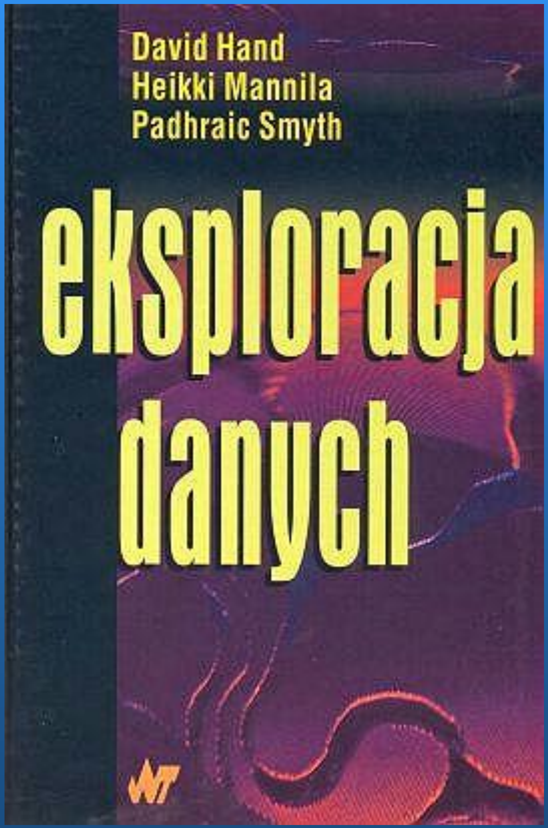
Eksploracja Danych

Cześć 1

Literatura

Książka z której korzystamy:



Definicja

Eksploracja danych jest analizą (często ogromnych) zbiorów danych obserwacyjnych w celu znalezienia nieoczekiwanych związków (zależności) i podsumowania danych w oryginalny sposób tak, aby były zarówno zrozumiałe jak i przydatne dla ich własciciela.

Zależności i podsumowania o których mowa nazywane są modelami lub wzorcami.

Przykłady:
równania liniowe, reguły, skupienia, grafy, struktury drzewiaste, wzorce rekurencyjne

Ogromne zbiory danych (10 lat temu)

- Biblioteka Kongresu – 20 TB (10^{12})
- World Data Centre for Climate – 220 TB
- YouTube – 47 TB
- China Mobile – 655 milionów klientów!!!
- "There was 5 exabytes (10^{18}) of information created between the dawn of civilization through 2003, but that much information is now created every 2 days, and the pace is increasing." Google CEO Eric Schmidt

Uwagi do definicji

Uwaga 1

Eksploracja danych nazwyane jest czasami „wtórną” analizą danych.

Dane, które są analizowane zostały wcześniej zgromadzone z przyczyn innych niż analizy prowadzące do wydobywania wiedzy np. *spisy rozmów telefonicznych*, *zapisy transakcji bakowych* etc.

Cele eksploracji danych nie odgrywają żadnej roli w strategii gromadzenia danych.

Uwagi do definicji

Uwaga 2

Dane poddawane eksploracji są ogromne. Stąd liczne problemy:

- jak wyznaczyć reprezentatywną próbkę danych?
- jak analizować dane w rozsądnycm czasie?
- jak uogólniać?
- jak decydować czy wykryta zależność jest jedynie przypadkowa i nie odzwierciedla jakiejś wewnętrznej rzeczywistości?

Uwagi do definicji

Uwaga 3

Zależności i struktury odkryte w zbiorze danych muszą być nowatorskie.

Nie ma sensu powtarzanie znanych już zależności lub zależności nieuniknionych (np. każda ciężarna jest kobietą)

Nowatorstwo zależy oczywiście od wcześniejszej wiedzy użytkowników

Techniki i metody

Techniki i metody wykorzystywane w eksploracji danych:

- ✓ metody statystyczne
- ✓ sieci neuronowe
- ✓ metody uczenia maszynowego
- ✓ metody ewolucyjne
- ✓ logika rozmyta
- ✓ zbiory przybliżone

Eksploracja a bazy danych

zapytania do bazy danych	*eksploracja danych*
ile sprzedano coli a ile wody mineralnej w poszczególne dni tygodnia	co kupowali klienci kupujący colę
jakie są najczęściej wybierane wycieczki	jakie strony odwiedziły osoby po obejrzeniu stron biura podróży
jaką najwyższą temperaturę ma pacjent z grypą	jakie są objawy grypy
ile jest zaobserwowanych Asteroid	kiedy uderzy w nas asteroida

Podział eksploracji danych

- eksploracyjna analiza danych (EDA)
- modelowanie opisowe:
  - modele całościowego rozkładu prawdopodobieństwa
  - danych (estymacja gęstości)
  - dzielenie przestrzeni na grupy (analiza skupień,
  - segmentacja)
  - modele opisujące związki między zmiennymi (modelowanie zależności)
- modelowanie przewidujące (predykcyjne):
  - klasyfikacja
  - regresja
- odkrywanie wzorców i reguł
- wyszukiwanie według zawartości

Poszukiwanie zależności

Proces poszukiwania zależności wymaga kilku kroków:

- ustalenie rodzaju i struktury używanej reprezentacji
- wybór funkcji oceny
- wybór algorytmy optymalizacji funkcji oceny
- decyzja, jakie zasady zarządzania danymi są wymagane, by algorytmy były wykonywane efektywnie

Poszukiwanie zależności

Przykład

Chcemy zbudować model przewidywania (predictive model), który pozwoli przewidywać roczne wydatki z karty kredytowej osób mając dany ich roczny dochód.

Model nasz ma wiązać zmienną przewidującą (predictor) X ze zmienną wynikową (response) Y np.

$$Y=aX+b$$

gdzie: X – dochód, Y- wydatki

Poszukiwanie zależności

W tym przypadku scenariusz byłby następujący:

- reprezentacją jest model z poprzedniego slajdu (zależność liniowa)
- funkcją oceny najcześciej stosowaną w takich przypadkach jest suma kwadratów rozbieżności między przewidywanym wydawaniem pieniedzy a wydawaniem realnym.

$$E(a,b) = \sum_{n=1}^{N} (y_n - (ax_n + b))^2$$

Poszukiwanie zależności

W tym przypadku scenariusz byłby następujący (cd):

- w przypadku regresji liniowej algorytm optymalizujący jest prosty np.: a i b mogą być wyrażone jako jawne funkcje zaobserwowanych wartości wydawania pieniędzy i dochodu.

$$\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sum_{n=1}^{N} x_n^2 & \sum_{n=1}^{N} x_n \\ \sum_{n=1}^{N} x_n & \sum_{n=1}^{N} 1 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{n=1}^{N} x_n y_n \\ \sum_{n=1}^{N} y_n \end{pmatrix}$$

Poszukiwanie zależności

- w przypadku regresji liniowej pojawia się niewiele problemów związanych z zarządzaniem danymi. Do obliczenia wartości a i b wystarczą proste podsumowania danych (sumy, sumy kwadratów, sumy iloczynów wartości X i Y)

$$\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sum_{n=1}^{N} x_n^2 & \sum_{n=1}^{N} x_n \\ \sum_{n=1}^{N} x_n & \sum_{n=1}^{N} 1 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{n=1}^{N} x_n y_n \\ \sum_{n=1}^{N} y_n \end{pmatrix}$$

Zbiory danych

Zbiory danych

Zbiór danych to zbiór pomiarów pobrany z pewnego środowiska lub procesu.

W najprostrzym przypadku dysponujemy kolekcją n obiektów i dla każdego obiektu mamy zbiór tych samych p pomiarów.

Zbiór pomiarów możemy zatem rozpatrywać jako macierz danych n x p.

Zbiory danych

Liczba n wierszy macierzy odpowiada n obiektom dla których zostały dokonane pomiary.

Wiersze takie mogą mieć różne nazwy: jednostki, instancje, encje, przypadki, rekordy.

Drugi wymiar naszej macierzy zawiera zbiór p pomiarów wykonanych dla każdego (lub nie) obiektu.

Kolumny takie mogą mieć różne nazwy: zmienne, cechy, atrybuty, pola.

Zbiory danych

Przykład

Example No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
			ed	No
			ed	Yes
			tic	No
			ed	No
			ed	Yes

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Zbiory danych

Przykład

Przykłady danych w zbiorach danych Public Use Microdata Sample

Identyfikator	Wiek	Płeć	Stan cywilny	Wykształcenie	Dochód
248	54	Mężczyzna	Żonaty	Absolwent szkoły średniej	100 000
249	??	Kobieta	Zamężna	Absolwent szkoły średniej	12 000
250	29	Mężczyzna	Żonaty	Szkoła pomaturalna	23 000
251	9	Mężczyzna	Stanu wolnego	Dziecko	0
252	85	Kobieta	Stanu wolnego	Absolwent szkoły średniej	19 798
253	40	Mężczyzna	Żonaty	Absolwent szkoły średniej	40 100
254	38	Kobieta	Stanu wolnego	Nieukończona pierwsza klasa	2691
255	7	Mężczyzna	??	Dziecko	0
256	49	Mężczyzna	Żonaty	11 klas	30 000
257	76	Mężczyzna	Żonaty	Stopień doktora	30 686

Zbiory danych

Przykład (cd)

Przykłady danych w zbiorach danych Public Use Microdata Sample

Identyfikator	Wiek	Płeć	Stan cywilny	Wykształcenie	Dochód
248	54	Mężczyzna	Żonaty	Absolwent szkoły średniej	100 000
249	??	Kobieta	Zamężna	Absolwent szkoły średniej	12 000
250	29	Mężczyzna	Żonaty	Szkoła pomaturalna	23 000
251	9	Mężczyzna	Stanu wolnego	Dziecko	0
252	85	Kobieta	Stanu wolnego	Absolwent szkoły średniej	19 798
253	40	Mężczyzna	Żonaty	Absolwent szkoły średniej	40 100
254	38	Kobieta	Stanu wolnego	Nieukończona pierwsza klasa	2691
255	7	Mężczyzna	??	Dziecko	0
256	49	Mężczyzna	Żonaty	11 klas	30 000
257	76	Mężczyzna	Żonaty	Stopień doktora	30 686

Uwagi:

- Różne typy zmiennych: ciągłe, kategoryczne
- Brakujące dane – w dużych zbiorach zjawisko powszechne
- Szum pomiarowy – czy dochód wynosi rzeczywiście 100000$?

Zbiory danych

Przykład (cd)

Przykłady danych w zbiorach danych Public Use Microdata Sample

Identyfikator	Wiek	Płeć	Stan cywilny	Wykształcenie	Dochód
248	54	Mężczyzna	Żonaty	Absolwent szkoły średniej	100 000
249	??	Kobieta	Zamężna	Absolwent szkoły średniej	12 000
250	29	Mężczyzna	Żonaty	Szkoła pomaturalna	23 000
251	9	Mężczyzna	Stanu wolnego	Dziecko	0
252	85	Kobieta	Stanu wolnego	Absolwent szkoły średniej	19 798
253	40	Mężczyzna	Żonaty	Absolwent szkoły średniej	40 100
254	38	Kobieta	Stanu wolnego	Nieukończona pierwsza klasa	2691

Typowym zadaniem dla takich danych jest znajdowanie różnego rodzaju zależności np:

- Zależnosć miedzy dochodem i innymi zmiennymi.
- Czy istnieją naturalnie wyodrębnione grupy ludzi czyli czy istnieją wartości na których zmienne często się pokrywają.
- Inne?

Typy pomiarów

Istnieją dwa typy pomiarów:

- ilościowe – zmienne ilościowe odmierzane są na skali numerycznej i mogą przyjąć (teoretycznie) każdą wartość

  Przykład: kolumny wiek i dochód

- kategoryczne – zmienne kategoryczne mogą przyjąć tylko pewne dyskretne wartości

  Przykład: kolumny płeć, wykształcenie, stan cywilny

  Dwa rodzaje: porządkowe (wykształcenie) lub symboliczne (płeć)

Typy pomiarów

<u>Uwaga</u>

Technika analizy danych właściwa dla jednego typu może nie być właściwa dla innego.

<u>Przykład 1</u>

Załóżmy, że stan cywilny reprezntujemy liczbami całkowitymi:

1 - panna/kawaler,
2 – zamężna/żonaty
3 – wdowa/wdowiec

Oczywiście nie ma sensu obliczać w tym przypadku średniej arytmetycznej.

Jeszcze o zbiorach danych

Przykład 2

Rozważmy zbiór danych dotyczących pacjentów. Może on zawierać:

- wielokrotne pomiary taj samej zmiennej (np. ciśnienie krwi) dokonywane w różne dni o różnych porach.
- dane obrazowe (np. prześwitlenia)
- dane w postaci tekstu (komentarze, opisy objawów)

Mogą istnieć zależności między pacjentami i lekarzami, szpitalami i położeniem geograicznym.

Jeszcze o zbiorach danych

Danych takich zwykle nie można zapisać w prostej macierzy n x p

Dużo informacji może być „spłaszczonych” do takiej macierzy.

W czasie wykładu zasadniczo będziemy przyjmowali że dane znajdują się w takiej macierzy… chyba, że zasygnalizujemy inaczej.

Inne określenia: zbiór danych, dane treningowe, próbka, baza danych.

Jeszcze o zbiorach danych

Przykład 1

Rozważmy zbiór dokumentów tekstowych np. stron WWW.

Czy możemy go rozpatrywać jako macierz?

- Rozwiązanie 1: wiersze – dokumenty, kolumny – słowa. Dany wpis (d,w) to 1 (słowo występuje) lub 0 (słowo nie występuje).
- Rozwiązanie 2: wiersze – dokumenty, kolumny – słowa. Dany wpis (d,w) to ilość wystąpień słowa w w dokumencie d.

Ale uwaga: tracimy informację o porządku słów!

Jeszcze o zbiorach danych

Przykład 2

Rozważmy teraz dziennik trasakcji w sieci WWW.

(użytkownik, strona WWW, czas)

Czy możemy go rozpatrywać jako macierz?

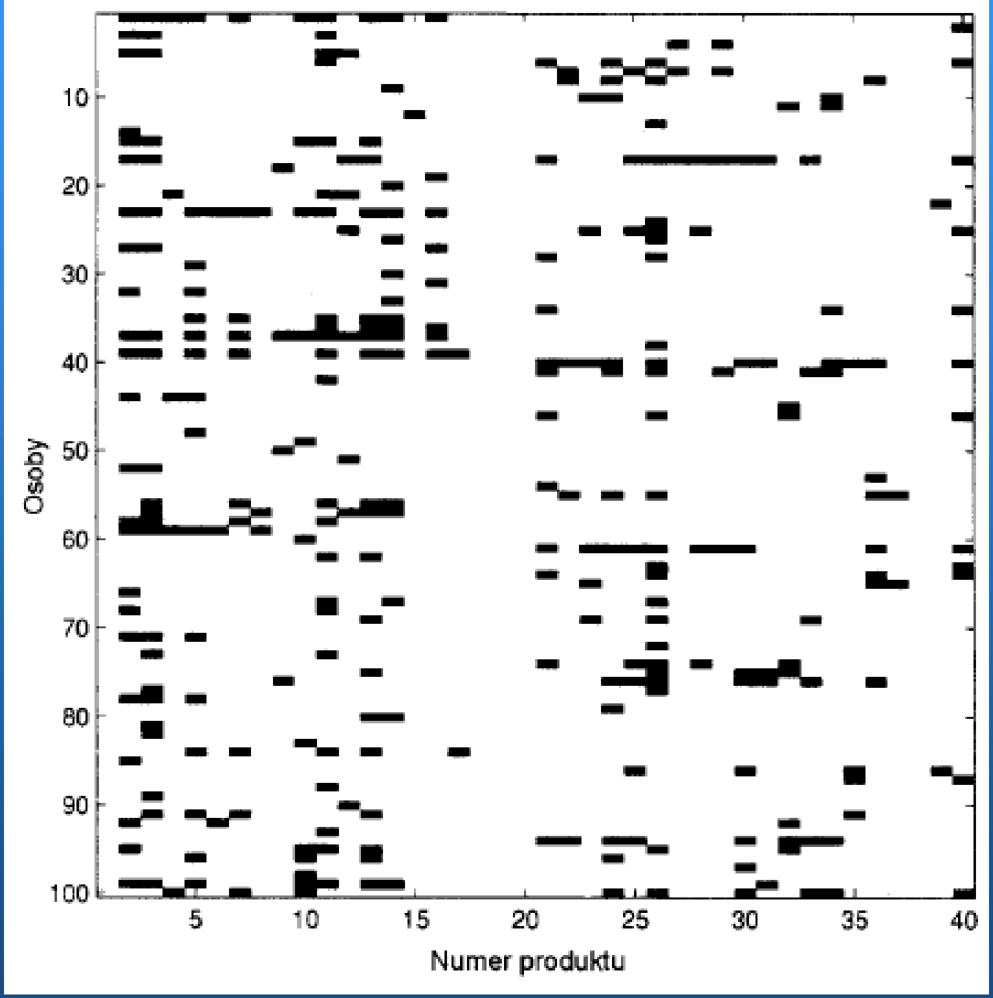
- Rozwiązanie 1: wiersze – pojedyncza osoba , kolumny – strona WWW. Dany wpis (o,w) to 1 (osoba odwiedziła stronę) lub 0 (nie odwiedziła).
- Rozwiązanie 2: wiersze – pojedyncza osoba , kolumny – strona WWW. Dany wpis (o,w) to liczba odwiedzin strony przez daną osobę.

Ale uwaga: tracimy informację o czasie!

Jeszcze o zbiorach danych

Przykład 3

Rozważmy dane transakcji w sklepie internetowym.



Modele

Efektem eksploracji danych mogą być modele.

Model – globalne podsumowanie zbioru danych. Model mówi coś o każdym punkcie przestrzeni pomiarowej.

Jeżeli wiersze macierzy danych rozważymy jako punkty w przestrzeni p-wymiarowej model może mówić coś o każdym takim punkcie. Nawet jeżeli brakuje niektórych pomiarów.

Prosty model: $Y = aX + b$ (X, Y – zmienne, a, b - parametry)

Modele i wzorce

	OPISOWY	*PREDYKCYJNY*
WZORZEC	tworzony w celu znalezienia nietypowych własności	pozwalający przewidzieć, które z obiektów będą miały niezwykłe właściwości
MODEL	podsumowujący dane	pozwalający formułować wnioski o populacji lub o prawdopodobnych wartościach przyszłych

Komponenty algorytmów ED

Algorytmy wydobywania wiedzy mają następujące cztery komponenty:

- Struktura modelu lub wzorca: ustalenie bazowej struktury lub postaci funkcyjnej
- Funkcja oceny: ocena jakości dopasowanego modelu
- Metoda optymalizacji i przeszukiwania: zoptymalizowane funkcje oceny i przeszukiwanie różnych struktur modeli i wzorców
- Strategia zarządzenia danymi: zapewnienie sprawnego dostępu do danych

Funkcje oceny

Funkcje oceny mierzą na ile dobrze model lub struktura parametryczna pasuje do danego zbioru danych.

Funcje oceny umożliwiają weryfikację czy jeden model jest lepszy niż inny. Pozwalają także dobrać wartości parametrów modeli.

Przykład

Sumaryczny błąd kwadratowy w naszym modelu:

$$\sum_{i=1}^{n}(y(i)-\hat{y}(i))^2,$$

Eksploracja a statystyka

- Eksploracja dotyczy ogromnych baz danych!
- Statystka może nie wystarczyć w przypadku tak ogromnych zbiorów danych (nie wszystko widać!).
- Cel eksploracji to wnioski dotyczące obiektów spoza dostępnej bazy danych.
- Powinniśmy unikać modeli lub wzorców, które są zbyt dokładnie dopasowane do bazy danych (nadmierne dopasowanie)

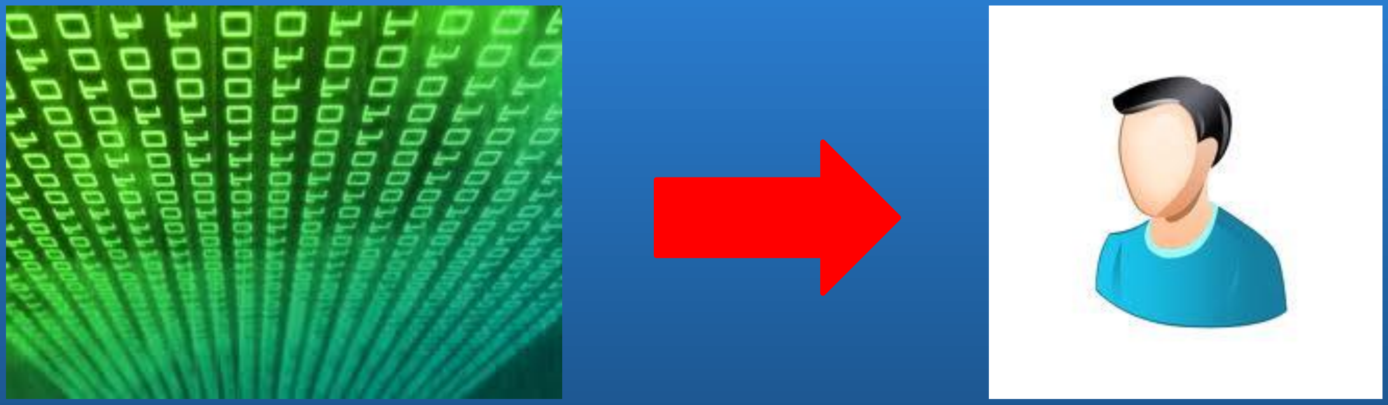
Ale uwaga…

Nie powinniśmy jednak spodziewać się otrzymania odpowiedzi na wszystkie pytania. Jak wszystkie procesy odkrywcze, udana eksploracja danych zawiera pierwiastek szczęśliwego trafu. Chociaż eksploracja danych dostarcza użytecznych narzędzi, nie oznacza to, że nieuchronnie prowadzi do ważnych, interesujących i cennych rezultatów. Musimy wystrzegać się zbytniego wyolbrzymiania prawdopodobnych wyników. Ale możliwości istnieją.

Badanie danych

Eksplorując dane chcemy odkrywać zależności w „rzeczywistym świecie” (naukowym, fizycznym, biznesowym etc.)

W tym calu badamy opisujące ten świat dane.



Można powiedzieć, że świat ten badamy od zewnątrz.

Badanie danych

Dane gromadzone są odwzorowując obiekty z dziedziny zainteresowania na reprezentację za pomocą procedury pomiarowej, która kojarzy wartość zmiennej z daną właściwością obiektu.

Zależności między obiektami modelowane są jako zależności między zmiennymi.

Nie interesują nas zależności będące artefaktami sposobu gromadzenia danych!

Miary odległości

Miary odległości

Wiele metod eksploracji danych opartych jest na miarach podobieństwa między obiektami.

Przyjmujemy, że miary podobieństwa mogą być uzyskane na dwa sposoby:

- bezpośrednio z obiektów – np. przeprowadzając badanie marketingowe na temat podobieństwa między parami produktów.
- niebezpośrednio z obiektów – z wektorów pomiarów lub właściwości opisujących każdy obiekt.

  W takim przypadku musimy jakoś zdefiniować „podobieństwo”.

Miary odległości

Warto zauważyć, że równie dobrze zamiast podobieństwa możemy operować niepodobieństwem.

Zależność między tymi tymi pojęciami może być zdefiniowana na różne sposoby np.:

$$d(i, j) = 1 - s(i, j)$$

$$d(i, j) = \sqrt{2(1 - s(i, j))}$$

W kontekście podobieństwa (niepodobieństwa) mówi się czasami o pojęciu bliskości.

Miary odległości

Pojęcie bliskości związane jest z kolei z pojęciami odległości i... metryki:

1) $d(i, j) \geq 0$ dla każdego i oraz j, ponadto $d(i, j) = 0$ wtedy i tylko wtedy, gdy $i = j$;
2) $d(i, j) = d(j, i)$ dla każdego i oraz j;
3) $d(i, j) \leq d(i, k) + d(k, j)$ dla każdego i, j oraz k.

Załóżmy, że mamy n obiektów a dla każdego z nich p pomiarów o wartościach rzeczywistych. Wektor obserwacji dla *i*-tego obiektu oznaczmy przez:

$$\mathbf{x}(i) = (x_1(i),\ x_2(i),\ \ldots,\ x_p(i)),\ 1 \leq i \leq n,$$

Miary odległości

Odległość Euklidesową definiujemy jako:

$$d_E(i,j) = \left(\sum_{k=1}^{p} (x_k(i) - x_k(j))^2 \right)^{\frac{1}{2}}.$$

Przykład

	osoba 1	osoba 2
waga	73	100
wzrost (cm)	170	180

odległość Euklidesa
28,79236

	osoba 1	osoba 2
waga	73	100
wzrost (mm)	1700	1800

odległość Euklidesa
103,5809

Zmiana jednostki powoduje zmianę ważności zmiennych!

Miary odległości

Z odległości Euklidesowej korzystamy w przypadku zmiennych współmiernych czyli np. o takiej samej jednostce (długości, wagi etc.).

W przypadku zmiennych niewspółmiernych odległość Euklidesa ma mniejszy sens.

W praktyce często mamy do czynienia ze zmiennymi niewspółmiernymi.

Jak wtedy obliczyć odległość?

Miary odległości

Popularnym rozwiązaniem jest normalizowanie danych w wyniku dzielenia każdej zmiennej przez odchylenie standardowe próbki. Dzięki temu wszystki próbki będą traktowane tak samo.

Odchylenie standardowe dla *k*-tej zmiennej X_k można obliczyć ze wzoru:

$$\hat{\sigma}_k = \left(\frac{1}{n} \sum_{i=1} (x_k(i) - \mu_k)^2 \right)^{\frac{1}{2}}$$

gdzie μ_k to wartość średnia dla zmiennej X_k

Miary odległości

Możemy ją obliczyć ze wzoru:

$$\bar{x}_k = \frac{1}{n}\sum_{i=1}^{n} x_k(i)$$

Zatem $x'_k = x_k / \hat{\sigma}_k$ eliminuje wpływ skali.

Przykład

	osoba 1	osoba 2
waga	73	100
wzrost(cm)	170	180

odległość Euklidesa

0,317326

	osoba 1	osoba 2
waga	73	100
wzrost(mm)	1700	1800

odległość Euklidesa

0,317326

Miary odległości

Odchylenie standardowe informuje nas o tym, na ile wyniki się "zmieniają” (czy rozrzut wyników wokół średniej jest mały czy wielki).

Przykład

Próbka 1 (średnie oceny): 1,1,2,2,3,3,4,4,5,5,6,6

Próbka 2 (średnie oceny): 3,3,3,3,3,3,4,4,4,4,4,4

Średnia w obu przypadkach wynosi 3.5. Średnia nie jest miarodajna bo wynika z niej, że klasy w ogóle się nie różnią.

Miary odległości

Przykład

Policzmy odchylenia standardowe dla próbek:

Próbka 1 = 1,707825

Próbka 2 = 0,5

Z obliczonych wartości odchylenia standardoego wynika, że próbka 1 jest bardziej zróżnicowana.

Miary odległości

Zmienne można też „ważyć”. Wówczas odległość obliczamy korzystając z formuły:

$$d_{WE}(i,j) = \left(\sum_{k=1}^{p} w_k (x_k(i) - x_k(j))^2 \right)^{\frac{1}{2}}$$

Odległość Euklidesowa (ważona i zwykła) są addytywne tzn. zmienne mają niezależny udział w mierzeniu odległości.

Nie zawsze jest to porządane!

Miary odległości

Kowariancję X i Y definiujemy następująco:

$$Cov(X,Y) = \frac{1}{n-1}\sum_{n=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$$

Kowariancja jest miarą tego jak X i Y różnią się od siebie.

- Duża wartość dodatnia oznacza, że duże (małe) wartości X wiążą się z dużymi (małymi) wartościami Y.
- Duża wartość ujemna oznacza zależność przeciwną.

Miary odległości

Wartość kowariancji zależy od zakresów wartości X i Y.

Zależność tę można wyeliminować dzieląc zmienne przez ich odchylenia standardowe. Otrzymujemy w ten współczynnik korelacji.

$$\rho(X,Y)=\frac{\sum_{i=1}^{n}(x(i)-\overline{x})(y(i)-\overline{y})}{\left(\sum_{i=1}^{n}(x(i)-\overline{x})^2\sum_{i=1}^{n}(y(i)-\overline{y})^2\right)^{\frac{1}{2}}}.$$

Dla p zmiennych możemy zbudować macierze kowariancji i korelacji (o wymiarach p x p).

Miary odległości

Wykres rozrzutu (wykres korelacji, diagram korelacji) służy do przedstawienia zależności między dwoma zmiennymi (cechami).

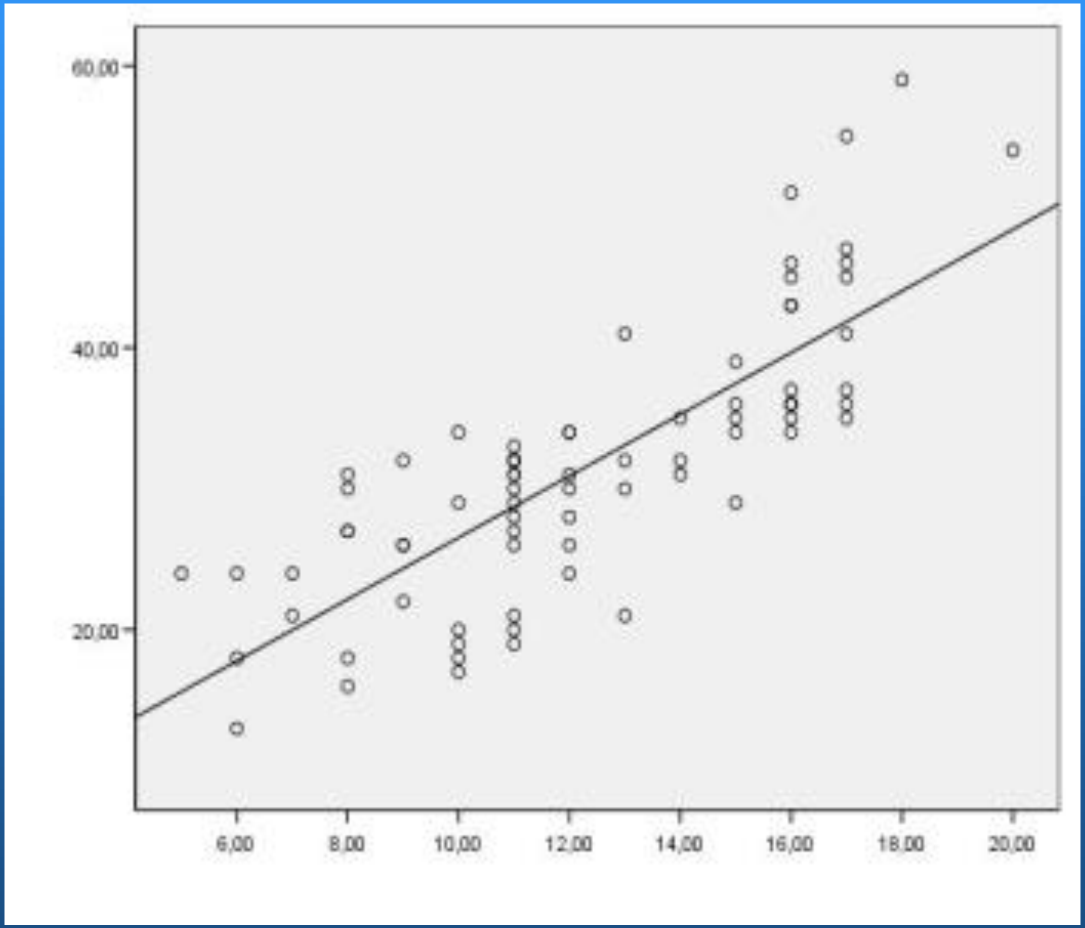
Analizując wykres rozrzutu możemy:

- stwierdzić, że zmienne są zależne
- ustalić kierunek związku
- ustalić siłę związku

Miary odległości

Przykład

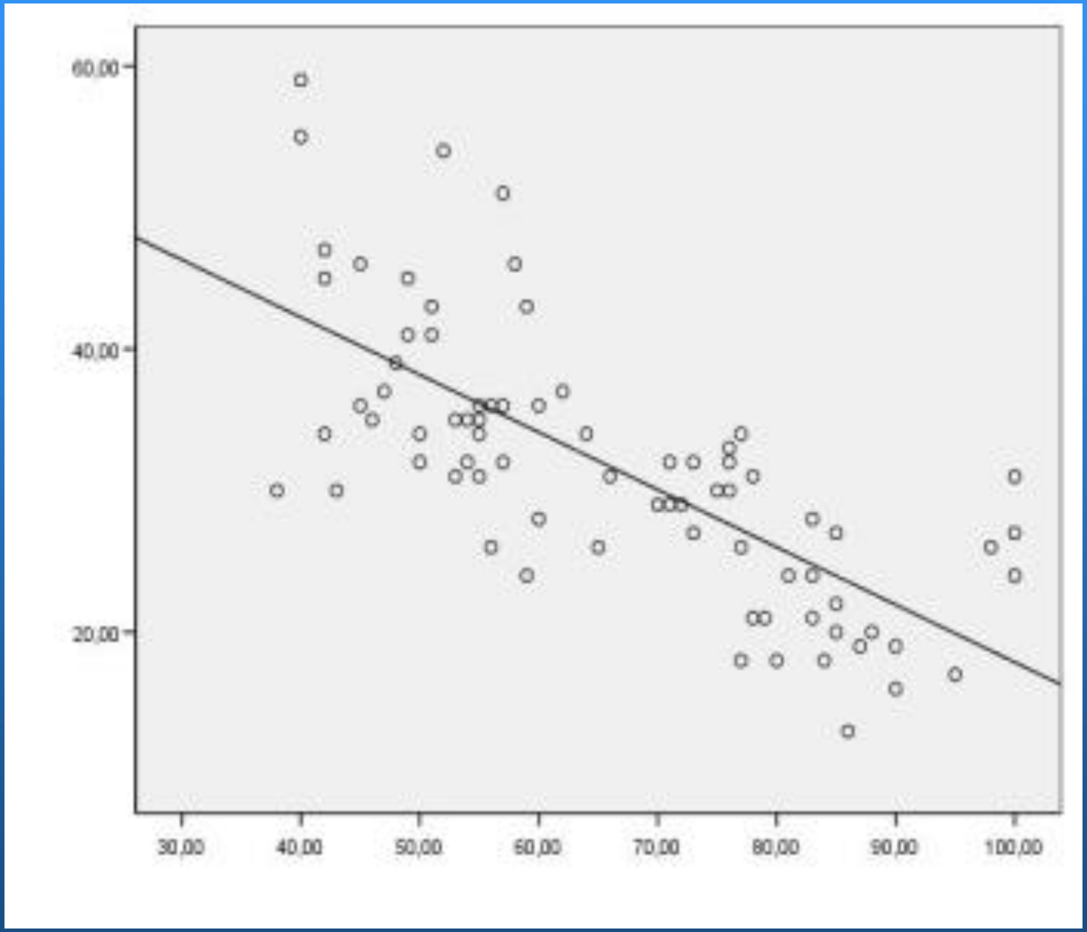
Korelacja dodatnia



Miary odległości

Przykład

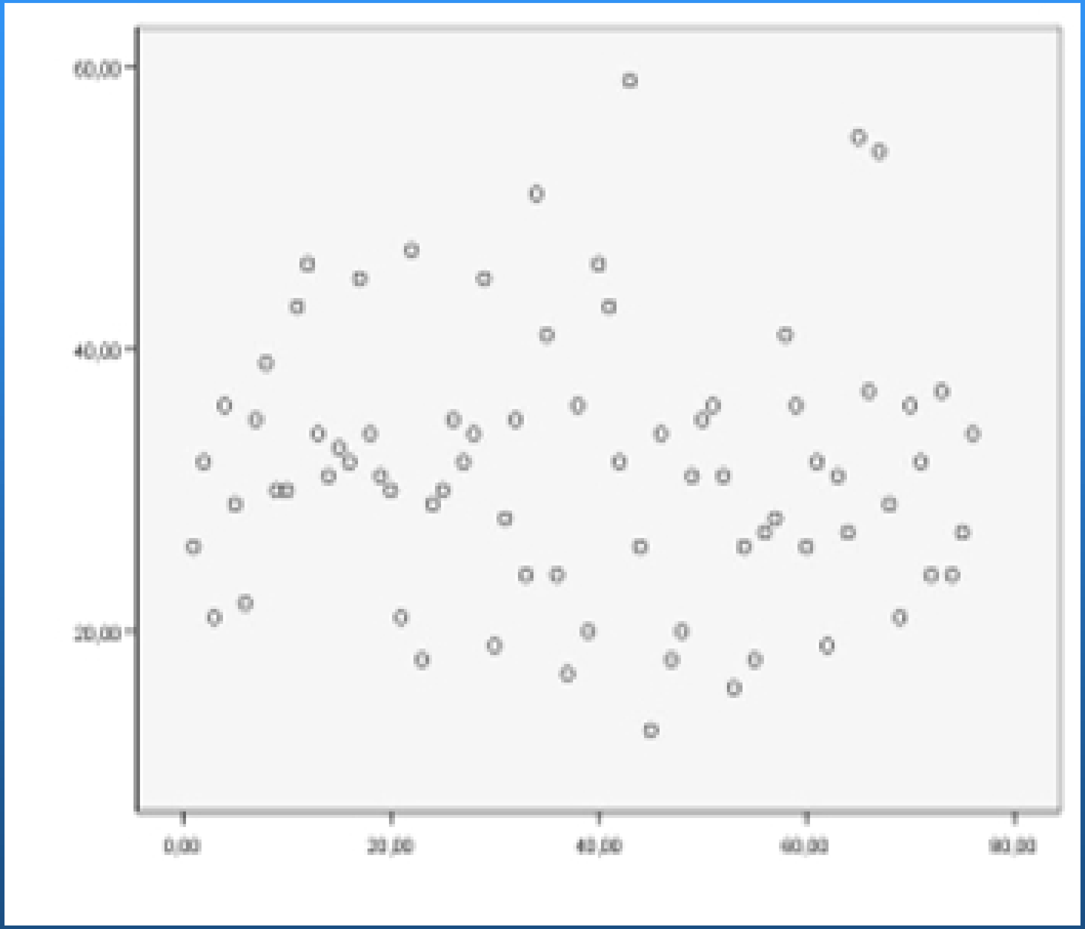
Korelacja
ujemna



Miary odległości

Przykład

Brak korelacji

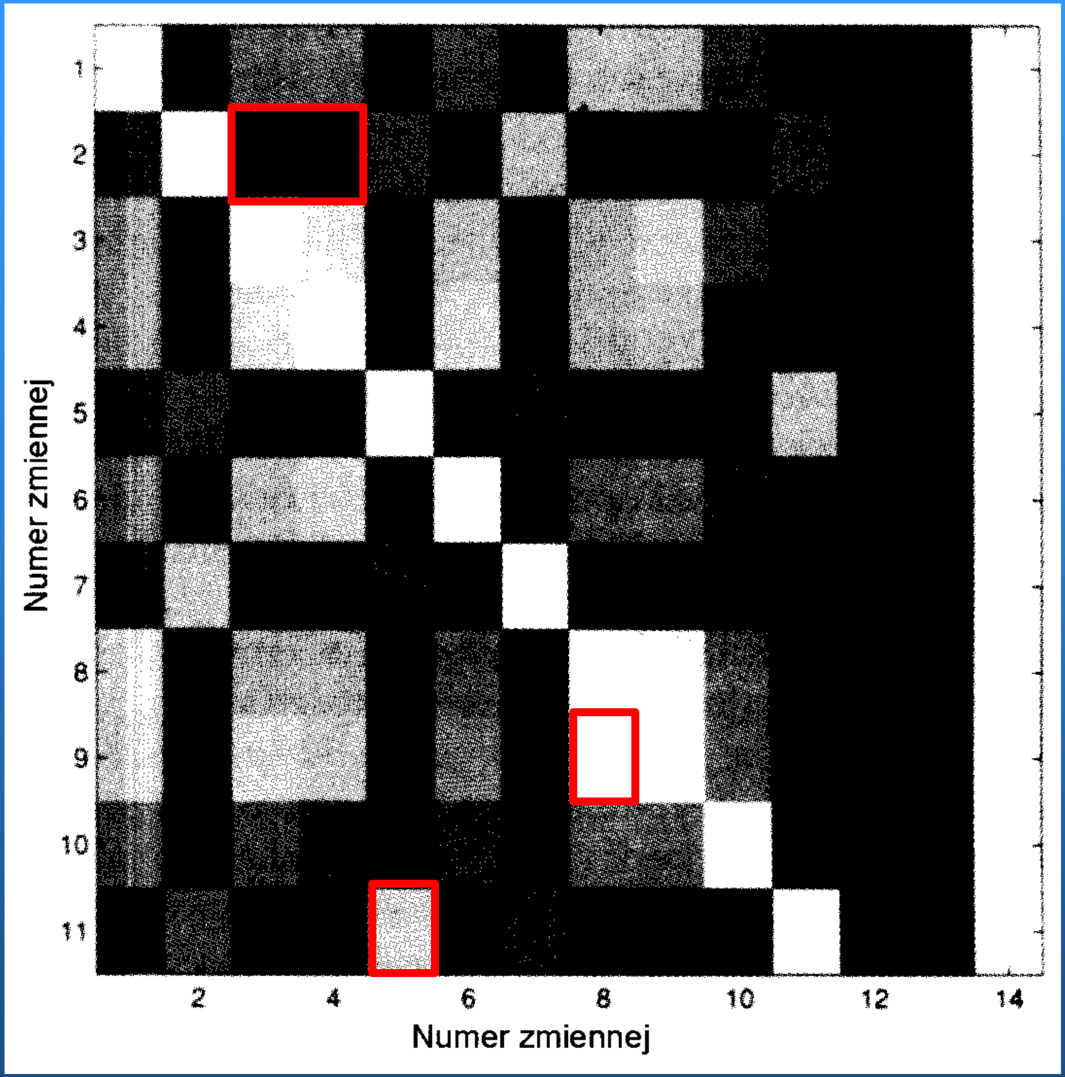


Miary odległości

Przykład

Macierz korelacji. Pola białe 1 , czarne -1.

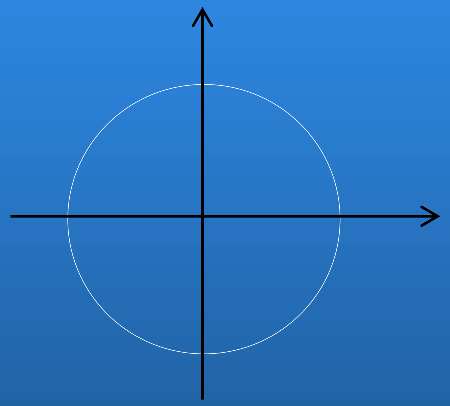
1 – wsp. przestępczości
2 – % terenu na duże działki
3 – % przedś. niedetaliczne
4 – stężenie tlenku azotu
5 – śr. liczba pomieszczeń
6 – % domów sprzed 1940
7 – odległość od centrum
8 – dostępność autostrady
9 – podatek od nieruchom.
10 – liczba uczniów
11 – wartość domu



Miary odległości

Kowariancja i korelacja identyfikują liniowe zależności między zmiennymi.

Rozważmy następującą zależność miedzy X i Y:



X i Y są oczywiście zależne, ale nie liniowo!

Zerowanie korelacji nie oznacza braku zależności!

Miary odległości

Istnieją inne sposoby uogólnienia metryki Euklidesowej.

Metryka Minkowskiego:

$$\left(\sum_{k=1}^{p}(x_k(i)-x_k(j))^{\lambda}\right)^{\frac{1}{\lambda}}, \qquad \lambda \geq 1$$

Metryka miejska:

$$\sum_{k=1}^{p}\left|x_k(i)-x_k(j)\right|.$$

Metryka L:

$$\max_{k}\left|x_k(i)-x_k(j)\right|.$$

Miary odległości

	Atrybuty									
A	0	1	1	1	0	1	0	1	0	0
B	1	1	0	1	1	0	0	1	0	1

W przypadku wielowymiarowych danych binarnych możemy policzyć na ilu zmiennych dwa obiekty przyjmują tą samą wartość.

W poniższej tabeli wszystkie p zmiennych dla obiektów i oraz j przyjmuje wartości ze zbioru {0, 1}.

	$j=1$	$j=0$
$i=1$	$n_{1,1}$	$n_{1,0}$
$i=0$	$n_{0,1}$	$n_{0,0}$

Miary odległości

Najprościej:

$$\frac{n_{1,1} + n_{0,0}}{n_{1,1} + n_{1,0} + n_{0,1} + n_{0,0}}$$

Oczywiście: $n_{1,1} + n_{1,0} + n_{0,1} + n_{0,0} = p$.

Współczynnik Jaccard'a:

$$\frac{n_{1,1}}{n_{1,1} + n_{1,0} + n_{0,1}}.$$

Współczynnik Dice'a:

$$2n_{1,1}/(2n_{1,1} + n_{1,0} + n_{0,1})$$

Miary odległości

Przykład

Rozważmy dane dotyczące obiektów A i B:

	Atrybuty									
A	0	1	1	1	0	1	0	1	0	0
B	1	1	0	1	1	0	0	1	0	1

Wówczas:

$n_{1,1}= 3$ $n_{0,1}= 3$ $n_{1,0}= 2$ $n_{0,0}= 2$

Statystyka

Mediana

W uporządkowanym szeregu liczbowym mediana to liczba, która jest w połowie szeregu w wypadku nieparzystej liczby elementów. Dla parzystej liczby elementów – średnia arytmetyczna dwóch środkowych liczb.

Mediana jako średnia jest bardziej odporna na elementy odstające niż średnia arytmetyczna.

Mediana

Przykład

Rozważmy szereg: 1, 3, 3, 5, 7, 10, 11, 12, 15, 16, 18, 89

mediana = (10+11)/2 = 10.5

średnia = 15.83

Inny szereg: 1, 3, 3, 5, 7, 10, 11, 12, 15, 16, 18

mediana = 10

średnia = 9.18

Dominanta

Dominanta (wartość modalna, moda, wartość najczęstsza) to jedna z miar tendencji centralnej, statystyka dla zmiennych o rozkładzie dyskretnym, wskazująca na wartość o największym prawdopodobieństwie wystąpienia, lub wartość najczęściej występująca w próbie.

Dla zmiennej losowej o rozkładzie ciągłym jest to wartość, dla której funkcja gęstości prawdopodobieństwa ma wartość największą.

Dominanta

Przykład

W zbiorze: 1, 3, 5, 7, 10, 7, 2, 2, 11, 12, 9
mamy dwie dominanty: 7, 2.

Przykład

W przypadku poniższej zmiennej losowej dominanta jest równa 2.

wartość	prawdopodobieństwo
1	0,2
2	0,3
3	0,1
4	0,11
5	0,29

Kwartyle

Kwartyle dzielą wszystkie nasze obserwacje na cztery równe co do ilości obserwacji grupy (w teorii).

- Kwartyl pierwszy (Q_1) - dzieli dzieli obserwacje w stosunku 25% - 75% tzn. 25% obserwacji jest niższa bądź równa wartości I-ego kwartyla, a 75% obserwacji jest równa bądź większa niż wartość I-ego kwartyla

- Kwartyl drugi (Q_2) - inaczej zwany medianą dzieli obserwacje na dwie części w stosunku 50%-50%

Kwartyle

- Kwartyl trzeci (Q_3) - dzieli obserwacje w stosunku 75% - 25% tzn. 75% obserwacji jest niższa bądź równa wartości III-ego kwartyla, a 25% obserwacji jest równa bądź większa niż wartość III-ego kwartyla.

Posortowany szereg statystyczny

25% jednostek		25% jednostek		25% jednostek		25% jednostek	
	↑ Kwartyl I		↑ Kwartyl II mediana		↑ Kwartyl III		↑ Kwartyl IV maksimum

Kwartyle

Procedura wyznaczania wartości dowolnego kwartyla.

- Uporzadkuj *n*-elementowy zbiór danych w kolejności rosnącej.
- Oblicz indeks *i* kwartyla k wg wzoru: $i = kn/4$.
  - Jeżeli *i* nie jest liczba całkowitą, zaokrąglij ją do następnej całkowitej. Znajdź tę pozycje w uporządkowanym zbiorze. Jej wartość jest poszukiwanym kwartylem.
  - Jeżeli *i* jest liczbą całkowitą, znajdź wartość średnią na pozycjach *i* i $i + 1$

Kwartyle

<u>Przykład</u>

Chcemy obliczyć medianę oraz kwartyle Q_1, Q_2, Q_3 i Q_4 dla następującego szeregu: 13, 11, 10, 13, 11, 10, 8, 12, 9, 9, 8, 9.

Szereg po posortowaniu: 8, 8, 9, 9, 9, 10, 10, 11, 11, 12, 13, 13.

Oblicz indeks *i* kwartyla k wg wzoru: $i = kn/4$.

- Kwartyl Q_1: $i = 1 \cdot 12/4 = 3$, $Q1 = (9+9)/2 = 9$
- Kwartyl Q_2: $i = 2 \cdot 12/4 = 6$, $Q2 = (10+10)/2 = 10$
- Kwartyl Q_3: $i = 3 \cdot 12/4 = 9$, $Q3 = (11+12)/2 = 11.5$
- Kwartyl Q_4: $i = 4 \cdot 12/4 = 12$, $Q4 = 13$

Rozstęp ćwiartkowy

Rozstęp ćwiartkowy (rozstęp kwartylowy, IQR. "interquartile range") to różnica pomiędzy 3 i 1 kwartylem.

Pomiędzy tymi kwartylami znajduje się z definicji 50% wszystkich obserwacji zatem im większa szerokość rozstępu ćwiartkowego, tym większe zróżnicowanie cechy.

Przykład

Dla szeregu z naszego przykładu otrzymujemy:

$$IQR = Q3 - Q1 = 11.5 - 9 = 2.5$$

Wykresy

Histogram

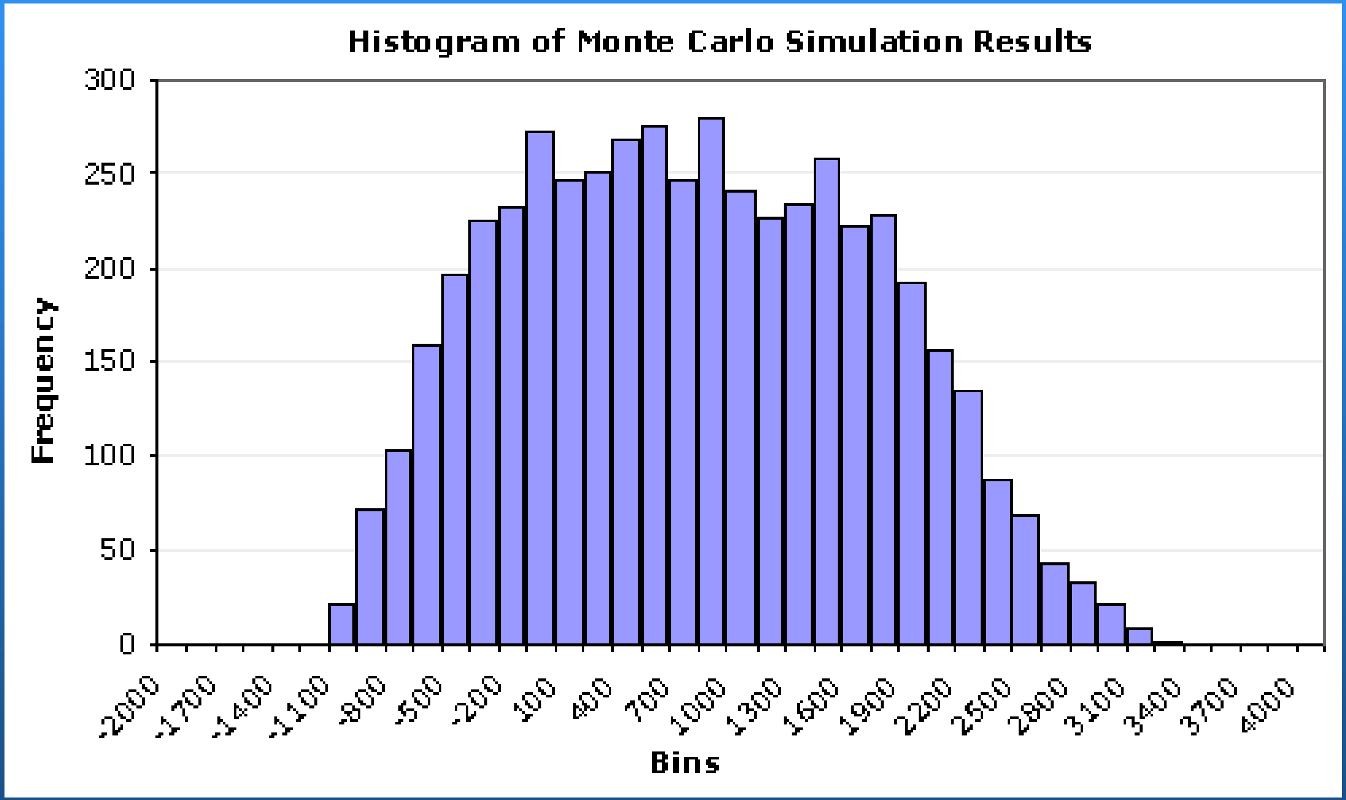
Histogram to jeden z graficznych sposobów przedstawiania rozkładu empirycznego cechy.

Składa się z szeregu prostokątów umieszczonych na osi współrzędnych.

Prostokąty te są z jednej strony wyznaczone przez przedziały klasowe wartości cechy, natomiast ich wysokość jest określona przez liczebności (lub częstości, ewentualnie gęstość prawdopodobieństwa) elementów wpadających do określonego przedziału klasowego.

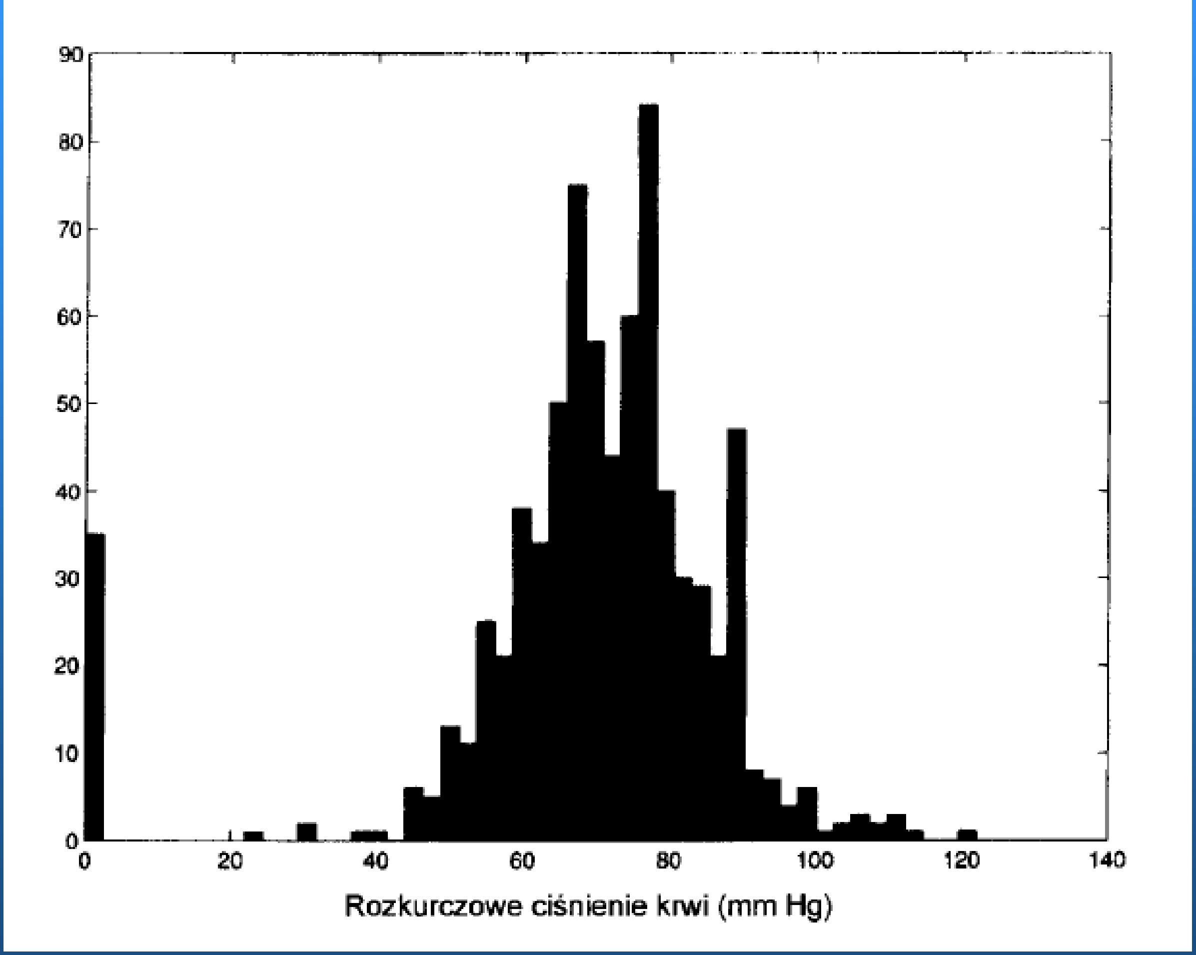
Histogram

Przykład



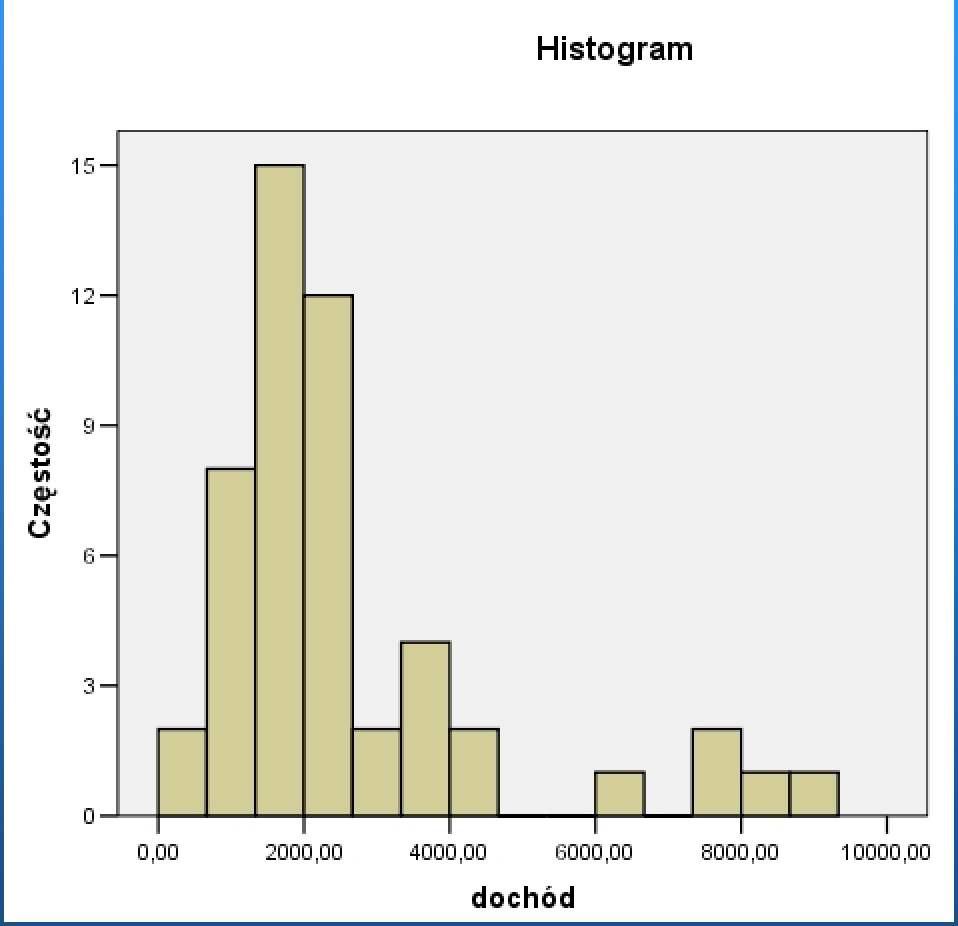
Histogram

Przykład



Histogram

Przykład



Skośność

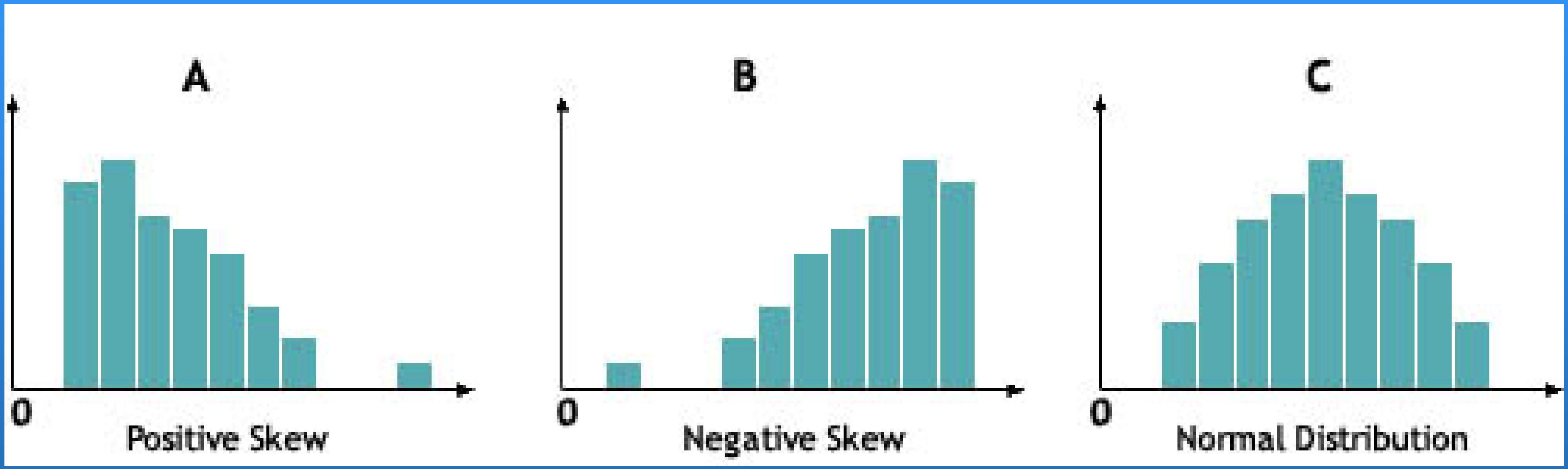
Skośność to miara asymetrii obserwowanych wyników.

Informuje nas o tym jak wyniki dla danej zmiennej kształtują się wokół średniej...

...czyli o tym czy większość zaobserwowanych wyników jest:

- po lewej strony wartości średniej
- blisko wartości średniej
- po prawej strony wartości średniej

Skośność



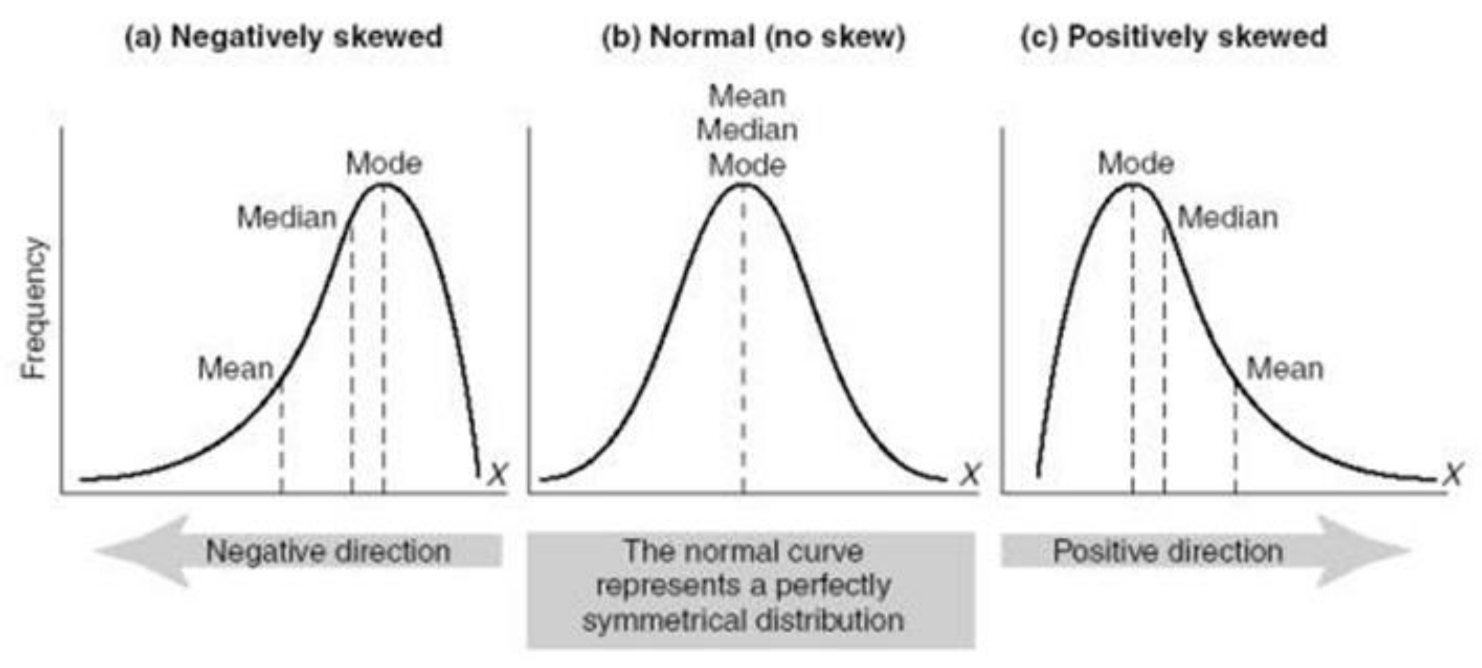
większość wyników
po lewej stronie
wartości średniej

większość wyników
po prawej stronie
wartości średniej

większość wyników
blisko
wartości średniej

Skośność

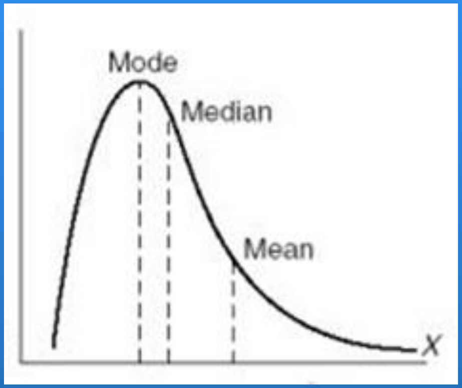
Skośność a mediana, moda i średnia:



Skośność

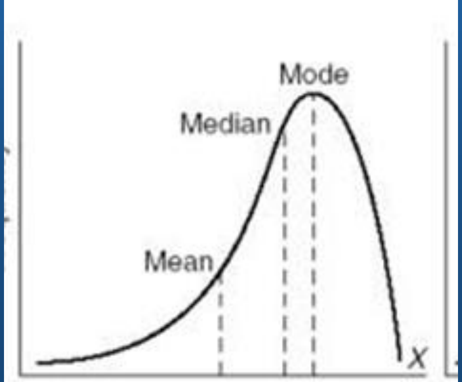
W rozkładzie o prawostronnej asymetrii (rozkład dodatnio skośny) zachodzi relacja:

Moda < Mediana < Średnia



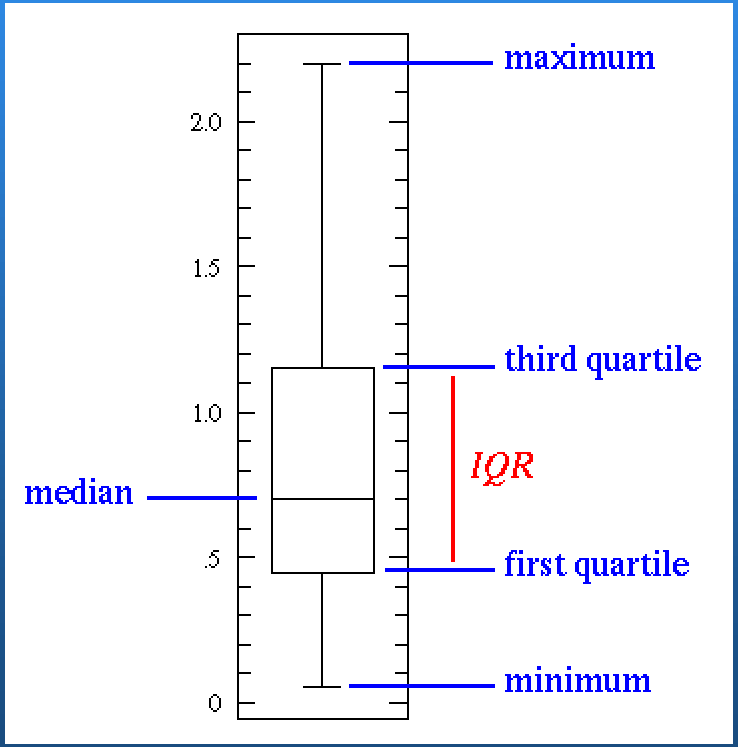
W rozkładzie o lewostronnej asymetrii (rozkład ujemno skośny zachodzi relacja:

Moda > Mediana > Średnia

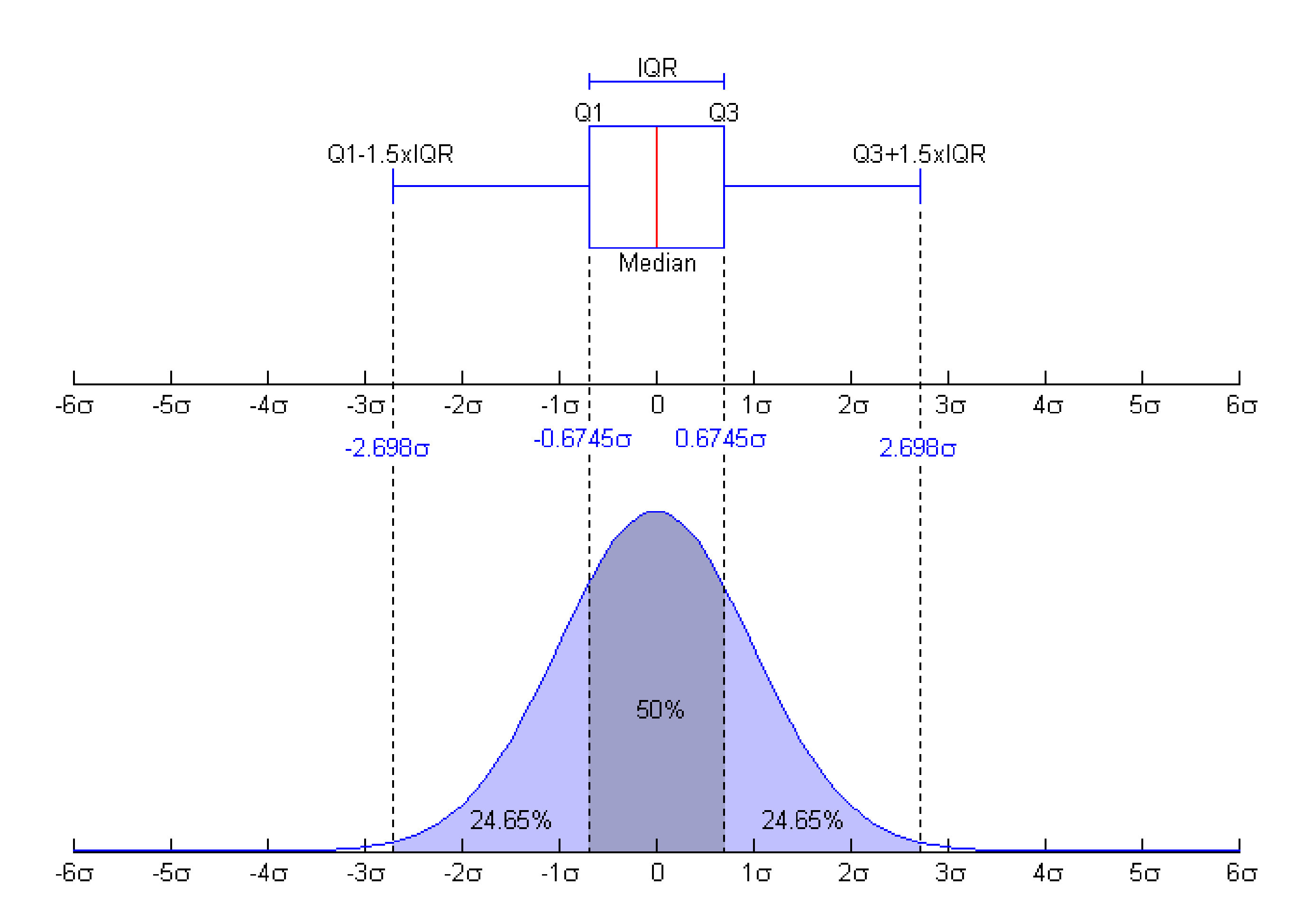


Wykres pudełkowy

Wykres pudełkowy to wygodny sposób opisu danych za pomocą kwartyli. Zasada tworzenia:



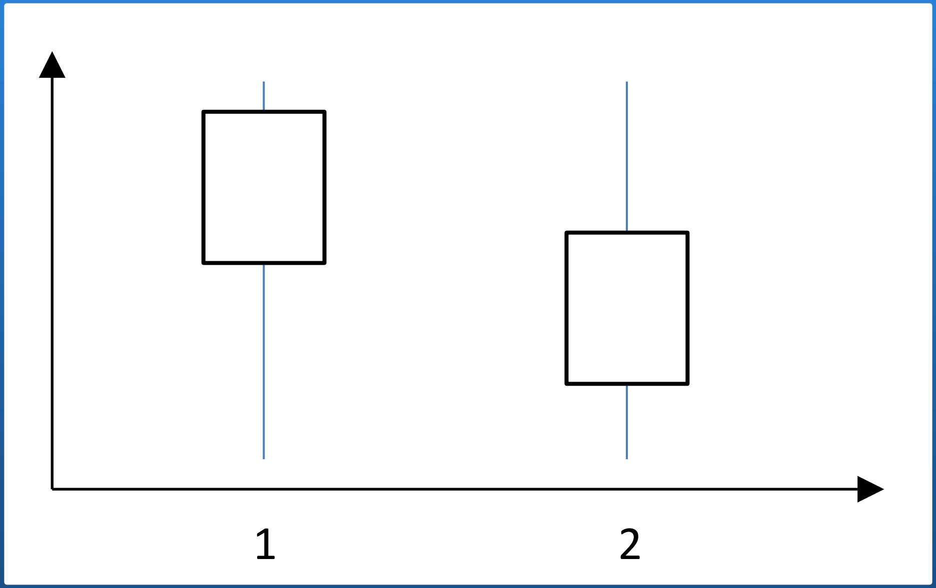
Wykres pudełkowy



Wykres pudełkowy

Ocena położenia

Im bardziej pudełko przesunięte do góry (w kierunku wyższych wartości), tym większy przeciętny poziom cechy.

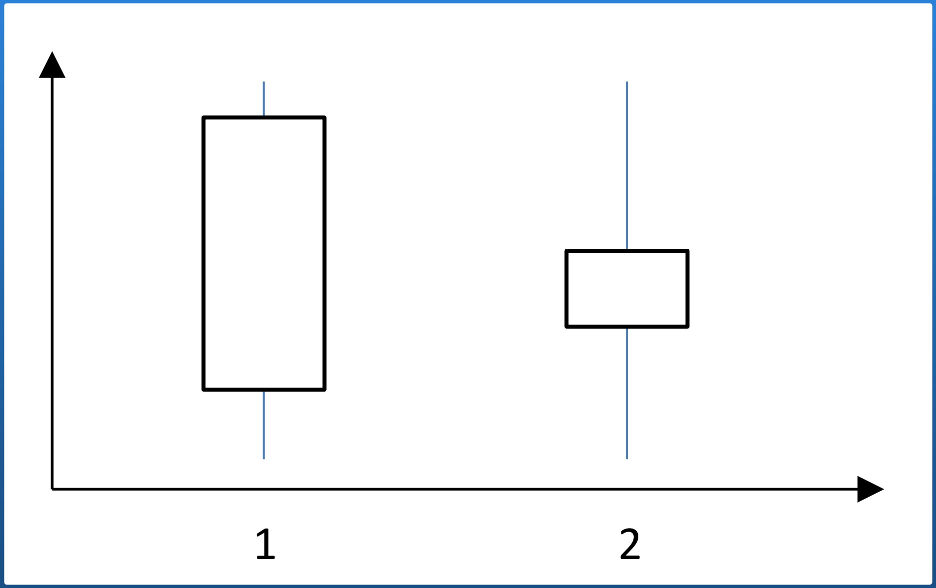


http://pl.wikipedia.org/wiki/Plik:Boxplot_vs_PDF.png

Wykres pudełkowy

Ocena dyspersji

Im szersze pudełko, tym większe zróżnicowanie wartości cechy w dwóch środkowych ćwiartkach rozkładu.

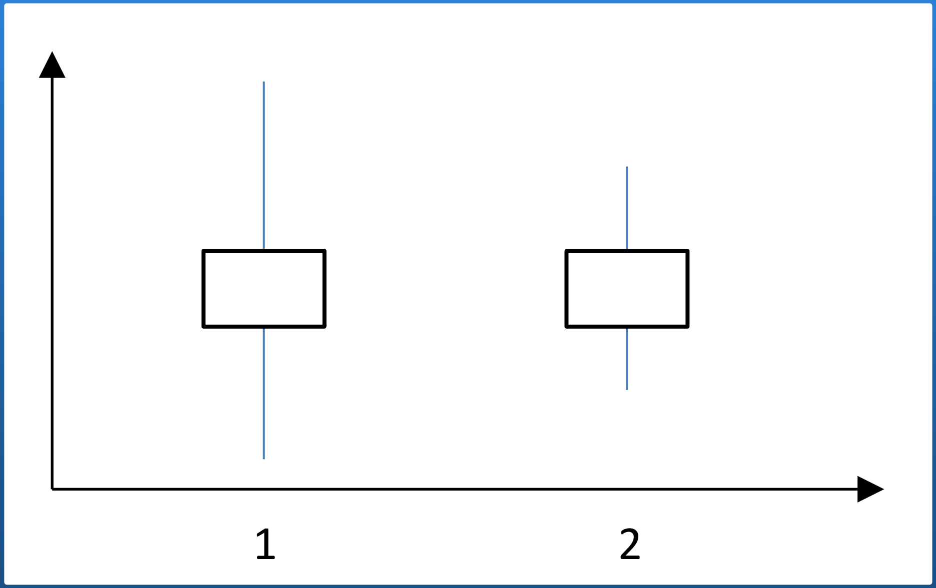


http://pl.wikipedia.org/wiki/Plik:Boxplot_vs_PDF.png

Wykres pudełkowy

Ocena dyspersji

Im dłuższe wąsy, tym większa dyspersja w skrajnych ćwiartkach rozkładu.

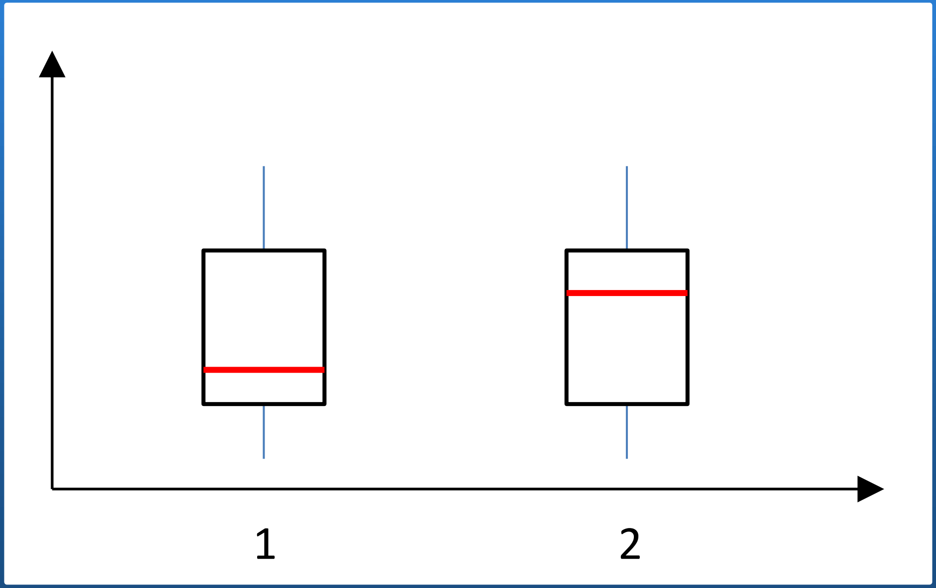


http://pl.wikipedia.org/wiki/Plik:Boxplot_vs_PDF.png

Wykres pudełkowy

Ocena asymetrii

Im bardziej linia podziału wyznaczona przez medianę odchyla się od środka pudełka, tym silniejsza jest asymetria w dwóch środkowych ćwiartkach rozkładu.

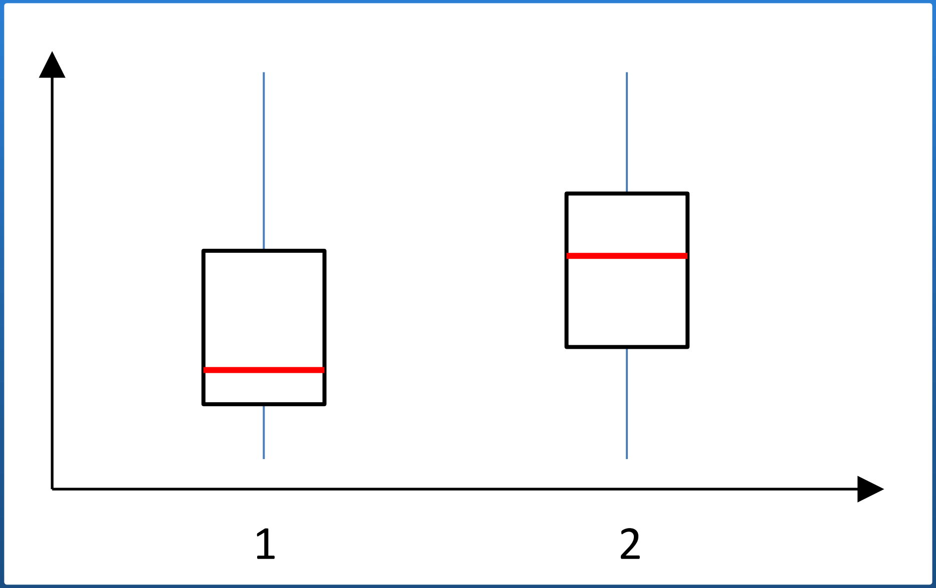


http://pl.wikipedia.org/wiki/Plik:Boxplot_vs_PDF.png

Wykres pudełkowy

Ocena asymetrii

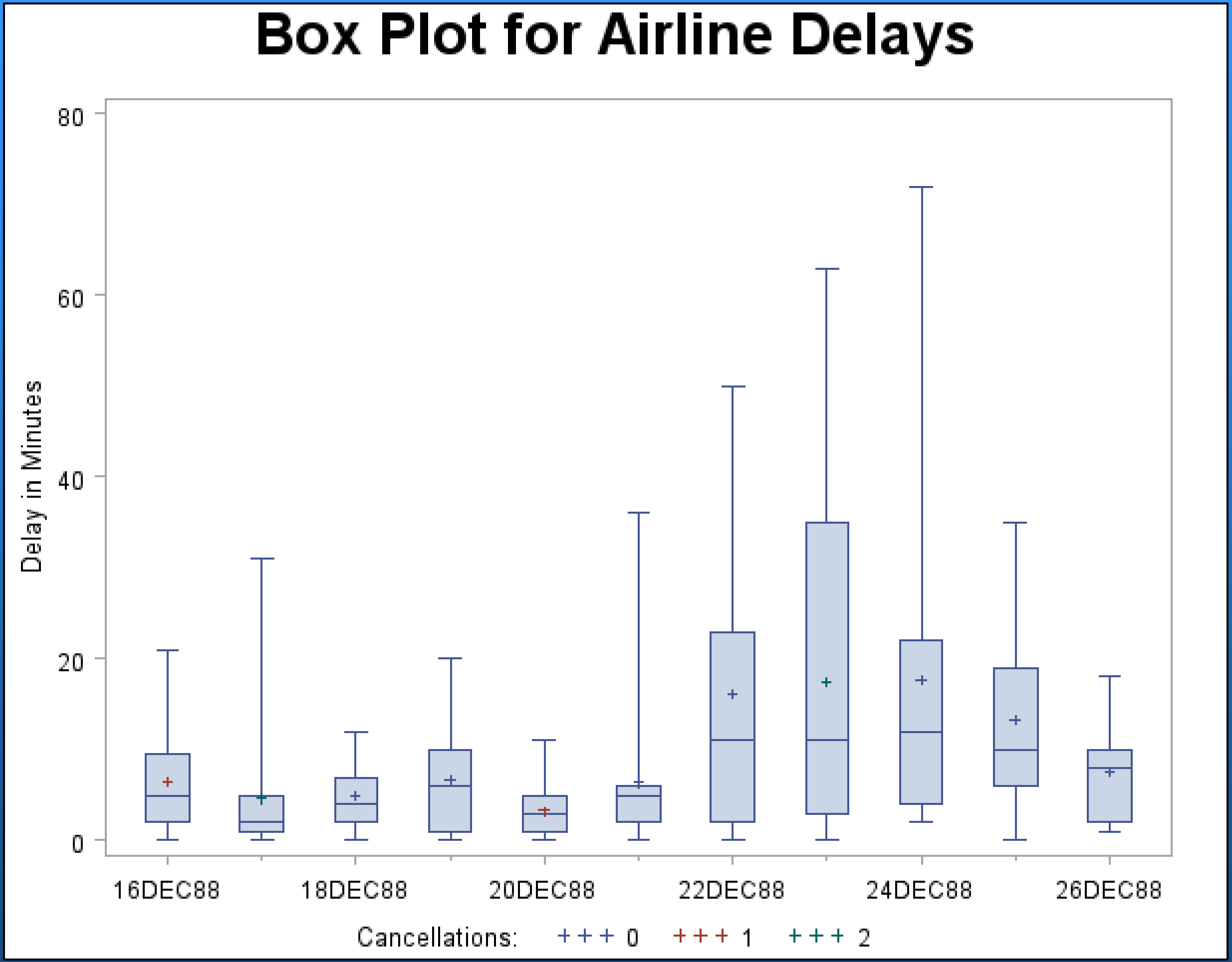
Im większa dysproporcja w długości wąsów, tym silniejsza asymetria w całym rozkładzie.



http://pl.wikipedia.org/wiki/Plik:Boxplot_vs_PDF.png

Wykres pudełkowy

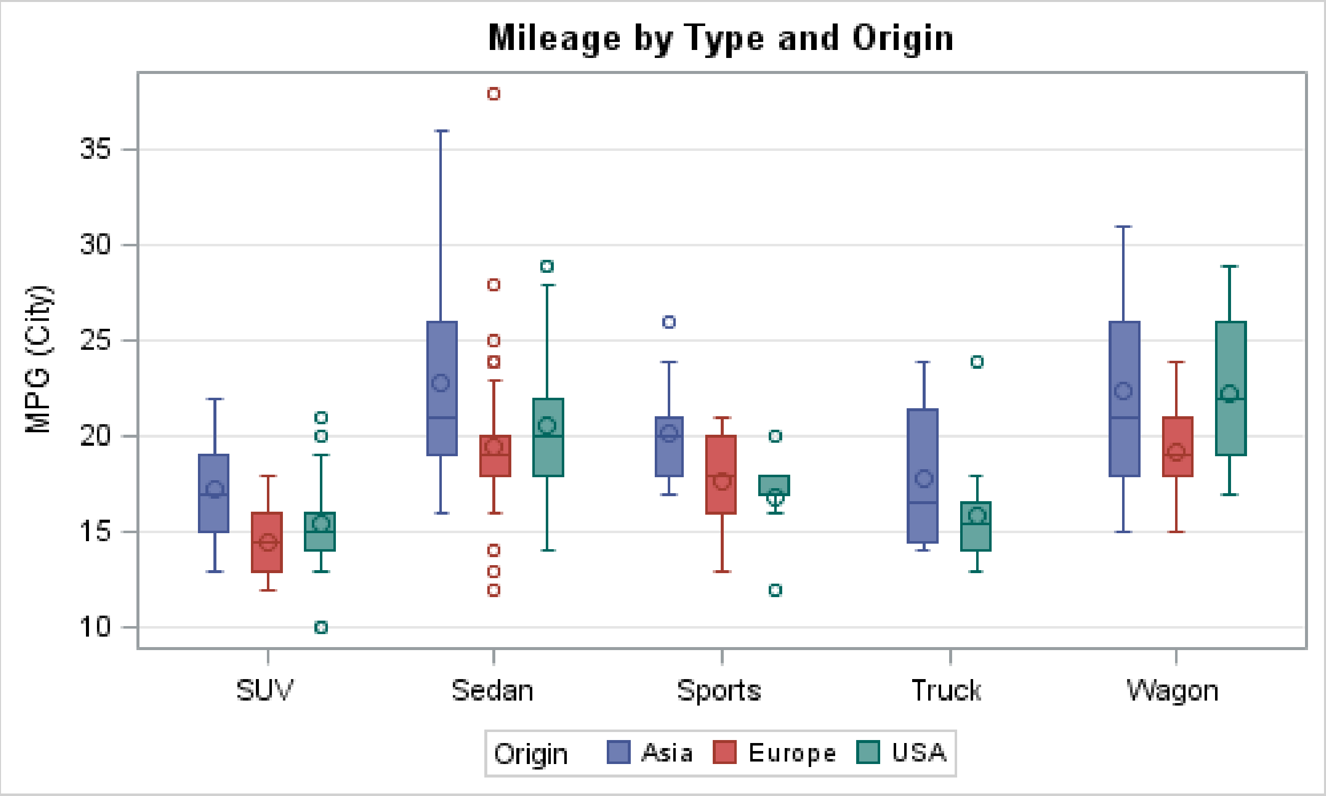
Przykład



http://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_boxplot_sect027.htm

Wykres pudełkowy

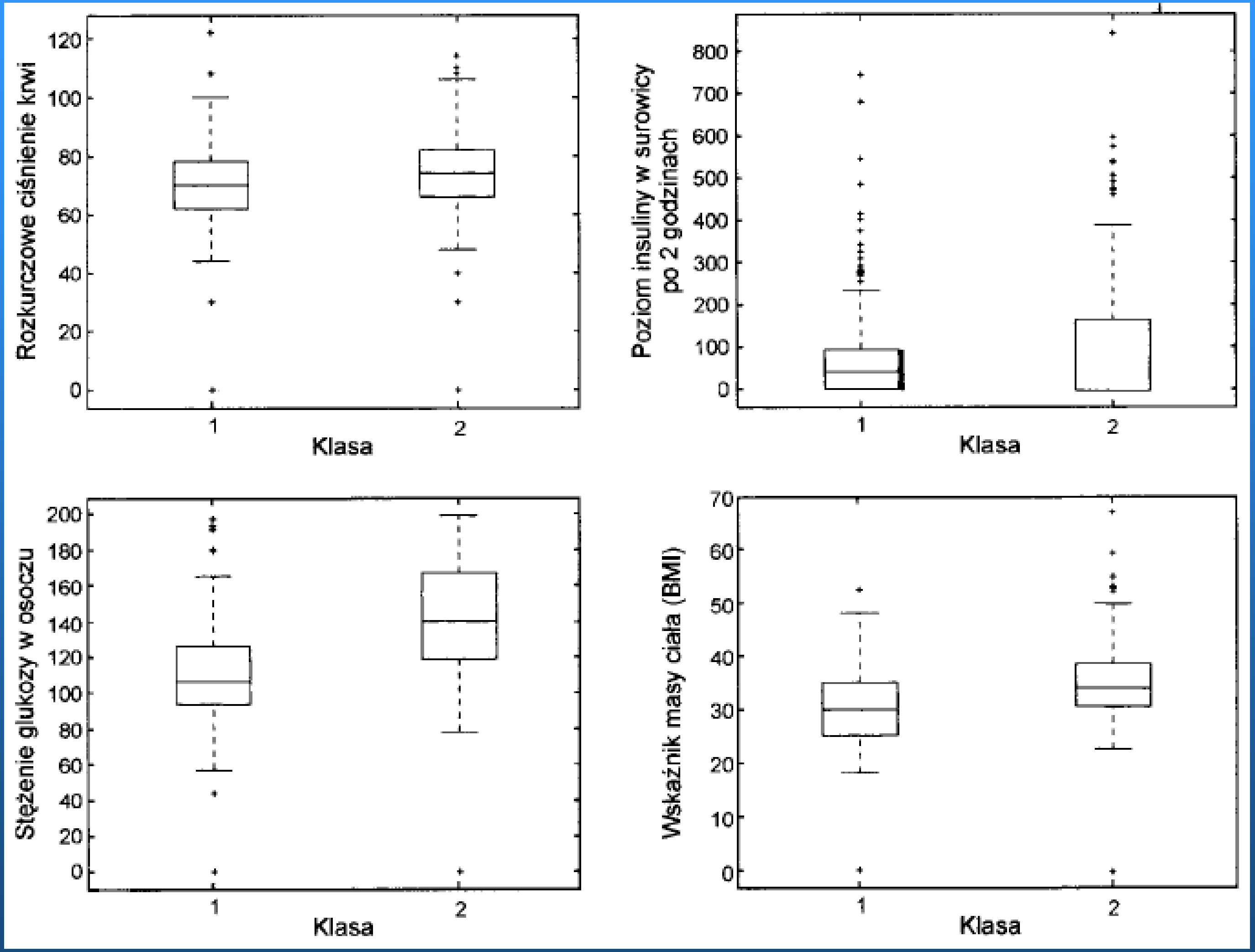
Przykład



Wykres pudełkowy

Przykład

1 – osoby zdrowe
2 – osoby chore



Wykresy rozrzutu

Wykres rozrzutu (wykres korelacji, diagram korelacji) służy do przedstawienia zależności między dwoma zmiennymi (cechami).

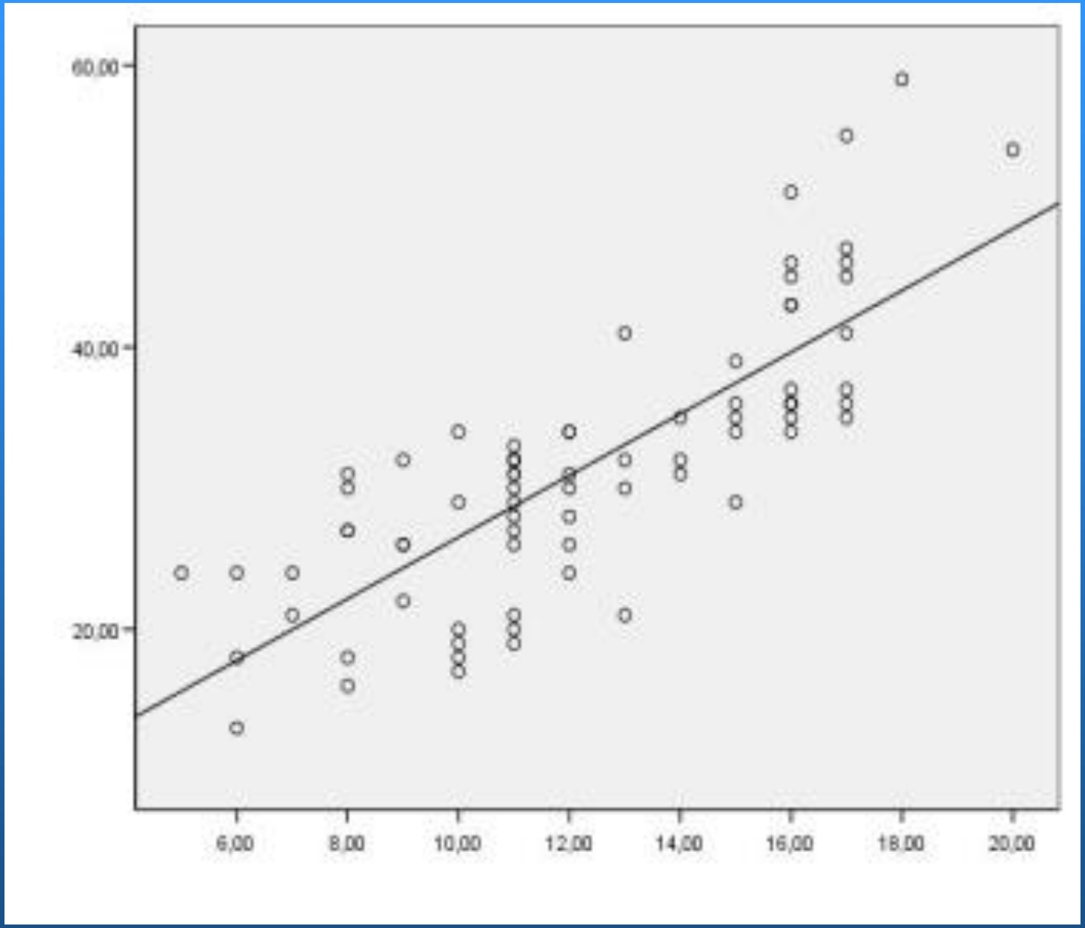
Analizując wykres rozrzutu możemy:

- stwierdzić, że zmienne są zależne
- ustalić kierunek związku
- ustalić siłę związku

Wykresy rozrzutu

Przykład

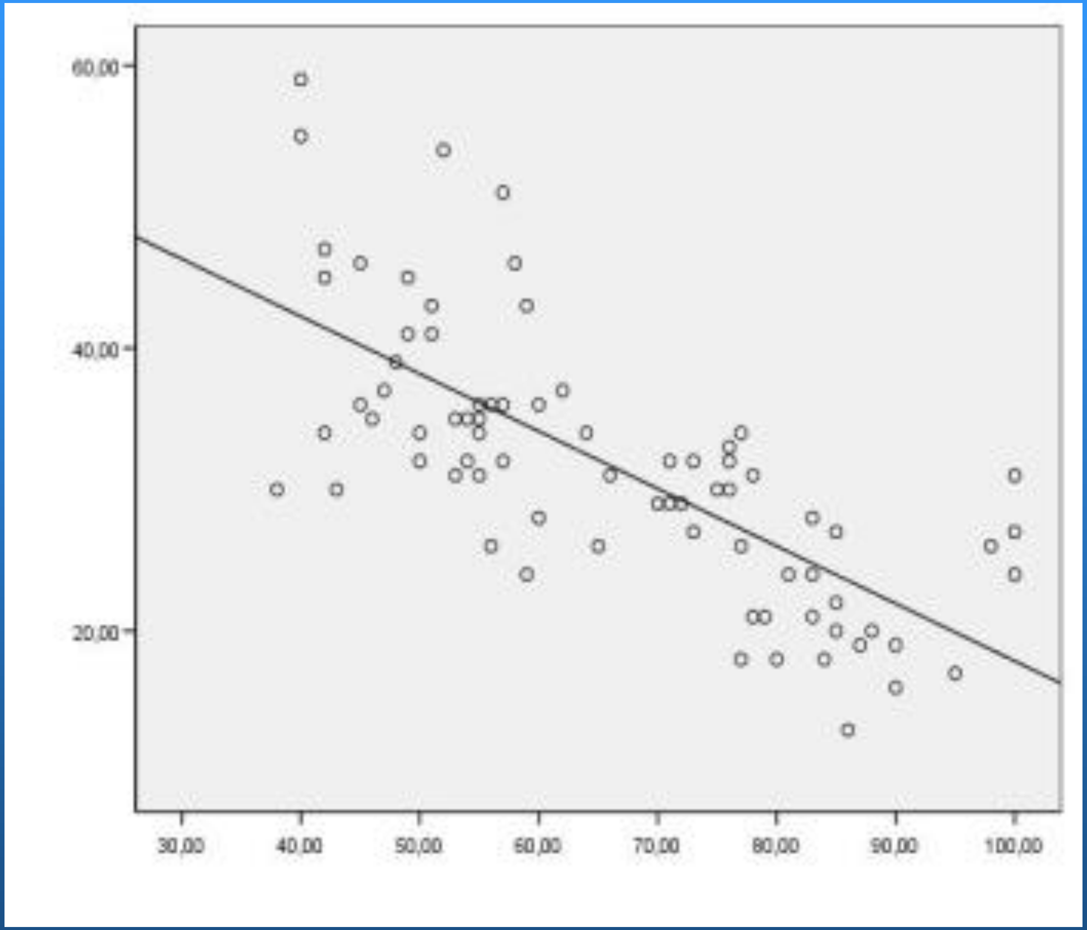
Korelacja dodatnia



Wykresy rozrzutu

Przykład

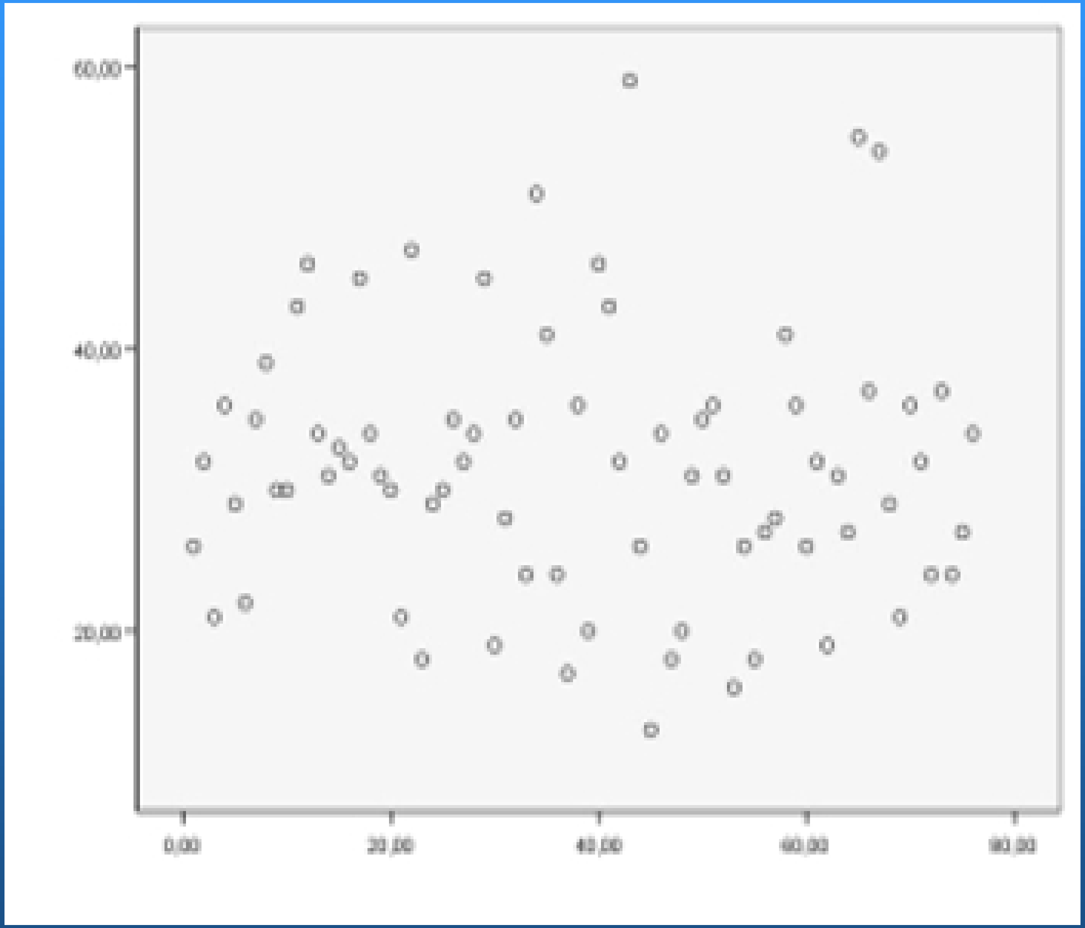
Korelacja ujemna



Wykresy rozrzutu

Przykład

Brak korelacji



Wykresy rozrzutu

Przykład

96 tyś punktów