Uczenie maszynowe w bioinformatyce

Wykład 4: podstawy statystycznego opisu danych (podejście jednowymiarowe)

Tymon Rubel

Zakład Elektroniki Jądrowej i Medycznej
Instytut Radioelektroniki i Technik Multimedialnych PW

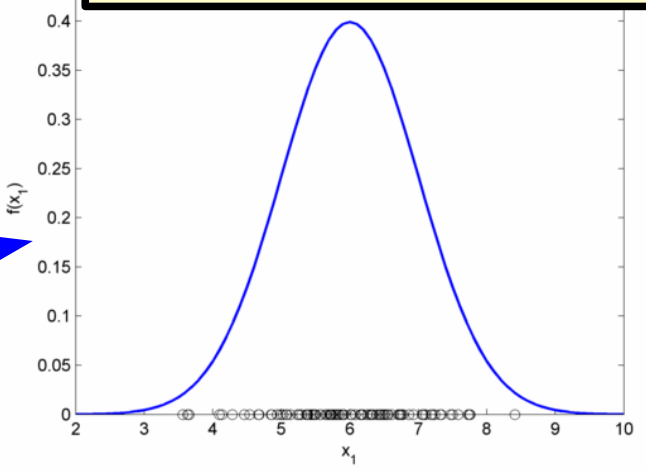
Statystyczny opis danych pomiarowych: rozkłady jednowymiarowe

Najprostszym sposobem opisu statystycznego danych z wielkoskalowych technik pomiarowych (i oczywiście również wielu innych rodzajów danych) jest założenie, że poszczególne atrybuty (np. geny, białka, …) są niezależnymi cechami ilościowymi badanej populacji. Każda z tych cech jest traktowana jako zmienna losowa charakteryzująca się jednowymiarowym ciągłym rozkładem prawdopodobieństwa.

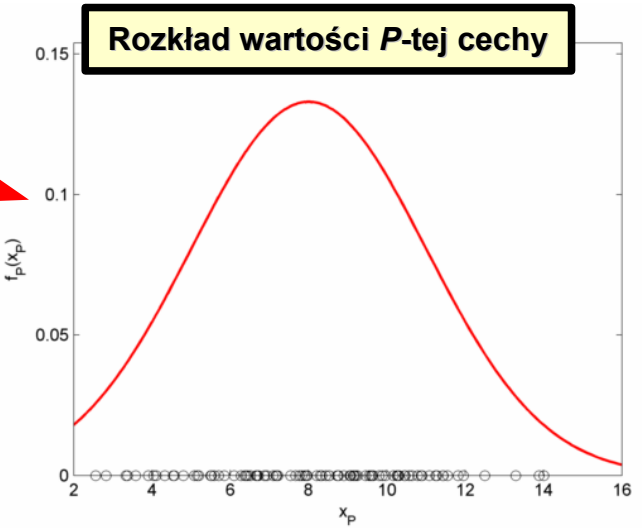
Macierz danych: *P* cech, *N* próbek

$$X = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1N} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2N} \\ x_{31} & x_{32} & x_{33} & \cdots & x_{3N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{P1} & x_{P2} & x_{P3} & \cdots & x_{PN} \end{bmatrix}$$

Rozkład wartości pierwszego pierwszej cechy



Rozkład wartości *P*-tej cechy

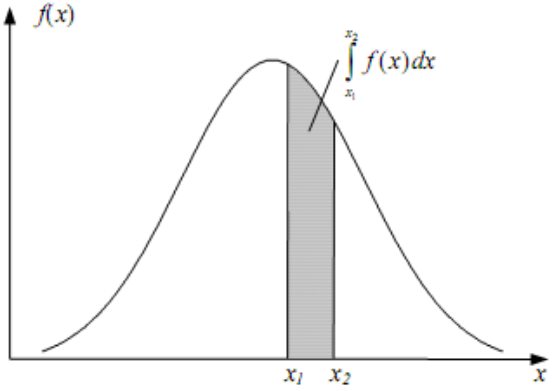


Rozkład prawdopodobieństwa cechy: opis za pomocą funkcji gęstości

Funkcją gęstości prawdopodobieństwa (PDF – *Probability Density Function*) ciągłej zmiennej losowej X nazywamy nieujemną funkcję spełniającą warunki:

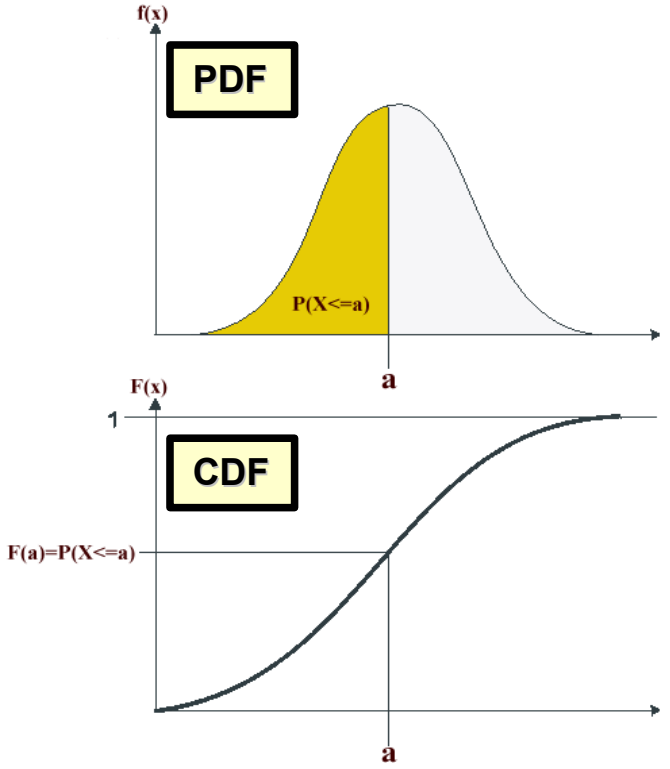
$$\Pr(x_1 < X < x_2) = \int_{x_1}^{x_2} f(x)\,dx$$

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



Dystrybuantą (CDF – *Cumulative Distribution Function*) ciągłej zmiennej losowej X jest niemalejąca, przynajmniej lewostronnie ciągła funkcja, dla której zachodzi:

$$F(x) = \Pr(X \le x) = \int_{-\infty}^{x} f(x)\,dx$$



	MATLAB	R
Wartości PDF:	`pdf`	-
	`normpdf`	`dnorm`
	`poisspdf`	`dpois`
	`tpdf`	`dt`
	…	…
Wartości CDF:	`cdf`	-
	`normcdf`	`pnorm`
	`poisscdf`	`ppois`
	`tcdf`	`pt`
	…	…

Rozkład prawdopodobieństwa cechy: opis za pomocą kwantyli

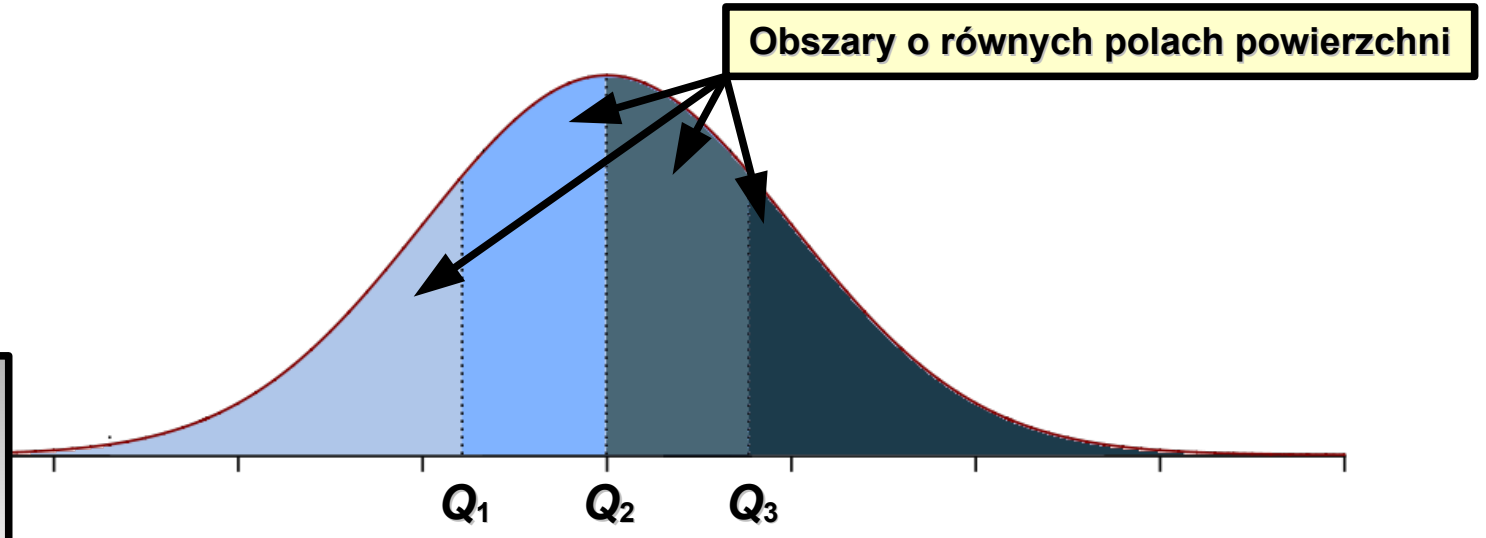
Cechą charakterystyczną rozkładu jest zbiór wartości jego kwantyli.

Kwantylem rzędu p ciągłego rozkładu zmiennej losowej X określa się wartość x_p, dla której zachodzi:

$$\Pr(X \le x_p) = F(x_p) = p$$

Szczególnymi przypadkami kwantyli są te, które dzielą pole powierzchni pod krzywą funkcji gęstości prawdopodobieństwa na równe części (a tym samym zakres wartości dystrybuanty na przedziały o równej długości). Przykładami mogą być:

- **percentyle – kwantyle rzędu** $1/100, 2/100, \ldots, 99/100$;
- **kwartyle (Q_1, Q_2, Q_3) – kwantyle rzędu** $1/4, 2/4, 3/4$;
- **mediana – kwantyl rzędu** $1/2$ **(drugi kwartyl).**



Kwantyle:	MATLAB	R
	icdf	-
	norminv	qnorm
	poissinv	qpois
	tinv	qt
	…	…

Rozkład prawdopodobieństwa cechy: opis za pomocą momentów

Liczbowymi charakterystykami kształtu rozkładu prawdopodobieństwa zmiennej losowej są wartości funkcji zwanych momentami.

Moment zwykły k-tego rzędu ($k = 1, 2, ...$) zmiennej losowej X jest definiowany jako wartość oczekiwana k-tej potęgi tej zmiennej. W przypadku zmiennej losowej typu ciągłego dany jest on zależnością:

$$\mu_k = \mathrm{E}[X^k] = \int_{-\infty}^{+\infty} x^k f(x)\,dx$$

Moment centralny k-tego rzędu ciągłej zmiennej losowej X jest definiowany jako:

$$\mu_k = \mathrm{E}[(X - \mathrm{E}[X])^k] = \int_{-\infty}^{+\infty} (x - \mathrm{E}[X])^k f(x)\,dx$$

Rozkład prawdopodobieństwa cechy: opis za pomocą momentów

Najczęściej wykorzystywanymi momentami są:

- **Wartość oczekiwana (pierwszy moment zwykły):** $\mathrm{E}[X]=\int_{-\infty}^{+\infty} x\, f(x)\, dx$

  Parametr ten określa położenie środka ciężkości rozkładu. Nie należy go mylić z wartością o największym prawdopodobieństwie występowania (modą) i wartością dzielącą rozkład na dwie części o równych polach (medianą), bo w przypadku wielu rozkładów wielkości te są różne.

- **Wariancja (drugi moment centralny):** $\mathrm{Var}[X]=\mathrm{E}[(X-\mathrm{E}[X])^2]=\int_{-\infty}^{+\infty} (x-\mathrm{E}[X])^2\, f(x)\, dx$

  Jest parametrem opisującym „szerokość" rozkładu, czyli tego jak duży jest rozrzut wartości cechy wokół jej wartości oczekiwanej. Pierwiastek wariancji określany jest odchyleniem standardowym.

- **Trzeci moment centralny związany jest z symetrią rozkładu – przyjmuje wartości niezerowe dla rozkładów asymetrycznych (odpowiednio ujemne i dodatnie dla rozkładów wydłużonym lewym i prawym ramieniu).**

Empiryczny rozkład prawdopodobieństwa cechy

Rzeczywisty rozkład prawdopodobieństwa cechy w populacji generalnej (czyli zbiorze wszystkich badanych obiektów) najczęściej nie jest znany.

W sytuacjach praktycznych zwykle dysponujemy jedynie n-elementową próbą losową x, czyli zbiorem n niezależnych wartości cechy, wybranych w sposób losowy z populacji generalnej. W naszym przypadku próbę losową stanowi n pomiarów ekspresji danego genu, wykonanych za pomocą n mikromacierzy (czyli pojedynczy wiersz macierzy danych):

$$x = \{x_1, x_2, \ldots, x_n\}$$

Inaczej mówiąc: nie możemy wnioskować o właściwościach cechy w oparciu o znajomość jej rzeczywistego rozkładu w populacji, a jedynie na podstawie empirycznego rozkładu prawdopodobieństwa, określanego z próby losowej.

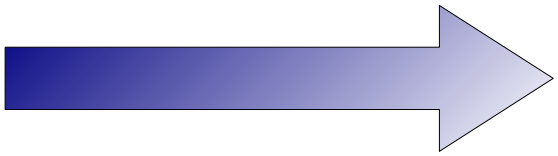
Empiryczny rozkład prawdopodobieństwa cechy

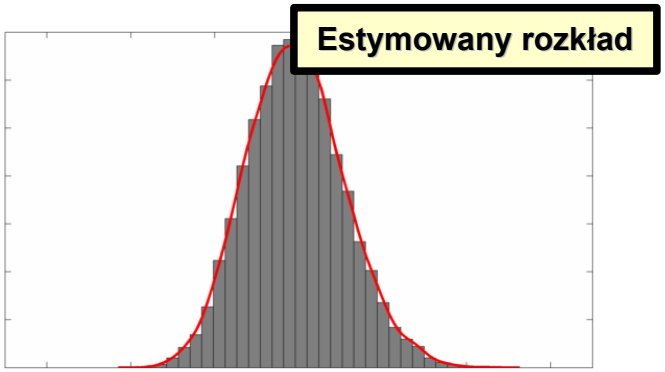
Możliwe są dwa ogólne podejścia do wyznaczania empirycznego rozkładu z próby:

- **nieparametryczne – bezpośrednia estymacja funkcji gęstości prawdopodobieństwa rozkładu na podstawie wartości tworzących próbę, bez wstępnych założeń co do jej funkcyjnej postaci. W tym celu używamy np. histogramy lub estymatory jądrowe;**

Próba losowa

$$x = \{x_1, x_2, \ldots, x_n\}$$



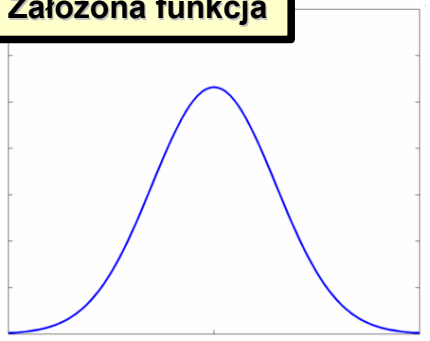


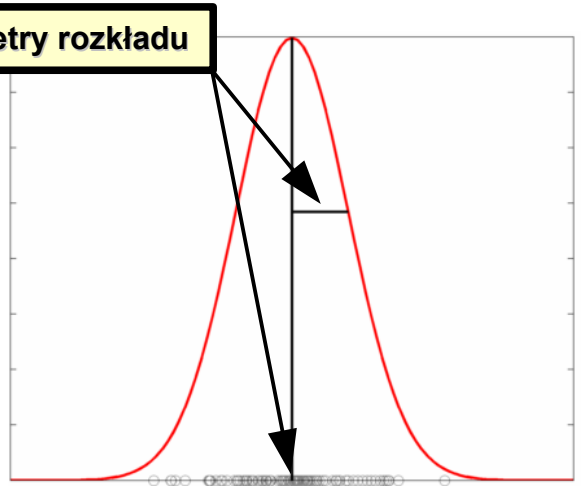
- **parametryczne – założenie znanej postaci funkcyjnej rozkładu (np. Gaussa) i ograniczenie się do określenia wartości jej parametrów (np. μ i σ) na podstawie próby losowej. W tym celu używamy metod estymacji parametrów rozkładu.**

Próba losowa

$$x = \{x_1, x_2, \ldots, x_n\}$$

Założona funkcja

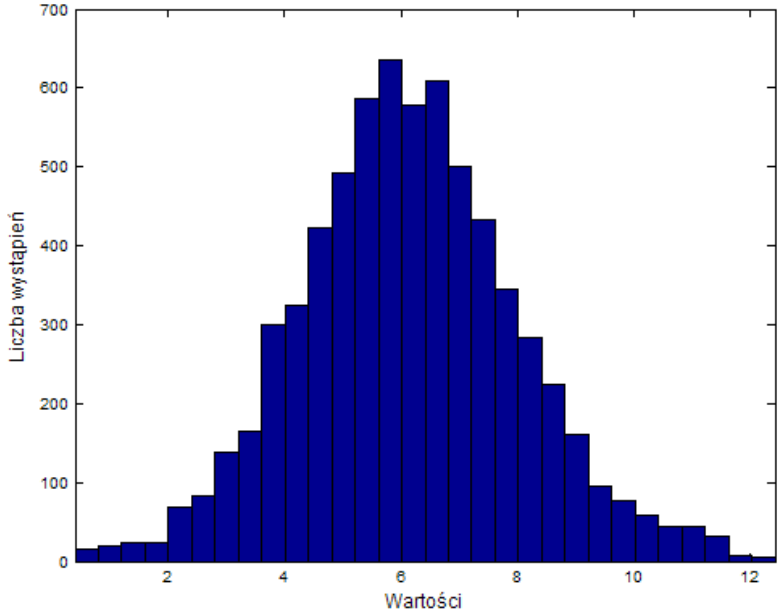




Empiryczny rozkład prawdopodobieństwa cechy: histogram

Najprostszą metodą estymacji empirycznego rozkładu prawdopodobieństwa cechy jest wyznaczenie histogramu.

Przedział zawierający wszystkie elementy próby losowej $\{x_1, x_2, \ldots, x_n\}$ dzielony jest na k podprzedziałów H_k o jednakowej szerokości h. Dla każdego podprzedziału H_k wartością histogramu staje się liczba elementów próby losowej, o wartościach należących do tego przedziału: $\#\{x_i \in H_k\}$



Aby histogram mógł posłużyć do estymacji funkcji gęstości prawdopodobieństwa cechy na podstawie wartości z próby musi zostać znormalizowany tak, aby jego całkowite pole jego pole powierzchni wynosiło 1. Dla każdego podprzedziału H_k:

$$\hat{f}(x) = \frac{\#\{x_i \in H_k\}}{hn}$$

	MATLAB	R
Histogram:	`hist`	`hist`

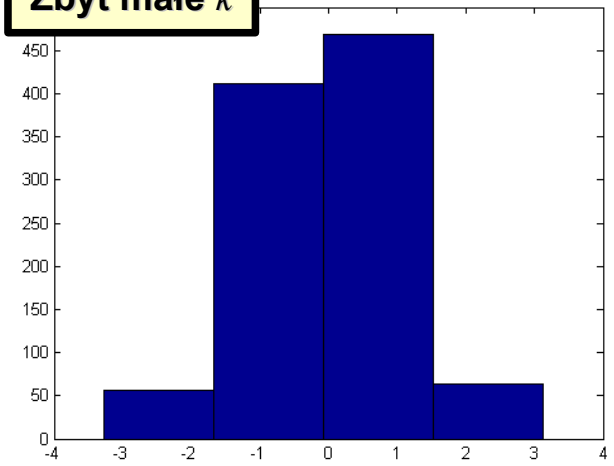
Empiryczny rozkład prawdopodobieństwa cechy: histogram

Problemem w przypadku tworzenia histogramu jest ustalenie właściwej liczby przedziałów k (lub szerokości przedziału h, co jest zadaniem równoważnym), na które dzielony jest zakres obserwowanych wartości cechy:

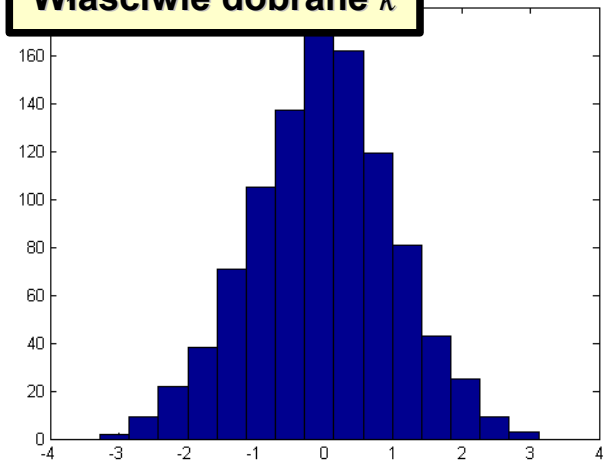
$$k = \left\lceil \frac{\max(x) - \min(x)}{h} \right\rceil = \mathrm{ceil}\left(\frac{\max(x) - \min(x)}{h} \right)$$

gdzie $\lceil \cdot \rceil = \mathrm{ceil}(\cdot)$ **oznacza zaokrąglenie do najbliższej liczby całkowitej.**

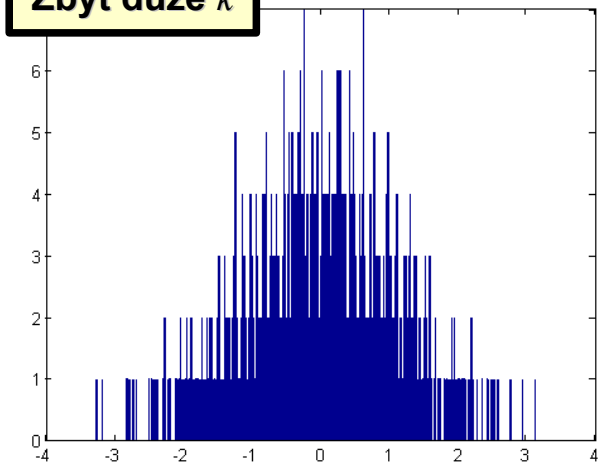
Zbyt małe k



Właściwie dobrane k



Zbyt duże k



W ogólności nie ma „optymalnej" wartości k, niezależnej od postaci rozkładu i celu analizy. Często jest ustalana na podstawie liczności n próby losowej, np. jako:

$$k = \sqrt{n}$$ **(wzór stosowany m.in. przez Excel)**

$$k = \left\lceil \log_2(n+1) \right\rceil$$ **(tzw. wzór Sturgesa)**

Empiryczny rozkład prawdopodobieństwa cechy: estymatory jądrowe

Estymator jądrowy funkcji gęstości prawdopodobieństwa (KDE – *Kernel Density Estimator*) definiowany jest jako:

$$\hat{f}(x)=\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)$$

gdzie: n – **liczność próby losowej**

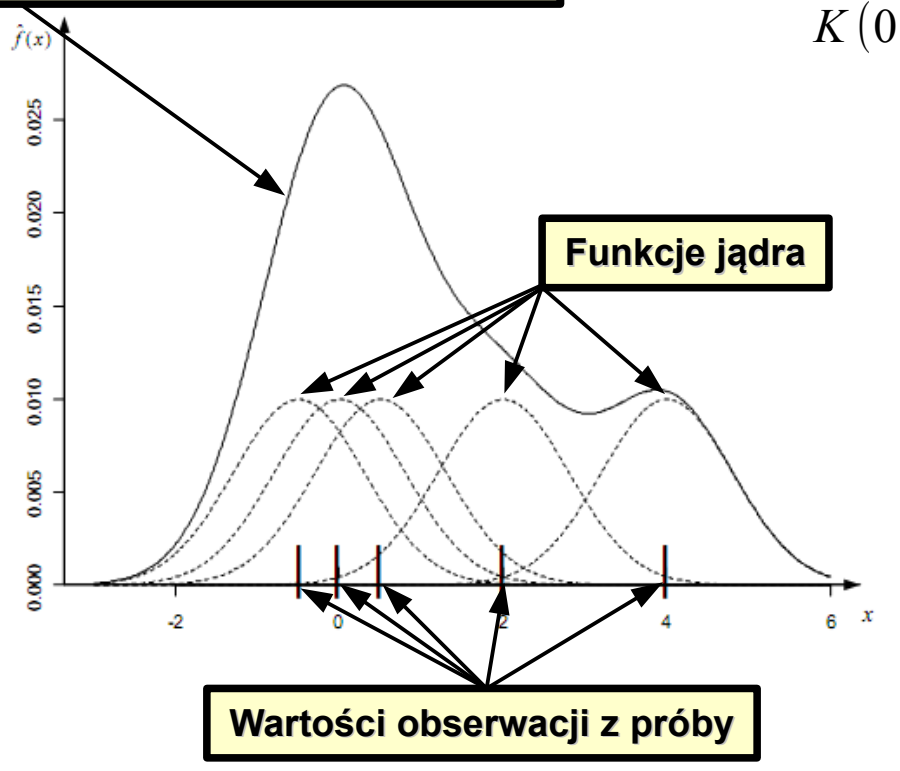
h – **parametr wygładzania** ($h>0$)

$K(x)$ – **funkcja jądra, spełniająca warunki:**

$$\int_{-\infty}^{+\infty} K(x)\,dx=1$$

$$K(x)=K(-x) \quad \text{dla każdego } x\in\mathbb{R}$$

$$K(0)\geq K(x)$$

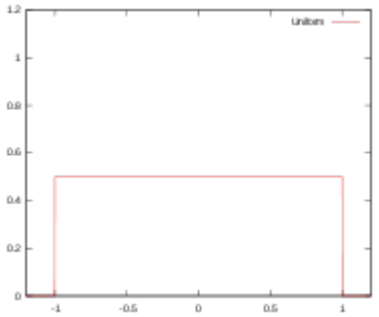


	MATLAB	R
KDE:	`ksdensity`	`density`

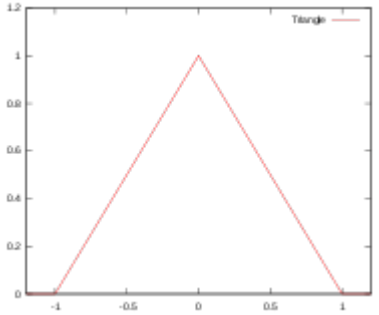
Empiryczny rozkład prawdopodobieństwa cechy: estymatory jądrowe

Przykładowe jądra:

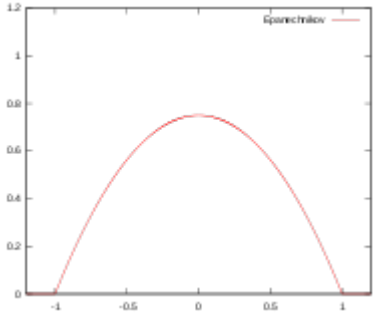
- **Jednostajne** $K(u) = \frac{1}{2}\,\mathbf{1}_{\{|u|\leq 1\}}$



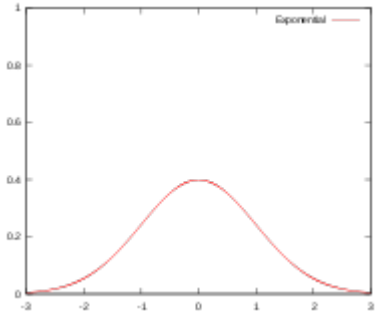
- **Trójkątne** $K(u) = (1-|u|)\,\mathbf{1}_{\{|u|\leq 1\}}$



- **Epanecznikowa** $K(u) = \frac{3}{4}(1-u^2)\,\mathbf{1}_{\{|u|\leq 1\}}$

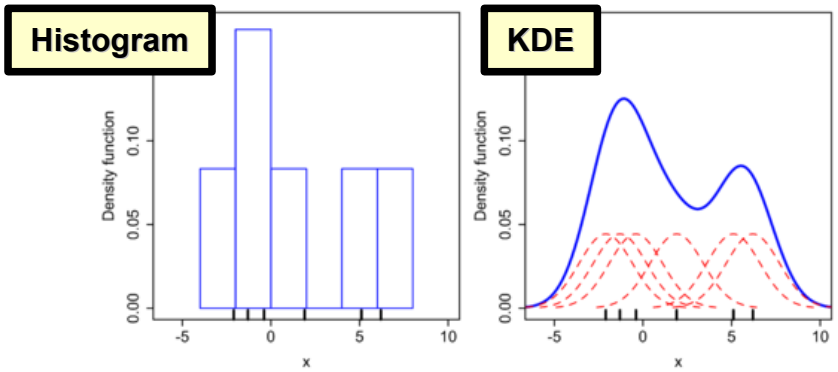


- **Gaussa** $K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}$

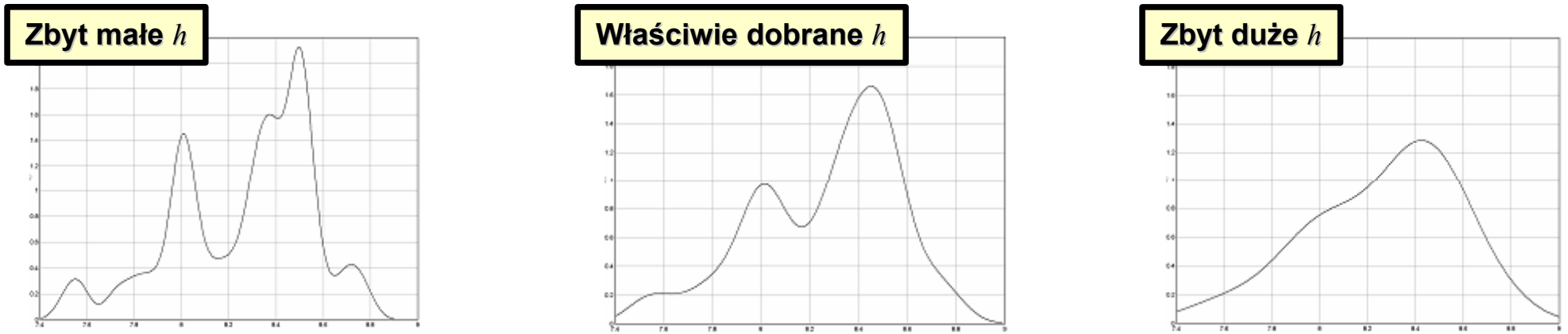


Empiryczny rozkład prawdopodobieństwa cechy: estymatory jądrowe

W porównaniu z histogramem estymatory jądrowe zapewniają gładszą i obarczoną mniejszym błędem estymatę funkcji gęstości prawdopodobieństwa.



Zasadnicze znaczenie dla jakości estymacji ma wartość parametru wygładzania h.



Dla Guassowskiej funkcji jądra i próby o liczebności n, pobranej z rozkładu normalnego można pokazać, że optymalna wartość parametru wygładzania wynosi:

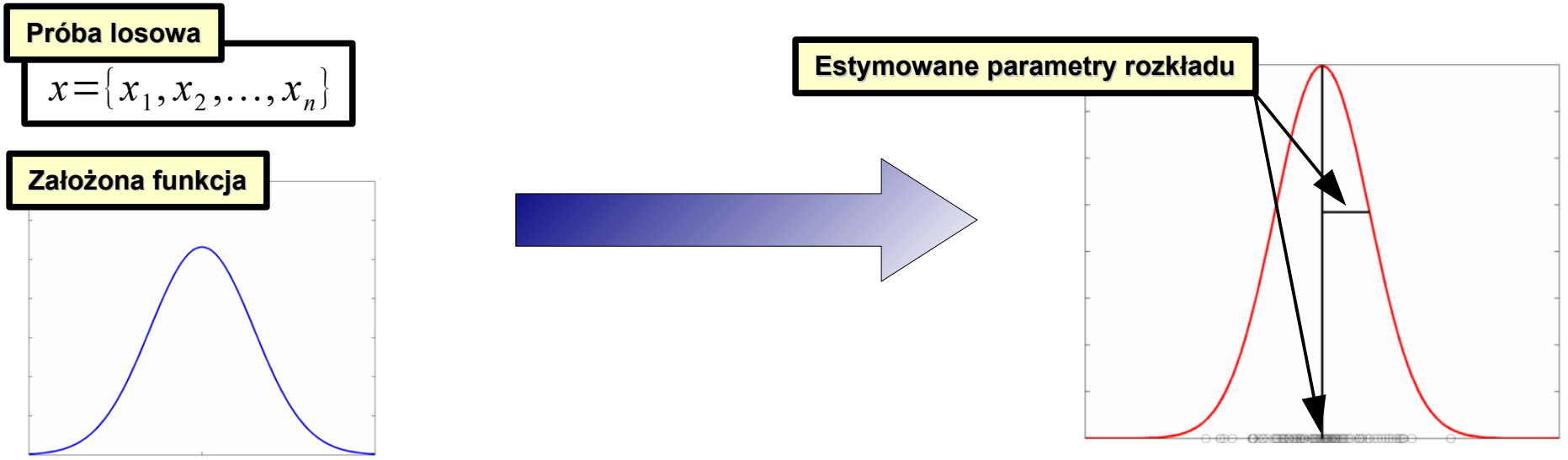
$$h = \left(\frac{4\, s^5}{3\mathrm{n}} \right)^{\frac{1}{5}} \quad ; \quad \mathrm{s} - \text{odchylenie standardowe z próby}$$

Empiryczny rozkład prawd. cechy: podejście parametryczne (z modelem)

Bezpośrednie wyznaczanie rozkładu prawdopodobieństwa na podstawie próby (za pomocą histogramu, czy też KDE) jest podejściem relatywnie rzadko stosowanym dla danych o dużej liczbie cech.

Częściej z góry zakłada się ogólną postać funkcji gęstości prawdopodobieństwa, dzięki czemu problem określenia rozkładu empirycznego sprowadza się do (znacznie prostszego) problemu estymacji parametrów wybranej funkcji na podstawie próby.

Przykładowo: zakłada się, że dane mogą być modelowane przez rozkład Gaussa, a na podstawie próby estymuje się jedynie jego dwa parametry: μ i σ.



Empiryczny rozkład prawd. cechy: podejście parametryczne (z modelem)

Podejście parametryczne wymaga:

- **wyboru modelu statystycznego**, czyli określenia rodziny rozkładów, które mogą posłużyć do możliwie dokładnego opisu wyników eksperymentalnych. W przypadku danych o ekspresji genów często zakłada się normalność rozkładu cech;

- **estymacji parametrów wybranego rozkładu** na podstawie wartości próby losowej. Przykładowo, przy założeniu, że cecha ma rozkład normalny trzeba wyznaczyć wartości parametrów μ i σ, decydujących o położeniu i szerokości tego rozkładu;

- **weryfikacji założeń**, czyli sprawdzenie czy wybrany rozkład teoretyczny opisuje dane w satysfakcjonujący sposób.

Empiryczny rozkład prawdopodobieństwa cechy: założony model $N(\mu,\sigma)$

Funkcja gęstości prawdopodobieństwa:

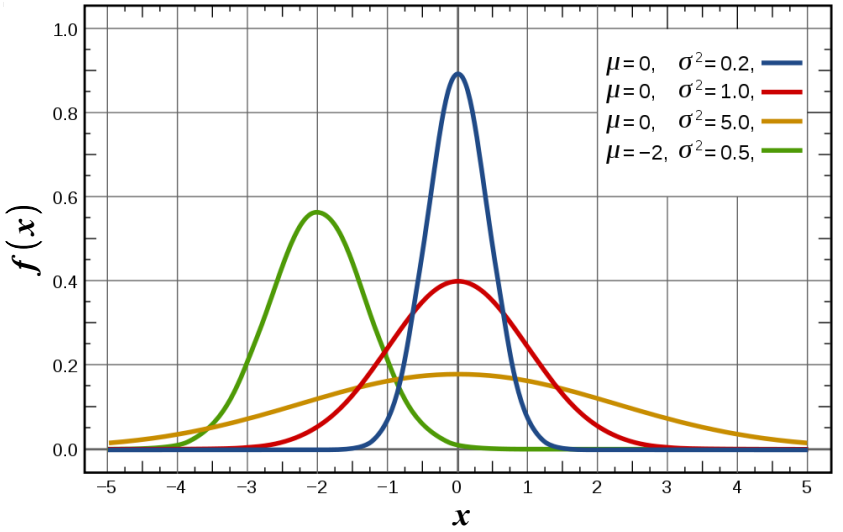
$$f(x;\mu,\sigma)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

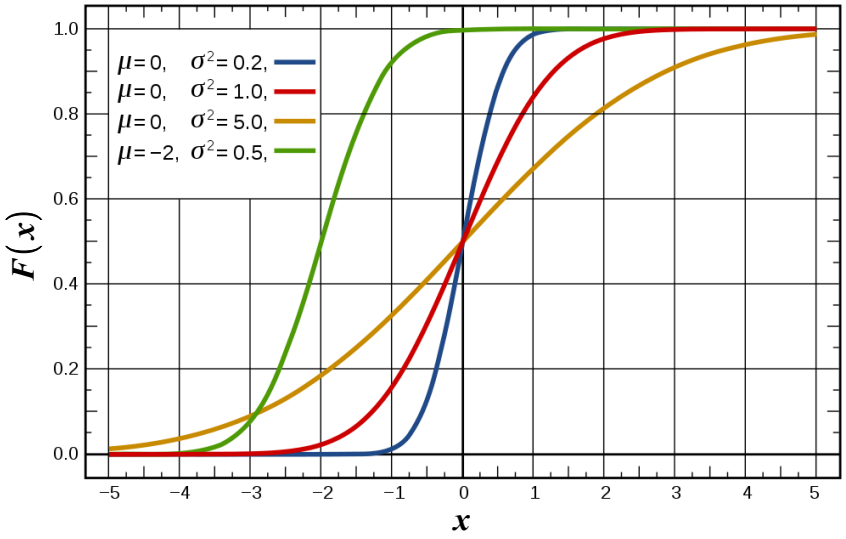
Dystrybuanta:

$$F(x;\mu,\sigma)=\frac{1}{2}\left[1+\operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$$

gdzie:

$$\operatorname{erf}(x)=\frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt$$





Wartość oczekiwana: μ

Wariancja: σ^2

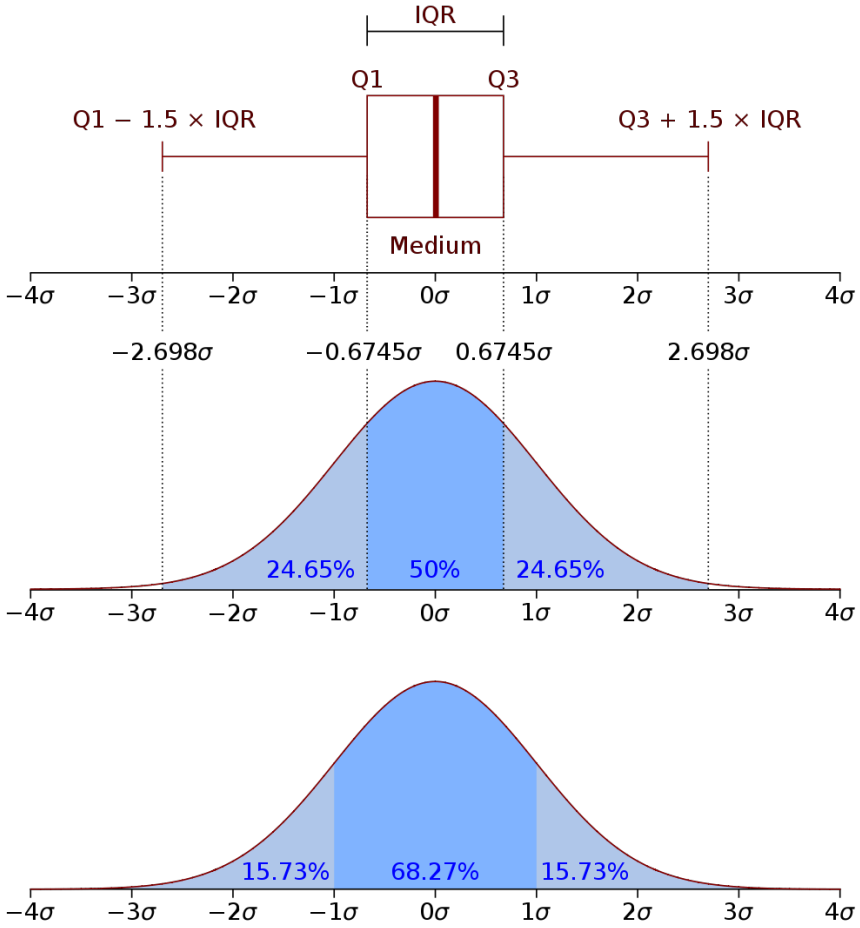
Odchylenie standardowe: σ

Zapis $X\sim N(\mu,\sigma)$ oznacza, że zmienna X ma rozkład normalny o parametrach μ i σ. Standardowym rozkładem normalnym określa się rozkład $N(0, 1)$ o zerowej wartości średniej i jednostkowej wariancji.

Empiryczny rozkład prawdopodobieństwa cechy: założony model $N(\mu,\sigma)$

Właściwości funkcji gęstości prawdopodobieństwa rozkładu normalnego:

- **ciągła i różniczkowalna dla** $x \in \mathrm{R}$;
- **unimodalna, symetryczna względem** μ**, z punktami przegięcia dla** x **równych** $\mu \pm \sigma$;
- **wartość średnia** μ **jest jednocześnie medianą oraz modą rozkładu;**
- **około** $68{,}3\%$ **pola pod krzywą znajduje się w przedziale** $\pm\sigma$ **wokół średniej, około** $95{,}5\%$ **w przedziale** $\pm 2\sigma$ **i około** $99{,}7\%$ **w przedziale** $\pm 3\sigma$**.**



	MATLAB	R
Funkcja gęstości:	`normpdf`	`dnorm`
Dystrybuanta:	`normcdf`	`pnorm`
Odwrotna dystrybuanta (kwantyle):	`norminv`	`qnorm`
Pobranie próby z rozkładu:	`normrnd/randn`	`rnorm`

Empiryczny rozkład prawdopodobieństwa cechy: estymacja parametrów

Estymator punktowy jest funkcją elementów próby losowej (tzw. statystyką) służącą do oszacowania nieznanej wartości wybranego parametru modelu statystycznego.

Estymator, jako funkcją zmiennej losowej, sam również jest zmienną losową o pewnym rozkładzie prawdopodobieństwa (w szczególności charakteryzujący się pewną wartością oczekiwaną i wariancją). Wartość estymatora dla konkretnej próby jest jedynie oceną, która prawie zawsze będzie różnić się od rzeczywistej wartości parametru oraz od ocen uzyskanych dla innych prób losowych.

O estymatorze $\hat{\Theta}$ parametru Θ mówimy, że jest:

- **zgodny**, gdy ze wzrostem liczebności próby rośnie również prawdopodobieństwo, że ocena będzie bliska prawdziwej wartości parametru;
- **nieobciążony**, jeśli wartość oczekiwana rozkładu estymatora jest równa wartości szacowanego parametru. Jeżeli obciążenie estymatora dąży do 0 przy liczebności próby dążącej do ∞, to estymator jest **asymptotycznie nieobciążony**;
- **najbardziej efektywny**, jeżeli ma najmniejszą wariancję ze wszystkich estymatorów nieobciążonych.

Empiryczny rozkład prawdopodobieństwa cechy: estymacja parametrów

Zgodnym i nieobciążonym estymatorem wartości oczekiwanej μ z próby losowej jest średnia arytmetyczna:

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Jeżeli próba pobrana została z populacji o rozkładzie normalnym $N(\mu, \sigma)$, to estymator $\hat{\mu}$ ma rozkład normalny o wartości oczekiwanej μ i odchyleniu standardowym:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Zgodnym i nieobciążonym (przy założeniu losowania ze zwracaniem elementów próby) estymatorem wariancji σ^2 z próby jest:

$$\hat{\sigma}^2 = s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2$$

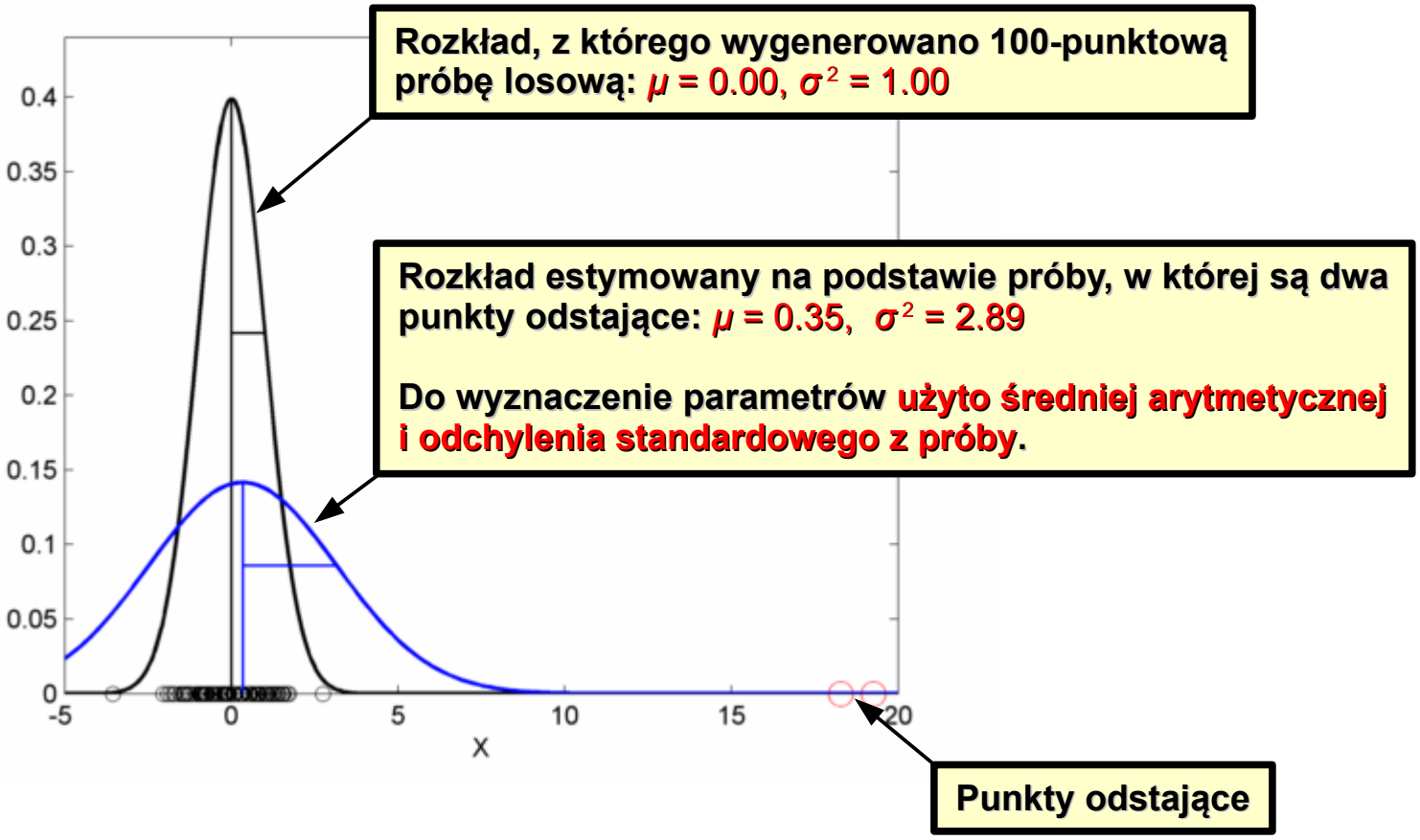
Zgodnym i asymptotycznie nieobciążonym estymatorem odchylenia standardowego σ z próby jest:

$$\hat{\sigma} = s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

	MATLAB	R
Średnia:	`mean`	`mean`
Wariancja:	`var`	`var`
Odch. std.:	`std`	`sd`

Empiryczny rozkład prawdopodobieństwa cechy: odporność estymatorów

Średnia arytmetyczna $\bar{x}$ i odchylenie standardowe z próby s nie są odporne wobec występowania w próbie losowej obserwacji odstających (*outliers*).



Przez obserwacje odstające rozumie się takie, których wartości znacząco obiegają reszty próby i są niezgodne z złożeniami stosowanego modelu statystycznego. Występowanie takich wartości może wynikać np. z grubych błędów pomiarowych lub niewłaściwego wprowadzania, bądź też przetwarzania danych (np. dzielenia przez 0).

Empiryczny rozkład prawdopodobieństwa cechy: odporność estymatorów

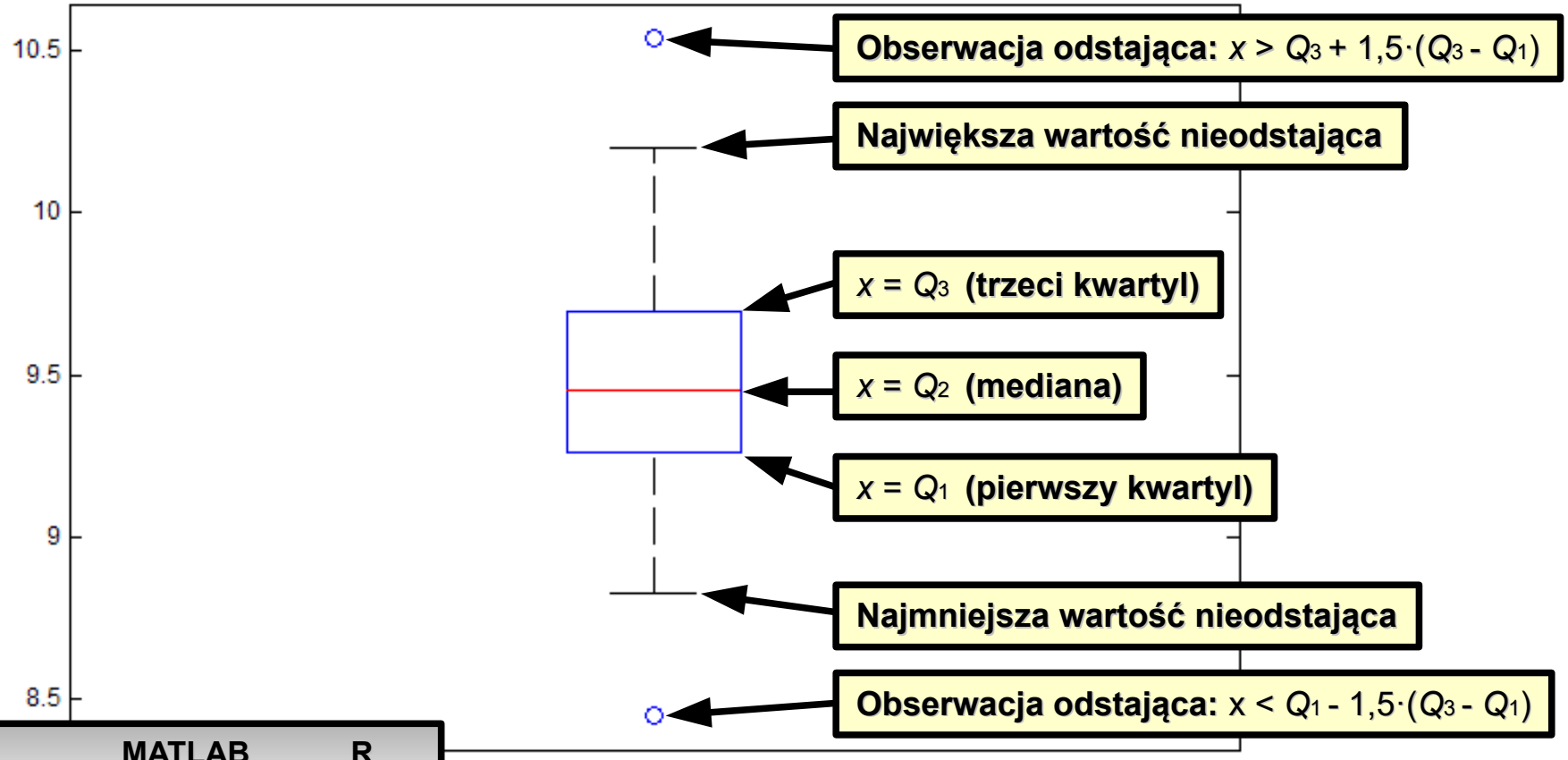
Za odstające zwykle uznaje się te obserwacje z próby, dla których zachodzi:

$$x > Q_3 + 1.5(Q_3 - Q_1)$$

$$x < Q_1 - 1.5(Q_3 - Q_1)$$

gdzie Q_1 i Q_3 to pierwszy i trzeci kwartyl (kwantyle rzędu $1/4$ i $3/4$).

Jako wizualny sposób oceny występowania odstających wartości może posłużyć wykres pudełkowy (*box-plot*).

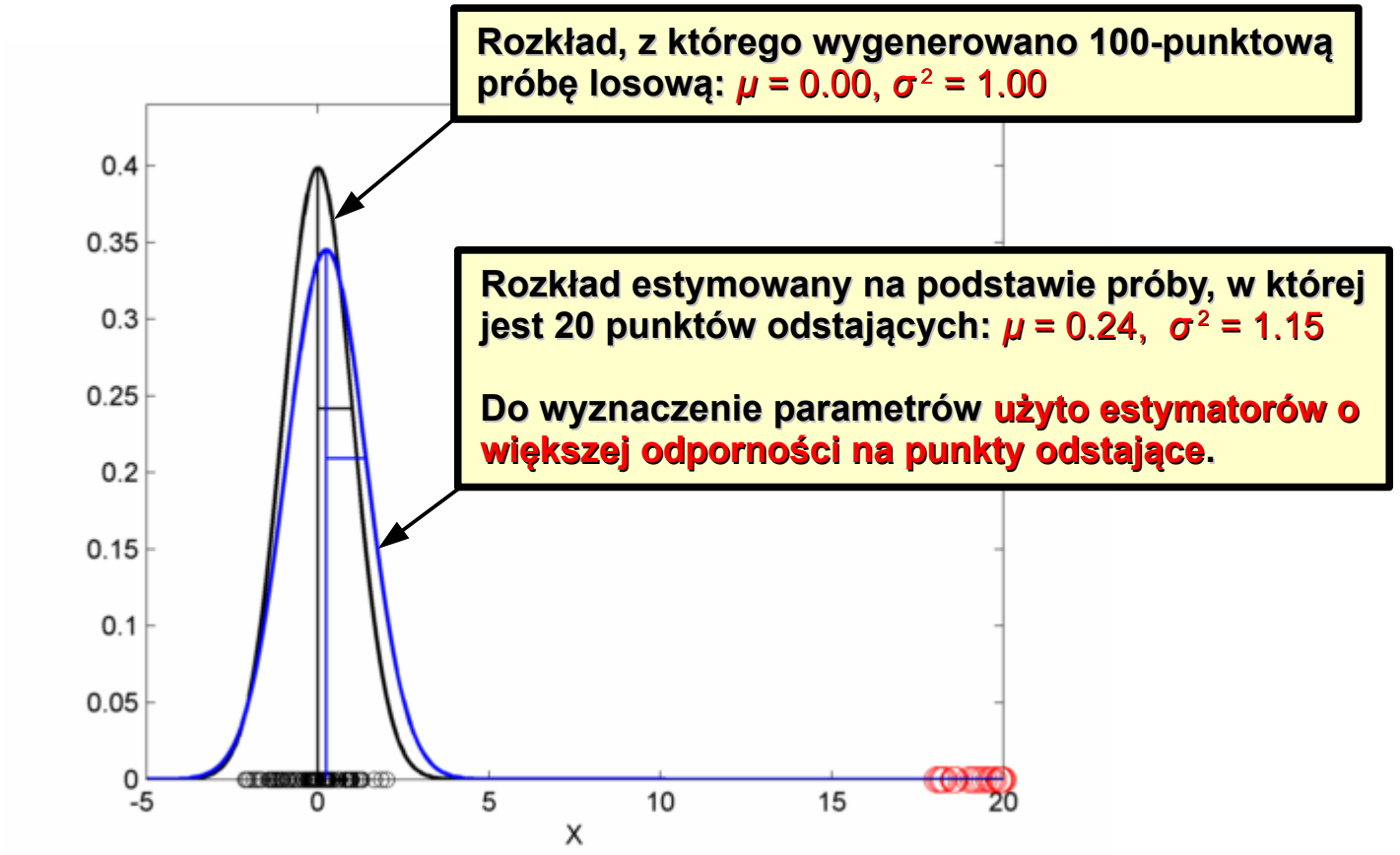


	MATLAB	R
Wykres *box-plot*:	`boxplot`	`boxplot`

Empiryczny rozkład prawdopodobieństwa cechy: odporność estymatorów

Miarą odporności estymatora jest jego punkt załamania (*breakdown point*), definiowany jako procent odstających obserwacji o dowolnie dużych wartościach (w szczególności równych ∞), których występowanie w próbie nie spowoduje, że estymator zacznie dawać dowolnie duże wyniki (w szczególności równe ∞).

Punkt załamania dla średniej arytmetycznej i odchylenia standardowego z próby jest równy 0%. Czyli nawet pojedynczy element próby może spowodować, że estymator przyjmie dowolnie dużą wartość. Istnieją jednak estymatory charakteryzujące się większą wartością punktu załamania (maksymalnie 50%).



Empiryczny rozkład prawdopodobieństwa cechy: odporność estymatorów

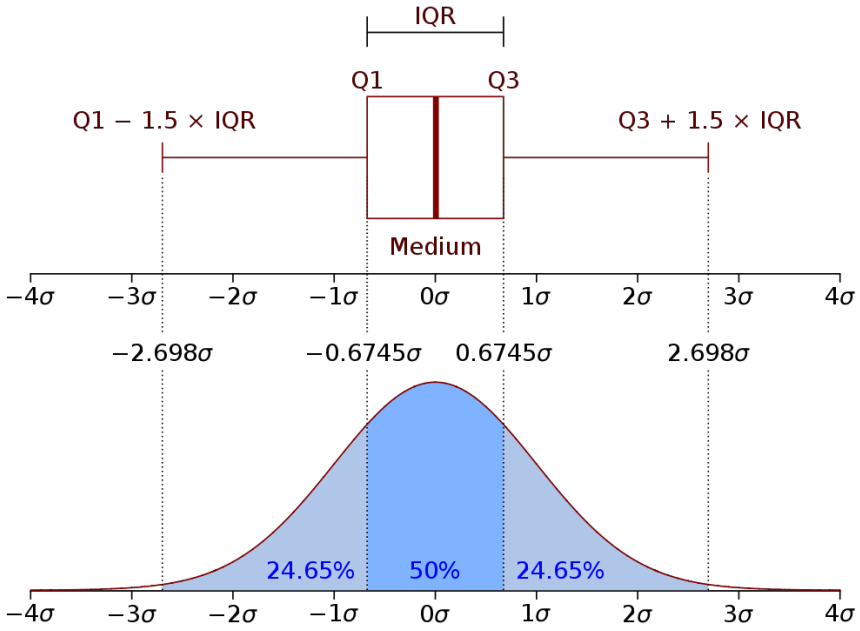
Estymatory wartości średniej o większej odporności wobec obserwacji odstających:

- **średnia obcięta próby (*truncated mean*, *trimmed mean*)** – średnia arytmetyczna nie uwzględniająca po $\alpha\%$ najmniejszych i największych wartości próby;

- **mediana próby** – wartość, poniżej której znajduje się 50% wartości próby.

$$\operatorname{med}(x) = \begin{cases} x_{(n+1)/2} & \text{gdy } n \text{ nieparzyste} \\ 0.5\left(x_{n/2} + x_{n/2+1}\right) & \text{gdy } n \text{ parzyste} \end{cases}$$

Dla danych o rozkładzie normalnym wielkości te są bezpośrednimi estymatorami μ.



	MATLAB	R
Średnia obcięta:	`trimmean`	`mean`
Mediana:	`median`	`median`

Empiryczny rozkład prawdopodobieństwa cechy: odporność estymatorów

Estymatory wariancji o większej odporności wobec obserwacji odstających:

- **odstęp między kwartylami próby (IQR – *InterQuartile Range*) – różnica pomiędzy trzecim a pierwszym kwartylem wartości próby losowej**

$$\mathrm{IQR}(x) = Q_3 - Q_1$$

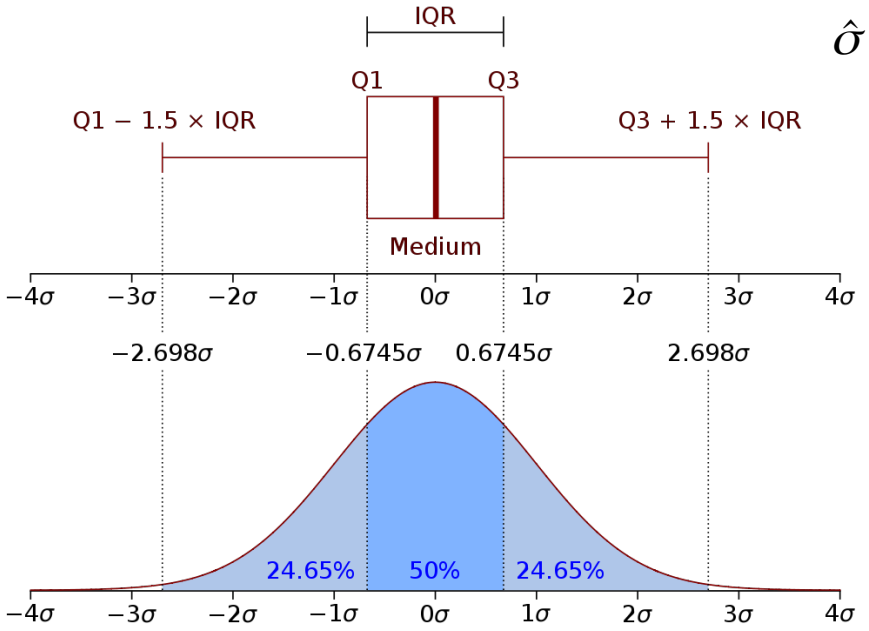
- **medianowe odchylenie bezwzględne próby (MAD – *Median Absolute Deviation*)**

$$\mathrm{MAD}(x) = \mathrm{med}\left(\left|x - \mathrm{med}(x)\right|\right)$$

Dla danych o rozkładzie normalnym σ można wyznaczyć jako:

$$\hat{\sigma} \approx \mathrm{IQR}(x)/1.3490$$

$$\hat{\sigma} \approx \mathrm{MAD}(x)/0.6745$$



	MATLAB	R
IQR:	`iqr`	`IQR`
MAD:	`mad`	`mad`

Empiryczny rozkład prawdopodobieństwa cechy: kwantyle z próby

Estymowany dla próby kwantyl rzędu $p \in (0, 1)$, to taka wartość q_p, że $p \cdot n$ elementów próby ma wartość mniejszą bądź równą q_p.

Algorytm wyznaczanie kwantyli próby:

- **uporządkowanie próby zgodnie z niemalejącymi wartościami jej elementów**

$$x^S = \mathrm{sort}(x)$$

- **wyznaczenie rzeczywistoliczbowego „indeksu” I**

$$I = np + 0.5$$

- **określenie q_p na podstawie elementów uporządkowanej próby, pomiędzy którymi znajduje się wartość I**

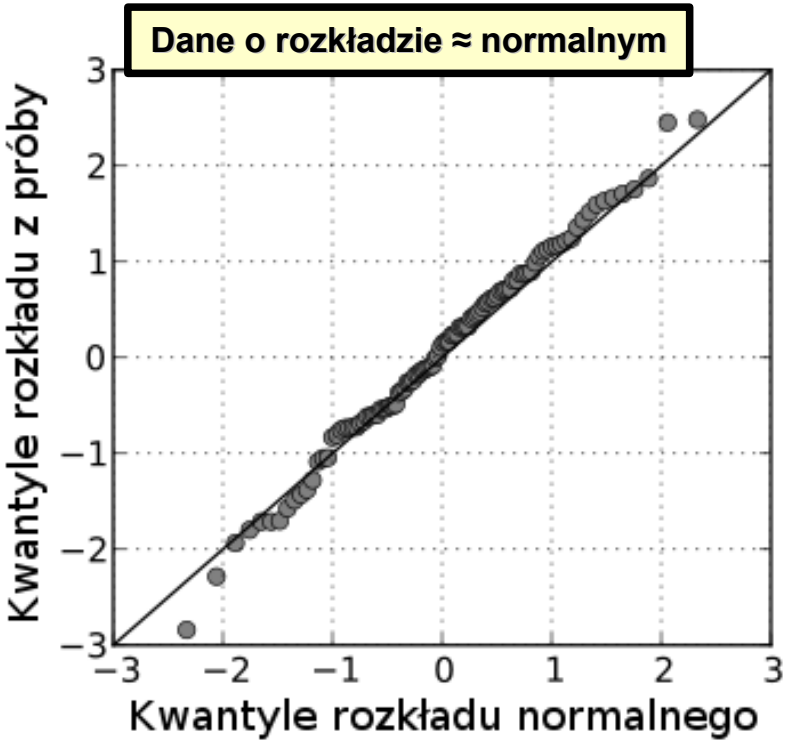
$$q_p = \begin{cases} x^S_1 & \text{gdy} \quad p < 0.5/n \\ x^S_n & \text{gdy} \quad p > (n-0.5)/n \\ x^S_{\lfloor I \rfloor} + (I - \lfloor I \rfloor)\left(x^S_{\lfloor I \rfloor + 1} - x^S_{\lfloor I \rfloor}\right) = x^S_{\mathrm{floor}(I)} + (I - \mathrm{floor}(I))\left(x^S_{\mathrm{floor}(I)+1} - x^S_{\mathrm{floor}(I)}\right) & \text{w pozostałych przypadkach} \end{cases}$$

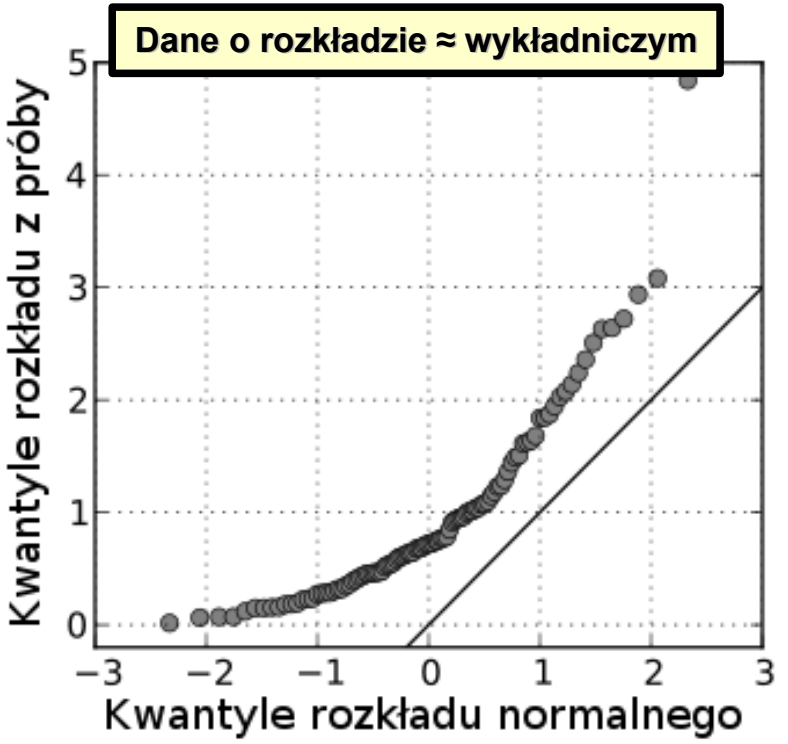
	MATLAB	R
Kwantyle z próby:	`quantile`	`quantile`

Empiryczny rozkład prawdopodobieństwa cechy: quantile-quantile plot

Wykres kwantyl-kwantyl (*quantile-quantile plot*) jest prostą metodą wizualnej oceny zgodności empirycznego rozkładu z teoretycznym rozkładem prawdopodobieństwa. Tworzy się go przez wykreślenie wybranych kwantyli (np. percentyli) empirycznego rozkładu wobec odpowiadającym im kwantyli rozkładu teoretycznego.

Jeżeli rozkłady są zgodne, to punkty wykresu będą leżeć na prostej. Jeśli parametry rozkładu zostały właściwie oszacowane, to prosta ta będzie leżeć na przekątnej.





	MATLAB	R
Wykresy Q-Q:	`qqplot/probplot`	`qqplot`

Uczenie maszynowe w bioinformatyce

Wykład 4: podstawy statystycznego opisu danych (podejście wielowymiarowe)

Tymon Rubel

Zakład Elektroniki Jądrowej i Medycznej
Instytut Radioelektroniki i Technik Multimedialnych PW

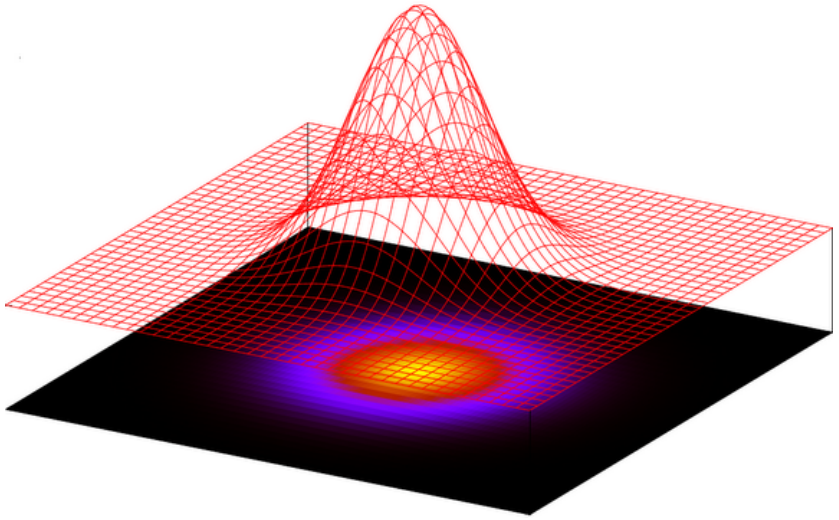
Statystyczny opis danych pomiarowych: rozkłady wielowymiarowe

W rzeczywistości wartości poszczególnych cech wcale nie muszą być są od siebie niezależne. Dlatego lepszym podejściem może okazać się traktowanie każdej próbki (pomiaru) jako realizacji *P*-wymiarowej zmiennej losowej. W takim przypadku każda kolumna macierzy danych traktowana jest jako wektor cech $\{x_1, x_2, ..., x_P\}$, którego elementy pochodzą ze jednego wielowymiarowego rozkładu prawdopodobieństwa.

Macierz danych: *P* cech, *N* próbek

$$X = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1N} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2N} \\ x_{31} & x_{32} & x_{33} & \cdots & x_{3N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{P1} & x_{P2} & x_{P3} & \cdots & x_{PN} \end{bmatrix}$$

***P*-wymiarowy rozkład wszystkich cech**



Kowariancja zmiennych losowych

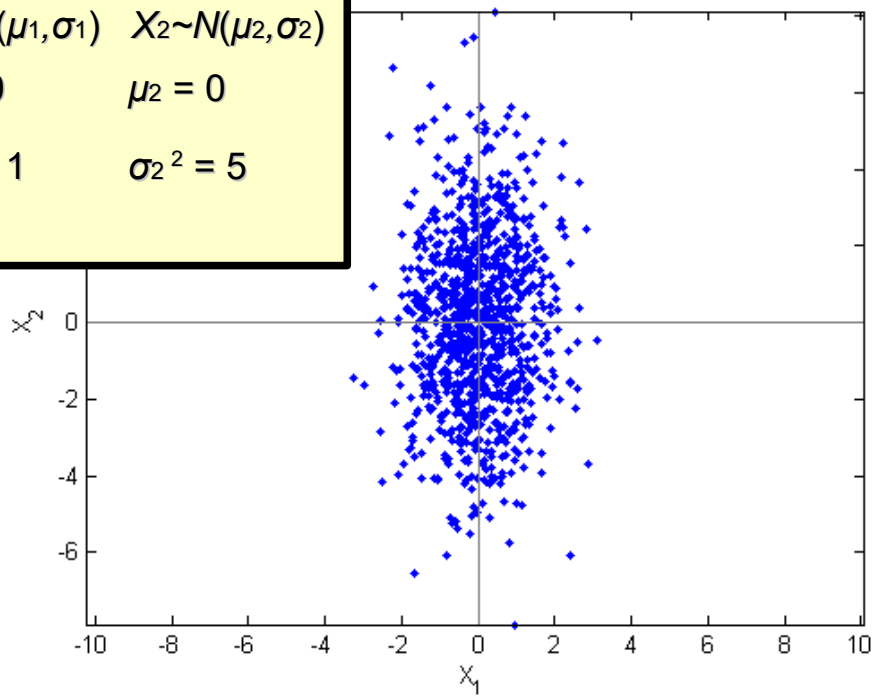
Przechodząc od przypadku jedno- do wielowymiarowego konieczne staje się uwzględnienie możliwości występowania pewnych zależności pomiędzy wartościami zmiennych losowych tworzących P-wymiarowy wektor cech.

$X_1 \sim N(\mu_1, \sigma_1)$ $X_2 \sim N(\mu_2, \sigma_2)$

$\mu_1 = 0$ $\mu_2 = 0$

$\sigma_1^2 = 1$ $\sigma_2^2 = 5$

?

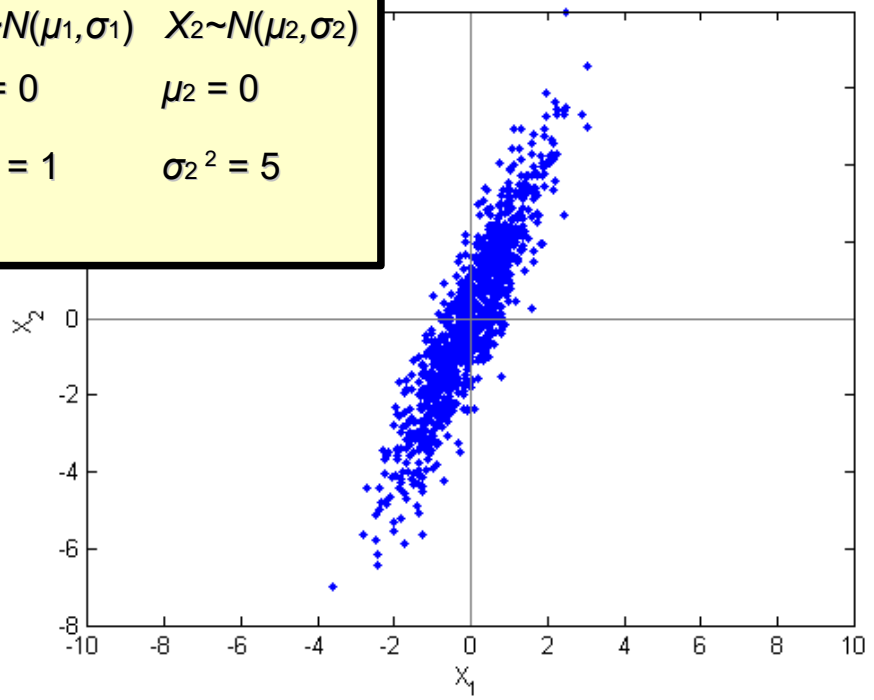


$X_1 \sim N(\mu_1, \sigma_1)$ $X_2 \sim N(\mu_2, \sigma_2)$

$\mu_1 = 0$ $\mu_2 = 0$

$\sigma_1^2 = 1$ $\sigma_2^2 = 5$

?



Kowariancja zmiennych losowych

Miarą zależności liniowej pomiędzy zmiennymi losowymi X_1 i X_2 jest kowariancja:

$$\mathrm{Cov}[X_1, X_2] = \mathrm{E}[(X_1 - \mathrm{E}[X_1])(X_2 - \mathrm{E}[X_2])] = \mathrm{E}[X_1 X_2] - \mathrm{E}[X_1]\mathrm{E}[X_2]$$

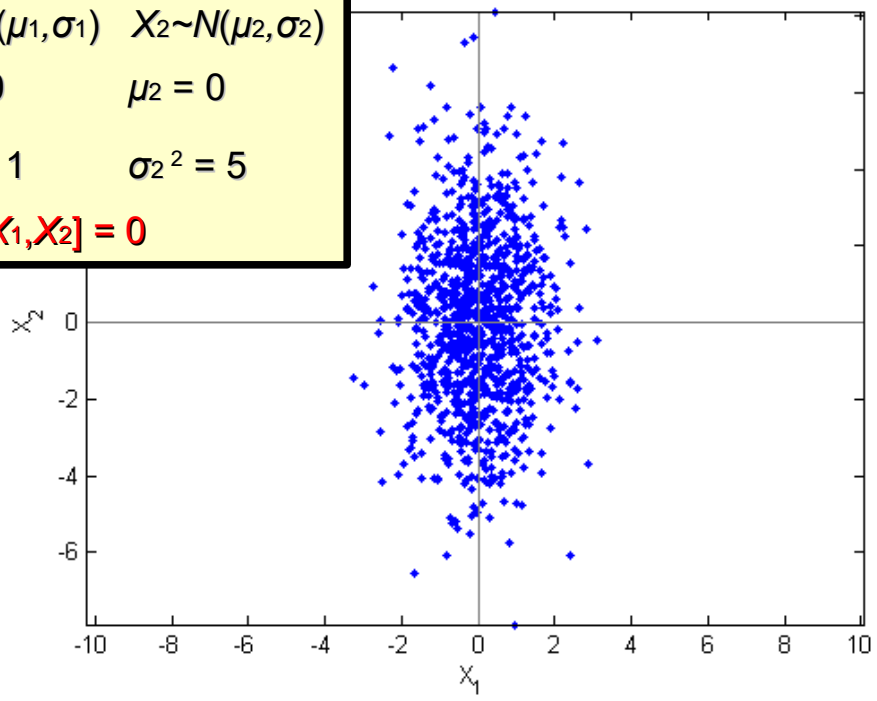
Przyjmuje ona niezerowe wartości dla zmiennych, których wartości są powiązane monotoniczną zależnością liniową, w przeciwnym razie wynosi 0.

$X_1 \sim N(\mu_1, \sigma_1)$ $X_2 \sim N(\mu_2, \sigma_2)$

$\mu_1 = 0$ $\mu_2 = 0$

$\sigma_1^2 = 1$ $\sigma_2^2 = 5$

$\mathrm{Cov}[X_1, X_2] = 0$

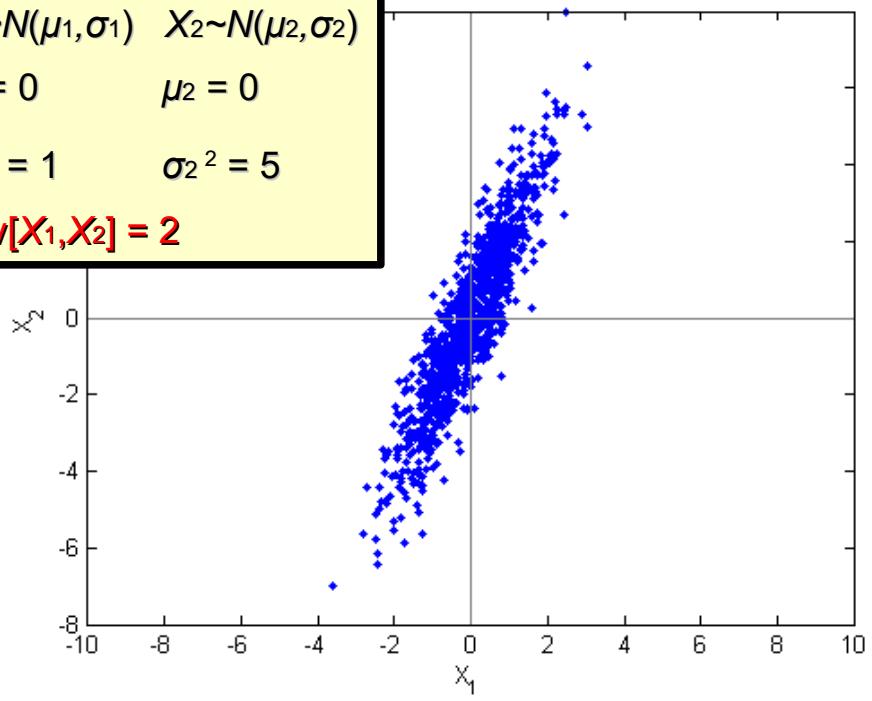


$X_1 \sim N(\mu_1, \sigma_1)$ $X_2 \sim N(\mu_2, \sigma_2)$

$\mu_1 = 0$ $\mu_2 = 0$

$\sigma_1^2 = 1$ $\sigma_2^2 = 5$

$\mathrm{Cov}[X_1, X_2] = 2$



Dla N-elementowej próby losowej kowariancja jest estymowana jako:

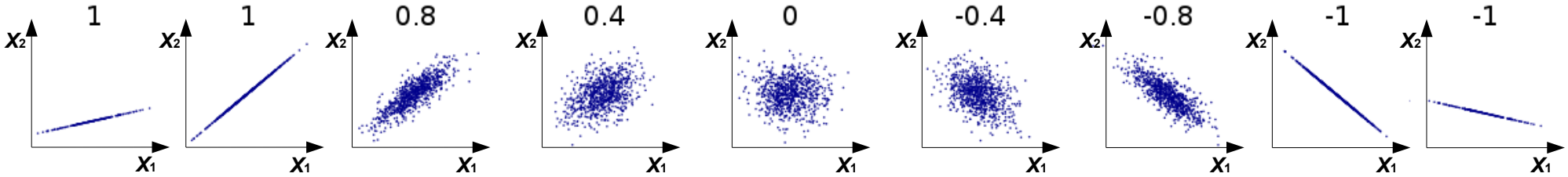
$$s^2_{X_1 X_2} = \frac{1}{N-1} \sum_{i=1}^{N} (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)$$

Współczynnik korelacji liniowej zmiennych losowych

Wartość kowariancji zależy od jednostek zmiennych losowych X_1 i X_2. Niezależną od jednostek miarą zależności liniowej jest współczynnik korelacji liniowej (Pearsona), równy kowariancji znormalizowanej wobec iloczynu odchyleń standardowych:

$$\rho_{X_1 X_2} = \frac{\mathrm{Cov}[X_1, X_2]}{\sqrt{\mathrm{Var}[X_1]}\sqrt{\mathrm{Var}[X_2]}}$$

Współczynnik korelacji przyjmuje wartości z przedziału $[-1, 1]$. Im większa jest jego wartość bezwzględna, tym silniejsza jest zależność liniowa między zmiennymi: $\rho = 0$ oznacza brak liniowej zależności, $\rho = 1$ oznacza dokładną dodatnią liniową zależność, $\rho = -1$ oznacza dokładną ujemną liniową zależność.



Współczynnik korelacji liniowej zmiennych losowych

Estymatorem współczynnika korelacji liniowej z próby dla zmiennych X_1 i X_2 jest:

$$r_{X_1X_2} = \frac{\sum_{i=1}^{N}(x_{1i}-\bar{x}_1)(x_{2i}-\bar{x}_2)}{\sqrt{\sum_{i=1}^{N}(x_{1i}-\bar{x}_1)^2}\sqrt{\sum_{i=1}^{N}(x_{2i}-\bar{x}_2)^2}} = \frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{x_{1i}-\bar{x}_1}{s_{x_1}}\right)\left(\frac{x_{2i}-\bar{x}_2}{s_{x_2}}\right)$$

gdzie to $\bar{x}$ i s to estymatory średniej i odchylenia standardowego z próby:

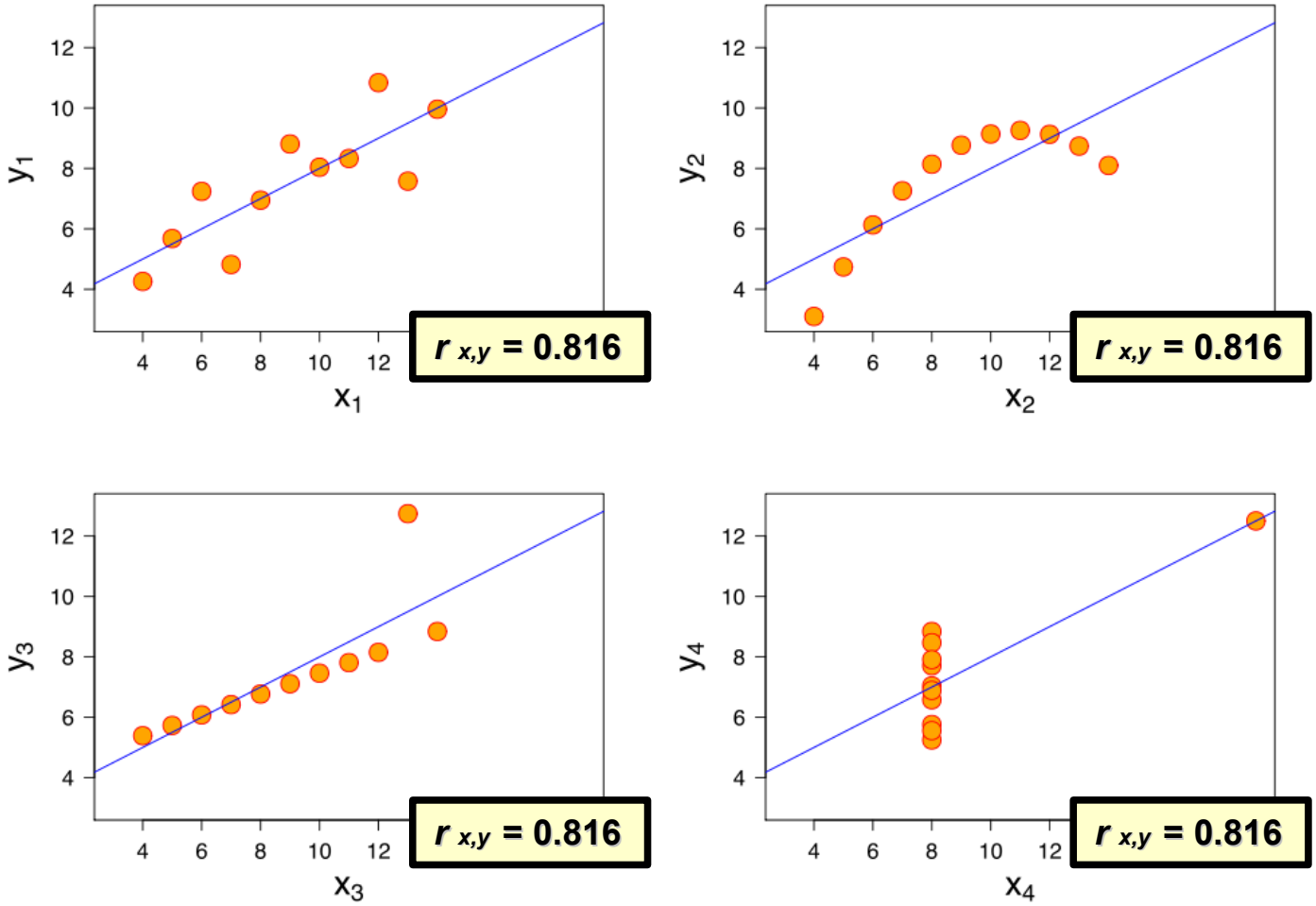
$$\bar{x} = \hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad s = \hat{\sigma} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i-\bar{x})^2}$$

	MATLAB	R
Wsp. korelacji liniowej:	`corcoeff` `corr`	`cor`

Współczynnik korelacji liniowej zmiennych losowych

Współczynnik korelacji Pearsona jest tylko miarą zależności liniowych pomiędzy zmiennymi losowymi. Ponadto charakteryzuje się on małą odpornością na występowanie w próbie wartości odstających (co wynika ze stosowania przy jego obliczaniu średnich arytmetycznych).

Za przykład obu tych właściwości może posłużyć tzw. kwartet Anscombe'a, złożony z czterech par cech o jednakowych współczynnikach korelacji, równych 0.816**.**

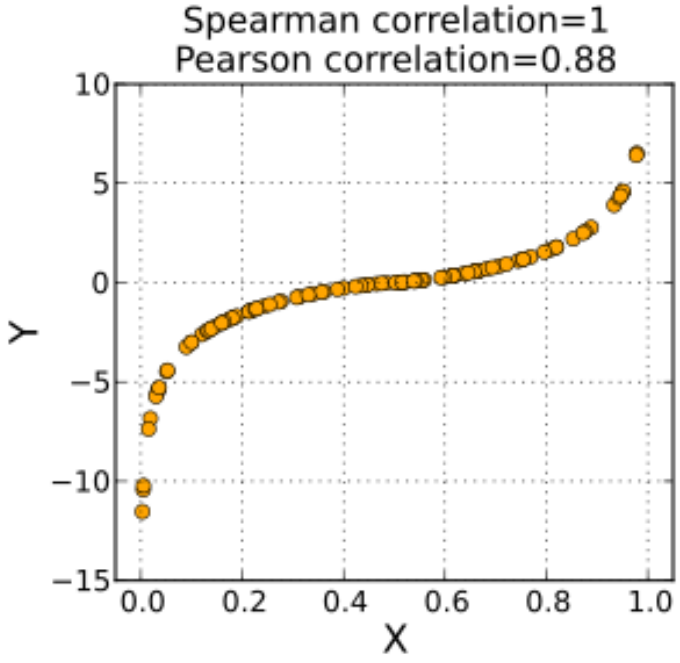


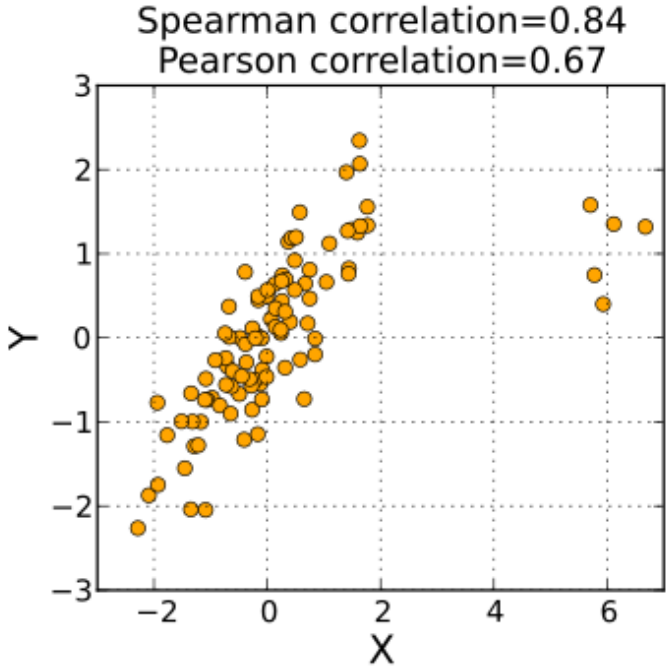
Rangowy współczynnik korelacji zmiennych losowych

Do badania dowolnych zależności monotonicznych, niekoniecznie liniowych, lepszy może okazać się rangowy współczynnik korelacji Spearmana, którego dodatkową zaletą jest większa odporność na występowanie w próbie obserwacji odstających

Współczynnik korelacji Spearmana zdefiniowany jest jako współczynnik korelacji liniowej rang wartości zmiennych, czyli pozycji w ciągu powstałym po posortowaniu niemalejąco wartości próby.

Zakres wartości współczynnika Spearmana (od -1 do 1) oraz jego interpretacja są takie same jak w przypadku współczynnika Pearsona.





Rangowy współczynnik korelacji zmiennych losowych

Algorytm wyznaczania współczynnika korelacji Spearmana

1. Wykonanie niezależnego rangowania porównywanych zmiennych, czyli:
   - posortowanie w kolejności niemalejącej wartości próby;
   - przypisanie każdej wartości x_i rangi Rx_i równej pozycji danej wartości w uporządkowanym ciągu (tzn. najmniejsza wartość otrzymuje rangę 1 itd.). Jeżeli w próbie występują powtórzenia tej samej wartości, to każde wystąpienie otrzymuje taką samą rangę, równą średniej arytmetycznej pozycji w uporządkowanym ciągu.

2. Wyznaczenie współczynnika korelacji liniowej rang za pomocą zależności:

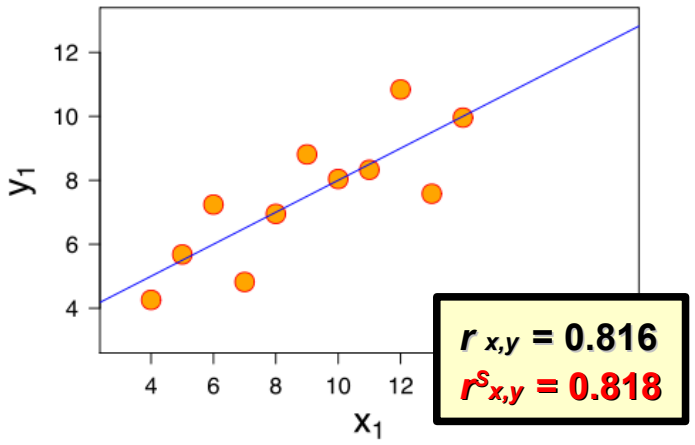
$$r^{S}_{X_1 X_2} = \frac{\sum_{i=1}^{n} \left(Rx_{1i} - \bar{Rx}_1\right)\left(Rx_{2i} - \bar{Rx}_2\right)}{\sqrt{\sum_{i=1}^{n} \left(Rx_{1i} - \bar{Rx}_1\right)^2} \sqrt{\sum_{i=1}^{n} \left(Rx_{2i} - \bar{Rx}_2\right)^2}}$$

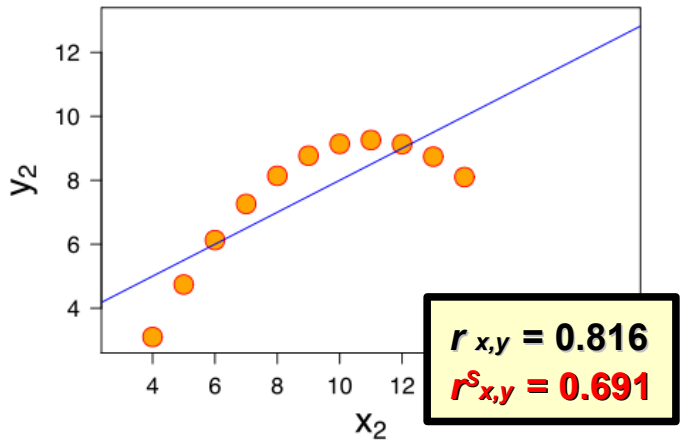
gdzie Rx_{1i} i Rx_{2i} to rangi i-tych elementów prób zmiennych losowych X_1 i X_2.

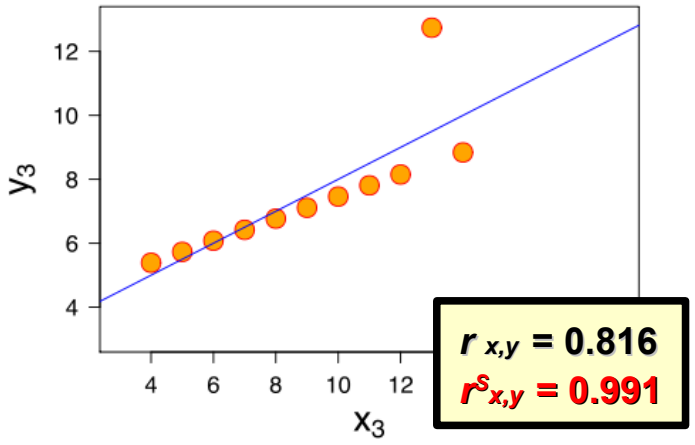
	MATLAB	R
Wsp. korelacji Spermana:	`corr`	`cor`

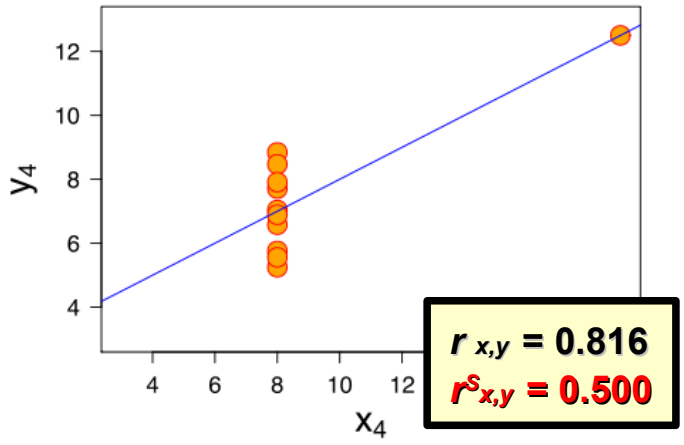
Rangowy współczynnik korelacji zmiennych losowych

Porównanie wartości współczynników korelacji Pearsona i Spearmana dla danych wchodzących w skład kwartetu Anscombe'a:









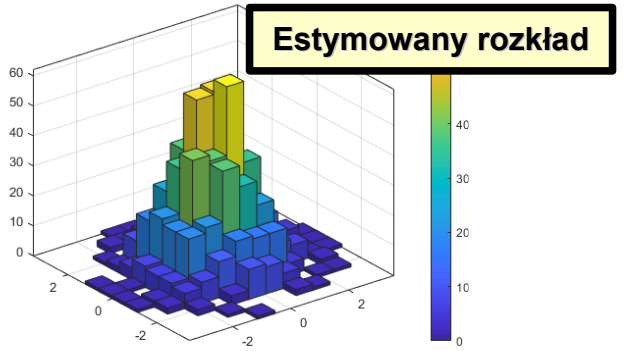
Empiryczny wielowymiarowy rozkład prawdopodobieństwa cech

W przypadku wielowymiarowym – podobnie jak dla zmiennych jednowymiarowych – możliwe jest podejście nieparametryczne, z pełną estymacją empirycznego rozkładu prawdopodobieństwa z próby za pomocą histogramów i estymatorów jądrowych.

Próba losowa

$$X = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1N} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2N} \end{bmatrix}$$

Estymowany rozkład

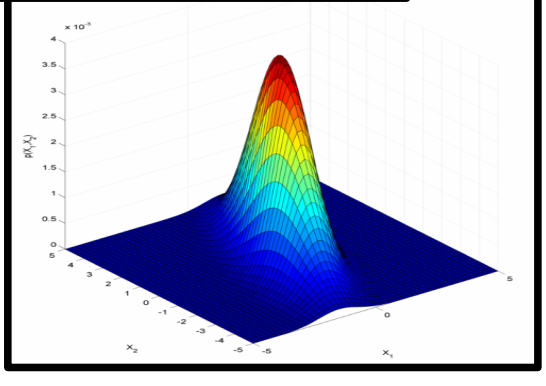


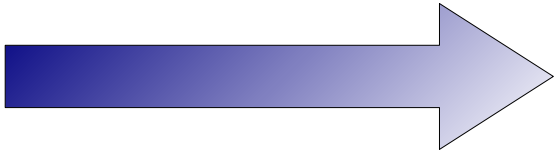
Jednak dla danych o dużej wymiarowości zdecydowanie częściej stosowane jest podejście parametryczne, w którym z góry zakłada się pewną funkcję rozkładu (np. Gaussa), a na podstawie próby estymuje się jedynie jej parametry.

Próba losowa

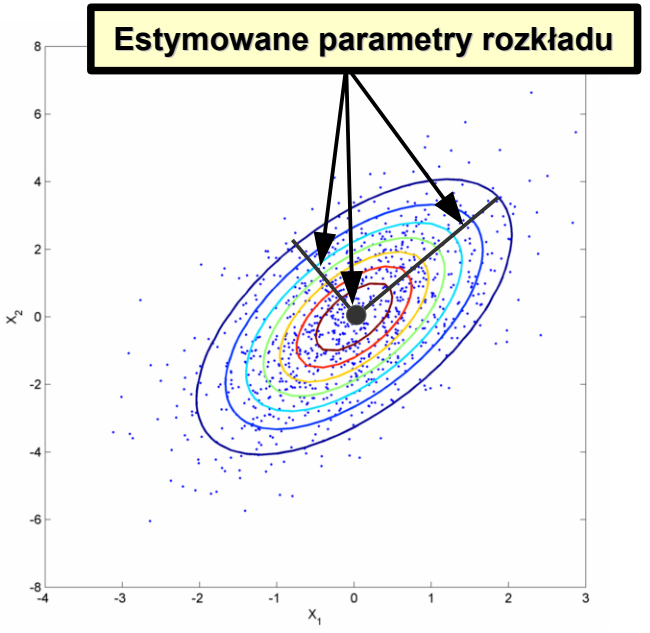
$$X = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1N} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2N} \end{bmatrix}$$

Założona funkcja rozkładu





Estymowane parametry rozkładu



Empiryczny wielowymiarowy rozkład prawd. cech: rozkład normalny

Funkcja gęstości P-wymiarowego rozkładu normalnego przyjmuje postać:

$$f(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})=\frac{1}{(2\pi)^{P/2}|\boldsymbol{\Sigma}|^{1/2}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

W powyższym wzorze μ jest wektorem wartości średnich, czyli P-wymiarowym wektorem, którego i-ty element jest wartością średnią i-tej cechy.

$$\boldsymbol{\mu}=\begin{bmatrix}\mu_1\\ \mu_2\\ \vdots\\ \mu_P\end{bmatrix}$$

Σ jest symetryczną macierzą kowariancji o wymiarze $P \times P$, której element (i,j) jest kowariancją cech i-tej i j-tej cechy ($|\Sigma|$ oznacza wyznacznik tej macierzy).

$$\boldsymbol{\Sigma}=\begin{bmatrix}\sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1P}\\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2P}\\ \vdots & \vdots & \ddots & \vdots\\ \sigma_{P1} & \sigma_{P2} & \cdots & \sigma_P^2\end{bmatrix}$$

$$\sigma_{ij}^2=\mathrm{Cov}[X_i,X_j]$$

$$\sigma_{ii}^2=\sigma_i^2=\mathrm{Cov}[X_i,X_i]=\mathrm{Var}[X_i]$$

	MATLAB	R
Funkcja gęstości:	`mvnpdf`	`dmvnorm`
Dystrybuanta:	`mvncdf`	-
Pobranie próby z rozkładu:	`mvnrnd`	`mvrnorm`

Empiryczny wielowymiarowy rozkład prawd. cech: rozkład normalny

Funkcja gęstości P-wymiarowego rozkładu normalnego przyjmuje postać:

$$f(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})=\frac{1}{(2\pi)^{P/2}|\boldsymbol{\Sigma}|^{1/2}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{T}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

W powyższym wzorze μ jes **wym wektorem, którego i-ty elemer**

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

$$f(x;\mu,\sigma)=\frac{1}{(2\pi)^{1/2}(\sigma^{2})^{1/2}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\mu)(\sigma^{2})^{-1}(x-\mu)\right)$$

Σ jest symetryczną macierzą jest kowariancją cech i-tej i j-tej ce

$$\boldsymbol{\Sigma}=\begin{bmatrix}\sigma_{1}^{2} & \sigma_{12} & \cdots & \sigma_{1P}\\ \sigma_{21} & \sigma_{2}^{2} & \cdots & \sigma_{2P}\\ \vdots & \vdots & \ddots & \vdots\\ \sigma_{P1} & \sigma_{P2} & \cdots & \sigma_{P}^{2}\end{bmatrix}$$

$$\sigma_{ij}=\mathrm{Cov}[X_i,X_j]$$

$$\sigma_{ii}^{2}=\sigma_{i}^{2}=\mathrm{Cov}[X_i,X_i]=\mathrm{Var}[X_i]$$

	MATLAB	R
Funkcja gęstości:	`mvnpdf`	`dmvnorm`
Dystrybuanta:	`mvncdf`	–
Pobranie próby z rozkładu:	`mvnrnd`	`mvrnorm`

Empiryczny wielowymiarowy rozkład prawd. cech: rozkład normalny

Wartości nieznanych parametrów μ i Σ wielowymiarowego rozkładu normalnego mogą być estymowane na podstawie N-elementowej próby losowej jako:

- **wektor średnich arytmetycznych**

$$\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}} = \begin{bmatrix} \hat{\mu}_1 \\ \vdots \\ \hat{\mu}_P \end{bmatrix} = \begin{bmatrix} \frac{1}{N}\sum_{i=1}^{N} x_{1i} \\ \vdots \\ \frac{1}{N}\sum_{i=1}^{N} x_{Pi} \end{bmatrix}$$

$$\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i$$

- **macierz kowariancji z próby**

$$\hat{\boldsymbol{\Sigma}} = \mathbf{S} = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1P} \\ s_{21} & s_{22} & \cdots & s_{2P} \\ \vdots & \vdots & \ddots & \vdots \\ s_{P1} & s_{P2} & \cdots & s_{PP} \end{bmatrix}$$

$$s_{kj} = \frac{1}{N-1}\sum_{i=1}^{N} (x_{ki} - \bar{x}_k)(x_{ji} - \bar{x}_j)$$

$$\hat{\boldsymbol{\Sigma}} = \mathbf{S} = \frac{1}{N-1}\sum_{i=1}^{N} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T = \frac{1}{N-1}(\mathbf{X} - \bar{\mathbf{x}})(\mathbf{X} - \bar{\mathbf{x}})^T$$

	MATLAB	R
Wektor wart. średnich:	`mean`	`mean`
Macierz kowariancji:	`cov`	`cov`

Empiryczny wielowymiarowy rozkład prawd. cech: rozkład normalny

Przykład dla $P = 2$:

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^{\,2} & \sigma_{12} \\ \sigma_{21} & \sigma_2^{\,2} \end{bmatrix} = \begin{bmatrix} 1 & 1.2 \\ 1.2 & 4 \end{bmatrix}$$