Statystyczna analiza danych

Wykład 1

STATYSTYCZNA ANALIZA DANYCH

III semestr studiów inżynierskich w PJATK, 2022/23

Prowadzący: dr hab. Elżbieta Ferenstein
dr Andrzej Sierociński

Cel wykładu - poznanie podstaw analizy danych

- **statystyka opisowa**
- **modelowanie probabilistyczne**
- **wnioskowanie statystyczne**

Tematyka wykładu SAD

- Metody graficzne prezentacji danych jakościowych i ilościowych. Statystyki próbkowe. Histogramy a gęstości prawdopodobieństwa, kwantyle, wykresy kwantylowe.

- Prawdopodobieństwo, niezależność zdarzeń, twierdzenie Bayes'a.

- Zmienne losowe, rozkłady prawdopodobieństwa i ich parametry, wybrane rozkłady prawdopodobieństwa.

- Podstawowe statystyki i ich własności, przedziały ufności, testy parametryczne dla średnich i wariancji jednej i dwu populacji, regresja liniowa jednowymiarowa.

Informacje praktyczne

Kontakt:

elaw@pjwstk.edu.pl
asier@pjwstk.edu.pl

Konsultacje: po umówieniu lub po (przed) wykładzie

Wykłady umieszczone są na

ftp/public/elaw/Informatyka dzienne

Ćwiczenia umieszczone są w katalogach Cx na

ftp/public/asier/Informatyka dzienne

Zaliczenie ćwiczeń: skala punktowa: 100 punktów = 90 punktów za 2 kolokwia plus 10 punktów za aktywność (m.in. obecności na ćwiczeniach)

Ocena z ćwiczeń: ≥ 91 pkt: bdb; ≥ 81pkt: db+; ≥ 71: db; ≥ 61: dost +; ≥ 51: dost.

Ocena dostateczna zalicza ćwiczenia i jest warunkiem dopuszczenia do egzaminu.

Ćwiczenia laboratoryjne - 30% czasu, 70% czasu – ćwiczenia rachunkowe.
Na ćwiczeniach obowiązuje znajomość materiału omawianego na wykładach.

Egzamin: zadania z zakresu wykładu i ćwiczeń.

Wymagania wstępne: Analiza I i II, Matematyka Dyskretna.
Software: Excel.

Literatura

Literatura podstawowa

- Jacek Koronacki, Jan Mielniczuk: *Statystyka dla studentów kierunków technicznych i przyrodniczych*, Wydawnictwa Naukowo-Techniczne 2001.
- David S. Moore, George P. McCabe: *Introduction to the Practice of Statistics*, W.H. Freeman&Co., 2000.
- Jay L. Devore: *Probability and Statistics for Engineers and the Sciences,* 1987.

Literatura uzupełniająca

- Janina Jóźwiak, Jarosław Podgórski: *Statystyka od podstaw*, PWE, Warszawa 2001(3), wyd. V (VI).
- Przemysław Grzegorzewski i inn.: *Rachunek prawdopodobieństwa i statystyka*, WSISiZ, Warszawa 2001.
- Amir D. Aczel: *Statystyka w zarządzaniu*, PWN, Warszawa 2000.
- K. Bobecka, P. Grzegorzewski, J. Pusz: *Zadania z rachunku prawdopodobieństwa i statystyki,* WSISiZ, Warszawa 2003.
- Mieczysław Sobczyk: *Statystyka*, PWN 2005.

Podręczniki w wersji elektronicznej (e-booki)

- http://www.stat.rice.edu/~dobelman/textfiles/DistributionsHandbook.pdf
- http://davidmlane.com/hyperstat/index.html

STATYSTYKA OPISOWA

Techniki wstępnej analizy danych i ich prezentacji:

- gromadzenie, przechowywanie danych, analiza danych surowych
- prezentacja danych: tabele, wykresy, parametry liczbowe obliczane dla danych.

Cel:

- charakteryzacja danych - w **zwięzłej formie** odzwierciedlająca pewne ich **cechy**, np. średni dochód, średnie zużycie paliwa, ..
- odnalezienie różnego rodzaju **regularności** (nieregularności) ukrytych w danych, **zależności** między podzbiorami danych.

- Obejrzenie danych surowych – nieprzetworzonych, niepogrupowanych, niezorganizowanych.
- Poznanie sposobu i celu zebrania danych:
  - jaką cechę mierzono (obserwowano) ?,
  - w jakich jednostkach ?,
  - ile wykonano obserwacji (liczebność zbioru danych), w jakich warunkach – czy nie zgubiono części danych, dane brakujące, czy jest możliwość przekłamań ?
  - czy celem zebrania danych ma być odpowiedź na konkretne pytania ?

- Cel badania statystycznego: poznanie charakterystyk dużej zbiorowości obiektów (osoby, przedmioty, zjawiska, możliwe wyniki eksperymentów ...) na podstawie obserwacji cech (danych) jedynie niektórych wylosowanych obiektów
- **Populacja:** zbiór obiektów badanych ze względu na określoną cechę nazywaną **zmienną**
- **Próbka:** zbiór cech zbadanych obiektów populacji

Rodzaje i przykłady cech statystycznych

- Ilościowe
  - Ciągłe : wzrost, waga itp.
  - Dyskretne : liczba dzieci, liczba reklamacji itp.
- Jakościowe
  - O kategoriach uporządkowanych: miasta (małe, średnie, duże), rodziny (bezdzietne, wielodzietne) itp.
  - Nominalne : grupa krwi, płeć, kolor oczu itp.

- **Badanie statystyczne pełne** (kompletne, całkowite, wyczerpujące) to badanie oparte o dane obejmujące wszystkie jednostki populacji.

- **Badanie statystyczne częściowe** (niekompletne, niepełne) to badanie oparte o dane obejmujące wybrane jednostki populacji.

- **Próba** to podzbiór populacji generalnej wykorzystywany w badaniu częściowym.

- **Próba reprezentatywna** to próba wybrana w sposób losowy i mająca dostateczną liczebność.

Aby wyniki badania próby można było odnieść do zbiorowości generalnej (uogólnić) próba <u>musi być</u> reprezentatywna.

Populacja	badana cecha (zmienna)	zebrane dane (próbka)
♦ zbiór detali	jakość detalu	zbiór jakości zbadanych detali
♦ zbiór komputerów w sieci	liczba awarii kompu-tera w danym okresie	zbiór liczb awarii wybranych komputerów w danym czasie
♦ zbiór projektów przysłanych na konkurs	ocena projektu	zbiór ocen wybranych projektów
♦ zbiór osób w zespole pracowników	staż pracy	zbiór staży pracy (lat pracy) wylosowanych osób

Opracowanie materiału statystycznego

- **Szereg szczegółowy** (**wyliczający**) – uporządkowany ciąg obserwowanych wartości badanej cechy statystycznej.
- **Szereg rozdzielczy** (**strukturalny**) – materiał statystyczny podzielony na grupy (klasy) według wybranego kryterium, zapisany w postaci tabelarycznej, z podaniem liczebności (lub częstości) każdej z wyodrębnionych grup,.
- Szeregi rozdzielcze są wynikiem operacji grupowania danych.
- W przypadku cechy mierzalnej z małą liczbą wariantów cechy tworzy się szeregi rozdzielcze **punktowe**.
- Gdy wariantów jest dużo buduje się szeregi rozdzielcze **przedziałowe**.
- Szereg rozdzielczy cechy mierzalnej opisuje **rozkład empiryczny** badanej cechy.

Przykład (szereg rozdzielczy punktowy)

Liczba pracowników w poszczególnych przedsiębiorstwach pewnego koncernu wynosi:

100; 125; 170; 144; 144; 235; 301; 100; 100; 170; 144; 235; 100; 301; 170; 301; 125; 125; 235, 125:125; 100; 144; 301; 144; 144; 170; 144; 144; 144.

Są to tzw. *dane surowe.* Opisują cechę mierzalną skokową.

Po uporządkowaniu danych (np. rosnąco) dostajemy szereg wyliczający

(zapisany 2 wierszach tabeli).

Ponieważ w zbiorze danych mamy tylko 5 wariantów cechy tworzymy szereg rozdzielczy punktowy postaci

Grupa	Liczebność
100	5
125	5
144	9
170	4
235	3
301	4
SUMA	30

Przykład (szereg rozdzielczy przedziałowy)

Powierzchnie użytkowe (w m^2) badanych sklepów przedstawia uporządkowany szereg wartości cechy:

76; 81; 83; 85; 87; 91; 93; 94; 95; 97; 99; 104;
111; 112; 113; 114; 116; 118; 119; 120; 121; 122; 123; 125;
126; 127; 128; 128; 129; 130; 131; 132; 133; 133; 135; 135;
136; 137; 138; 138; 141; 141; 141; 141; 143; 144; 146; 146;
148; 148; 152; 155; 158; 159; 161; 162; 163; 165; 166; 167;
178; 179; 179; 182; 184; 184; 193, 198; 200.

Powierzchnia jest cechą mierzalną ciągłą, dlatego przeprowadzimy grupowanie statystyczne danych tworząc szereg rozdzielczy, z przedziałami klasowymi o rozpiętości 20 m^2 i początkiem pierwszego przedziału klasowego równym 70 m^2.

Otrzymany szereg rozdzielczy (liczebności) ma postać:

przedział	70-90	90-110	110-130	130-150	150-170	170-190	190-210
liczebność	5	7	17	21	10	6	3

(*przyjęto przedziały lewostronnie domknięte, prawostronnie otwarte*)

Szereg rozdzielczy częstości uzyskujemy zastępując liczebności przez odpowiadające im częstości (częstości względne)

częstość = (liczebność grupy) / (liczebność łączna) ($w_i = \frac{n_i}{N}$)

Szereg rozdzielczy częstości dla prezentowanych danych ma postać

przedział	70-90	90-110	110-130	130-150	150-170	170-190	190-210
częstość	0,07	0,10	0,25	0,30	0,14	0,09	0,04

w ujęciu procentowym

przedział	70-90	90-110	110-130	130-150	150-170	170-190	190-210
częstość	7%	10%	25%	30%	14%	9%	4%

Szeregi rozdzielcze skumulowane

przedział	70-90	90-110	110-130	130-150	150-170	170-190	190-210
liczebność skumulowana	5	12	29	50	60	66	69

przedział	70-90	90-110	110-130	130-150	150-170	170-190	190-210
częstość skumulowana	0,07	0,17	0,42	0,72	0,87	0,96	1,00

przedział	70-90	90-110	110-130	130-150	150-170	170-190	190-210
częstość skumulowana (%)	7%	17%	42%	72%	87%	96%	100%

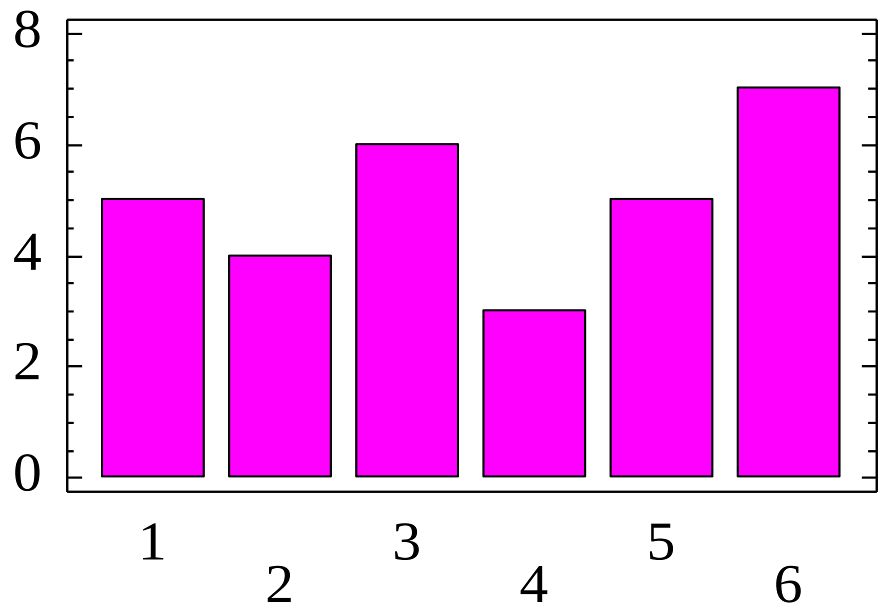
<u>Przykład.</u> W 30 rzutach kostką sześcienną otrzymano liczby oczek:

3 5 6 1 4 6 2 3 5 6 2 6 5 3 5 4

6 6 5 1 5 2 4 3 6 1 1 2 1 3 3 6

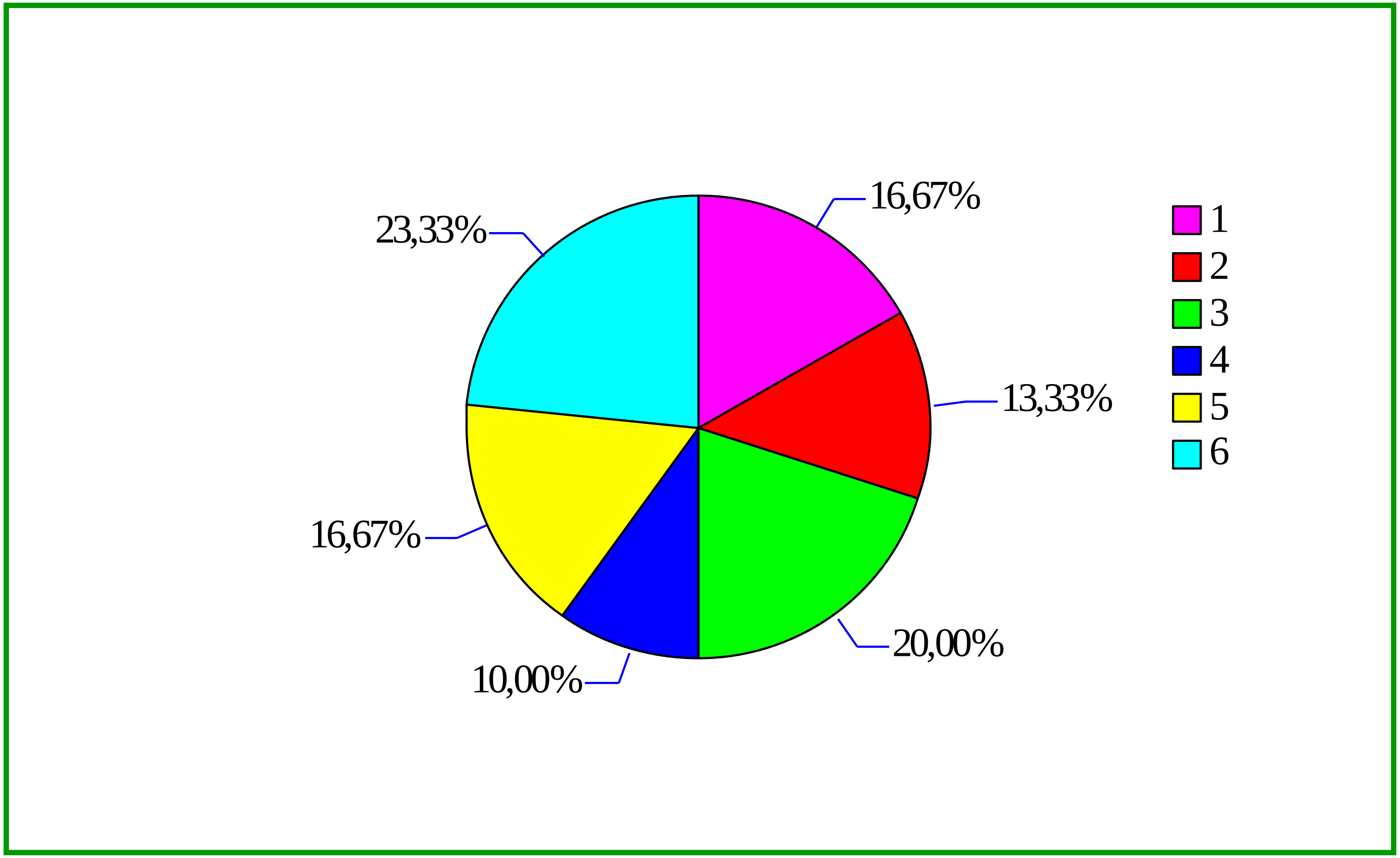
wartość (liczba oczek)	**1**	**2**	**3**	**4**	**5**	**6**
liczność (liczba wystąpień)	**5**	**4**	**6**	**3**	**5**	**7**
częstość	$\frac{5}{30}$	$\frac{4}{30}$	$\frac{6}{30}$	$\frac{3}{30}$	$\frac{5}{30}$	$\frac{7}{30}$

Diagram liczebności
Wykres słupkowy



Liczba oczek

Wykres kołowy



Metody opisu danych jakościowych

wykres słupkowy, wykres kołowy

Przykład. Liczby studentów w kraju na różnych kierunkach studiów w roku ak. 1990/91 oraz 1997/98 podane są w tabeli. Wykonamy:

- wstępną analizę danych
- wykresy słupkowe (procentowe, ilościowe)
- wykresy kołowe

Prezentacja materiału statystycznego

Tablica danych

Grupa kierunków	rok 1990/91		rok1997/98	
	liczba	%	liczba	%
pedagogiczne	99552	18,3	91100	7,2
humanistyczne	69088	12,7	110565	8,7
prawne i nauki społeczne	133824	24,6	566475	44,8
nauki ścisłe i przyrodnicze	144704	26,6	292110	23,1
medyczne	81600	15,0	95550	7,6
pozostałe	15232	2,8	109200	8,6
ogółem	544000	100,0	1265000	100,0

Prezentacja materiału statystycznego

Opis danych surowych:

- 2 próbki o licznościach n_1 = 544000 oraz n_2 = 1265000
- cecha jakościowa: grupa kierunków studiów
- 6 kategorii (atrybutów) cechy
- atrybuty: grupa kierunków pedagogicznych, humanistycznych, medycznych, ...

Najliczniejsze grupy kierunków:

- nauki ścisłe i przyrodnicze w 1990/91 roku
- prawo i nauki społeczne w 1997/98 roku

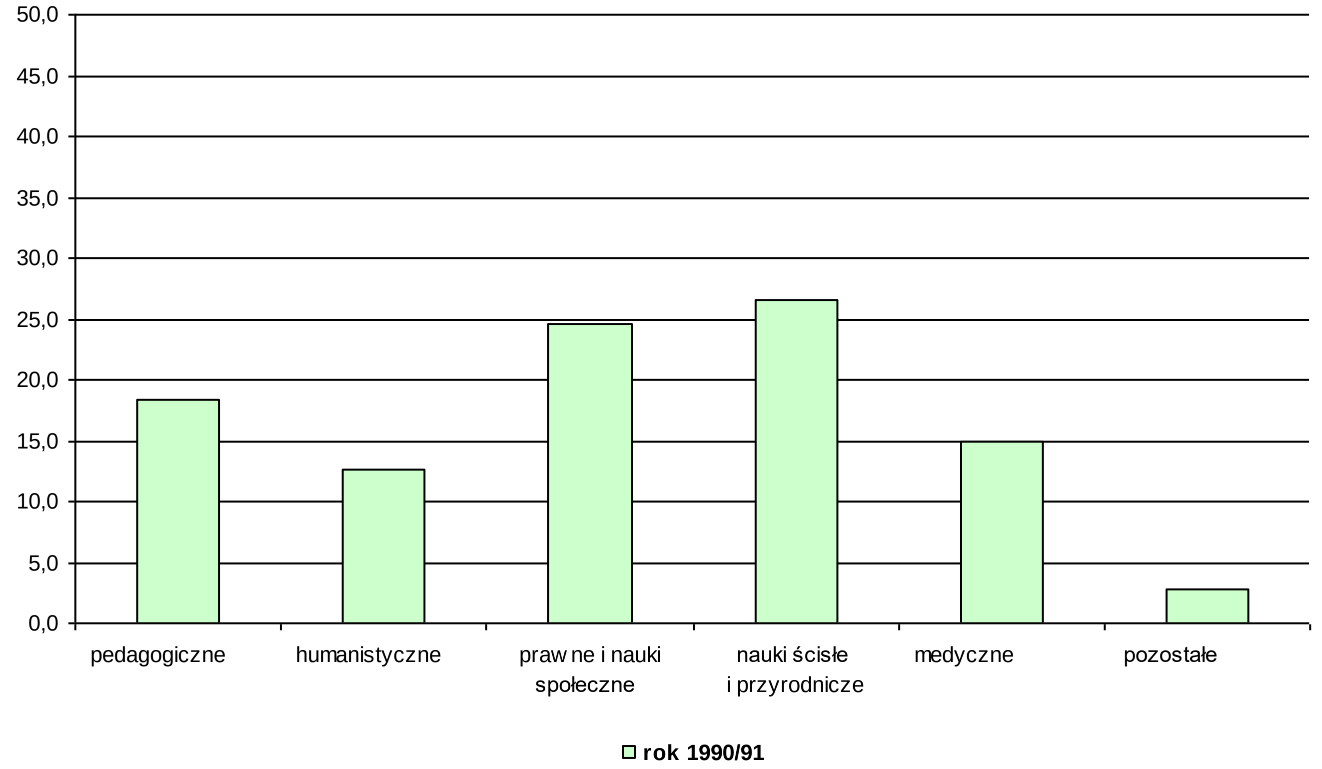
Procentowy udział klasy

(liczność klasy / liczność próbki) * 100% = częstość * 100%

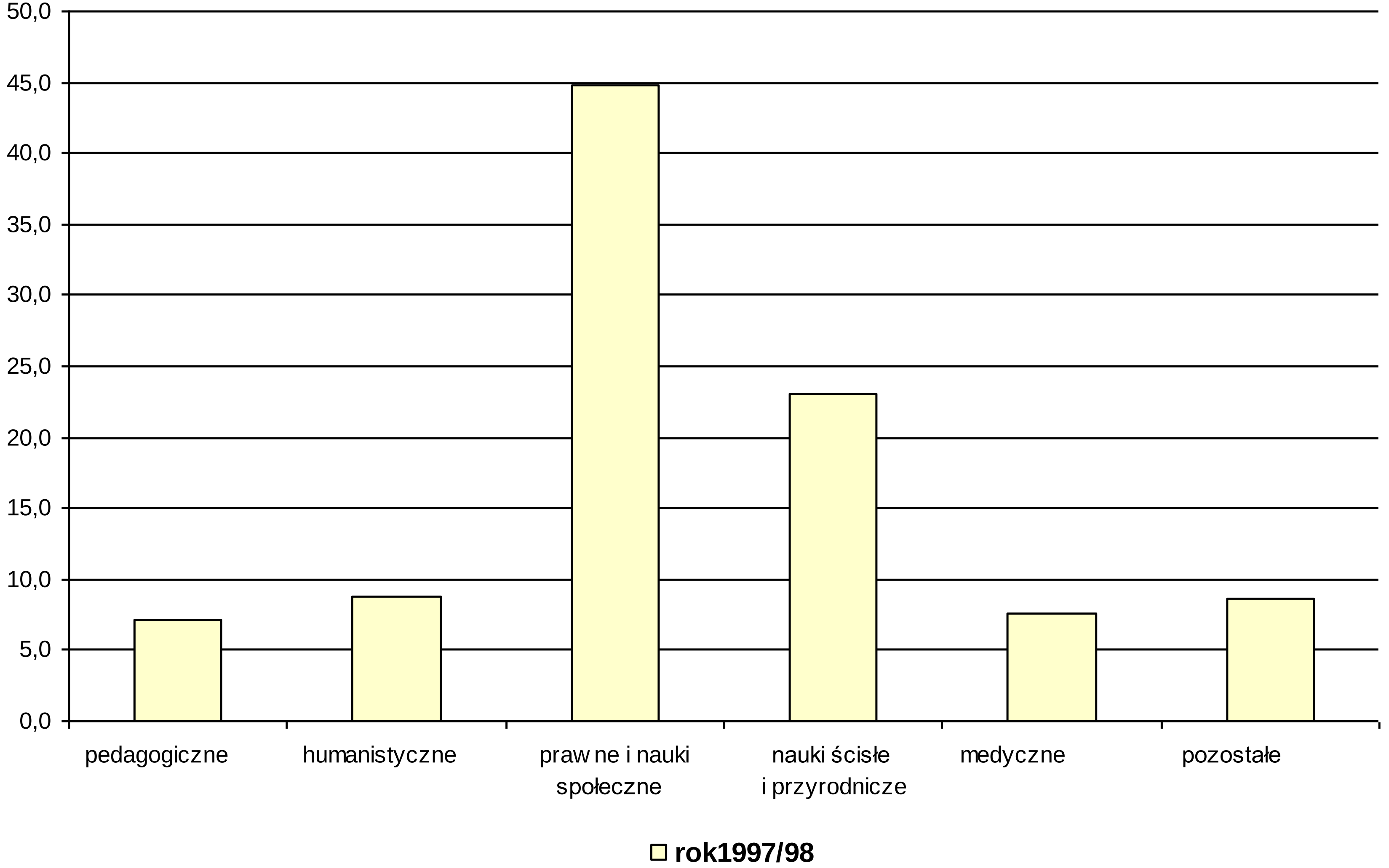
Prezentacja materiału statystycznego

Wykres słupkowy

Wykres słupkowy procentowego udziału grup kierunków studiów
w roku akad. 1990/91

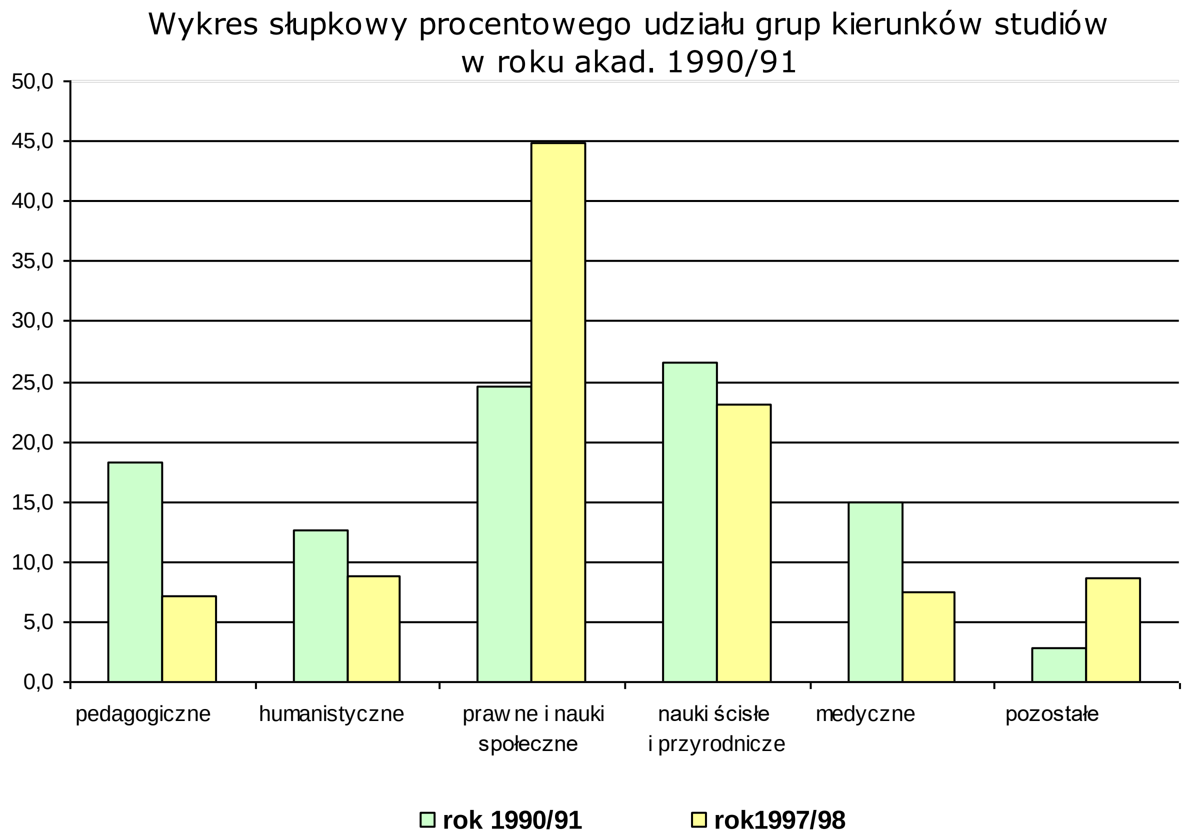


Wykres słupkowy procentowego udziału grup kierunków studiów
w roku akad. oraz 1997/98



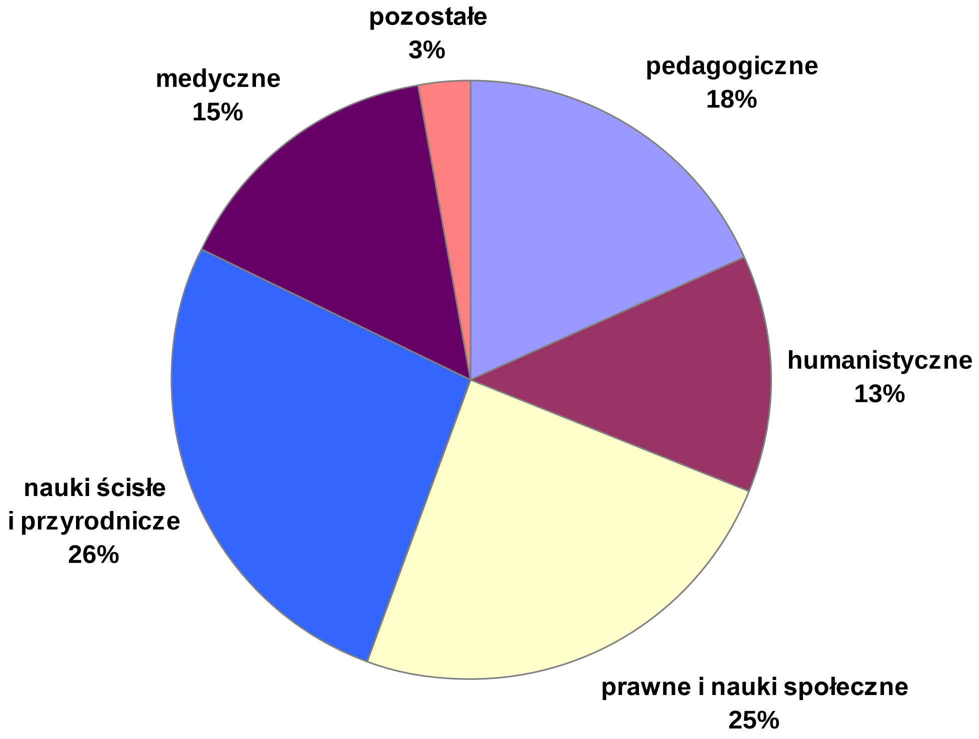
Prezentacja materiału statystycznego

Połączony wykres słupkowy



Wykres kołowy

Wykres kołowy procentowego udziału grup kierunków studiów w roku akad. 1990/91



Kąt wycinka koła dla grupy humanistycznej =

$$0{,}127 \times 360^{o} = 45{,}72^{o}$$

Kąt wycinka koła odpowiadającego określonej kategorii =

Liczebność kategorii / liczebność próbki)$\times 360^{o}$.

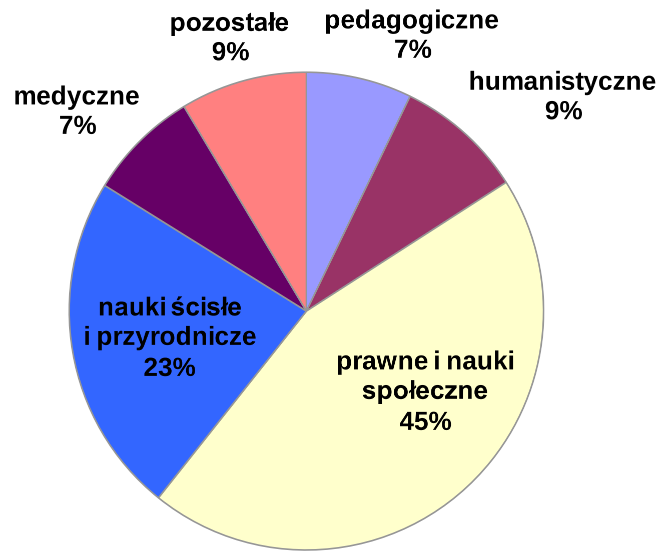
częstość kategorii x 100% =

= (pole wycinka / pole koła) x 100%

Prezentacja materiału statystycznego

Wykres kołowy

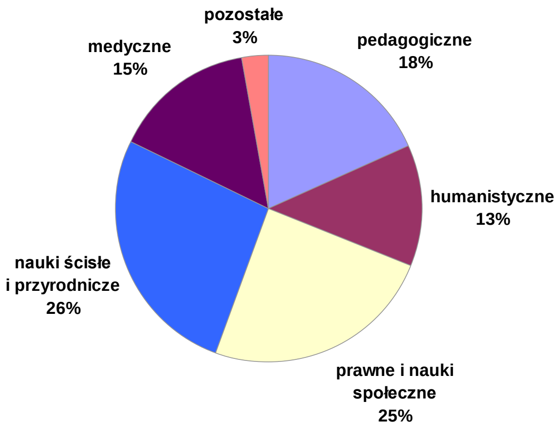
Wykres kołowy procentowego udziału grup kierunków studiów w roku akad. 1997/98



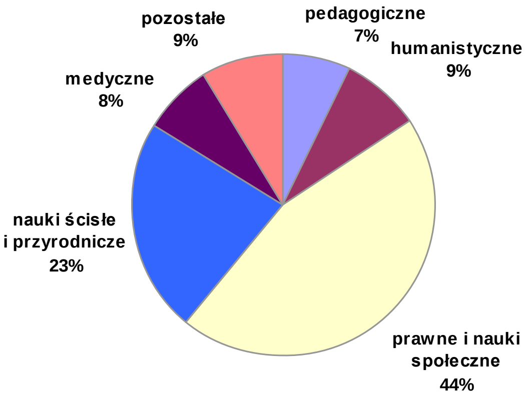
Prezentacja materiału statystycznego

Wykresy kołowe

Wykres kołowy procentowego udziału grup kierunków studiów w roku akad. 1990/91



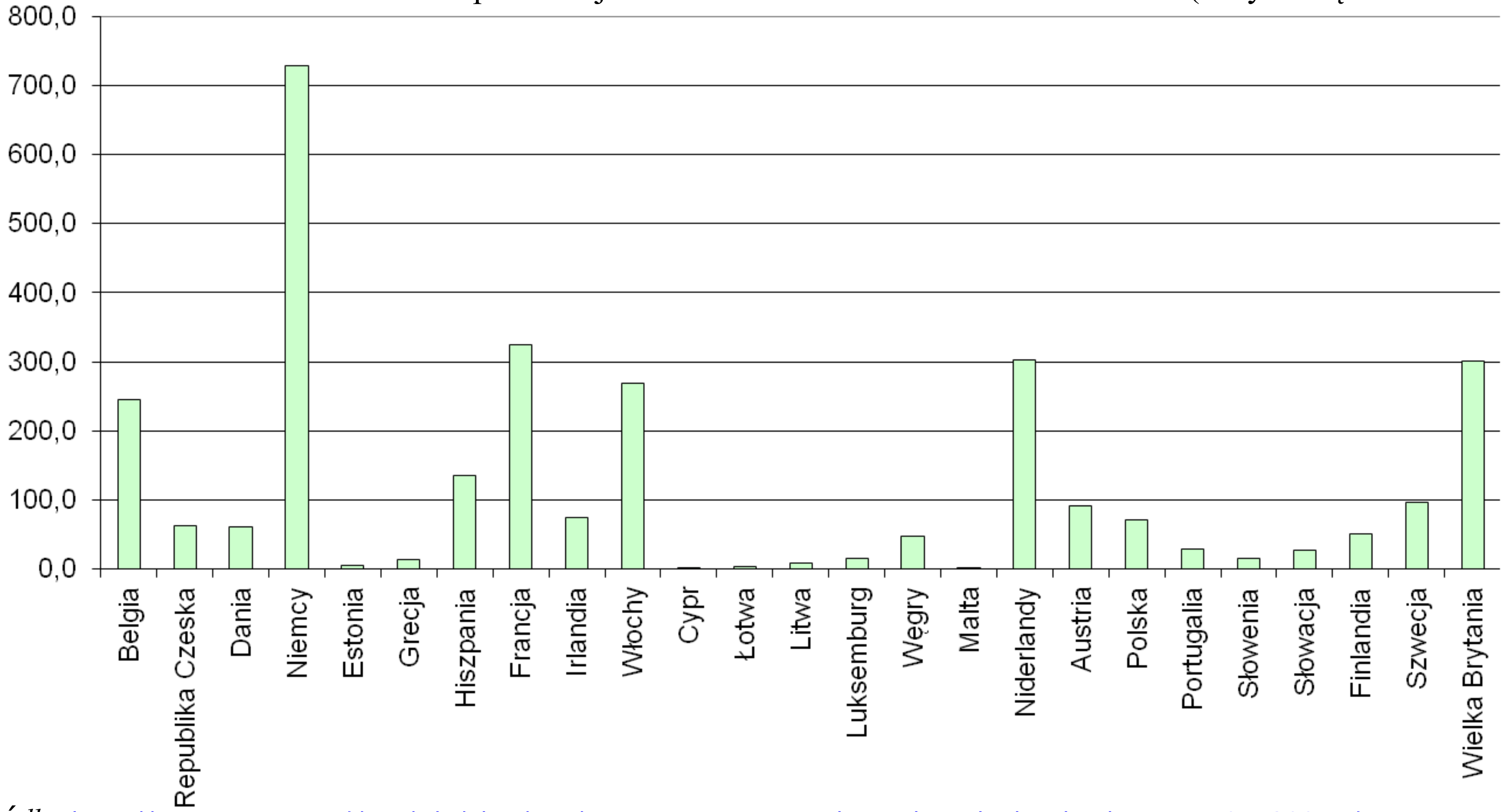
Wykres kołowy procentowego udziału grup kierunków studiów w roku akad. 1997/98



Wykres słupkowy

Przykład

Tablica xx. Wartość eksportu krajów członkowskich UE w okresie2006 I-X (ceny bieżące w mld EUR)

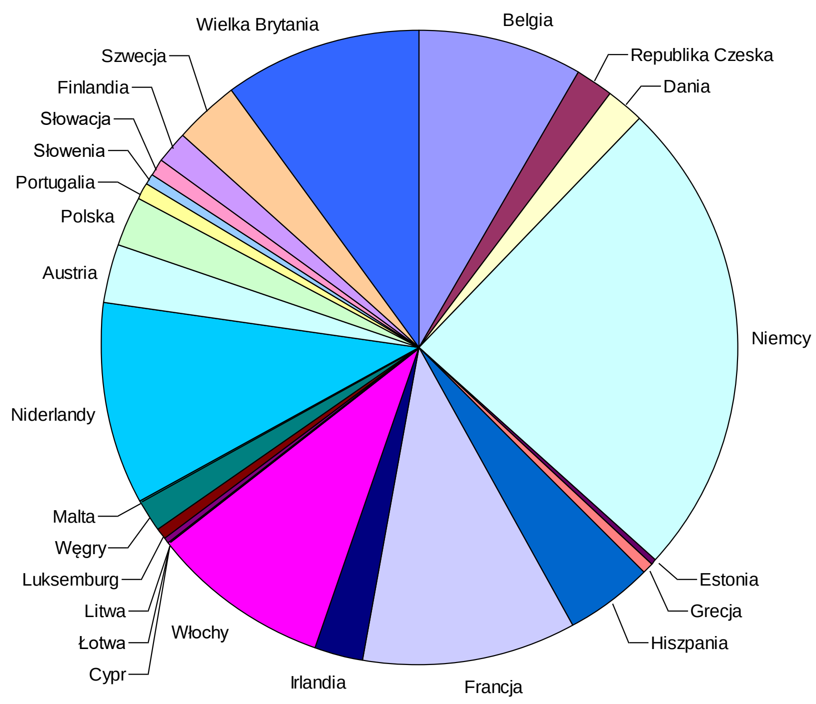


Źródło: http://www.stat.gov.pl/cps/rde/xbcr/gus/PUBL_unia_europejska_wskazniki_krotkookresowe_01_2007.xls

Wykres kołowy

Przykład

Tablica xx. Wartość eksportu krajów członkowskich UE w okresie2006 I-X (ceny bieżące w mld EUR)



Źródło:
http://www.stat.gov.pl/cps/rde/xbcr/gus/PUBL_unia_europejska_wskazniki_krotkookresowe_01_2007.xls

Ograniczenia wykresów kołowych:

- można przedstawić jedynie dane procentowe
- w próbce musi być co najmniej 1 obserwacja każdej kategorii (bo łączna suma pól wycinków musi stanowić 100 % pola koła)
- mało czytelne przy dużej liczbie kategorii
- analiza dwóch wykresów kołowych bardziej kłopotliwa niż połączonego wykresu słupkowego.

METODY OPISU DANYCH ILOŚCIOWYCH SKALARNYCH

Wykresy: **diagramy, histogramy, łamane częstości, wykresy przebiegu.**

Przykład. W stu kolejnych rzutach kostką sześcienną otrzymano wyniki (próbkę cechy dyskretnej o liczności 100):

5 2 2 6 3 2 5 3 1 2 5 3 6 2 5 4 4 6 1 6 4 5 5 2 4 6 1 4 4 3 4 2 4 2 4 4
1 1 4 5 3 1 5 6 5 6 1 5 6 2 4 5 5 2 5 4 5 5 1 1 2 2 5 5 2 6 3 5 5 4 1 4
5 5 1 4 3 2 1 2 6 1 2 1 6 5 1 3 6 1 5 6 6 2 2 3 5 5 2 4

Rozkład liczby oczek w próbce

Wartość (l. oczek)	1	2	3	4	5	6
Liczność (l. wystąpień)	16	19	9	17	25	14

Rozkład częstości liczby oczek w próbce

Wartość (l. oczek)	1	2	3	4	5	6
Częstość	0,16	0,19	0,09	0,17	0,25	0,14

Zwięzły opis próbki: **rozkład cechy w próbce**, tzn. zapisanie jakie wartości wystąpiły w próbce i ile razy, lub z jaką częstością.

Diagram liczebności **Diagram częstości**

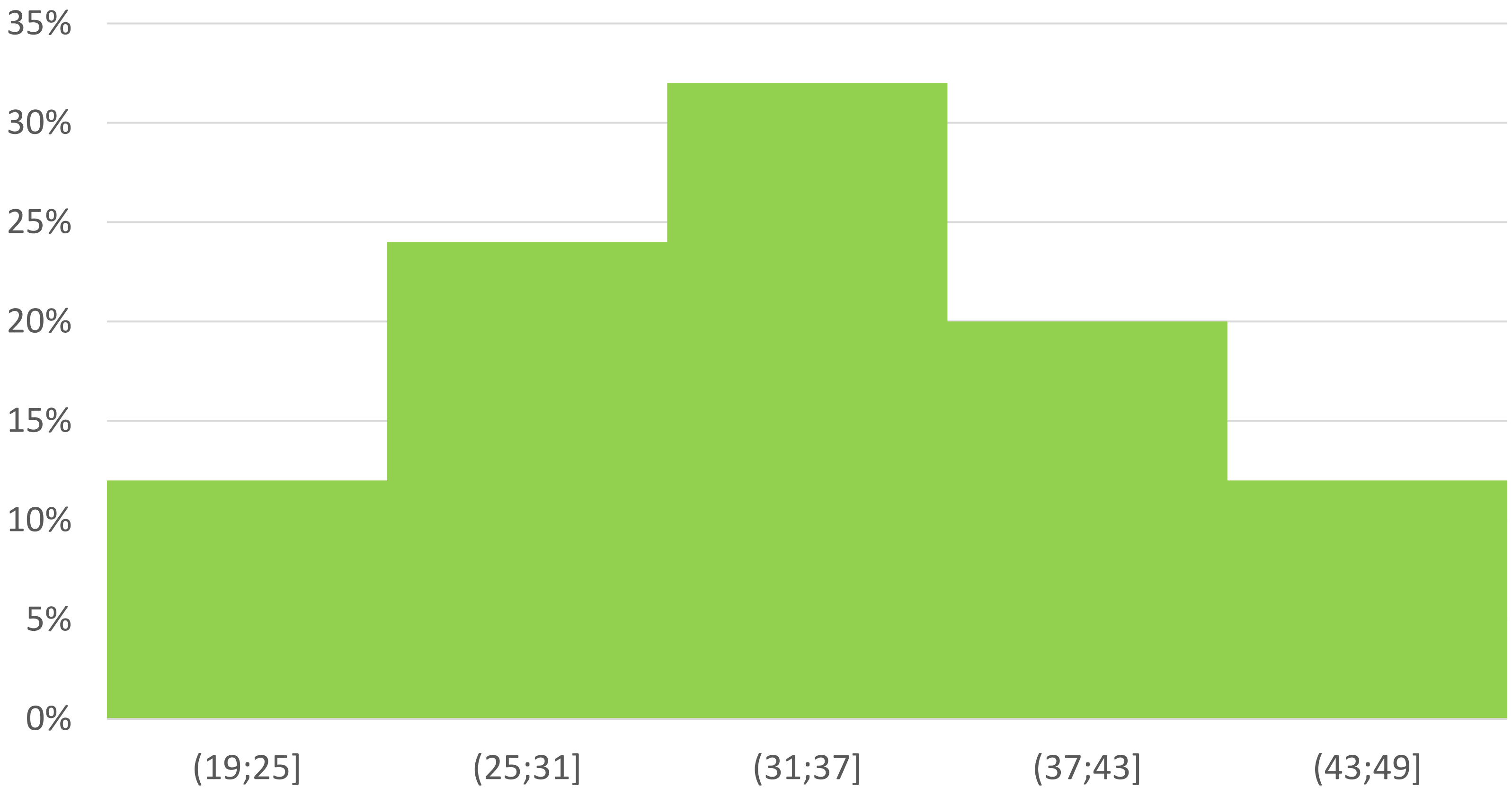
Przykład. Wiek **25** osób, które ubezpieczyły się w III filarze emerytalnym w pewnym zakładzie pracy: 30, **49**, 33, 35, 37, **20**, 31, 30, 36, 46, 39, 40, 38, 41, 35, 37, 24, 27, 36, 43, 45, 25, 32, 29, 28.

- **21 różnych wartości:** diagram rozkładu lat nieczytelny.
- **Agregacja danych:** przedziały wiekowe zawierające wszystkie obserwacje, liczba obserwacji w tych przedziałach.

Tabela liczności i częstości

Przedział	Obserwacje	Liczność	Częstość
(19;25]	20,24,25	3	12%
(25;31]	27,28,29,30,30,31	6	24%
(31;37]	32,33,35,35,36,36,37,37	8	32%
(37;43]	38,39,40,41,43	5	20%
(43;49]	45,46,49	3	12%

Histogram częstości wieku



Na osiach poziomych: granice klas wiekowych (przedziałów)

wysokości słupków = procentowy udział każdej klasy w próbce

Wysokość słupka = częstość klasy x **100%.**

Pole słupka =

stała długość przedziału x **częstość** x **100**

Histogram **liczebności:** wysokość słupka = **liczność klasy**

Histogram **częstości**: wysokość słupka = **częstość klasy**

Niech $x_1, x_2, \dots, x_n$ będzie $n-$elementową próbką.

Rozstępem z próbki nazywamy liczbę

$$R = x_{max} - x_{min} := x_{(n)} - x_{(1)}.$$

Przy większej liczności próbki $(n > 30)$, w celu ułatwienia analizy danych, wartości liczbowe próbki grupuje się w ***klasach*** (przedziałach najczęściej o jednakowej długości), przyjmując uproszczone założenie, że wszystkie wartości znajdujące się w danej klasie są identyczne ze środkiem przedziału. W wyniku grupowania otrzymujemy szereg rozdzielczy i histogram.

SZEREG ROZDZIELCZY i HISTOGRAM

Szeregiem rozdzielczym nazywamy ciag par

$$(\bar{x}_i, n_i), i = 1, \dots, k,$$

gdzie $\bar{x}_i$ jest środkiem *i* – tej klasy.

Ciąg $\{n_i\}$ nazywamy rozkładem liczności przy danej liczbie klas.

Początkowy wybór długości przedziałów:

$$h = 2{,}64 \cdot IQR \cdot n^{-1/3}$$

$n =$ liczność próbki

$IQR =$ rozstęp międzykwartylowy = zakres 50% środkowych wartości w próbce (po uporządkowaniu rosnąco)

Długość klasy - h

$h \cong \frac{R}{k}$, bo musi zachodzić $h \cdot k \geq R$.

Punkty stanowiące granice poszczególnych klas ustala się z dokładnością do $\frac{d}{2}$, gdzie d jest dokładnością pomiaru (lub przyjmuje się przedziały jednostronnie otwarte, aby każdy element próbki należał tylko do jednej klasy).

Oznaczmy przez n_i liczność $i-$tej klasy. Stąd

$$\sum_{i=1}^{k} n_i = n.$$

Liczba klas - k.

Liczność próbki *n*	Liczba klas *k*
30 - 60	6 – 8
60 – 100	7 –10
100 - 200	9 –12
200 – 500	11 – 17
500 – 1500	16 – 25

Na ogół nie stosuje się liczby klas *k* większej od 30.

Praktyczna metoda ustalania liczby klas **k** i szerokości **h** w próbie o liczności **n** i rozstępie **R**

- $k = \lfloor \sqrt{n} \rfloor \ \ lub \ \lceil \sqrt{n} \rceil$
- $h = \left\lceil \frac{R}{k} \right\rceil$
- Zalecane jest stosowanie przedziałów grupowania postaci (;] (otwarto - domkniętych)
- Pierwszy przedział może, w razie potrzeby, być domknięty – [;].

Mała długość przedziału to : **nieregularność** histogramu

Duża długość przedziału to: za duże **wygładzenie** histogramu

Przy ustaleniu kompromisu pomiędzy zbyt dużym wygładzeniem histogramu (redukcją informacji) a dużą nieregularnością histogramu pomocne są dodatkowe informacje o naturze obserwowanego zjawiska, np. obserwacje z kilku różnych populacji mogą dawać histogramy wielomodalne.

- **Początek histogramu**: najmniejsza obserwacja stanowi środek pierwszego przedziału. Uśredniając kilka histogramów o nieznacznie przesuniętych początkach można uniezależnić się od wpływu początku histogramu na jego kształt.

Przykład. Wkładka topikowa bezpiecznika o natężeniu znamionowym 20A winna, zgodnie z normą, wytrzymać bez przepalenia się natężenie 28A w ciągu 1 godziny. W celu sprawdzenia zgodności z normą, z partii wkładek topikowych tego typu pobrano losowo 40 sztuk i zanotowano czasy przepalenia się wkładki przy natężeniu prądu 28A. Otrzymano następujące wyniki w minutach:

51 58 64 69 61 56 41 48 56 61
75 55 46 57 70 55 47 62 55 60
54 57 65 60 53 54 49 58 62 59
53 50 58 63 64 59 52 51 65 60

Dla przedstawionej próbki zbudować szereg rozdzielczy oraz narysować histogram i łamaną częstości.

Rozwiązanie.

Zauważmy, że $x_{min} = 41$ oraz $x_{max} = 75.$

Stąd rozstęp z próbki: $R = 75 - 41 = 34.$

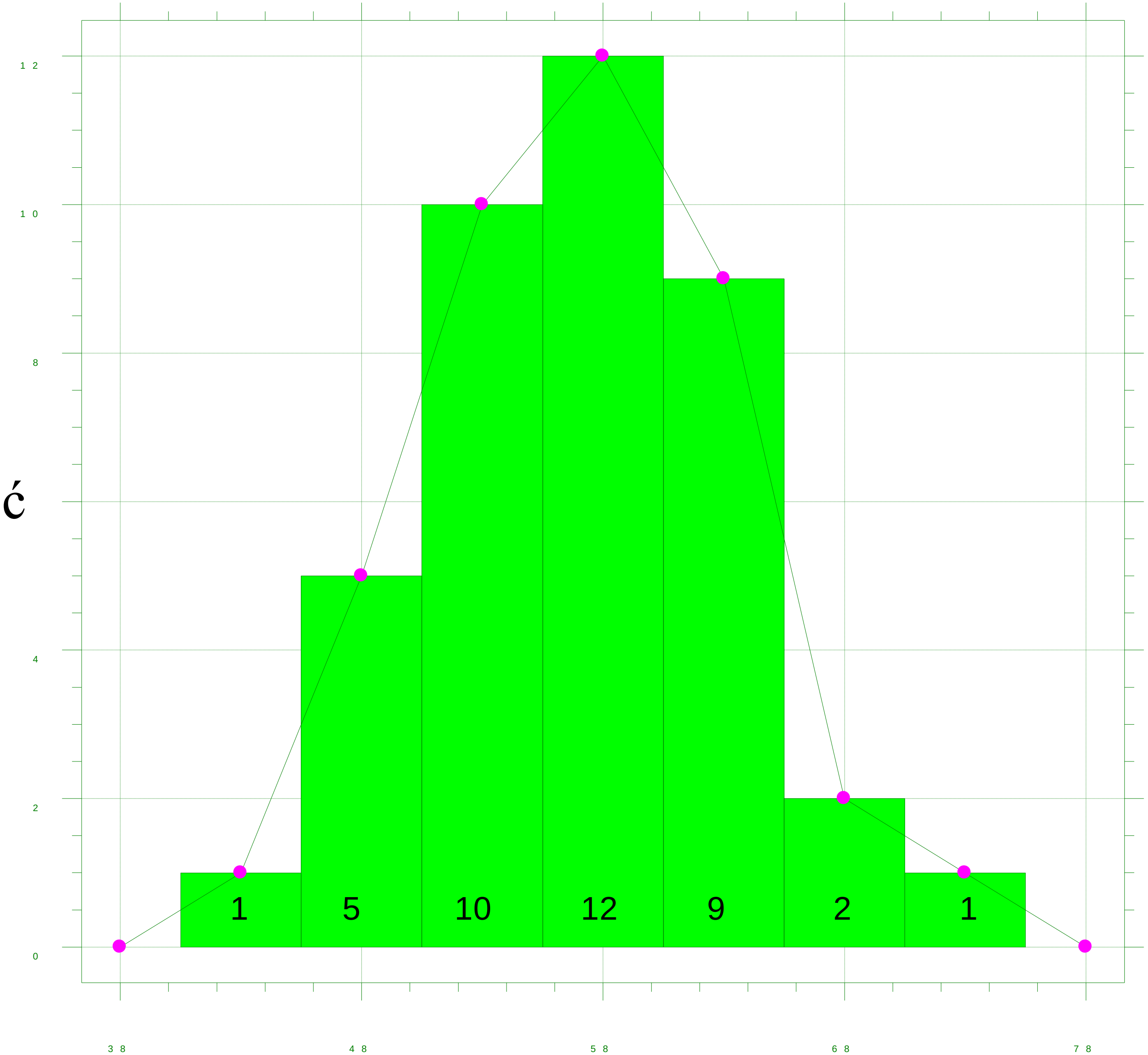
Ponieważ liczność próbki $n = 40$, to wygodnie jest przyjąć liczbę klas $k = 7$ oraz szerokość klasy $h = 5.$

Tym samym otrzymujemy szereg rozdzielczy:

Nr klasy i	Klasa		$\overline{x}_i$	n_i	w_i	N_i	W_i
1	40,5	45,5	43	1	0,025	1	0,025
2	45,5	50,5	48	5	0,125	6	0,150
3	50,5	55,5	53	10	0,250	16	0,400
4	55,5	60,5	58	12	0,300	28	0,700
5	60,5	65,5	63	9	0,225	37	0,925
6	65,5	70,5	68	2	0,050	39	0,975
7	70,5	75,5	73	1	0,025	40	1,000

Histogram oraz łamana liczności

liczność



WSKAŹNIKI SUMARYCZNE

WSKAŹNIKI POŁOŻENIA (**miary położenia**, **parametry położenia**) charakteryzują najbardziej reprezentatywne dane, centralną „tendencję” danych, określają „środek” próbki:

Niech **:** $x_1, x_2, \ldots, x_n$ - próbka o liczności n.

Wartość średnia w próbce (**średnia próbkowa, średnia próbki**)

$$\boxed{\bar{x} = \frac{1}{n}(x_1 + x_2 + \ldots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i}$$

Przykład. Miesięczny dochód 10-ciu osób (w tys. PLN):

Dochód (PLN)	[1, 1,5]	(1,5, 2]	(2, 2,5]	(2,5, 3]
Liczba osób	2	2	4	2

Średnia na podstawie danych zgrupowanych:

$$\overline{x} = \sum_{i=1}^{k} \frac{n_i \tilde{x}_i}{n} = \frac{2 \times 1{,}25 + 2 \times 1{,}75 + 4 \times 2{,}25 + 2 \times 2{,}75}{10} = 2{,}05$$

Mediana w próbce (mediana próbki, mediana próbkowa)

Niech $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n-1)} \le x_{(n)}$

uporządkowane w sposób rosnący wartości próbki:

$x_{(1)} = \min\{x_1, x_2, \ldots, x_n\}, \ldots, x_{(n)} = \max\{x_1, x_2, \ldots, x_n\}$

$\boldsymbol{x_{med}}$	=	$x_{((n+1)/2)}$,	gdy n jest nieparzyste
$\boldsymbol{x_{med}}$	=	$\frac{1}{2}(x_{(n/2)} + x_{(n/2+1)})$,	gdy n jest parzyste.

Przykład. Miesięczny dochód 11-tu osób:

Dochód (PLN)	2000	2500	3500	19000
Liczba osób	4	4	2	1

Średnie wynagrodzenie tej grupy osób to:

$$\bar{x} = \frac{1}{11}(4 \times 2000 + 4 \times 2500 + 2 \times 3500 + 19000) = \mathbf{4000}$$

2000, 2000, 2000, 2000, 2500, 2500, 2500, 2500, 3500, 3500, 19000

Mediana = 2500

Średnia wrażliwa na obserwacje odstające:

$\bar{x} = 4000 > 3500 = x_{(10)}, \; x_{(11)} = 19000$ - **średnia nie odzwierciedla „typowego" dochodu.**

Mediana odporna (mało wrażliwa) na obserwacje odstające:

$x_{med} = x_{(6)} = 2500$ - **mediana jest lepszą miarą przeciętnego wynagrodzenia niż średnia**

Średnia ucinana (**ucięta**) (z parametrem k)

$$\boxed{\bar{x}_{tk} = \frac{1}{n-2k} \sum_{i=k+1}^{n-k} x_{(i)}},$$

stosowana gdy wartości odstające są wynikiem błędu

(błędne przetworzenie danych lub błędy przyrządów

pomiarowych).

Ostrzeżenie: obserwacje odstające mogą być bardzo istotne, np. są wynikiem rozregulowania procesu produkcji

Średnia winsorowska (z parametrem k)

$$\bar{x}_{wk} = \frac{1}{n}\left[(k+1)x_{(k+1)} + \sum_{i=k+2}^{n-k-1} x_{(i)} + (k+1)x_{(n-k)}\right]$$

Stosowana w sytuacjach gdy wartości skrajne (k najmniejszych lub k największych) niepewne co do ich prawdziwych wartości (np. zostały utracone z bazy danych; nie mogły być zaobserwowane w przypadku badania czasu życia lub czasu bezawaryjnej pracy urządzenia gdy eksperymentator ma ograniczony czas obserwowania zjawiska.

Moda – najczęściej występująca wartość (lub wartości) w próbce.

WSKAŹNIKI ROZPROSZENIA (**miary rozproszenia**, **parametry rozproszenia**) charakteryzują rozrzut danych, rozproszenie wartości próbki wokół parametru położenia.

Rozstęp próbki

$$\boxed{R = x_{(n)} - x_{(1)}},$$

Wariancja próbki (w próbce)

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

Przykład. Miesięczny dochód 10-ciu osób (w tys. PLN):

Dochód (PLN)	[1, 1,5]	(1,5, 2]	(2, 2,5]	(2,5, 3]
Liczba osób	2	2	4	2

Wariancja na podstawie danych zgrupowanych:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{k} n_i(\widetilde{x}_i - \bar{x})^2 = 0{,}2889.$$

Odchylenie standardowe w próbce (próbki)

$$\boxed{s = \sqrt{s^2}}$$

Odchylenie przeciętne od wartości średniej

$$\boxed{d_1 = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - \bar{x}\right|}$$

Dolny (pierwszy) kwartyl

Q_1 = mediana podpróbki składającej się z

elementów próbki „mniejszych" od mediany x_{med} .

Górny (trzeci) kwartyl

Q_3 = mediana podpróbki składającej się z elementów próbki „większych” od mediany (w próbce uporządkowanej rosnąco są to elementy występujące na pozycjach po pozycji mediany.

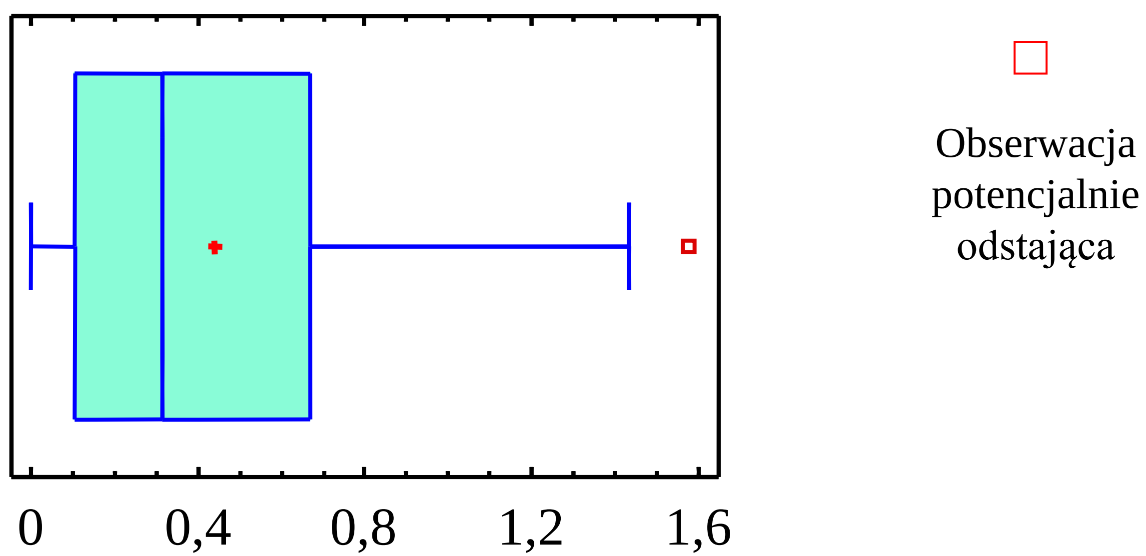
Rozstęp międzykwartylowy

$$IQR = Q_3 - Q_1.$$

WYKRES RAMKOWY (pudełkowy)

ilustruje wzajemne położenie pięciu wskaźników sumarycznych:

$$x_{(1)} = x_{min}, \quad Q_1, \quad x_{med}, \quad Q_3, \quad x_{(n)} = x_{max}.$$



Z wykresu odczytujemy następujące wskaźniki:

- Q_1 = 0,1 = rzut na oś poziomą lewego boku prostokąta
- Q_3 = 0,7 = rzut na oś poziomą prawego boku prostokąta
- Q_2 = 0,3 = rzut na oś poziomą pionowego odcinka wewnątrz prostokąta
- IQR = długość podstawy prostokąta

Wąsy wykresu ramkowego = linie po obu stronach prostokąta.

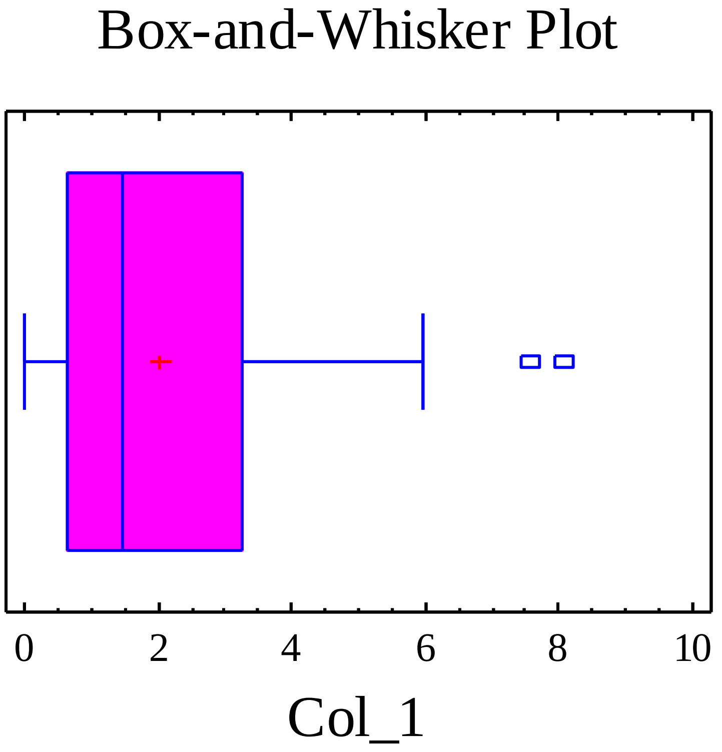
Rzut lewego wąsa na oś poziomą = przedział $[x_*, Q_1]$, gdzie

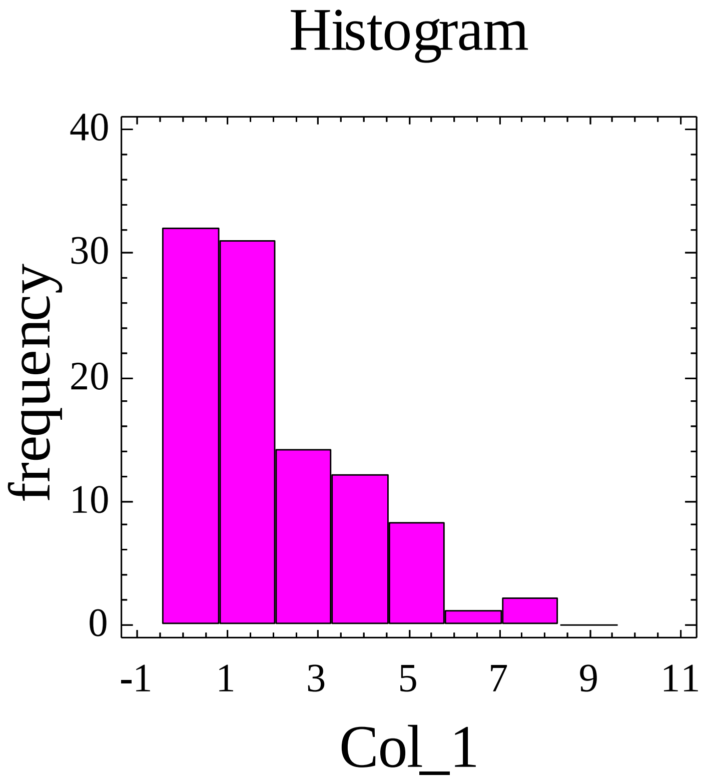
$x_* = \min\{ x_k\colon\ Q_1 - 3/2 \cdot IQR \le x_k \le Q_1 \}$,

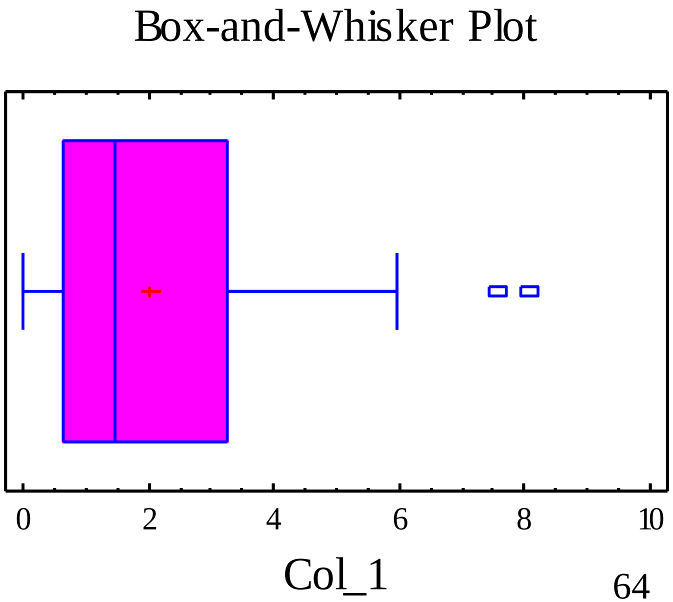
podobnie określamy rzut prawego wąsa = przedział $[Q_3, x^*]$, gdzie

$x^* = \max\{ x_k\colon\ Q_3 \le x_k \le Q_3 + 3/2 \cdot IQR \}$

Count = 100
Average = 2,02544
Median = 1,46467
Variance = 3,16395
Standard deviation = 1,77875
Minimum = 0,0150559
Maximum = 8,05684
Range = 8,04179
Lower quartile = 0,638618
Upper quartile = 3,23695
Interquartile range = 2,59833
Coeff. of variation = 87,8206%

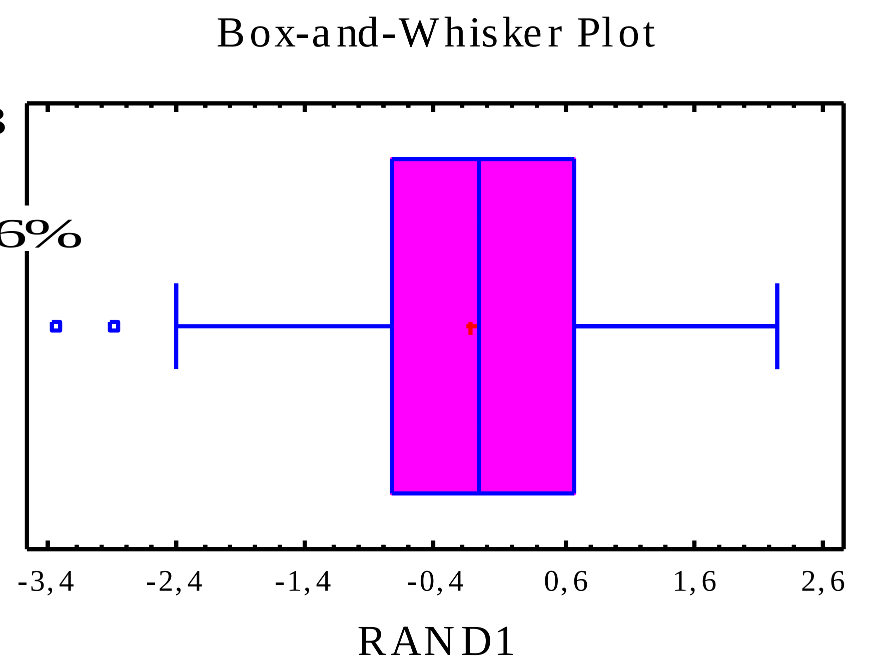



Histogram
40
30
20
10
0
frequency
-1
1
3
5
7
9
11
Col_1

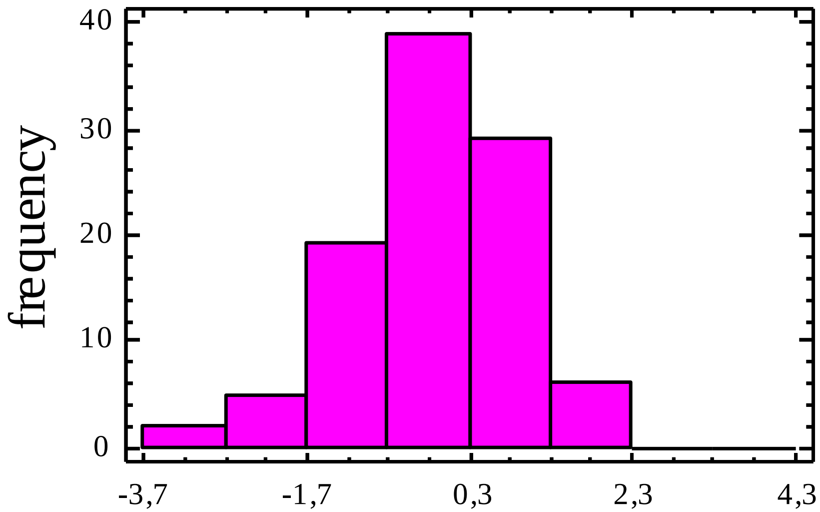

Box-and-Whisker Plot
0
2
4
6
8
10
Col_1

Summary Statistics for RAND1

Count = 100
Average = -0,110696
Median = -0,0516888
Variance = 1,07775
Standard deviation = 1,03815
Minimum = -3,36516
Maximum = 2,26235
Range = 5,62751
Lower quartile = -0,726224
Upper quartile = 0,680553
Interquartile range = 1,40678
Stnd. skewness = -1,86072
Coeff. of variation = -937,836%



Histogram



Box-and-Whisker Plot