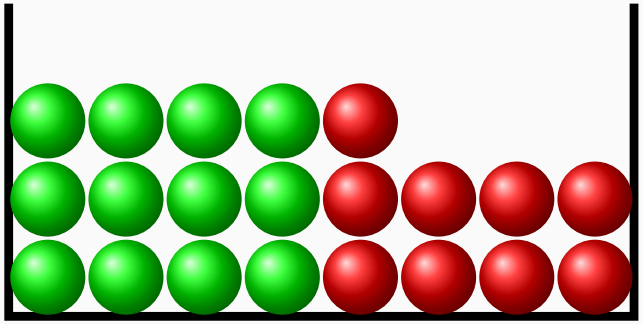


$\frac{6}{9} > \frac{9}{14}$



A jak połączymy zawartość koszyków?





Spektakularne i nieco mniej zastosowania statystyki

Przewidywanie wyników wyborów prezydenckich, USA 1936

Magazyn *Literary Digest* wysyła 10 milionów kartek do głosowania pozyskując adresy z książek telefonicznych i rejestrów właścicieli samochodów.

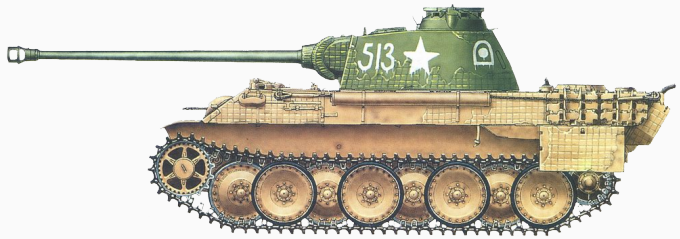
Powraca 2.3 miliona kartek. Wynik 57 – 43 na korzyść Landona.

W tzw. międzyczasie instytut założony przez G. Gallupa na podstawie 5 tysięcy ankiet (wysłanych do odpowiednio wyselekcjonowanej grupy osób) trafnie przewiduje

1. Wynik ankiety *Literary Digest*,
2. Zdecydowaną wygraną w wyborach ... Roosvelta (61 - 37)!.

Wyższość statystyki nad wywiadem wojskowym

An empirical Approach to Economic Intelligence in World War II, Journal of the American Statistical Association, Vol. 42, No. 237 (Mar., 1947), pp. 72–91.



Produkcja czołgów (dane historyczne):

Miesiąc	est. statystyków	est. agentów	dane prod.
Czerwiec 1940	169	1000	122
Czerwiec 1941	244	1550	271
Sierpień 1942	327	1550	342

Jak liczyć gołębie

- Chwytamy (losowo), oznaczamy i wypuszczamy $M = 100$ gołębi.
- Po pewnym czasie chwytamy (losowo) $C = 1000$ gołębi i sprawdzamy, że $R = 10$ z nich jest oznaczonych.
- Estymujemy liczbę N gołębi wg formuły:

$$N \overset{✎}{=} M/\frac{R}{C} = \frac{MC}{R} = 10000.$$

Metoda ta nazywana jest metodą wielokrotnych złowień. Funkcjonują też nazwy: *capture-recapture, capture-mark-recapture (CMR), mark-recapture, sight-resight, multiple systems estimation, band recovery, Petersen method, Lincoln method.*

Powtórka z RP

Prawdopodobieństwo

Ω przestrzeń probabilistyczna, $A \subset \Omega$ zdarzenie.
Prawdopodobieństwo, to funkcja $\mathbb{P}$ przyporządkowująca zdarzeniom liczby rzeczywiste w taki sposób, że:

- $\mathbb{P}(A) \geq 0$ dla każdego zdarzenia A,
- $\mathbb{P}(\Omega) = 1$,
- Dla każdych rozłącznych zdarzeń $A_1, A_2, A_3, \ldots$

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

Własności prawdopodobieństwa

- $\mathbb{P}(\varnothing) = 0$,
- $A \subset B \to \mathbb{P}(A) \leq \mathbb{P}(B)$,
- $0 \leq \mathbb{P}(A) \leq 1$,
- $\mathbb{P}(A^C) = 1 - \mathbb{P}(A)$,
- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

Niezależność zdarzeń

Zdarzenia A i B są niezależne, jeśli

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B).$$

Zdarzenia A_i, $i \in I$ są niezależne, jeśli dla każdego skończonego podzbioru indeksów $J \subset I$:

$$\mathbb{P}\left(\bigcap_{i \in J} A_i\right) = \prod_{i \in J} \mathbb{P}(A_i).$$

Zmienne losowe i ich dystrybuanty

Zmienna losowa to funkcja

$$X\colon \Omega \to E \subset \mathbb{R}.$$

Jej dystrybuanta:

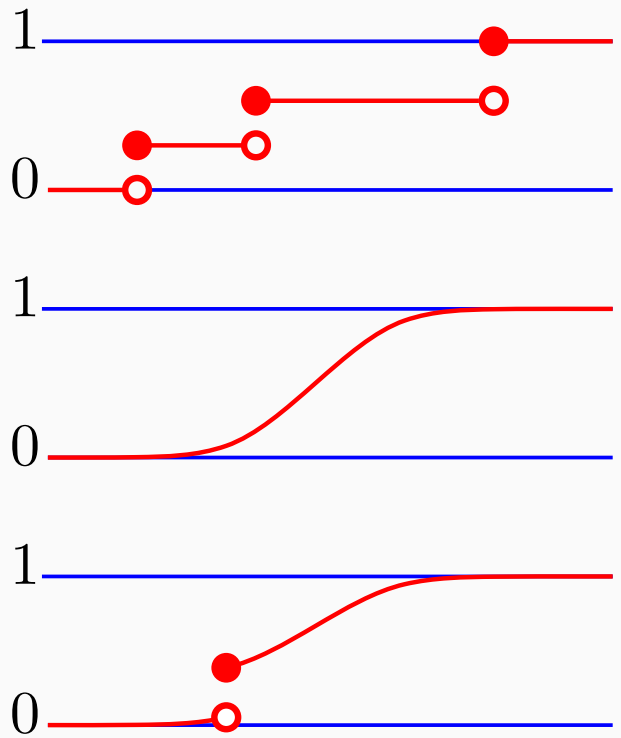
$$F_X\colon \mathbb{R} \to [0,1], \qquad F_X(x) = \mathbb{P}(X \leq x).$$

Duże litery — zmienne losowe, małe litery — ich wartości.

Własności dystrybuant

Każda dystrybuanta $F(x)$ jest funkcją

- niemalejącą,
- dążącą do 1 dla $x \to +\infty$,
- dążącą do 0 dla $x \to -\infty$,
- prawostronnie ciągłą,
- posiadającą lewostronne granice,
- różniczkowalną prawie wszędzie.



Dyskretne zmienne losowe

Zmienna losowa X

$$X: \Omega \to E \subset \mathbb{R}.$$

jest dyskretną zmienną losową, jeśli zbiór E jest co najwyżej przeliczalny ($E = \{x_1, x_2, \ldots, x_N\}$ lub $E = \{x_1, x_2, \ldots\}$).

Funkcja prawdopodobieństwa zmiennej losowej X:

$$f_X: E \to \mathbb{R}, \qquad f_X(x) = \mathbb{P}(X = x).$$

$$F_X(x) = \mathbb{P}(X \leq x) = \sum_{x_i \leq x} f_X(x_i).$$

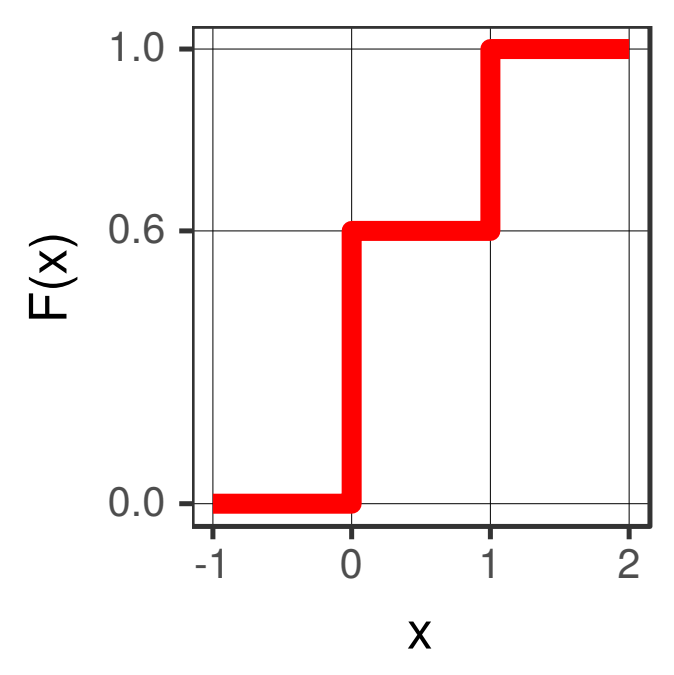
Przykłady dyskretnych zmiennych losowych

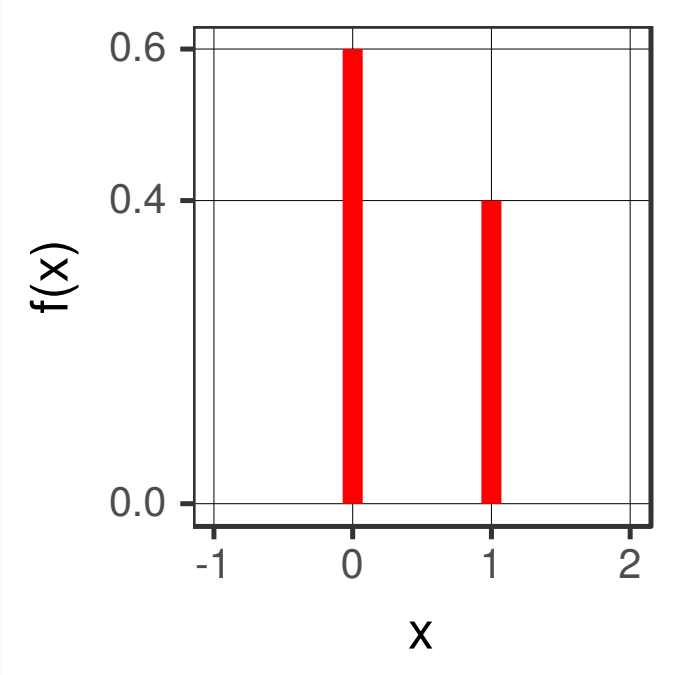
X ma rozkład Bernoulliego (ewent. dwupunktowy) z parametrem p, $X \sim \text{Bern}(p)$, jeśli

$$X : \Omega \to \{0, 1\}$$

oraz

$$\mathbb{P}(X = 1) = p, \qquad \mathbb{P}(X = 0) = 1 - p.$$





Przykłady dyskretnych zmiennych losowych

X ma rozkład dwumianowy (*ang. binomial*) z parametrami n, p, $X \sim \text{Binom}(n, p)$, jeśli

$$X : \Omega \to \{0, 1, \ldots, n\}$$

oraz

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

Suma n niezależnych zmiennych o rozkładzie $\text{Bern}(p)$ ma rozkład $\text{Binom}(n, p)$.

Przykłady dyskretnych zmiennych losowych

X ma rozkład Poissona z parametrem λ, $X \sim \mathrm{Pois}(\lambda)$, jeśli

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

Rozkład Poissona modeluje liczbę zdarzeń, które występują w ciągu jednostkowego odcinka czasu przy założeniu, że

- liczby zdarzeń występujących na rozłącznych podprzedziałach są niezależne,
- prawdopodobieństwa wystąpienia zdarzenia na „krótkim” podprzedziale o długości h szacuje się przez $\lambda h + \mathrm{o}(h)$,
- prawdopodobieństwo wystąpienia więcej niż jednego zdarzenia na „krótkim” podprzedziale o długości h szacuje się przez $\mathrm{o}(h)$.

Ciągłe zmienne losowe

Zmienna losowa $X: \Omega \to E \subset \mathbb{R}$ jest zmienną ciągłą jeśli istnieje funkcja gęstości prawdopodobieństwa $f_X: E \to \mathbb{R}$ taka, że

- $f_X(x) \geq 0$,
- $\int_{-\infty}^{+\infty} f_X(x)\mathrm{d}x = 1$,
- $\mathbb{P}(a < X < b) = \int_a^b f_X(x)\mathrm{d}x, \qquad \forall a \leq b.$

$$F_X(x) = \int_{-\infty}^{x} f_X(x)\mathrm{d}x, \qquad f_X(x) = F_X'(x).$$

Przykłady ciągłych zmiennych losowych

X ma rozkład jednostajny (*ang. uniform*) na przedziale $[a, b]$, $X \sim \text{Unif}([a, b])$, jeśli

$$X : \Omega \to [a, b]$$

oraz

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

Przykłady ciągłych zmiennych losowych

X ma rozkład normalny z parametrami μ, σ^2, $X \sim N(\mu, \sigma^2)$, jeśli

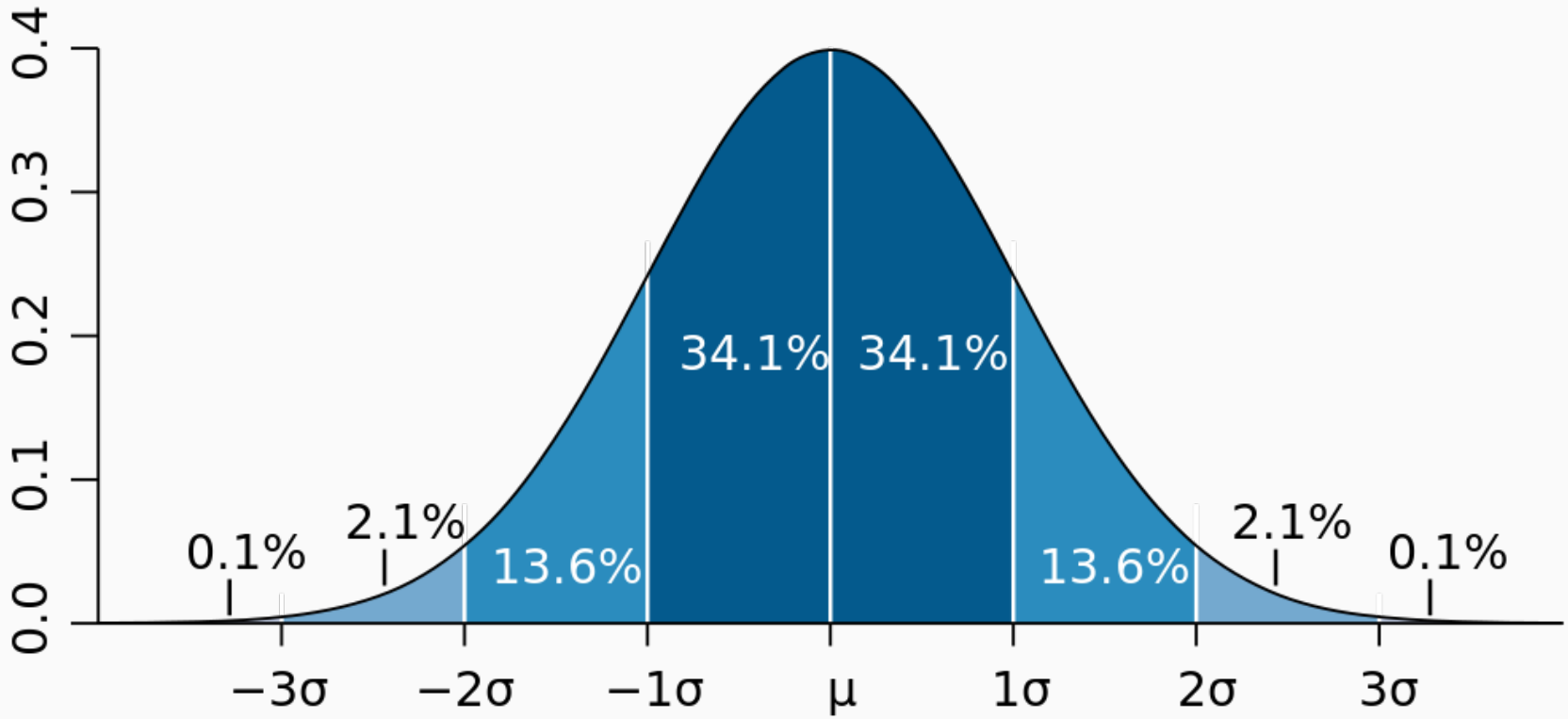
$$X : \Omega \to \mathbb{R}$$

oraz

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

Funkcję gęstości oraz dystrybuantę zmiennej o rozkładzie $N(0,1)$ oznacza się, odpowiednio, literami ϕ oraz Φ, a kwantyl rzędu p — z_p.

Reguła trzech sigm (reguła 68–95–99.7)



Odwrotna dystrybuanta (funkcja kwantylowa)

Jeśli X jest zmienną losową o dystrybuancie F, to odwrotną dystrybuantą tej zmiennej nazywamy funkcję

$$F^{-1}\colon [0,1] \to \mathbb{R}, \quad F^{-1}(q) = \inf\{x\colon F(x) \geq q\}.$$

$F^{-1}(\frac{1}{4})$ pierwszy (dolny) kwartyl rozkładu,
$F^{-1}(\frac{1}{2})$ mediana rozkładu,
$F^{-1}(\frac{3}{4})$ trzeci (górny) kwartyl.

Zmienne wielowymiarowe

Zmienna wielowymiarowa, to zmienna postaci:

$$X = (X_1, \ldots, X_k) : \Omega \to E \subset \mathbb{R}^k, \quad k > 1.$$

Podobnie jak w przypadku jednowymiarowym definiujemy funkcje prawdopodobieństwa (dla zm. dyskretnych) i funkcje gęstości prawdopodobieństwa (dla zm. ciągłych).

Niezależność zmiennych losowych

Zmienne losowe X i Y nazywane są niezależnymi, jeśli

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B), \qquad \forall A, B.$$

Jeśli $X_1, \ldots, X_N$ zmienne losowe z łączną funkcją prawdopodobieństwa (lub gęstości prawdopodobieństwa) f, to zmienne te są niezależne, jeśli

$$f(x_1, \ldots, x_N) = \prod_{i=1}^{N} f_{X_i}(x_i), \qquad \forall x_1, \ldots, x_N.$$

Próba losowa

Definicja

Jeśli $X_1, \ldots, X_N$ są niezależnymi zmiennymi losowymi o tym samym rozkładzie, to wektor $(X_1, \ldots, X_N)$ nazywamy próbą losową rozmiaru N z tego rozkładu.

Jeśli $X_1, \ldots, X_N$ próba losowa i X_i zadane dystrybuantą F lub funkcją (gęstości) prawdopodobieństwa f, to piszemy także

$$X_1, \ldots, X_N \sim F \text{ lub } X_1, \ldots, X_N \sim f$$

Statystyka jako funkcja

Definicja

Statystyka to dowolna funkcja próby losowej:

$$T\text{: próba losowa} \mapsto x \in \mathbb{R}$$

Przykłady

- Średnia (arytmetyczna): $(x_1, x_2, \ldots, x_n) \mapsto \frac{x_1+x_2+\cdots+x_n}{n}$,
- Średnia geometryczna: $(x_1, x_2, \ldots, x_n) \mapsto \sqrt[n]{x_1 x_2 \ldots x_n}$,
- Średnia harmoniczna: $(x_1, x_2, \ldots, x_n) \mapsto \frac{n}{\frac{1}{x_1}+\frac{1}{x_2}+\cdots+\frac{1}{x_n}}$,
- Mediana, kwantyle, kwartyle, percentyle,
- Statystyki pozycyjne.

Przykład statystyk

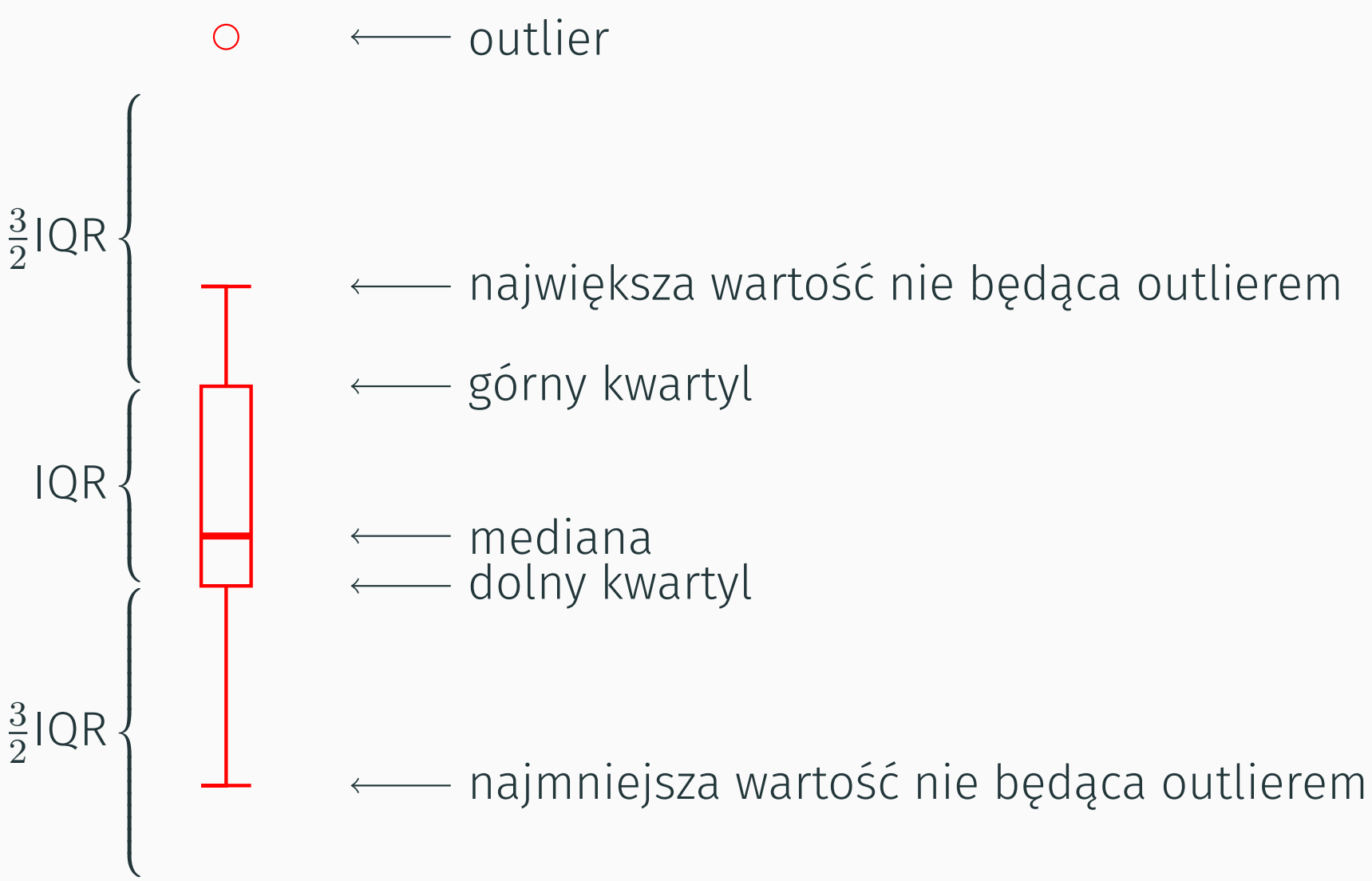
Badamy jak w ciągu roku zmieniła się cena dwóch różnych produktów, które początkowo kosztowały 100 [zł].

Jeden z nich podrożał dwukrotnie i kosztuje 200, a cena drugiego dwukrotnie spadła (50).

Czy łączny koszt tych artykułów wzrósł, czy zmalał?

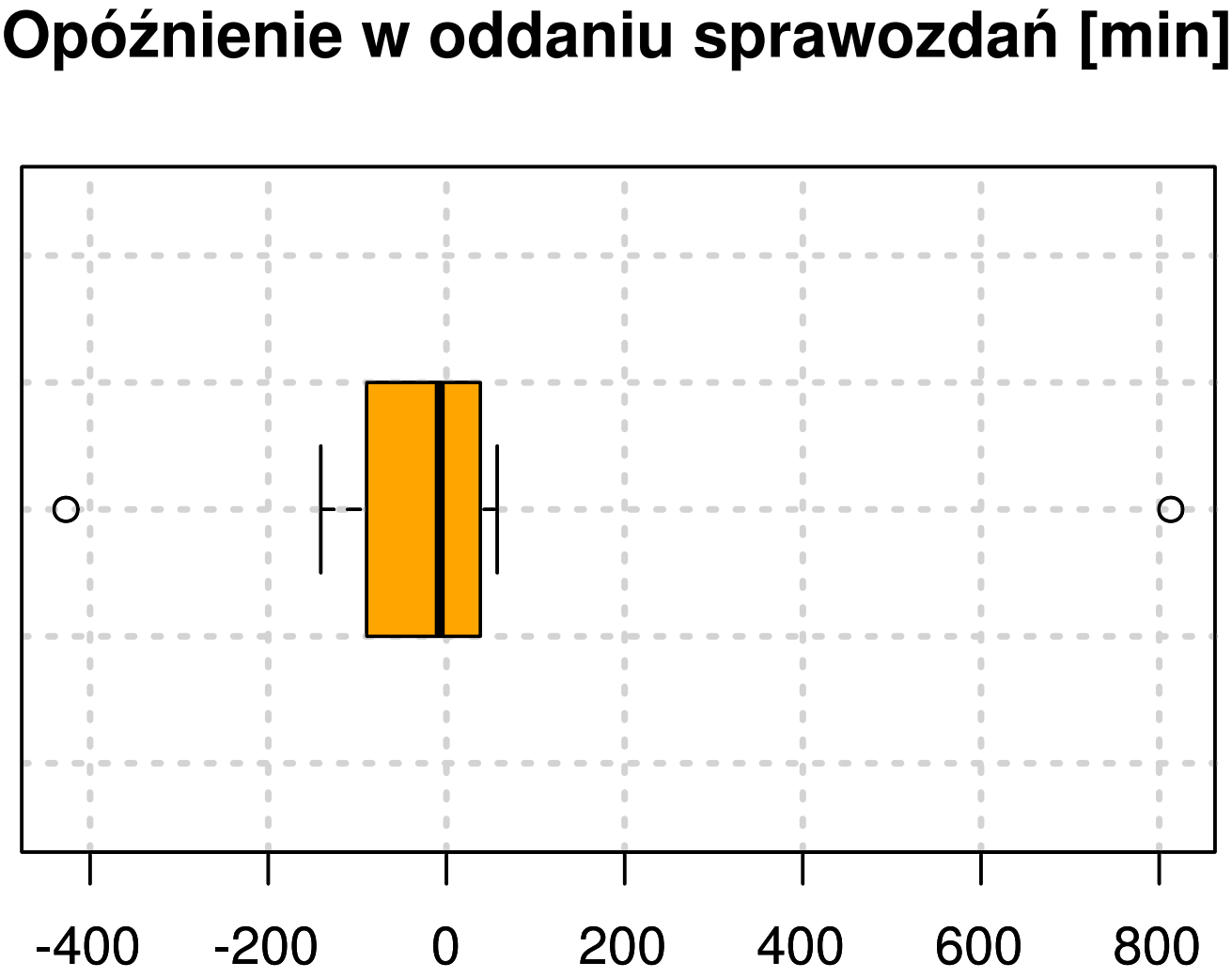
Wzrósł — wystarczy spojrzeć na średnią cenę: $100 < 125$.
Zmalał — spójrzmy na średnią harmoniczną: $100 > 80$.
Nie zmienił się — spójrzmy na śr. geometryczną: $100 = 100$.

Wykres pudełkowy



IQR to tzw. rozstęp ćwiartkowy (rozstęp kwartylowy).

Wykres pudełkowy — „punktualność studentów”



Dwa ciekawe i użyteczne fakty

Fakt

X ciągła zmienna losowa o dystrybuancie F. Wówczas

$$Z = F(X) \sim Unif([0,1]).$$

Fakt

$U \sim Unif([0,1])$ i $F\colon \mathbb{R} \to [0,1]$ prawostronnie ciągła funkcja niemalejąca spełniająca warunki

$$\lim_{t \to +\infty} F(t) = 1, \qquad \lim_{t \to -\infty} F(t) = 0,$$

to zmienna $X = F^{-1}(U)$ opisana jest dystrybuantą F.

Wartość oczekiwana

Wartość oczekiwana zmiennej losowej X, to liczba

$$\mu_X = \mathbb{E}(X) = \begin{cases} \sum_x x f(x), & \text{jeśli } X \text{ dyskr.} \\ \int x f(x) \mathrm{d}x, & \text{jeśli } X \text{ ciągła} \end{cases}$$

Wartość oczekiwana istnieje wtw, gdy (odpowiednio dla zm. dyskretnej i ciągłej)

$$\sum_x |x| f(x) < \infty, \qquad \int |x| f(x) \mathrm{d}x < \infty.$$

Wartość oczekiwana — przykłady

$$X \sim \text{Bern}(p) \quad \Rightarrow \quad \mathbb{E}(X) = p.$$

(rozkład Cauchy'ego): $f_X(x) = \dfrac{1}{\pi(1+x^2)} \Rightarrow \mathbb{E}(X)$ nie istnieje!

Wartość oczekiwana funkcji zmiennej losowej

$$Y = h(X), \quad \mathbb{E}(Y) = ?$$

$$\mathbb{E}(Y) = \int h(x) f_X(x) \mathrm{d}x.$$

Wartość oczekiwana — własności

$X_1, \ldots, X_n$ zmienne los. (niekoniecznie niezależne)

$$\mathbb{E}(a_1 X_1 + \cdots + a_n X_n) = a_1 \mathbb{E}(X_1) + \cdots + a_n \mathbb{E}(X_n).$$

Jeśli $X_1, \ldots, X_n$ niezależne, to

$$\mathbb{E}(X_1 X_2 \ldots X_n) = \mathbb{E}(X_1)\mathbb{E}(X_2) \ldots \mathbb{E}(X_n).$$

Wariancja

Wariancją zmiennej losowej X nazywamy liczbę

$$\mathbb{V}X = \sigma_X^2 = \mathbb{E}(X - \mathbb{E}X)^2 = \mathbb{E}(X - \mu_X)^2 = \mathbb{E}X^2 - (\mathbb{E}X)^2.$$

Odchyleniem standardowym zmiennej losowej X nazywamy liczbę

$$\sigma_X = \sqrt{\mathbb{V}X}.$$

Wariancja — własności

$$\mathbb{V}(aX + b) = a^2 \mathbb{V} X, \qquad \forall a, b \in \mathbb{R}.$$

Jeśli $X_1, \ldots, X_n$ niezależne zm. los. oraz $a_1, \ldots, a_n$ pewne stałe, to

$$\mathbb{V}(a_1 X_1 + \cdots + a_n X_n) = a_1^2 \mathbb{V} X_1 + \cdots + a_n^2 \mathbb{V} X_n.$$

Średnia z próby

Definicja

Średnią z próby losowej $X_1, \ldots, X_n$ nazywamy zmienną losową

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Jeśli próba z rozkładu o średniej μ ($\mathbb{E}(X_i) = \mu$) i wariancji σ^2 ($\mathbb{V} X_i = \sigma^2$), to

$$\mathbb{E}\overline{X}_n = \mu, \quad \mathbb{V}\overline{X}_n = \frac{\sigma^2}{n}.$$

Wariancja z próby

Definicja

Wariancją z próby losowej $X_1, \ldots, X_n$ nazywamy zmienną losową

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

Jeśli próba z rozkładu o średniej μ i wariancji σ^2, to

$$\mathbb{E} S_n^2 = \sigma^2.$$

Kowariancja

Kowariancją zmiennych losowych X i Y nazywamy liczbę

$$\mathbb{C}(X,Y) = \mathbb{E}\left((X-\mu_X)(Y-\mu_Y)\right) = \mathbb{E}XY - \mathbb{E}X\mathbb{E}Y.$$

Współczynnikiem korelacji tych zmiennych nazywamy liczbę

$$\rho_{X,Y} = \rho(X,Y) = \frac{\mathbb{C}(X,Y)}{\sigma_X\sigma_Y}, \qquad -1 \leq \rho_{X,Y} \leq 1.$$

Kowariancja — własności

$$\begin{aligned}
\mathbb{C}(X, X) &= \mathbb{V}X, \\
\mathbb{V}(X + Y) &= \mathbb{V}X + \mathbb{V}Y + 2\mathbb{C}(X, Y), \\
\mathbb{V}(X - Y) &= \mathbb{V}X + \mathbb{V}Y - 2\mathbb{C}(X, Y),
\end{aligned}$$

$$\mathbb{V}\left(\sum_i (a_i X_i)\right) = \sum_i a_i^2 \mathbb{V}X_i + 2\sum_{i<j} a_i a_j \mathbb{C}(X_i, X_j).$$

Jeśli X, Y niezależne, to $\mathbb{C}(X, Y) = 0$.

Jeśli $Y = aX + b$, to $\rho_{X,Y} = \operatorname{sgn} a$.

Zbieżność zmiennych losowych

$X, X_1, X_2, \ldots$ zm. los. o dystrybuantach $F, F_1, F_2, \ldots$

Ciąg $X_1, X_2, \ldots$ zbiega do X według prawdopodobieństwa, $X_n \xrightarrow{\mathbb{P}} X$, jeśli

$$\mathbb{P}(|X_n - X| > \epsilon) \xrightarrow{n\to\infty} 0 \qquad \forall \epsilon > 0.$$

Ciąg $X_1, X_2, \ldots$ zbiega do X według rozkładu, $X_n \xrightarrow{d} X$, jeśli

$$\lim_{n\to\infty} F_n(t) = F(t)$$

dla każdego punktu t, w którym F jest ciągłe.

Twierdzenie ((Słabe) prawo wielkich liczb (PWL))

Jeśli $X_1, X_2, \ldots$ są niezależnymi zmiennymi losowymi o tym samym rozkładzie i skończonej wartości oczekiwanej ($|\mathbb{E}X_1| < \infty$), to

$$\overline{X}_n = \frac{X_1 + \cdots + X_n}{n} \xrightarrow{\mathbb{P}} \mathbb{E}X_1.$$

Twierdzenie (Centralne twierdzenie graniczne (CTG))

Jeśli $X_1, X_2, \ldots$ są niezależnymi zmiennymi losowymi o tym samym rozkładzie, $\mathbb{E}X_1 = \mu$, $\mathbb{V}X_1 = \sigma^2$, to

$$\sqrt{n}\left(\overline{X}_n - \mu\right) \xrightarrow{d} N(0, \sigma^2).$$

Równoważnie

$$\frac{\sqrt{n}\left(\overline{X}_n - \mu\right)}{\sigma} \xrightarrow{d} N(0,1),$$

$$\frac{\overline{X}_n - \mathbb{E}\overline{X}_n}{\sqrt{\mathbb{V}\overline{X}_n}} \xrightarrow{d} N(0,1),$$

Przy założeniach CTG także

$$\frac{\sqrt{n}\left(\overline{X}_n - \mu\right)}{S_n} \xrightarrow{d} N(0,1),$$