Podstawy Biostatystyki

Wydział Nauki o Zdrowiu

Zakład Profilaktyki Zagrożeń Środowiskowych i Alergologii

Marta Zalewska

Treść wykładu

- **W1-W2** Statystyka opisowa. Podstawowe pojęcia statystyki. Prezentacja i podsumowanie danych. Miary położenia i dyspersji, pojecie wartości średniej, kwantyla, mediany, wartości modalnej. Dane pogrupowane. Szeregi rozdzielcze – obliczenia z danych pośrednich, histogramy. Tablice kontyngencji. Zasady losowania próby z populacji.
- **W3** Przypomnienie wybranych treści z rachunku prawdopodobieństwa w kontekście medycznym. Zdarzenia niezależne i zależne. Prawdopodobieństwo warunkowe. Wzór na prawdopodobieństwo całkowite oraz wzór Bayesa.

Treść wykładu

- **W4** Rozkład wartości cechy w populacji. Pojęcie zmiennej losowej, rodzaje zmiennych losowych, rozkład zmiennej losowej. Parametry rozkładu prawdopodobieństwa. Podstawowe rozkłady prawdopodobieństwa i ich własności (rozkład Bernoulliego, normalny).
- **W5** Estymacja punktowa. Podstawy estymacji przedziałowej. Przedział ufności dla średniej. Przedział ufności dla wskaźnika struktury.
- **W6** Wprowadzenie do testowania hipotez statystycznych. Hipoteza zerowa, alternatywna, błędy pierwszego i drugiego rodzaju. Poziom istotności testu.

Treść wykładu

- **W7** Testy istotności. Test t-Studenta. Porównanie z normą. Porównanie dwóch populacji. Porównanie wielu populacji. Testy nieparametryczne.
- **W8** Testy zgodności z rozkładem: test chi-kwadrat zgodności i test Kołmogorowa.
- **W9** Zależność między cechami. Współczynnik korelacji Pearsona i Spearmana. Wprowadzenie do analizy regresji. Test chi-kwadrat niezależności. Czułość i specyficzność w testach medycznych. Ryzyko względne i ryzyko przypisane.
- **W10** Przykłady analiz statystycznych przy użyciu pakietów statystycznych.

Cele

- Prezentacja danych przy użyciu statystyki opisowej i odpowiednich wykresów
- Podstawowe zrozumienie statystyki matematycznej

  - Właściwy dobór podstawowych testów statystycznych do danych empirycznych

  - Prawidłowa interpretacja wyników podstawowych analiz

Literatura:

- Łomnicki A. Wprowadzenie do statystyki dla przyrodników. PWN, Warszawa, 2003.
- **Stanisz A. Pod redakcją. Biostatystyka. Wydawnictwo Uniwersytetu Jagiellońskiego, Kraków, 2005.**
- Watała C. Biostatystyka – wykorzystanie metod statystycznych w pracy badawczej w naukach biomedycznych. -medica press, Bielsko-Biała, 2002.
- Zieliński W. Wybrane testy statystyczne. Fundacja „Rozwój SGGW”. Warszawa, 1999.

Metody oceny: kolokwium, aktywność na zajęciach, kartkówki.

Egzamin: po I semestrze.

Konsultacje

- Wtorki p.112 blok F
- Proszę o wcześniejszy kontakt mailowy
- zalewska.marta@gmail.com

Biostatystyka

- Biostatystyka zajmuje się zbieraniem, przetwarzaniem, przedstawianiem oraz wnioskowaniem na podstawie danych biomedycznych.
- Dane obejmują obserwacje jednej lub wielu zmiennych.
- Dane statystyczne dotyczą pewnej zbiorowości, zwanej populacją.
- Obserwuje się lub bada elementy tej zbiorowości, czyli jednostki badania tworzące próbkę.
- Interesują nas pewne cechy jednostek.

Populacja

- Pacjenci, lekarze, szpitale, przychodnie, studenci, uczelnie
- Populacja (jednostki badawcze) może być zdefiniowana jako:
  - przychodnie w województwie mazowieckim
  - przychodnie w Warszawie
  - przychodnie w Polsce
- Badacz definiuje populację, w stosunku do której będzie odnosił uzyskane wnioski z przeprowadzonych badań
- Z ustalonej populacji wybieramy próbkę (część populacji)
- Często, choć nie zawsze jest to próbka losowa

Próbka losowa

- Próbka reprezentatywna
- Dobór jednostek z populacji do próbki
- Różne modele losowania

Rachunek prawdopodobieństwa

Wnioskowanie statystyczne

- Uogólnianie informacji zawartych w analizowanych danych (wartości pewnej cechy lub zestawu cech) na całą populację.

- Wnioskowanie o całej populacji na podstawie losowej próbki wymaga metod rachunku prawdopodobieństwa.

Organizacja badań i zbieranie danych

- Pełne – obejmuje całą populację (wyników nie uogólnia się; nie używa się metod wnioskowania statystycznego).
- Reprezentacyjne – dysponujemy danymi dla części populacji (metoda reprezentacyjna).

Organizacja badań biomedycznych (wybór próbki):

- Przekrojowe (Cross-sectional). Badania podstawowe
- Kohortowe, Prospektywne (Cohort) - Potwierdzające
- Przypadek-kontrola, Retrospektywne (Case-Control), Kliniczno kontrolne - Badawcze
- Porównywanie testów medycznych

Zmienne (cechy) dzielimy na:

- Jakościowe – opisowo określane własności jednostek, kategorie (niemierzalne)
- Ilościowe – wielkości liczbowe (mierzalne)

- Jakościowa - kategorialna (kategoryczna)
  - Nominalna (grupa krwi, stan cywilny)
  - Porządkowa (stan zaawansowania choroby, stopień otyłości)

- Numeryczna, ilościowa
  - dyskretna (liczba dni choroby w ciągu roku)
  - ciągła (masa w kg, wzrost w cm, dochody,

Dane numeryczne (ilościowe)

- dyskretne (skokowe) - skończona liczba wartości (liczba posiadanych dzieci)
- ciągłe - dochody, wzrost, masa ciała

<u>Uwaga</u>

Dane numeryczne ciągłe wprowadzamy do baz danych z tą samą dokładnością z jaką zostały zmierzone, wszystkie w tych samych jednostkach np.

masa ciała w [kg]

W zależności od potrzeby badacza:

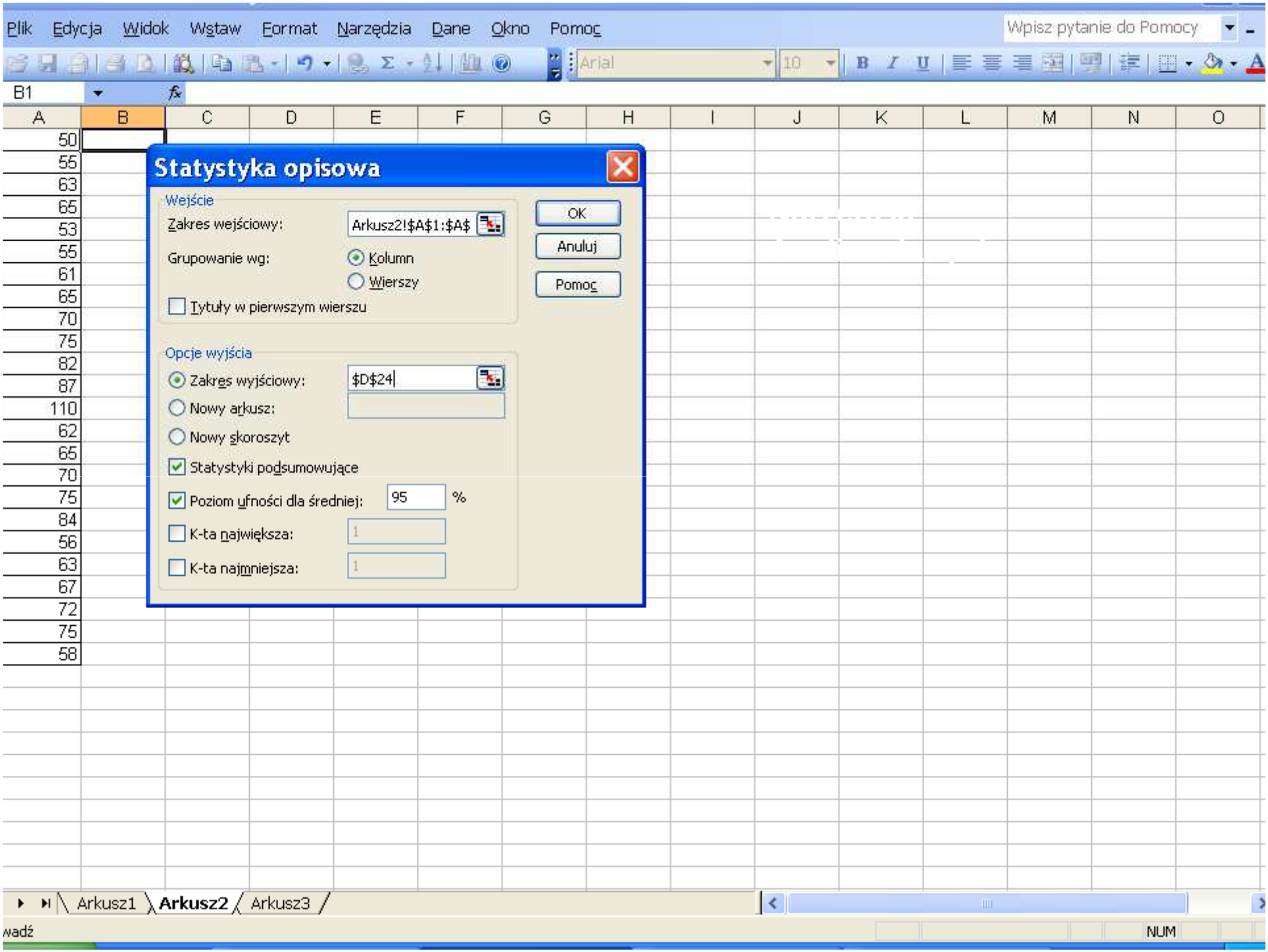
- Zawał – „czy był” – cecha jakościowa (1,0) dychotomiczna (dwie wartości), dyskretna.
- Zawał - „jaki był” (lekki, średni, silny) – porządkowa.
- Zawał - „ile zawałów” - to cecha ilościowa (dyskretna).

Statystyka opisowa

- Metody pozwalające na określenie częstości występowania danej wartości cechy, wartości średniej oraz rozrzutu danej cechy bez użycia rachunku prawdopodobieństwa
- Przetwarzamy posiadane dane (dotyczące badanych zmiennych) nie wnikając w to, czy dotyczą one całej populacji, czy też tylko próbki z populacji

Statystyka opisowa

- Scharakteryzować badaną (obserwowaną) grupę podać wskaźniki sumaryczne
- Charakterystyki podstawowe:
  - Miary położenia
  - Miary rozproszenia
- Charakterystyki uzupełniające:
  -Współczynnik zmienności
  - Miary skośności i kurtozy


Plik Edycja Widok Wstaw Format Narzędzia Dane Okno Pomoc
Wpisz pytanie do Pomocy
Arial
10
B1
Statystyka opisowa
Wejście
Zakres wejściowy:
Arkusz2!A1:A
Grupowanie wg:
Kolumn
Wierszy
Tytuły w pierwszym wierszu
OK
Anuluj
Pomoc
Opcje wyjścia
Zakres wyjściowy:
D24
Nowy arkusz:
Nowy skoroszyt
Statystyki podsumowujące
Poziom ufności dla średniej:
95
%
K-ta największa:
1
K-ta najmniejsza:
1
Arkusz1
Arkusz2
Arkusz3
NUM

Wyniki statystyki opisowej

	Kolumna1
Średnia	68,25
Błąd standardowy	2,70550298
Mediana	65
Tryb	65
Odchylenie standardowe	13,2542036
Wariancja próbki	175,673913
Kurtoza	3,027149576
Skośność	1,398332022
Zakres	60
Minimum	50
Maksimum	110
Suma	1638
Licznik	24
Poziom ufności(95,0%)	5,596759298

Prezentacja danych medycznych

- Tabele
- Wykresy
- Diagramy

Tabela1.Zbiorcze zestawienie danych szpitali klinicznych Warszawskiego Uniwersytetu Medycznego 2008 rok

Szpitale WUM	Baza łóżkowa 2008
1	**1 048**
2	**52**
3	**384**
4	**714**
5	**215**
	2 413

LEGENDA

1	SP Centralny Szpital Kliniczny ul. Banacha 1a, 02-097 Warszawa
2	SP Kliniczny Szpital Okulistyczny ul. Sierakowskiego 13, 03-709 Warszawa
3	SP Dziecięcy Szpital Kliniczny ul. Marszałkowska 24, 00-570 Warszawa
4	Szpital Kliniczny Dzieciątka Jezus - CLO ul. Lindleya 4, 02-005 Warszawa
5	Szpital Kliniczny im. ks. Anny Mazowieckiej ul. Karowa 2, 00-315 Warszawa

Biuro ds. Szpitali i Bazy Klinicznej, WUM.
23.03.2009 r

Przykład

W populacji studentów WUM interesują nas cechy: ocena z biostatystyki, wiek studenta, płeć. Cechy ilościowe i jakościowe.

student	ocena	wiek	płeć
1	5	19	K
2	3	21	M
3	4	22	K
…	…	…	…

Ogólna postać danych

Obiekt (id)	cecha X	cecha Y	cecha Z
1	x_1	y_1	z_1
2	x_2	y_2	z_2
3	x_3	y_3	z_3
…	…	…	…

Jeśli cechy X i Y są ilościowe to x_1,… i y_1,… są liczbami. Jaśli cecha Z- jakościowa to wartości z_i traktujemy jak nazwy lub umowne symbole,
można również używać symboli liczbowych: M=1, K=2

Kodowanie informacji-przykłady

-„wzrost” (kod 1-niski, 2-średni, 3-wysoki)

-„oddział” (1-internistyczny, 2-urazowy,…)

- „występowanie choroby” – (1-tak , 0-nie)

Prezentacja danych (jedna zmienna)

Rozkłady częstości (oparte o liczebność lub częstość względną).

Dane kategoryczne, dyskretne

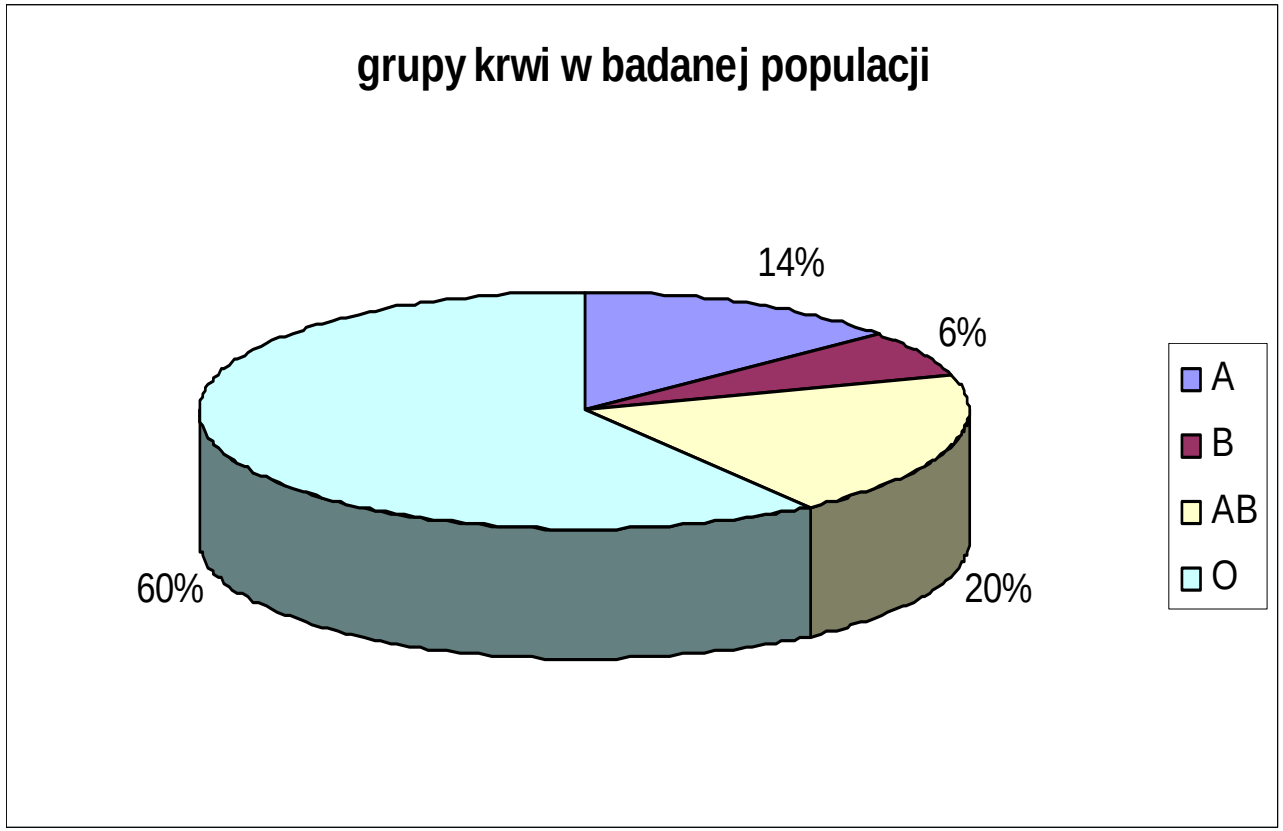
- Wykresy słupkowe lub kolumnowe
- Wykresy kołowe

Dane ciągłe

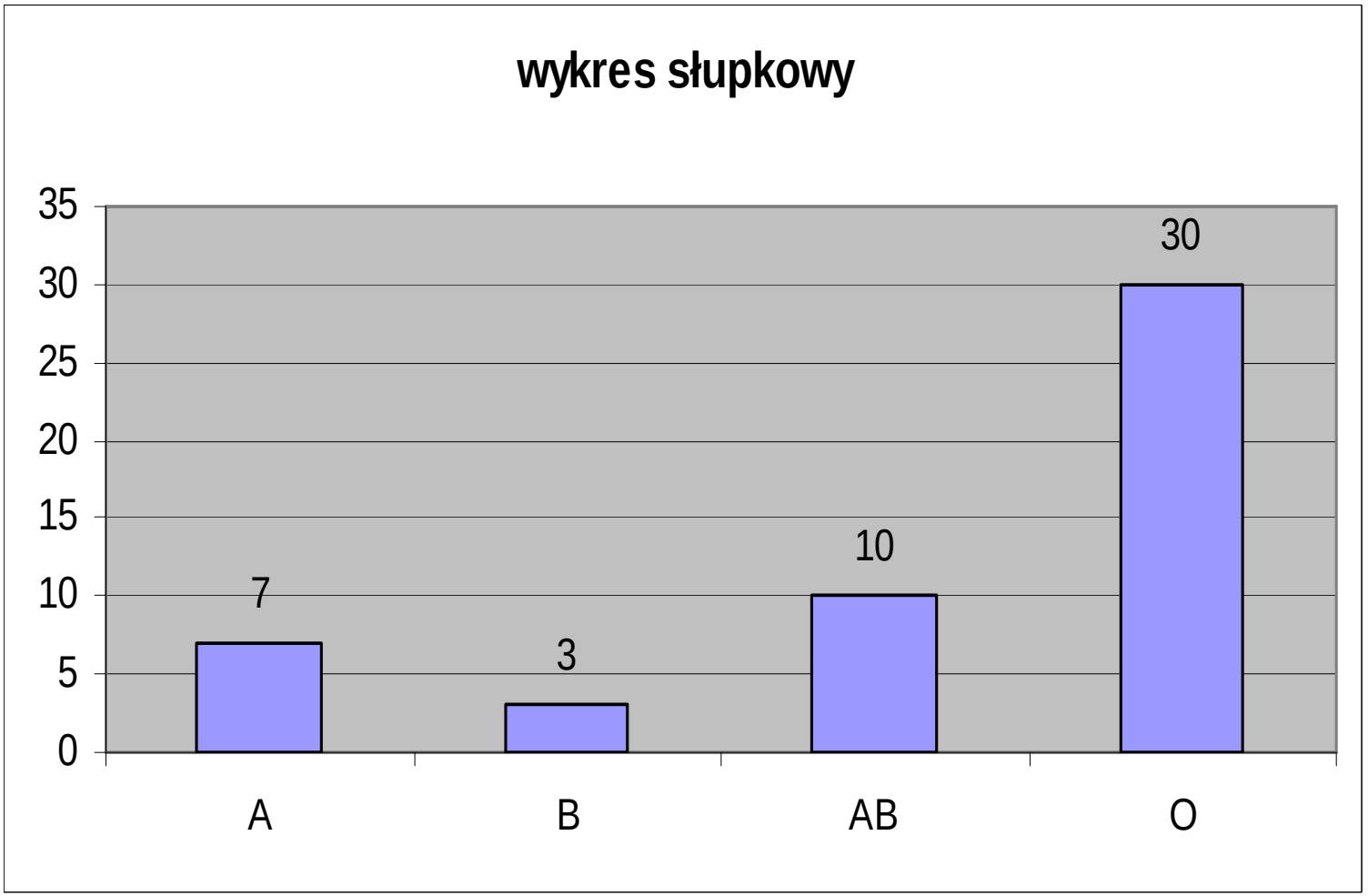
- Histogram
- Wykres punktowy
- Łodyga i liście
- Wykres pudełkowy

Cecha jakościowa, dyskretna

Grupa krwi	Liczba osób
A	7
B	3
AB	10
0	30



Cecha jakościowa, dyskretna c.d.



Cecha ciągła- histogram dla podobny do wykresu słupkowego (odległości między słupkami=0) do badania kształtu rozkładu

Obserwujemy rozkłady pojawiania się różnych wartości cechy (wiek)

Wartości cechy

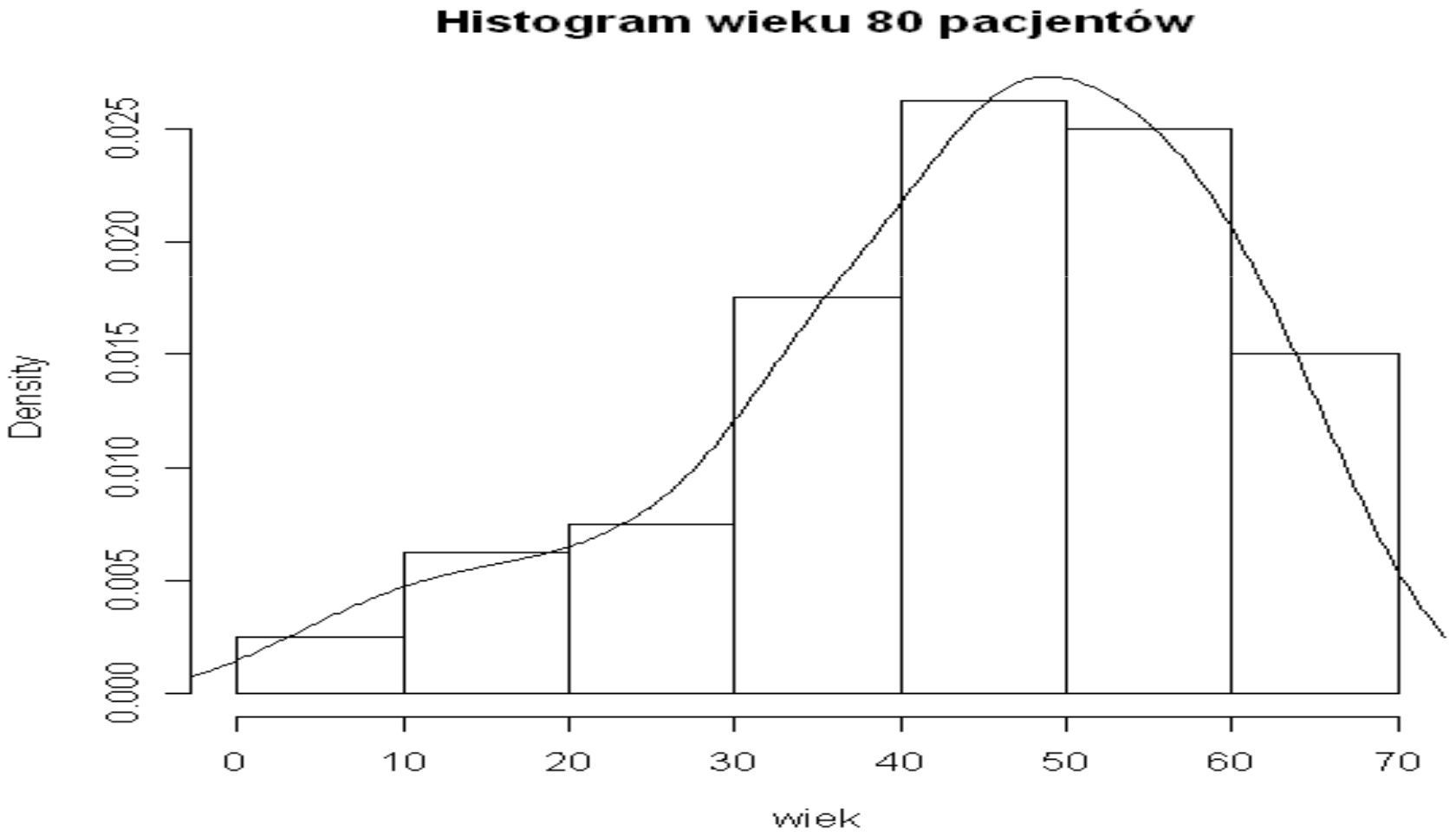
- Liczności
- Częstości względne
- Częstości względne (%)

100*2/80

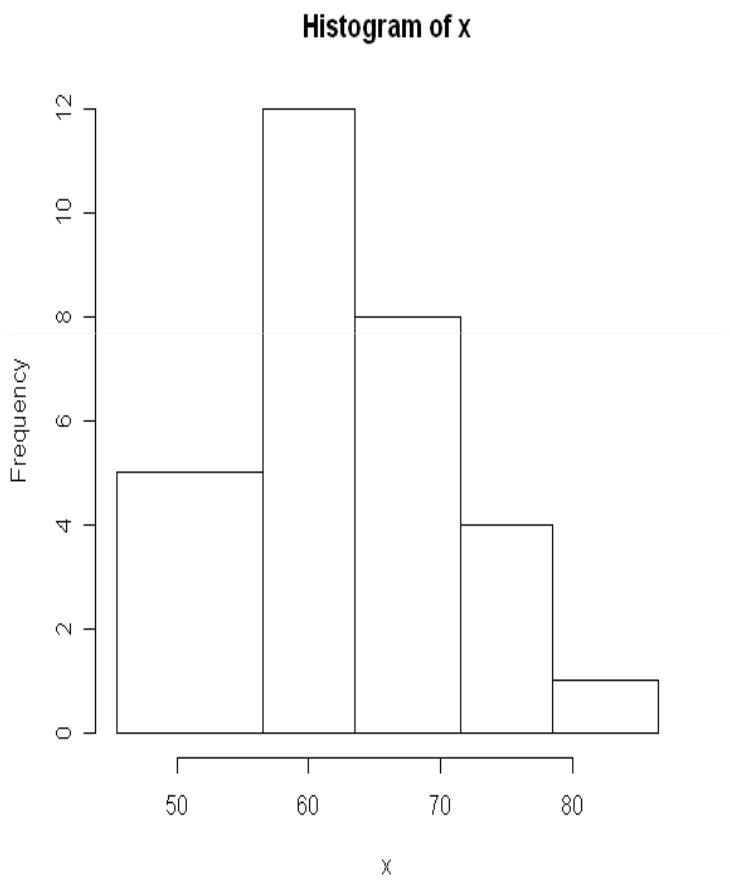
Lata	Liczność	Częstość
0 - 9	2	2.50
10-19	5	6.25
20-29	6	7.50
30-39	14	17.50
40-49	21	26.25
50-59	20	25.00
>=60	12	15.00

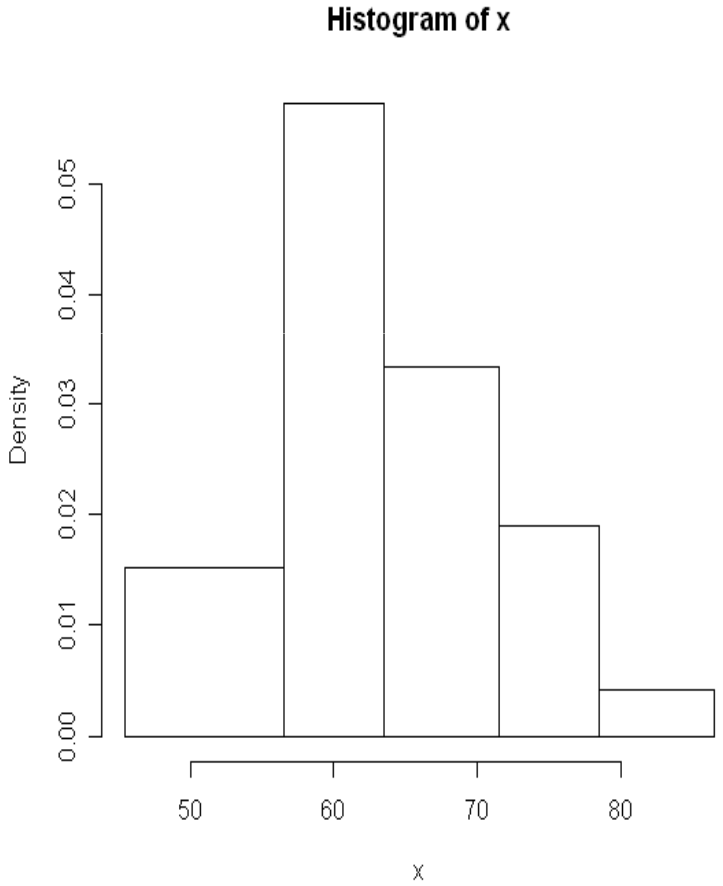
Histogram

Lata	Liczność	Częstość
0 - 9	2	2.50
10–19	5	6.25
20–29	6	7.50
30–39	14	17.50
40–49	21	26.25
50–59	20	25.00
>=60	12	15.00



HISTOGRAM (liczebności i częstości względne)





Wykres pudełkowy wieku 80 pacjentów

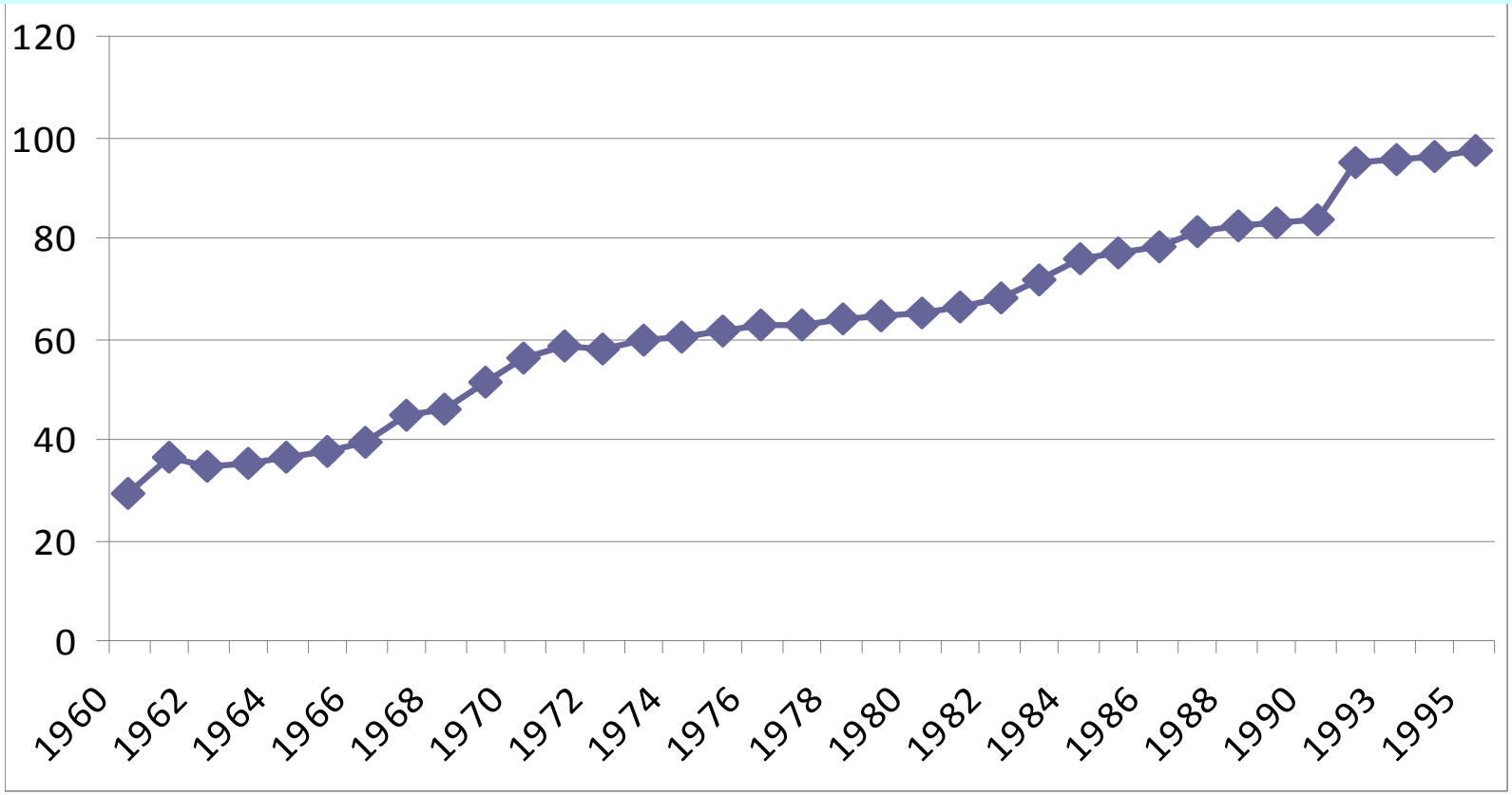
Wykres „łodyga i liście”

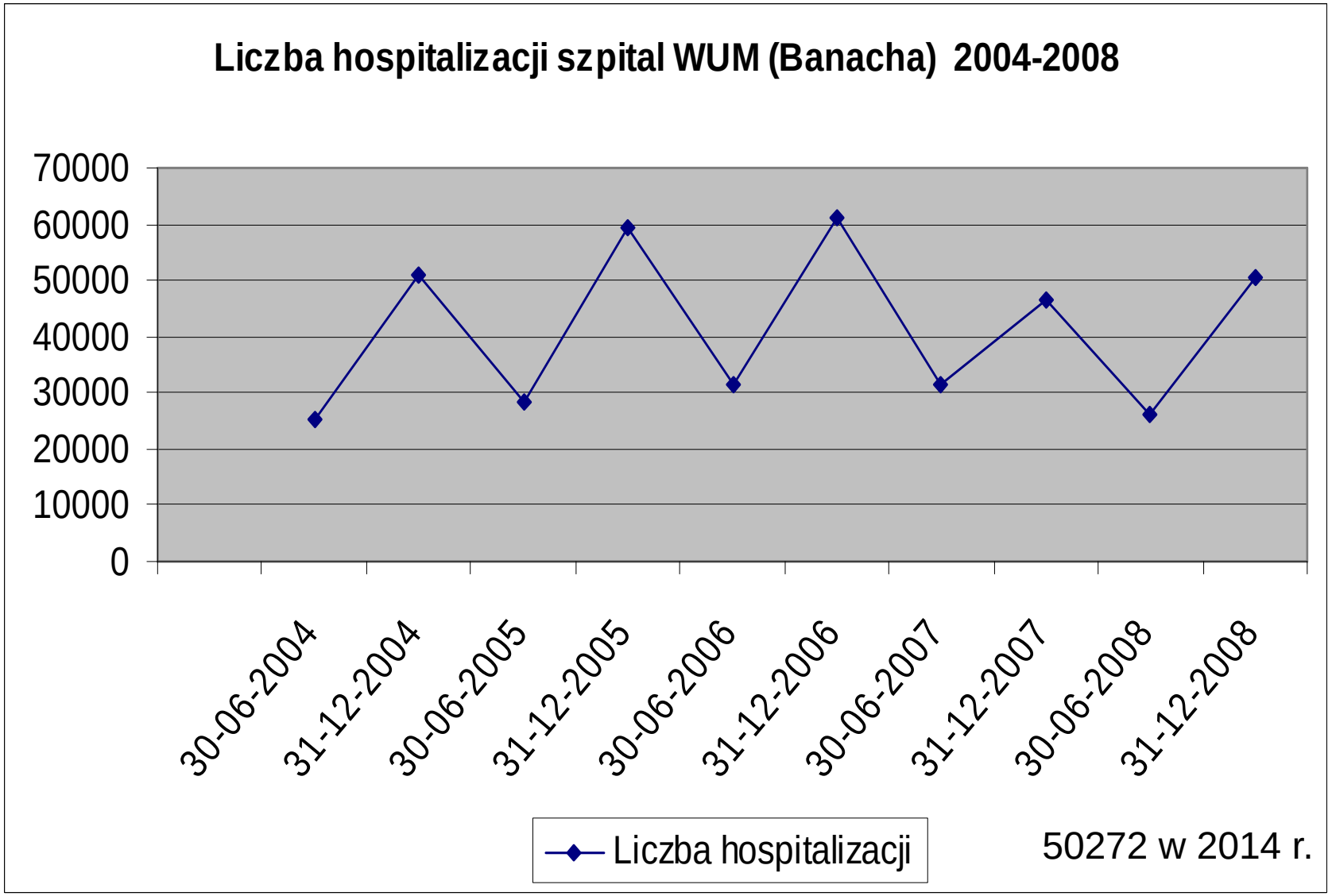
stem(wiek)

0 | 46
1 | 12378
2 | 224689
3 | 12335566677888
4 | 122235555666667777788
5 | 111122345555556667788
6 | 111122335567

4, 6, 11, 12, 13, 17, 18, 22, 22, 24, 26, 28, 29, 31, 32, 33, 33, 35, ….,67

Wykres przebiegu zachorowań na nowotwory złośliwe w latach 1960-1995 w tys.




Liczba hospitalizacji szpital WUM (Banacha) 2004-2008
70000
60000
50000
40000
30000
20000
10000
0
30-06-2004
31-12-2004
30-06-2005
31-12-2005
30-06-2006
31-12-2006
30-06-2007
31-12-2007
30-06-2008
31-12-2008
Liczba hospitalizacji
50272 w 2014 r.

Podstawowe pojęcia statystyczne

- Wprowadzimy przy założeniu, że obserwujemy czy mierzymy jedną cechę

  w pewnej grupie *n* - elementowej

Pojedyncza cecha ilościowa

- Interesuje nas tylko jedna cecha ilościowa. Dane mają postać ciągu liczb:

$$x_1, x_2, \ldots, x_n ,$$

gdzie n jest liczbą zbadanych (zaobserwowanych) jednostek (obiektów, pacjentów) zaś x_i oznacza wartość cechy X dla i-tej spośród tych jednostek.

ŚREDNIA (wartość przeciętna)

- Najprostszym sposobem „streszczenia” danych jest obliczenie średniej
  - średnia arytmetyczna
  - średnia geometryczna
  - średnia harmoniczne
  - średnia ważona

Średnia (lub wartość przeciętna) to liczba:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Zauważmy:

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

oraz

$$\sum (x_i - \bar{x}) = 0$$

Przykład

Ilość zużytego składnika wypełnienia (w gramach) w ciągu 10 kolejnych dni wyniosła:

12.0, 10.5, 17.3, 21.1, 14.7,
18.0, 11.5, 12.7, 10.9, 9.3.

Sumaryczne zużycie składnika w ciągu 10 dni ma wartość 138.

Średnia dzienna wartość zużycia składnika jest równa 138/10= 13.8

Od interpretacji danych zależy, czy obliczanie średniej arytmetycznej jest uzasadnione, czy nie.

Bardzo ważną rolę będzie odgrywać średnia ważona

Średnia ważona

- Definicja. Średnią ważoną liczb $x_1, x_2, \ldots, x_k$ z odpowiadającymi im wagami $w_1, w_2, \ldots, w_k$ nazywamy liczbę

$$\overline{x}_w = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_k x_k}{w_1 + w_2 + \cdots + w_k}$$

Jeśli wagi są jednakowe $w_1 = w_2 = \ldots = w_k$, to średnia ważona jest po prostu średnią arytmetyczną

Przykład

100 kg pewnej mieszanki zawiera 3 składniki:

składnik	A	B	C
(w) ilość (kg)	50	30	20
(x) cena (zł /kg)	15	20	30

$$\overline{x}_w = \sum_i^n (w_i \cdot x_i) / \sum_i^n w_i$$

Ile wynosi cena mieszanki za 1 kg?

Cena 1 kg mieszanki jest równa 1950/100=19.5 zł.

KWANTYLE

Rozważmy ciąg n niemalejących liczb (niektóre liczby w tym ciągu mogą się powtarzać)

$$x_1, x_2, \ldots, x_n$$

Kwantylem rzędu q nazywamy taką liczbę ξ_q, że na lewo od tej liczby znajduje się ok. q*100% danych, a na prawo około (1-q)*100% danych.

Kwantyl rzędu **0.15** znaczy, że na lewo od niego znajduje się ok. **15%** danych, a na prawo **85%** danych.

KWARTYLE

Kwantyl rzędu 0.25 (**dolny kwartyl - Q_1**), na lewo od niego znajduje się 25% danych, a na prawo 75% danych.

Mediana to kwantyl rzędu 0.50 (**drugi kwartyl - Q_2**) co znaczy, że dzieli dane na połowy (w uporządkowanej próbce jest to ta liczba od której około połowa danych jest nie większa i połowa nie mniejsza)

Kwantyl rzędu 0.75 (**górny kwartyl Q_3**).

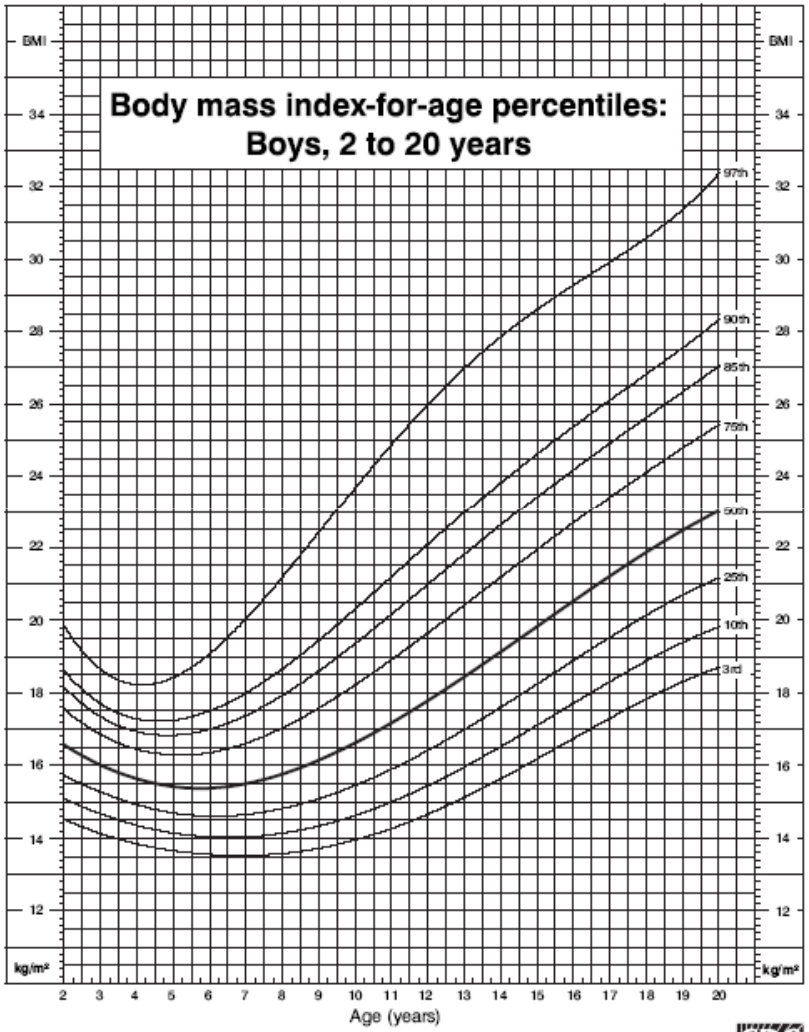
Medianą jest wartość środkowa w uporządkowanej próbce nieparzystej.

W uporządkowanej próbce parzystej medianą jest wartość średniej arytmetycznej z dwóch środkowych danych.

Inne nazwy kwantyli:
Decyle – podział danych na porcje 10%, czyli kwantyl rzędu 0.1 to 1 decyl
Centyle – podział danych na porcje 100%, czyli kwantyl rzędu 0.01 to 1 centyl

Ważne:
Kwantyl rzędu 0.05 – 5% danych jest na lewo od niego (jest nie większych) i 95% danych jest na prawo (jest nie mniejszych)

CDC Growth Charts: United States



Published May 30, 2000.
SOURCE: Developed by the National Center for Health Statistics in collaboration with the National Center for Chronic Disease Prevention and Health Promotion (2000).

SAFER · HEALTHIER · PEOPLE™

Przykład. Wartość sprzedaży (w tys. zł) w pewnej aptece w 10 kolejnych dniach wyniosła:

12.0, 10.5, 17.3, 21.1, 14.7, 18.0, 11.5, 12.7, 10.9, 9.3

$$\overline{x} = 13\ .8$$

Wyzaczamy kwartyle:
Uporządkujmy dane w kolejności rosnącej:

9.3, 10.5, 10.9, 11.5, 12.0, 12.7, 14.7, 17.3, 18.0, 21.1,

Liczba elementów mniejszych od 10.9 jest = 2
Liczba elementów mniejszych lub równych od 10.9 jest =3
2/10 < 0.25
3/10 >0.25
7/10<0.75
8/10>0.75
m=med= (12.0+12.7)/2=12.3

Rozpatrzymy ciąg 11 liczb powstały z poprzedniego przez dołączenie liczby 62:

9.3, 10.5, 10.9, 11.5, 12.0, 12.7, 14.7, 17.3, 18.0, 21.1

9.3, 10.5, 10.9, 11.5, 12.0, 12.7, 14.7, 17.3, 18.0, 21.1, 62.

Teraz średnia = 18.18 mediana= 12.7

Średnia jest wrażliwa na ekstremalne wartości danych (wyjątkowo duże lub małe).

Mediana jest „bardziej odporna”.

Moda (dominanta)

Definicja. Modą ciągu liczb $x_1, x_2, \ldots, x_n$ nazywamy taką wartość *m*, która powtarza się w tym ciągu najwięcej razy.

Przykład: 5 jest modą w następującym ciągu liczb:

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

Środek zakresu

Definicja. Środkiem zakresu ciągu liczb $x_1, x_2, \ldots, x_n$ nazywamy liczbę

$$\frac{x_{\max} + x_{\min}}{2}$$

Przykład: 3.5 jest środkiem zakresu

(6+1)/2=3.5

Miary położenia - podsumowanie

- Średnia (arytmetyczna lub ważona)
- Mediana
- Moda
- Środek zakresu

Każda z tych miar w inny sposób precyzuje

„wokół jakiej liczby dane się koncentrują”

Miary rozproszenia (rozrzutu danych)

- Wariancja
- Odchylenie standardowe
- Odchylenie przeciętne
- Rozstęp międzykwartylowy (IQR)
- Zakres danych

Wariancja

Wariancją danych $x_1, x_2, \cdots x_n$ nazywamy liczbę

$$S^2 = \frac{1}{n-1}[(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2]$$

Przykład. Obliczamy wariancję sprzedaży (w tys. zł) na podstawie danych z 10 kolejnych dni.

12.0, 10.5, 17.3, 21.1, 14.7, 18.0, 11.5, 12.7, 10.9, 9.3

Obliczmy wariancję:

$$S^2 = \frac{1}{10-1}[(12.0-13.8)^2 + (10.5-13.8)^2$$
$$+ (17.3-13.8)^2 + (21.1-13.8)^2$$
$$+ (14.7-13.8)^2 + (18.0-13.8)^2$$
$$+ (11.5-13.8)^2 + (12.7-13.8)^2$$
$$+ (10.9-13.8)^2 + (9.3-13.8)^2] = 14.81$$

Odchylenie standardowe

Odchyleniem standardowym nazywamy pierwiastek z wariancji

(Odchylenie standardowe wartości sprzedaży wyrażone jest w tych samych jednostkach co sprzedaż)

$$S = \sqrt{S^2}$$

Zauważmy, wariancja jest wyrażona w „jednostkach kwadratowych”
W naszym przykładzie wyniosła 14.81(tys. zł)^2
(sprzedaż podana jest w tys. zł)
Obliczmy odchylenie standardowe

$$S = \sqrt{14.81} = 3.85$$

Odchylenie standardowe jest wyrażone w tys zł., czyli jest równe 3850 zł.
Łatwiej jest interpretować odchylenie standardowe. Jest to mówiąc bardzo nieprecyzyjnie „typowa” wartość rozrzutu danych wokół średniej.
Odchylenie standardowe wyrażone jest w jednostkach badanej cechy

Odchylenie przeciętne

- Odchyleniem przeciętnym ciągu danych

$x_1, x_2, \cdots x_n$ nazywamy liczbę

$$D = \frac{1}{n}\left[\left|x_1 - m\right| + \left|x_2 - m\right| + \cdots + \left|x_n - m\right|\right]$$

$$m = med\,(x_1, x_2, \ldots . x_n)$$

Rozstęp międzykwartylowy

- Rozstępem międzykwartylowym nazywamy liczbę -

$$\xi_{0.75} - \xi_{0.25}$$

Inne oznaczenie kwartyli:

$$Q_3 - Q_1$$

trzeci kwartyl - pierwszy kwartyl

Przykład. Obliczmy rozstęp międzykwartylowy:

9.3, 10.5, 10.9, 11.5, 12.0, 12.7, 14.7, 17.3, 18.0, 21.1,

Rozstęp międzykwartylowy sprzedaży jest równy:

17.3 - 10.9 = 6.4

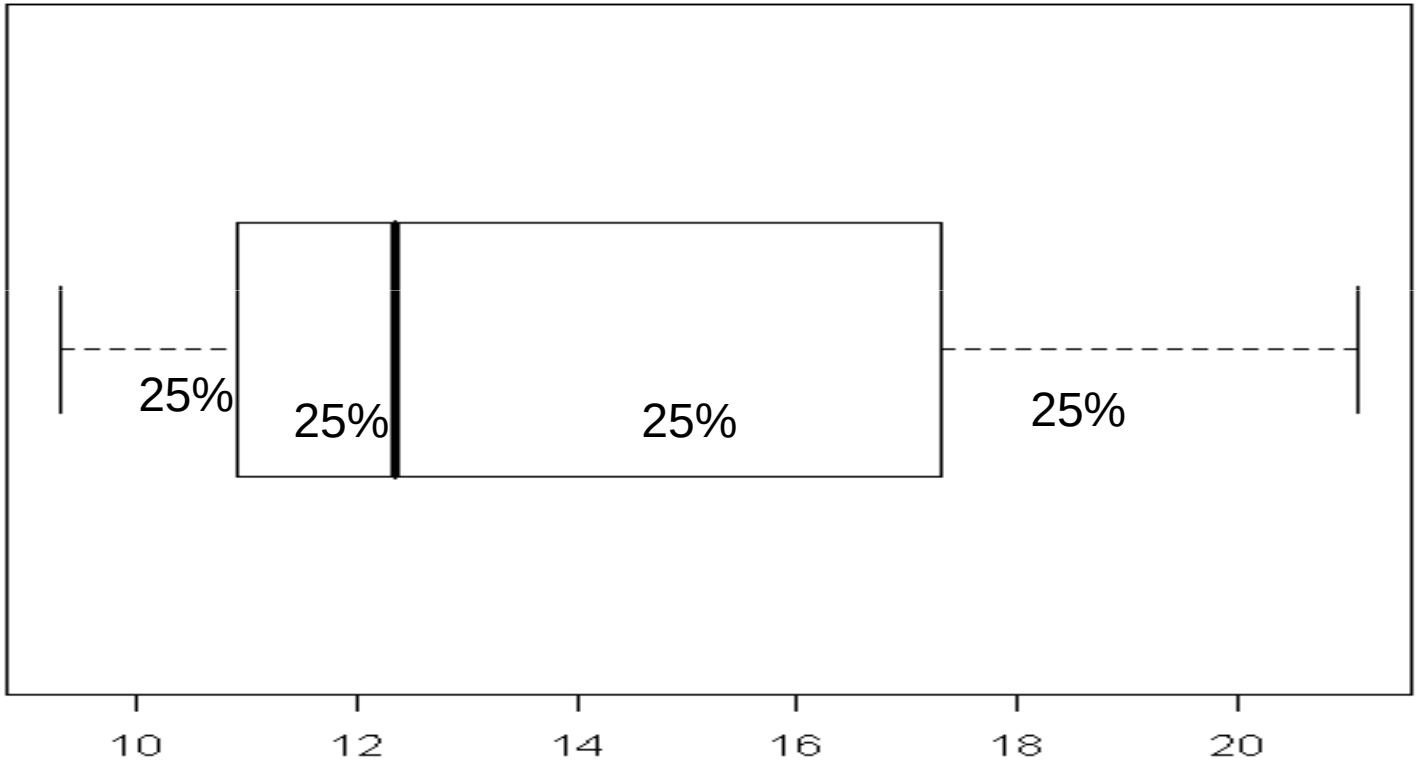
Zakres danych

- Zakres danych to liczba

$$x_{max} - x_{min}$$

Różnica między największą i najmniejszą wartością danych

Wykres „pudełkowy”.



min Q_1 med Q_3 max

Podsumowanie:

<u>komendy w programie R</u>

```
x=c(12.0,  10.5,  17.3,  21.1,  14.7,  18.0,
11.5, 12.7,  10.9,  9.3)
summary(x)
```

```
Min.    1st Qu.  Median    Mean  3rd Qu.
   Max.  9.30   11.05   12.35        13.80
   16.65     21.10
```

Ogólnie,

zarówno wybór miary położenia jak i rozproszenia zależy od tego jaką informację o danych chcemy przekazać

Własności miar położenia i rozproszenia

1. Jeżeli do wszystkich danych dodamy jakąś liczbę, to średnia wyliczona z danych zwiększy się o tą samą liczbę. Wariancja i odchylenie standardowe pozostaną takie same.
2. Jeżeli pomnożymy wszystkie nasze dane przez stałą a to średnia będzie równa a razy średnia, Wariancja zmieni się a^2 razy. Odchylenie standardowe zmieni się a razy tak jak średnia.

Tablica kontyngencji

Często dane w postaci „tablicy kontyngencji”, czyli „tablicy powtórzeń”. Ogólnie, taka tablica ma postać:

wartość cechy	x_1	x_2	…	x_k	razem
liczba jednostek	n_1	n_2	…	n_k	n

Zauważmy, że k oznacza liczbę możliwych wartości cechy zaś n liczbę jednostek
Oczywiście $n_1+n_2+\ldots+n_k = n$

Przykład. W grupie składającej się z 20 studentów, oceny ze statystyki były następujące:

2,3,3.5,4,4.5,4,5,3,3,3,3,4,3,3.5,3.5,2,4,3.5,3.5,5

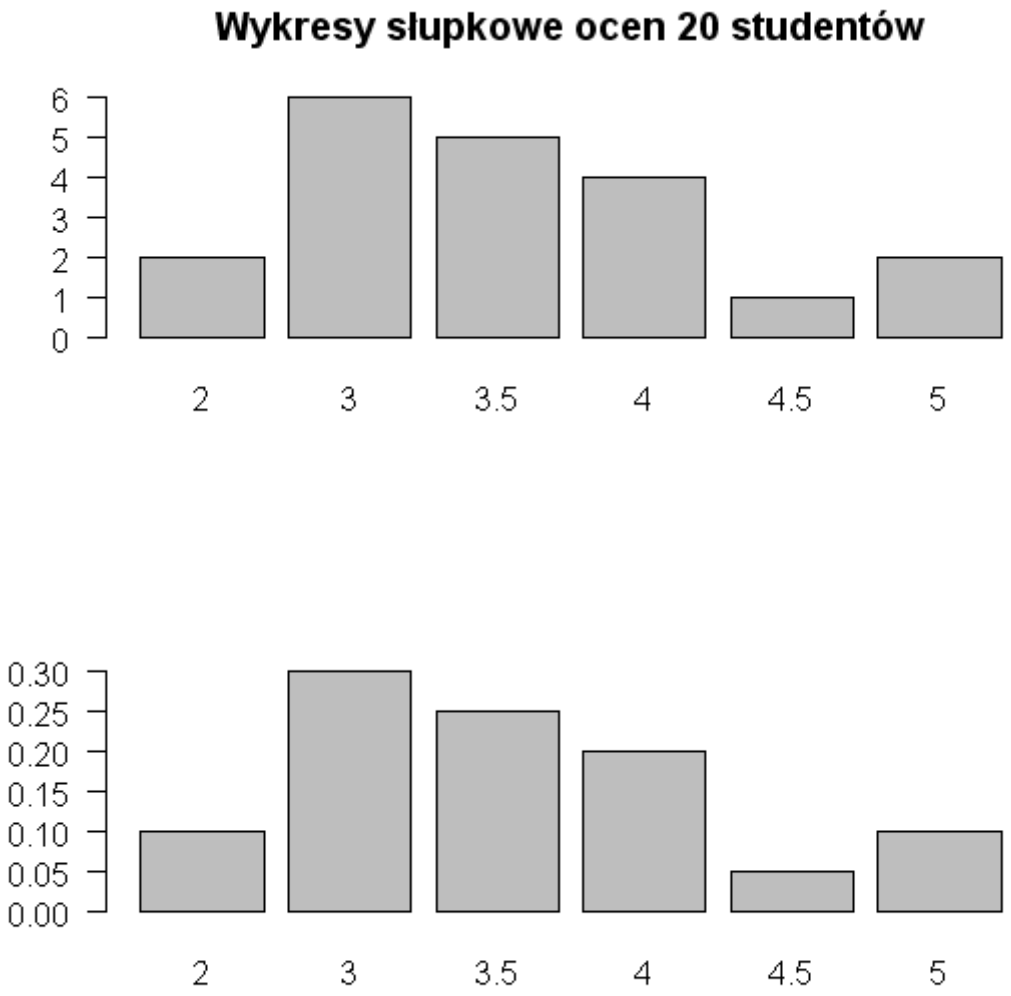
Dane można zapisać w skróconej postaci, notując ile razy powtórzyły się poszczególne wartości:

ocena	2	3	3.5	4	4.5	5	razem
liczba studentów	2	6	5	4	1	2	20

ocena	2	3	3.5	4	4.5	5	razem
liczba studentów	2	6	5	4	1	2	20

Możemy podać w podobnej tabeli odpowiednie ułamki (procenty) całkowitej liczby studentów (2/20)*100=10, (6/20)*100=30

ocena	2	3	3.5	4	4.5	5	razem
procent studentów	10	30	25	20	5	10	100


Wykresy słupkowe ocen 20 studentów
6
5
4
3
2
1
0
2
3
3.5
4
4.5
5
0.30
0.25
0.20
0.15
0.10
0.05
0.00
2
3
3.5
4
4.5
5

Częstość względna

$$w_i = \frac{n_i}{n} \qquad n = \sum n_i$$

- Średnia ważona na podstawie częstości w próbce:

$$\overline{x}_w = \frac{n_1 x_1 + n_2 x_2 + \cdots + n_k x_k}{n_1 + n_2 + \cdots + n_k} = \frac{n_1}{n} x_1 + \cdots + \frac{n_k}{n} x_k$$

n_i liczność i- tej grupy

Obliczenia na podstawie tablicy kontyngencji

$$\bar{x} = \frac{2 \cdot 2 + 6 \cdot 3 + 5 \cdot 3.5 + 4 \cdot 4 + 1 \cdot 4.5 + 2 \cdot 5}{20} = 3.5$$

Zauważmy, że średnia arytmetyczna wyjściowych 20 ocen jest tym samym, co średnia ważona 6 różnych możliwości ocen z wagami odpowiadającymi liczbie powtórzeń. Ten oczywisty fakt wyjaśnia dlaczego w statystyce często posługujemy się średnią ważoną

Przykład. Wariancja jest ważoną średnią kwadratów odchyleń od średniej:

$$S^2 = \frac{2}{20-1}(2-3.5)^2 + \frac{6}{20-1}(3-3.5)^2$$
$$+ \frac{5}{20-1}(3.5-3.5)^2 + \frac{4}{20-1}(4-3.5)^2$$
$$+ \frac{1}{20-1}(4.5-3.5)^2 + \frac{2}{20-1}(5-3.5)^2 = 0.658$$

Odchylenie standardowe:

$$S = \sqrt{0.658} = 0.81$$

ocena	2	3	3.5	4	4.5	5	razem
liczba studentów	2	6	5	4	1	2	20

Medianą ocen jest 3.5 bo liczba studentów o ocenie mniejszej niż 3.5 czyli 8 nie przekracza połowy,
zaś liczba studentów o ocenie mniejszej
lub równej 3.5, czyli 13 przekracza połowę.

Podsumujmy:

$$\bar{x}_w = \frac{n_1 x_1 + n_2 x_2 + \cdots + n_k x_k}{n_1 + n_2 + \cdots + n_k} = \frac{n_1}{n} x_1 + \cdots + \frac{n_k}{n} x_k$$

$$S_w^{\,2} = \frac{n_1}{n-1}(x_1 - \bar{x})^2 + \cdots + \frac{n_k}{n-1}(x_k - \bar{x})^2$$

$$S_w = \sqrt{S_w^{\,2}}$$

Szereg przedziałowy

cecha X	liczba jednostek
x_0-x_1	n_1
x_1-x_2	n_2
.....	...
x_{k-1}-x_k	n_k
razem	n

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i \dot{x}_i$$

$\dot{x}_i$ środek przedziału $\dot{x}_i = \frac{x_{i-1} + x_i}{2}$

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

suma po wszystkich przedziałach (tyle mamy jednostek)

Wzór na obliczanie wariancji:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{k} n_i (\dot{x}_i - \bar{x})^2$$

Przykład. Wielkość mieszkań w pewnym osiedlu (w m^2) zostały pogrupowane w przedziałach wielkości:

przedział wielkości	liczba mieszkań
(30,40]	**10**
(40,50]	**20**
(50,60]	**30**
(60,70]	**15**
(70,80]	**12**
(80,90]	**7**
(90,100]	**2**
(100,110]	**2**
(110,120]	**2**
Razem	**100**

Przykład c.d. Obliczmy średnią:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i \dot{x}_i$$

średnia=(10*35+20*45+30*55+15*65+12*75+7*85+2*95+2*105+2*115)/100= 60

Obliczmy wariancję:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{k} n_i (\dot{x}_i - \bar{x})^2$$

$$S^2 = \frac{1}{100-1}[10(35-60)^2 + 20(45-60)^2 + 30(55-60)^2 + 15(65-60)^2 + 12(75-60)^2 + 7(85-60)^2 + 2(95-60)^2 + 2(105-60)^2 + 2(115-60)^2]$$

Średnia, mediana i kwartyle

Średnia = 60

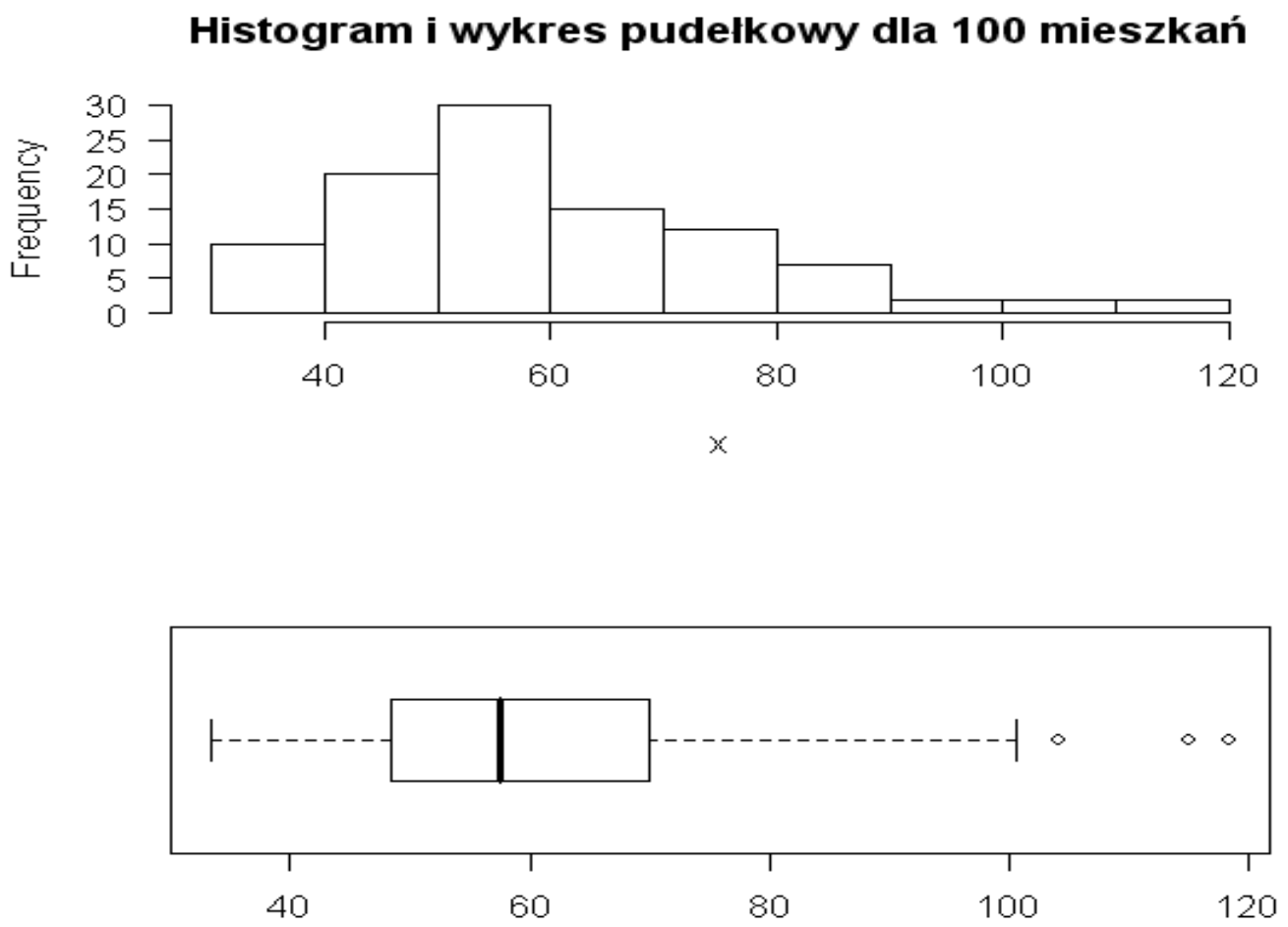
Wróćmy do tabelki

- Mediana z przedziału **(50,+60]**
- Q_1 z przedziału **(40,50]**
- Q_3 z przedziału **(60,70]**

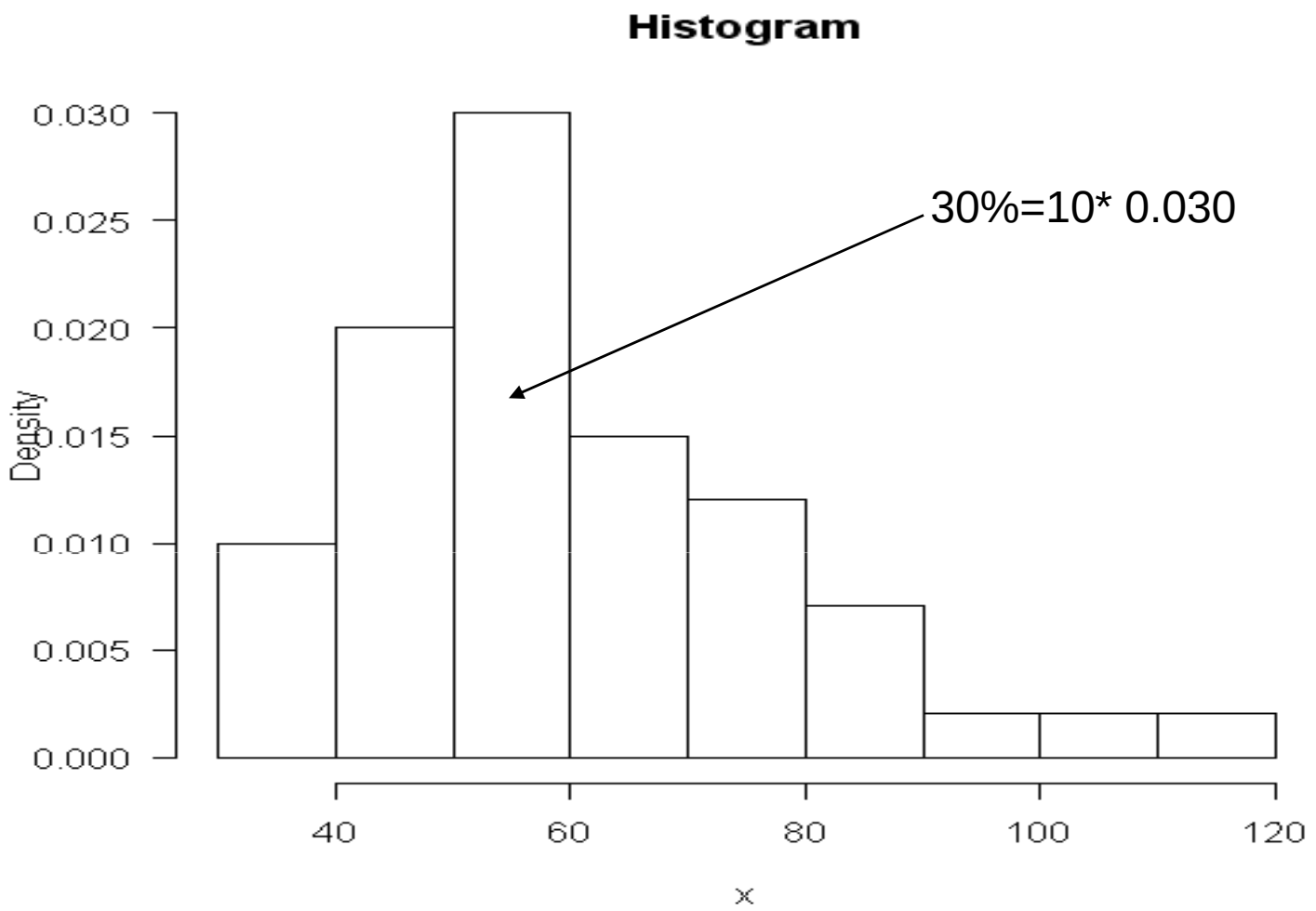
Pełne dane, na podstawie których powstała tabelka

```
 [1]  33.5  34.1  34.6  35.5  35.2  37.4  38.6  38.2  38.3  39.0  40.2  40.4
[13]  41.1  42.6  42.1  43.4  43.5  44.7  44.7  44.2  45.2  46.4  47.1  47.3
[25]  48.3  48.4  48.5  48.5  49.9  49.4  50.9  50.1  50.2  50.4  50.4  51.7
[37]  51.8  51.9  51.2  55.9  55.1  55.2  55.2  56.3  56.4  56.6  56.3  56.4
[49]  57.4  57.5  57.6  57.8  57.8  58.9  58.8  58.3  59.1  59.3  59.7  59.8
[61]  60.3  61.6  63.9  64.3  66.4  68.8  64.8  64.9  64.9  65.7  65.9  66.2
[73]  67.3  67.4  68.4  71.3  71.7  72.8  72.9  73.9  73.6  75.6  75.2  77.2
[85]  78.7  78.2  79.5  80.1  81.3  84.4  85.2  86.7  86.8  88.3  90.3  93.6
[97] 100.6 104.1 115.1 118.3
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
33.50	48.38	57.55	60.30	69.42	118.30



Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
33.46	48.39	57.59	60.32	69.40	118.30

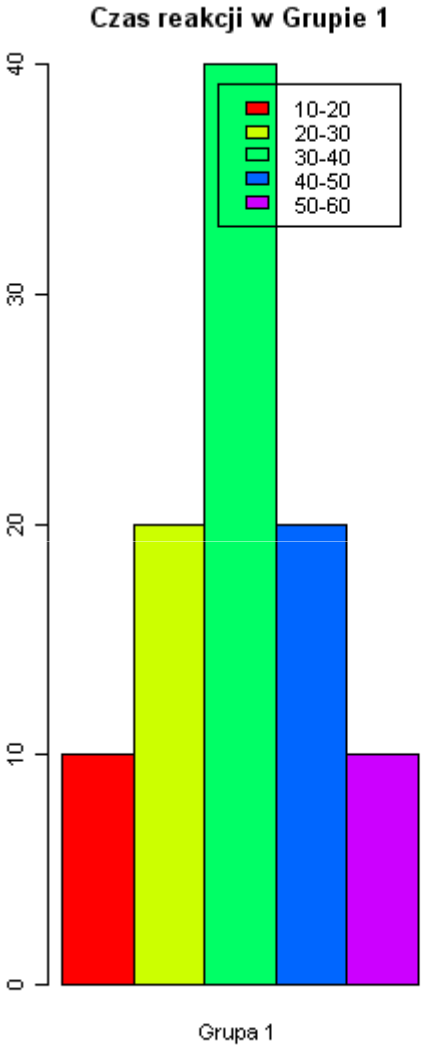


0.010	0.020	0.030	0.015	0.012	0.007	0.002	0.002	0.002
10	10	10	10	10	10	10	10	10

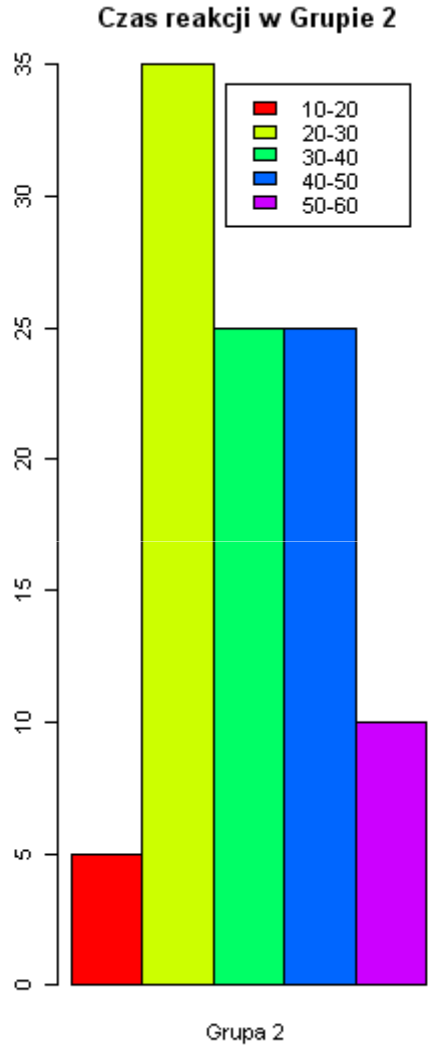
Podsumowanie

Postać i interpretacja szeregu przedziałowego są podobne jak dla tablicy kontyngencji. Zwróćmy uwagę na istotną różnicę. Podsumowując dane w postaci szeregu przedziałowego tracimy część informacji.
Z tabelki nie możemy się dowiedzieć na przykład, ile jest mieszkań o metrażu 30-35 (na podstawie pełnych danych wiemy, że jest ich 3). Na podstawie tej tabelki nie możemy dokładnie obliczyć średniej oryginalnych danych.

Porównanie wielu próbek

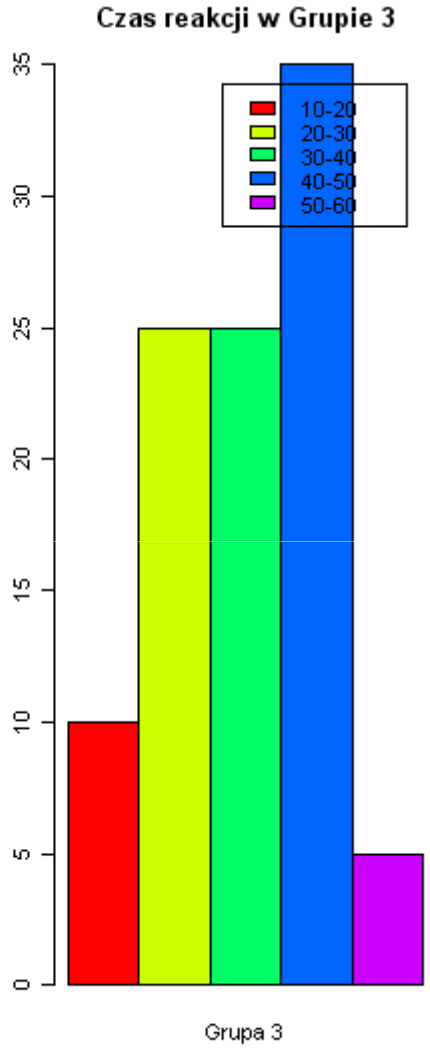


Moda=średniej



Prawoskośny

Moda<średniej



Lewoskośny

Moda>średniej

Porównanie wielu próbek

- Współczynnik zmienności:

$$V = \frac{S}{\overline{x}} \cdot 100\%$$

Miara asymetrii (skośności)

- Współczynnik asymetrii

$$A_1 = \frac{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^3}{n}}{S^3}$$

Pakiety używane w analizie danych statystycznych

- Excel: funkcje statystyczne oraz moduł Analiza Danych
- Pakiety statystyczne: SAS, SPSS, Stata, Statgraphics, Statistica, S+, R