

Statystyka Matematyczna

Anna Janicka

wykład I, 22.02.2016

STATYSTYKA OPISOWA, cz. I

Kwestie techniczne

- Kontakt: ajanicka@wne.uw.edu.pl
- Dyżur:
- strona z materiałami z przedmiotu:
wne.uw.edu.pl/**azylicz**
akson.sgh.waw.pl/~aborata
- Podręczniki: Niemiro, Kornacki & Mielniczuk, Jóźwiak & Podgórski
- zbiory zadań: **Boratyńska**

+info w sylabusie



WARSAW UNIVERSITY
Faculty of Economic Sciences

Reguły zaliczenia

1. Dowolny z wykładów. Obecność **ZALECANA**. Za obecności (3/4 sprawdzonych) dodatkowy punkt na egzaminie.
2. na egzaminie zakres materiału **z wykładu + ćw.**
3. Ćwiczenia **OBOWIĄZKOWE** (max. 3 nieobecności)
4. Zaliczenie ćwiczeń: połowa punktów z 2 kolokwiów i 3 niezapowiedzianych kartkówek.
5. Ocena z ćwiczeń: punkty + ew. aktywność (w gestii ćwiczeniowca)
6. Do egzaminu dopuszczeni wszyscy, którzy **chodzili na ćw.**
7. Egzamin: 8 zadań, każde po 2 pkt. Ocena z egzaminu = liczba zdobytych punktów (tu punkt za obecność) / 3
8. Ocena końcowa = $\max\{\text{ocena z egzaminu}, 1/3 * \text{ocena z ćwiczeń} + 2/3 * \text{ocena z egzaminu}\}$, zaokrąglona. Przy czym osoba mająca 2 z ćwiczeń musi mieć ≥ 9 pkt z egz.
9. Gdy w I terminie ≥ 7 pkt z egz., zaliczone ćwiczenia.



Czego można się spodziewać

- ☐ Materiały do wykładu, przykłady etc. na stronie www
- ☐ Pod koniec semestru info, jaki zakres materiału obowiązuje na egzaminie



Tematyka zajęć – plan semestru

- ☐ *Indeksy*
- ☐ Statystyka opisowa
- ☐ Model statystyczny, wnioskowanie statystyczne, pojęcie statystyki
- ☐ Estymacja, własności estymatorów
- ☐ Weryfikacja hipotez statystycznych, różne rodzaje testów
- ☐ Estymacja Bayesowska



Plan na dzisiaj

1. Wstęp

2. Statystyka opisowa:

- podstawowe pojęcia
- prezentacja danych
- charakterystyki próbkowe
- miary
 - położenia



Czym się różni Statystyka od Statystyki Matematycznej?

Statystyka: nauka, której przedmiotem jest pozyskiwanie i analizowanie danych dotyczących zjawisk *masowych*

historycznie: już starożytność, potem różne spisy, opis stanu państwa

Statystyka Matematyczna: statystyka z matematycznego punktu widzenia, tj. dział matematyki stosowanej w którym opisuje się zjawiska przy pomocy narzędzi m.in. rachunku prawdopodobieństwa

historycznie: wraz z początkami rachunku pstwa: Pascal,

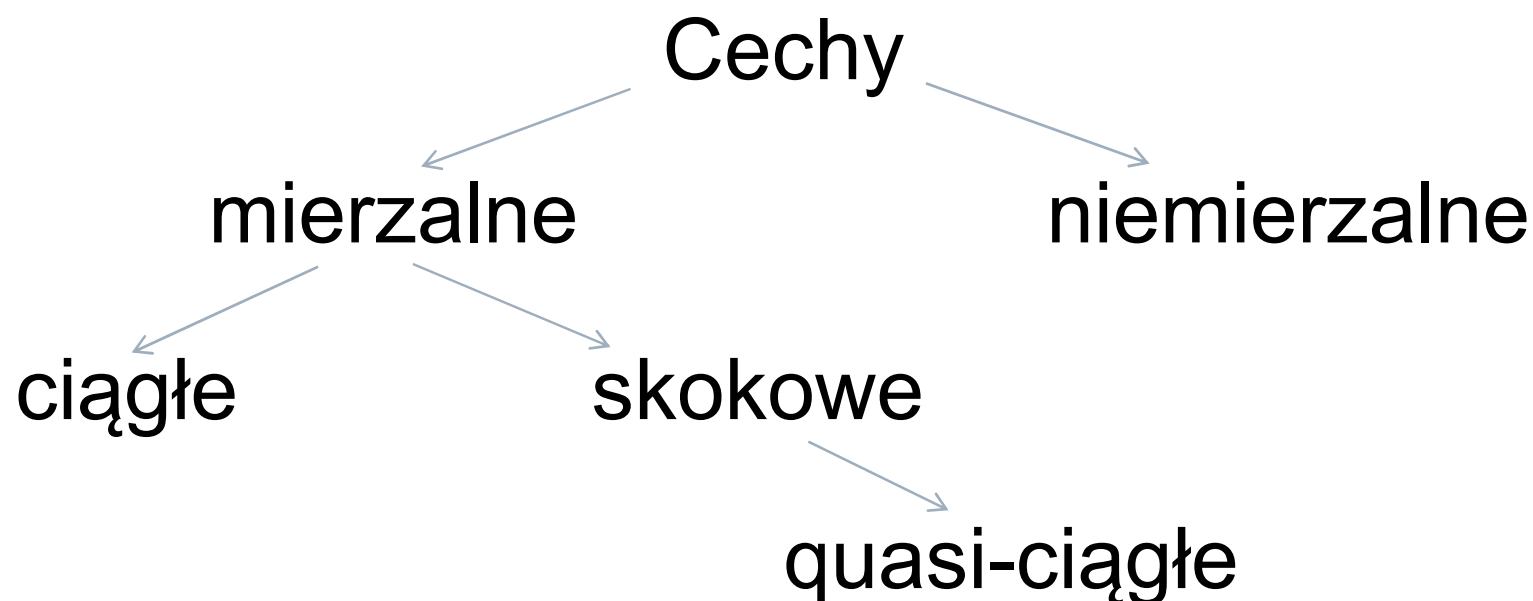
Fermat, Gauss



Statystyka opisowa

Opis danych w kategoriach ilościowych.

Dane dotyczą **próby** z pewnej **populacji**,
badanych pod kątem **cechy** (lub cech)



Badanie

- ❑ **pełne** – dotyczy całej populacji
- ❑ **reprezentacyjne** – dotyczy wybranej części populacji; próba $\neq$ populacja
w takim przypadku wnioskowanie wymaga założeń oraz wykorzystania narzędzi rachunku prawdopodobieństwa



Prezentacja danych

- ☐ Cel: jak największa przejrzystość
- ☐ Zależy od rodzaju cechy

- ☐ tabelaryczna
- ☐ graficzna



Przykład 1 – cecha skokowa

Oceny z egzaminu z RP w roku 2015/2016
(168 osób piszących w I terminie)

3,5 3 3 2 2 3 2 3 2 2 2 3 2 3,5 3 3,5 4 2 2 2 3,5 2 3
3 2 2 3 3 2 4,5 4,5 3 2 2 2 2 2 4,5 3 3,5 3 2 4 2
4,5 3 2 3,5 4 2 3,5 2 3 2 3,5 2 3 3,5 3,5 2 2 4 3 2
3,5 3 2 3 3,5 2 2 2 2 3 3,5 3,5 2 3,5 3,5 2 4 3,5 3 3
2 2 3 5 3,5 2 3,5 3,5 4,5 3 3,5 5 5 3 4 2 3 2 2 4,5 2
2 3 2 3 3 2 3 2 3,5 2 3 2 3 2 2 3,5 3 4 3 2 2 2 2
3,5 4 2 3,5 4 2 3,5 2 3,5 2 2 2 3 5 2 2 2 3 4,5 4 3
3,5 3,5 3 2 4 2 3 3,5 3 3,5 2 3 3,5 2 2 3 2



Tablica kontyngencji

Szereg rozdzielczy punktowy

<i>Wartość cechy</i>	<i>Liczebność</i>	<i>Częstość</i>
x_1	n_1	$f_1 = n_1/n$
x_2	n_2	$f_2 = n_2/n$
x_3	n_3	$f_3 = n_3/n$
$\dots$	$\dots$	$\dots$
x_k	n_k	$f_k = n_k/n$
Razem	n	1



Przykład 1 – cd.

<i>Ocena</i>	<i>Liczebność</i>	<i>Częstość</i>
2	72	0,429
3	42	0,250
3,5	32	0,190
4	11	0,065
4,5	7	0,042
5	4	0,024
Razem	168	1,000

[Średnia arytmetyczna – przykłady](#)

[Mediana – przykłady](#)

[Moda – przykłady](#)

[Kwartyle – przykłady](#)

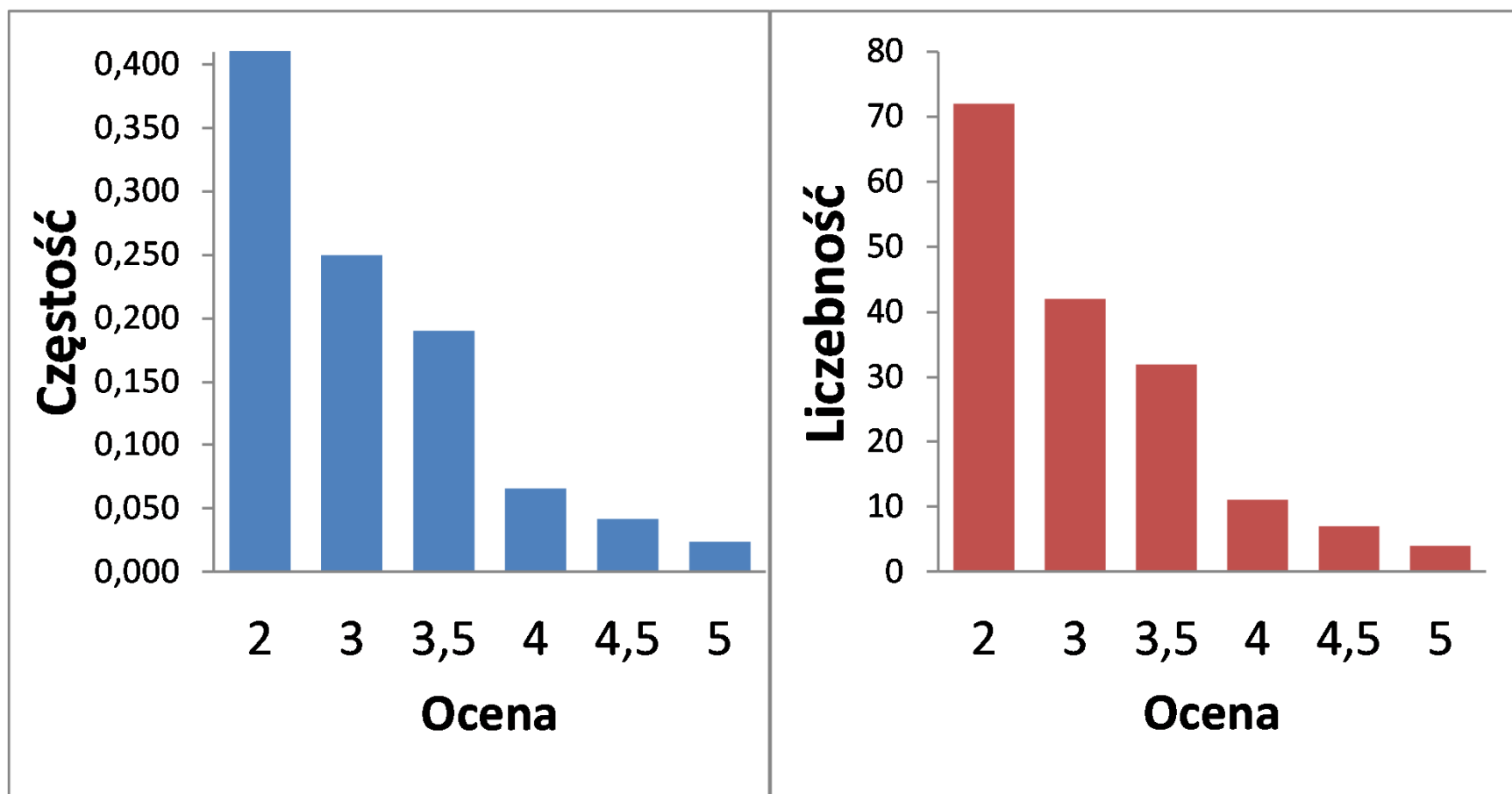
[Wariancja – przykłady](#)



WARSAW UNIVERSITY
Faculty of Economic Sciences

Przykład 1 – cd. (2).

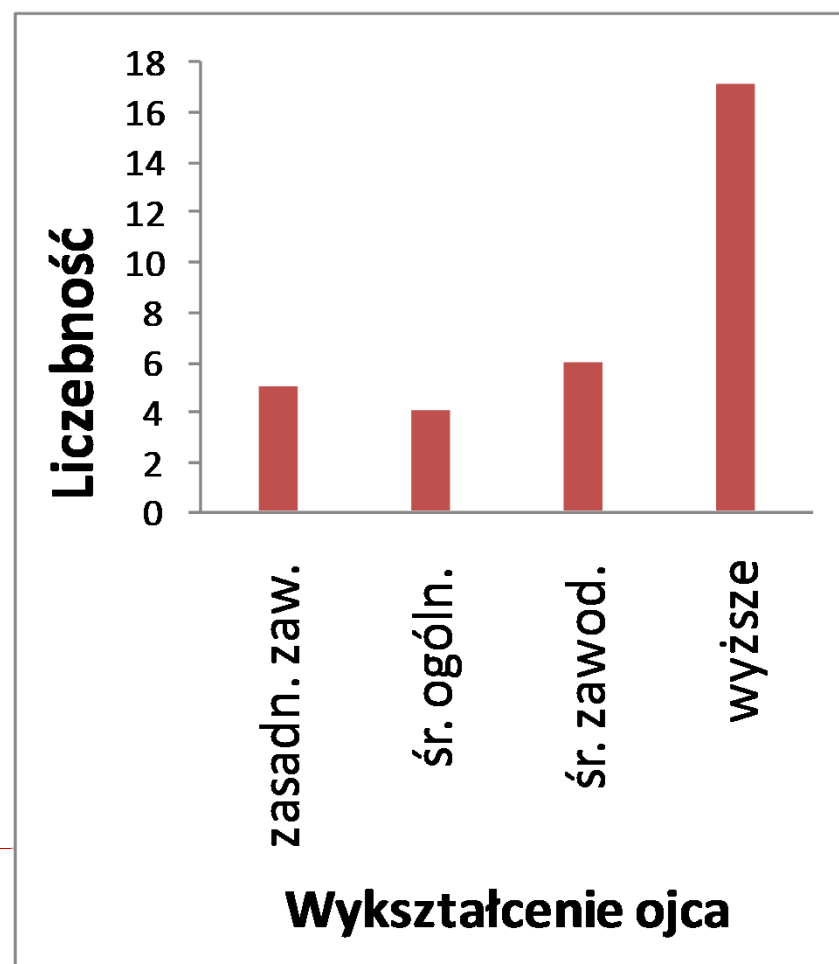
Wykresy słupkowe liczebności i częstości



Przykład 2 – cecha niemierzalna

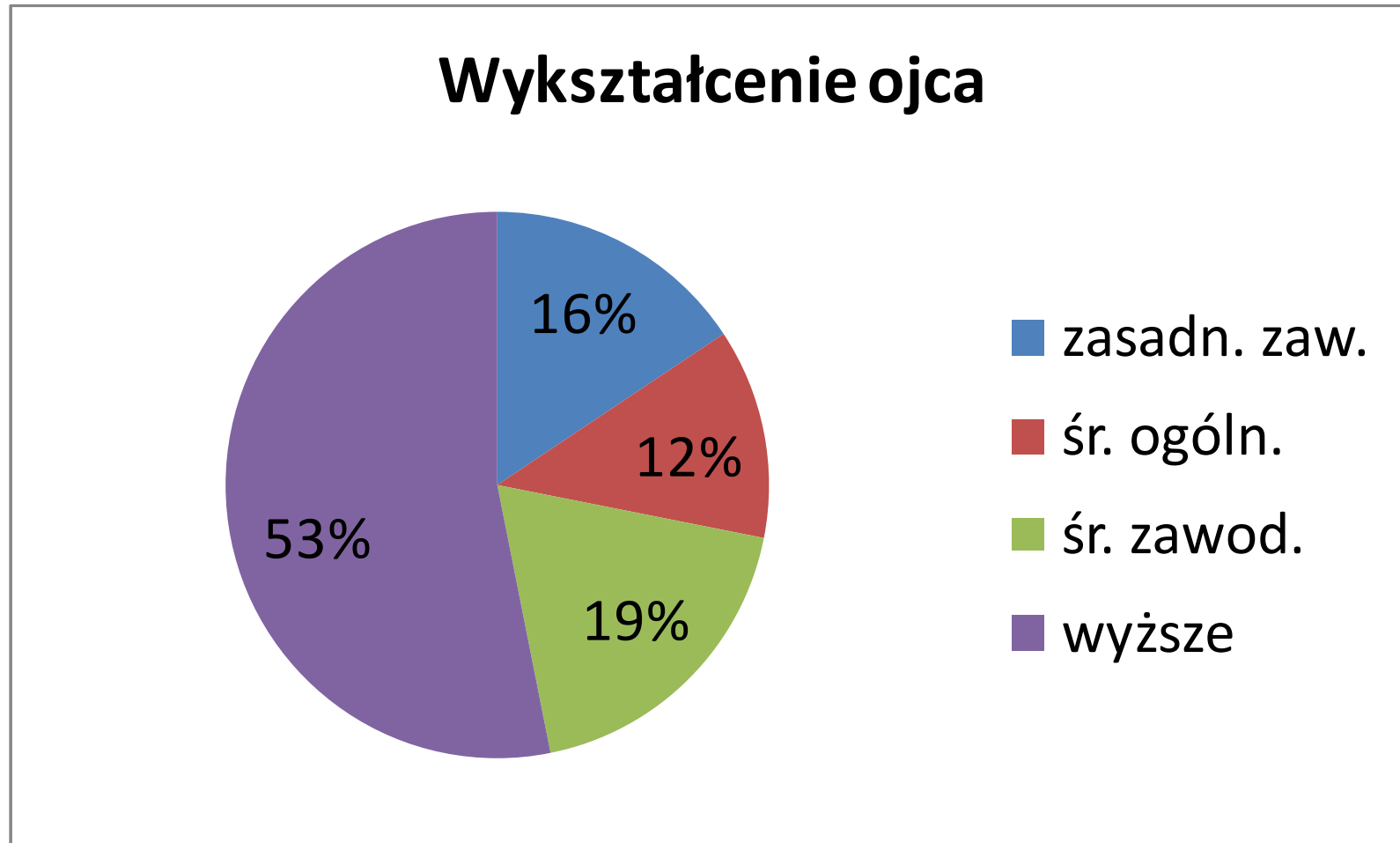
Wykształcenie ojca dla próby 32 studentów
III roku WNE

Wykształcenie ojca	Liczebność	Częstość
zasadnicze zawodowe	5	0,16
średnie ogólnokształcące	4	0,13
średnie zawodowe	6	0,19
wyższe	17	0,53
Razem	32	1,00



Przykład 2 – cd.

Wykres kołowy



Przykład 3 – cecha ciągła lub quasi-ciągła

Powierzchnia mieszkań na pewnym osiedlu, $n=100$

32,45	33,21	34,36	35,78	37,79	38,54	38,91	38,96	39,5	39,67
39,8	41,45	41,55	42,27	42,4	42,45	44,25	44,5	44,7	44,83
44,9	45,1	45,9	46,52	47,65	48,1	48,55	48,9	49	49,24
49,55	49,65	49,7	49,9	50,9	51,4	51,5	51,65	51,7	51,8
51,98	52	52,1	52,3	53,65	53,89	53,9	54	54,1	55,2
55,3	55,56	55,62	56	56,7	56,8	56,9	56,95	57,13	57,45
57,7	57,9	58	58,5	58,67	58,8	59,23	63,4	63,7	64,2
64,3	64,6	65	66,29	66,78	67,8	68,9	69	69,5	73,2
76,8	77,1	77,8	78,9	79,5	82,7	83,4	84,5	84,9	85
86	89,1	89,6	93	96,7	98,78	103	107,9	112,7	118,9



Szereg rozdzielczy przedziałowy

<i>Przedział</i>	<i>Środek przedziału</i>	<i>Liczebność</i>	<i>Częstość</i>	<i>Liczebność skumulowana</i> cn_i	<i>Częstość skumulowana</i> cf_i
$(c_0, c_1]$	$\bar{c}_1$	n_1	$f_1 = n_1/n$	n_1	f_1
$(c_1, c_2]$	$\bar{c}_2$	n_2	$f_2 = n_2/n$	$n_1 + n_2$	$f_1 + f_2$
$(c_2, c_3]$	$\bar{c}_3$	n_3	$f_3 = n_3/n$	$n_1 + n_2 + n_3$	$f_1 + f_2 + f_3$
...		...	...		
$(c_{k-1}, c_k]$	$\bar{c}_k$	n_k	$f_k = n_k/n$	$\sum n_i = n$	$\sum f_i = 1$
Razem		n	1		

Dobór klas (granic przedziałów): najczęściej jednakowej długości lub zbliżonej liczności

Przykład 3 – cd.

<i>Przedział</i>	<i>Środek przedziału</i>	<i>Liczebność</i>	<i>Częstość</i>	<i>Liczebność skumulowana</i>	<i>Częstość skumulowana</i>
(30,40]	35	11	0,11	11	0,11
(40,50]	45	23	0,23	34	0,34
(50,60]	55	33	0,33	67	0,67
(60,70]	65	12	0,12	79	0,79
(70,80]	75	6	0,06	85	0,85
(80,90]	85	8	0,08	93	0,93
(90,100]	95	3	0,03	96	0,96
(100,110]	105	2	0,02	98	0,98
(110,120]	115	2	0,02	100	1
Razem		100	1		

[Średnia arytmetyczna – przykłady](#)

[Mediana – przykłady](#)

[Moda – przykłady](#)

[Kwartyle – przykłady](#)

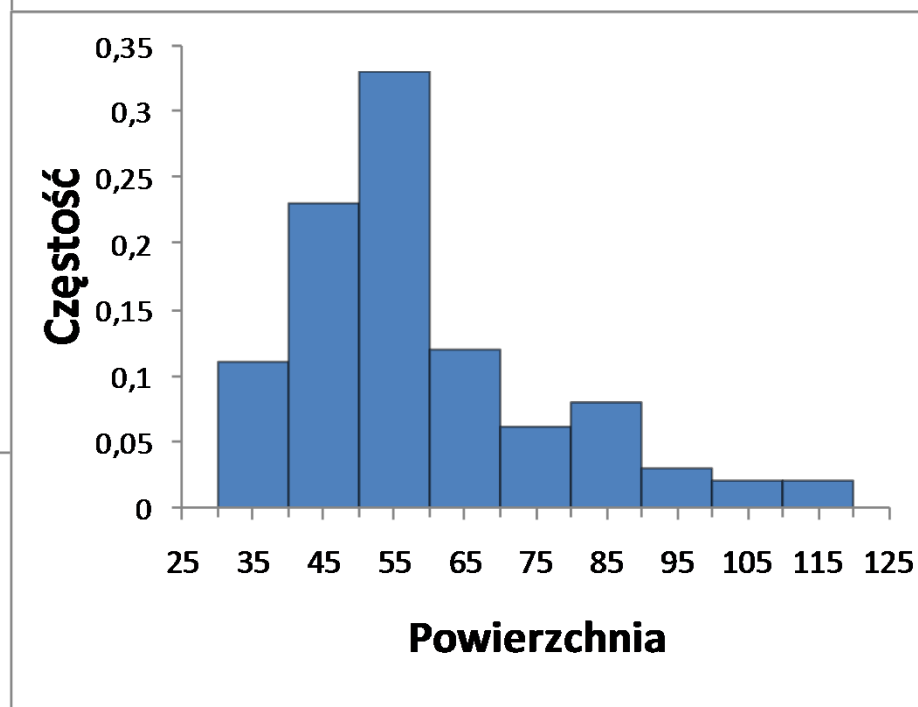
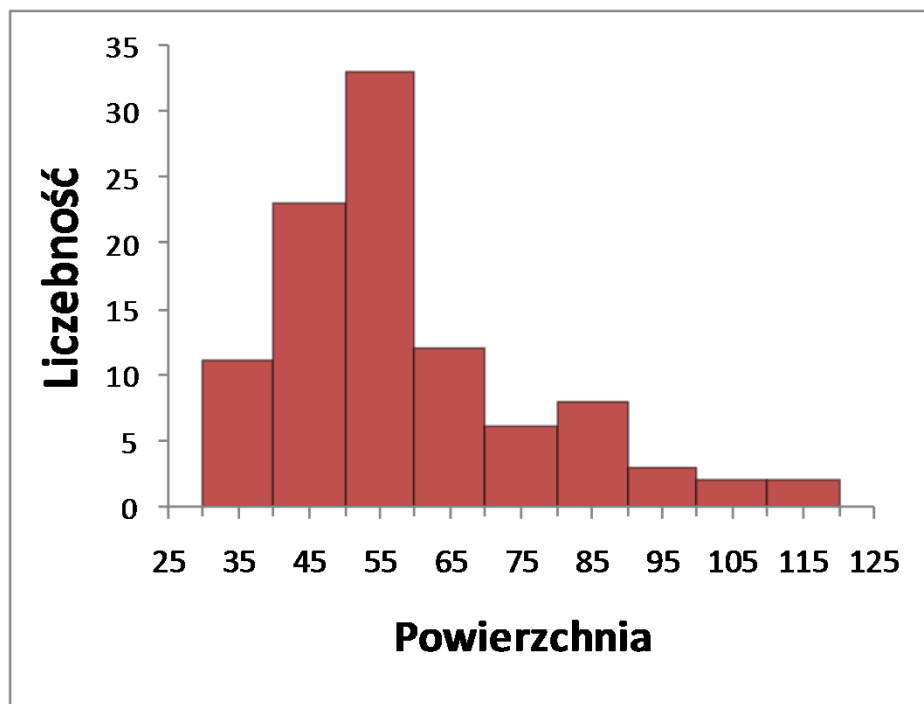
[Wariancja – przykłady](#)



WARSAW UNIVERSITY
Faculty of Economic Sciences

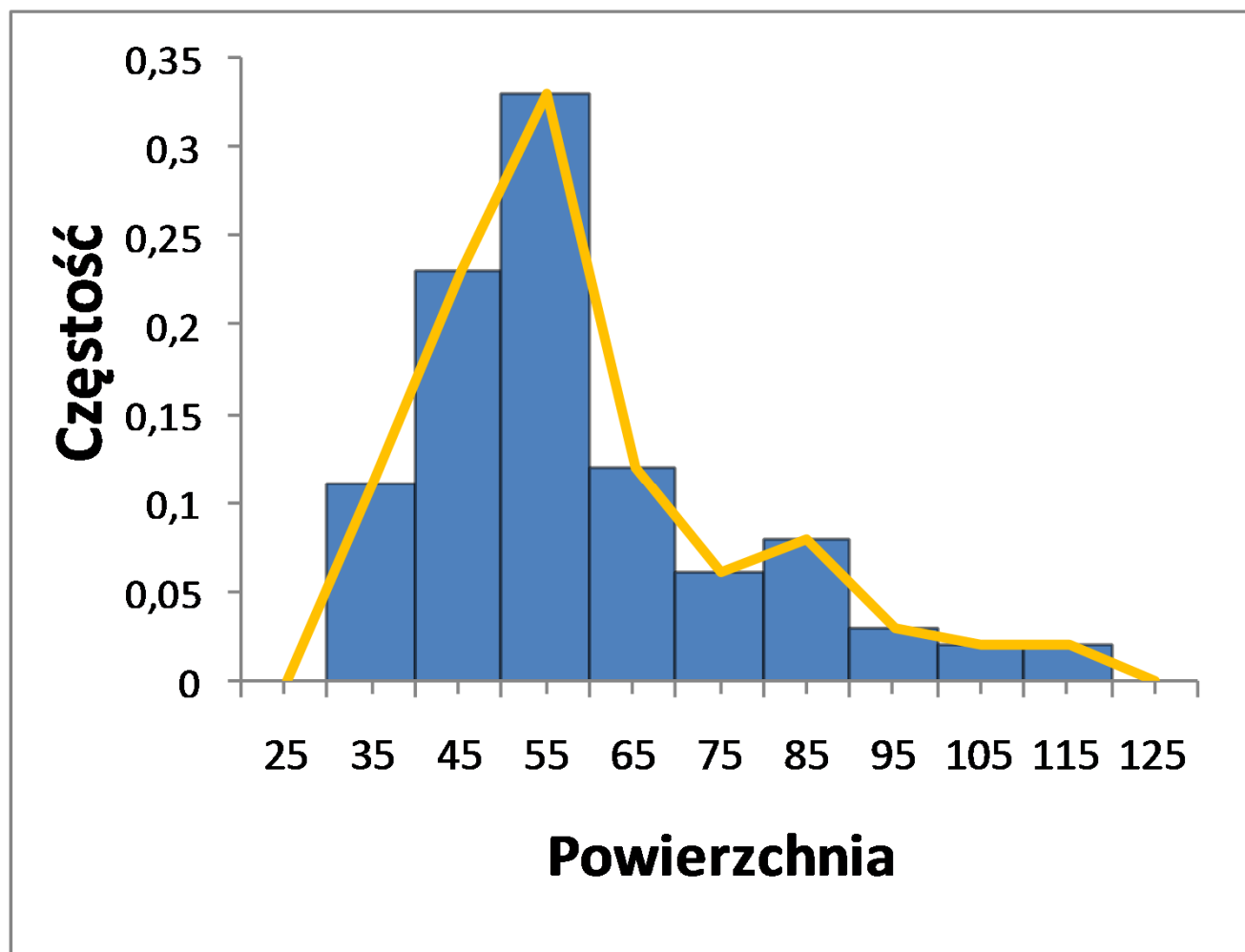
Przykład 3 – cd. (2)

Histogram liczebności i histogram częstości

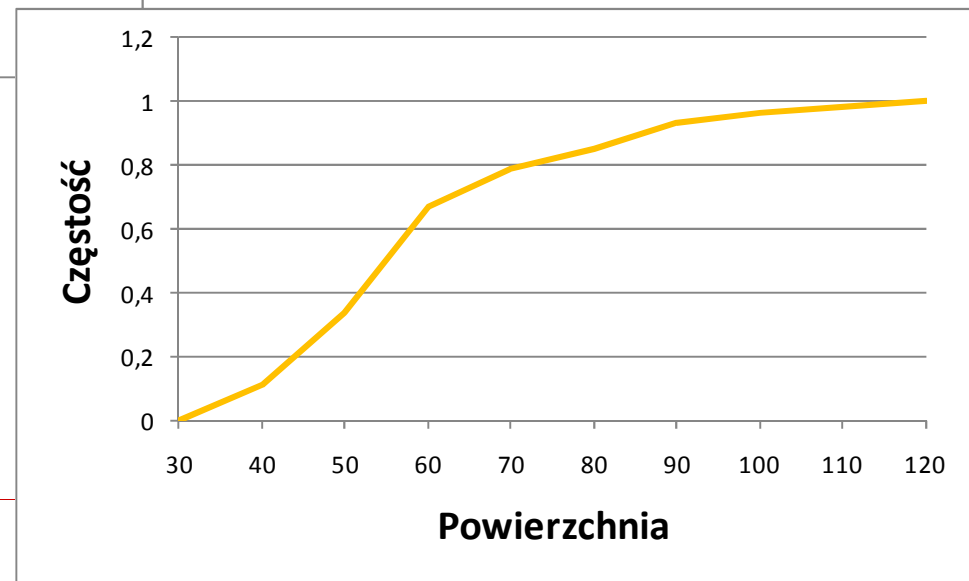
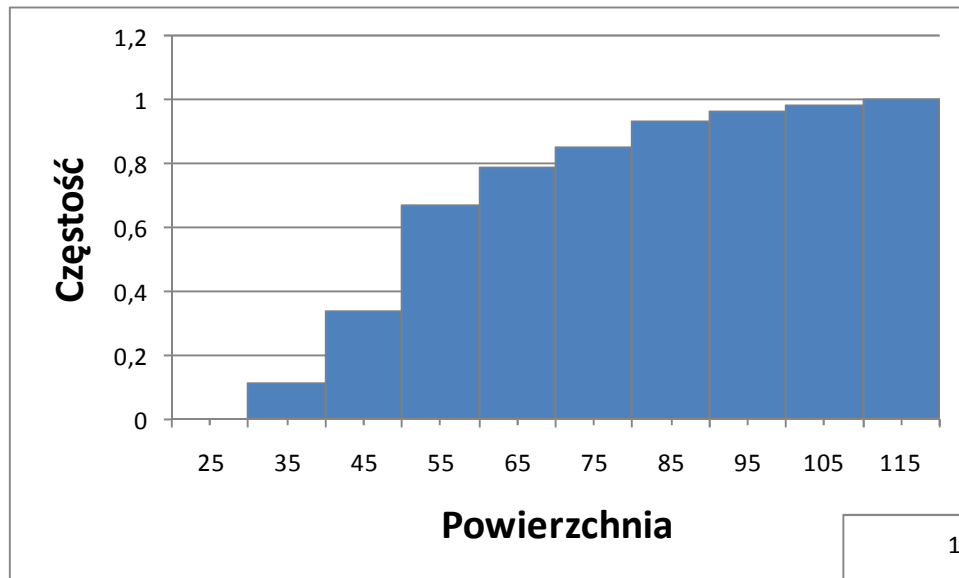


Przykład 3 – cd. (3)

Histogram częstości i łamana częstości

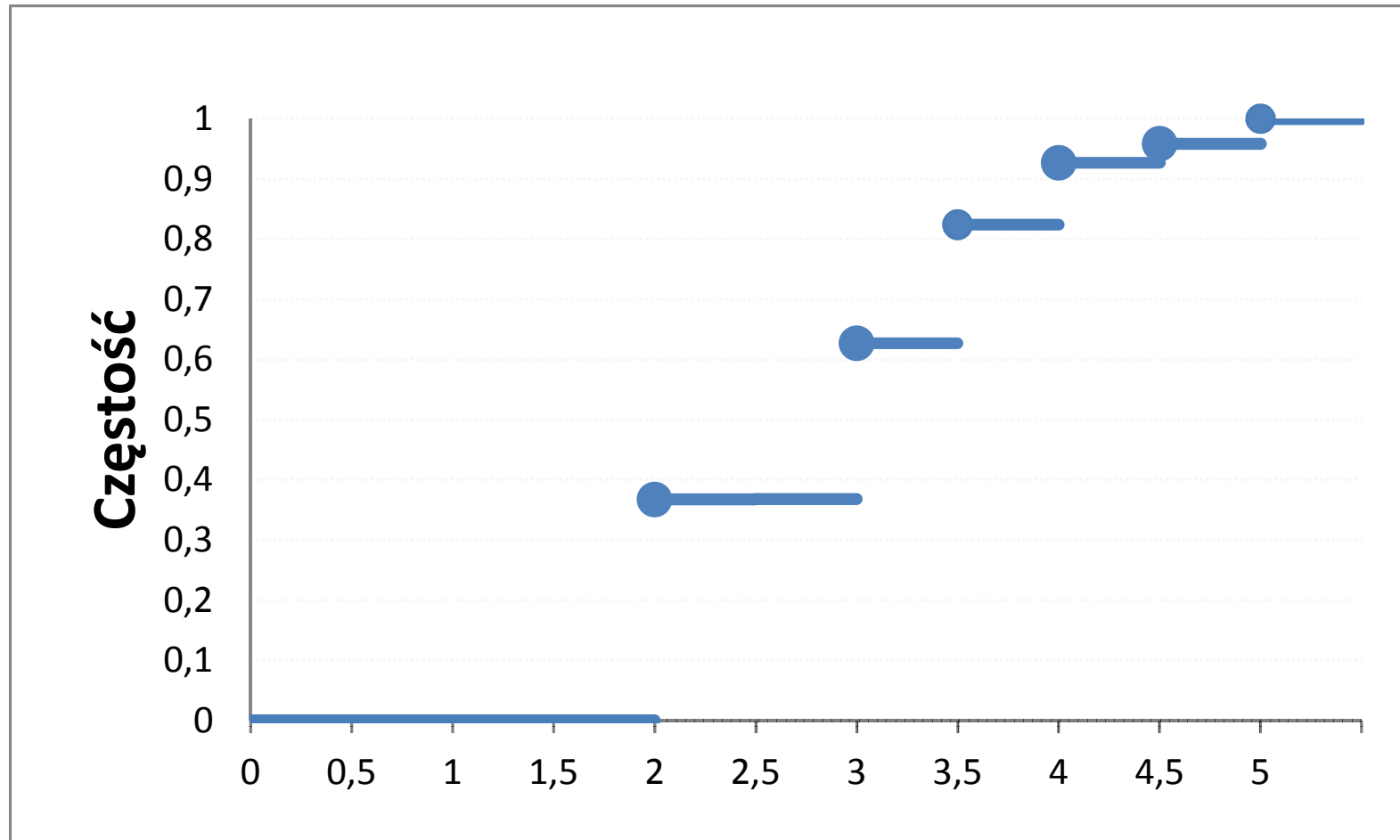


Przykład 3 – cd. (4) Histogram częstości skumulowanej i łamana częstości skumulowanej



Przykład 1 – cd. (3)

Dystrybuanta empiryczna



Charakterystyki próbkowe

Opisują różne własności rozkładów mierzalnych
Miary

- położenia
- zróżnicowania (zmienności, rozproszenia, dyspersji)
- asymetrii
- koncentracji

Typy:

- klasyczne (oparte o momenty)
- pozycyjne



Miary położenia

☐ Klasyczne:

- średnia arytmetyczna

☐ Pozycyjne:

- mediana
- moda (dominanta)
- kwantyle



Średnia arytmetyczna

□ dane surowe:
$$\bar{X} = \frac{1}{n} \sum_{i=1}^n x_i$$

□ dane z szeregu rozdzielczego punktowego:
$$\bar{X} = \frac{1}{n} \sum_{i=1}^k x_i \cdot n_i$$

□ dane z szeregu rozdzielczego przedziałowego:
$$\bar{X} \approx \frac{1}{n} \sum_{i=1}^k \bar{c}_i \cdot n_i$$



Średnia arytmetyczna – przykłady

[Przykład 1 – cd.](#)

Przykład 1:

$$\bar{X} = \frac{2 \cdot 72 + 3 \cdot 42 + 3,5 \cdot 32 + 4 \cdot 11 + 4,5 \cdot 7 + 5 \cdot 4}{168} \approx 2,842$$

Przykład 3:

[Przykład 3 – cd.](#)

$$\bar{X} \cong$$

$$\cong \frac{35 \cdot 11 + 45 \cdot 23 + 55 \cdot 33 + 65 \cdot 12 + 75 \cdot 6 + 85 \cdot 8 + 95 \cdot 3 + 105 \cdot 2 + 115 \cdot 2}{100}$$

$$= 58,7$$

a naprawdę: $\bar{X} = 59,58$

tylko jeśli nie ma
danych surowych!



WARSAW UNIVERSITY
Faculty of Economic Sciences

Mediana

Mediana (wartość środkowa)

taka wartość, że co najmniej 50% wartości w próbie nie mniejsze i co najmniej 50% wartości w próbie nie większe od niej

□ dane surowe, szereg rozdzielczy punktowy:

$$Med = \begin{cases} X_{\frac{n+1}{2}:n} & n \text{ nieparzyste} \\ \frac{1}{2} (X_{\frac{n}{2}:n} + X_{\frac{n}{2}+1:n}) & n \text{ parzyste} \end{cases}$$

gdzie $X_{i:n}$ jest ***i*-tą statystyką pozycyjną**, tj. *i*-tą wartością uszeregowanej rosnąco próby



Mediana – cd.

□ dla szeregu rozdzielczego przedziałowego:

$$Med \cong c_L + \frac{b}{n_M} \left(\frac{n}{2} - \sum_{i=1}^{M-1} n_i \right)$$

gdzie:

M – numer klasy mediany

c_L – dolny koniec klasy mediany

b – rozpiętość (długość) klasy mediany



Mediana – przykłady

[Przykład 1 – cd.](#)

Przykład 1:

$$Med = \frac{1}{2} X_{84:168} + \frac{1}{2} X_{85:168} = 3$$

Przykład 3:

[Przykład 3 – cd.](#)

$$M=3, \quad n_3=33, \quad c_L=50, \quad b=10$$

$$Med \cong 50 + \frac{10}{33} (50 - 34) \approx 54,85$$

a naprawdę: $Med = 55,25$



Moda

Moda (dominanta, wartość modalna)

wartość najczęściej pojawiająca się w próbie

□ dla szeregu rozdzielczego punktowego:

M_o = wartość najczęstsza

□ dla szeregu rozdzielczego przedziałowego

$$M_o \cong c_L + \frac{n_{M_o} - n_{M_o-1}}{(n_{M_o} - n_{M_o-1}) + (n_{M_o} - n_{M_o+1})} \cdot b$$

gdzie

n_{M_o} – liczebność klasy dominanty,

c_L, b – dla dominanty analogicznie do mediany



Moda – przykłady

[Przykład 1 – cd.](#)

Przykład 1:

$$Mo = 2$$

[Przykład 3 – cd.](#)

Przykład 3:

przedział dominanty to $(50,60]$, o liczebności 33

$$n_{Mo} = 33, c_L = 50, b = 10, n_{Mo-1} = 23, n_{Mo+1} = 12$$

$$Mo \cong 50 + \frac{33 - 23}{(33 - 23) + (33 - 12)} \cdot 10 \approx 53,23$$



Którą miarę stosować?

- ☐ Średnia arytmetyczna: do szeregów typowych (jedno max, częstości monotoniczne)
- ☐ Dominanta: do szeregów typowych, danych pogrupowanych, długości przedziału dominanty i sąsiednich powinny być równe
- ☐ Mediana: nie ma ograniczeń. Najbardziej odporna na zaburzenia, niedokładności pomiaru, zmiany, wartości odstające



Kwantyle, kwartyle

- p -ty kwantyl (kwantyl rzędu p): odsetek wartości nie większych niż on wynosi co najmniej p , a wartości nie mniejszych co najmniej $1-p$
- Q_1 : Pierwszy kwartyl = kwantyl rzędu $\frac{1}{4}$
- Drugi kwartyl = mediana = kwantyl rzędu $\frac{1}{2}$
- Q_3 : Trzeci kwartyl = kwantyl rzędu $\frac{3}{4}$



Kwantyle – cd.

Kwantyl próbkowy rzędu p :

$$Q_p = \begin{cases} \frac{X_{np:n} + X_{np+1:n}}{2} & np \in Z \\ X_{[np]+1:n} & np \notin Z \end{cases}$$



Kwartyle – cd.

- Kwantyle dla $p = 1/4$ i $p = 3/4$.
- Dla szeregu rozdzielczego przedziałowego: wzór jak dla mediany

$$Q_k \cong c_L + \frac{b}{n_{M_k}} \left(\frac{k \cdot n}{4} - \sum_{i=1}^{M_k-1} n_i \right)$$

dla $k=1$ lub 3 , odpowiednio

gdzie M_1, M_3 – numer klasy kwartyla

b – szerokość klasy kwartyla

c_L – dolny koniec klasy kwartyla



Kwartyle – przykłady

[Przykład 1 – cd.](#)

Przykład 1:

$$168 \cdot \frac{1}{4} = 42 \quad 168 \cdot \frac{3}{4} = 126$$

a więc

[Przykład 3 – cd.](#)

$$Q_1 = \frac{1}{2}(X_{42:168} + X_{43:168}) = 2, \quad Q_3 = \frac{1}{2}(X_{126:168} + X_{127:168}) = 3,5$$

Przykład 3:

$$100 \cdot \frac{1}{4} = 25 \quad 100 \cdot \frac{3}{4} = 75$$

$M_1 = 2, \quad M_3 = 4$ a więc

$$Q_1 \cong 40 + \frac{10}{23}(25 - 11) \approx 40,09 \quad Q_3 \cong 60 + \frac{10}{12}(75 - 67) \approx 66,67$$





WARSAW UNIVERSITY
Faculty of Economic Sciences