

Statystyka Matematyczna

Anna Janicka

wykład II, 29.02.2016

**STATYSTYKA OPISOWA, cz. II
WSTĘP DO STATYSTYKI MATEMATYCZNEJ**

Plan na dzisiaj

1. Statystyka opisowa, cz. II:

- miary położenia – dokończenie
- miary zróżnicowania
- miary asymetrii
- wykres pudełkowy

2. Wstęp do statystyki matematycznej

- model statystyczny



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



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



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		

Moda – przykłady

[Przykład 1 – cd.](#)

Przykład 1:

$$Mo = 2$$

Przykład 3:

[Przykład 3 – cd.](#)

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

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

Przykład 3:

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

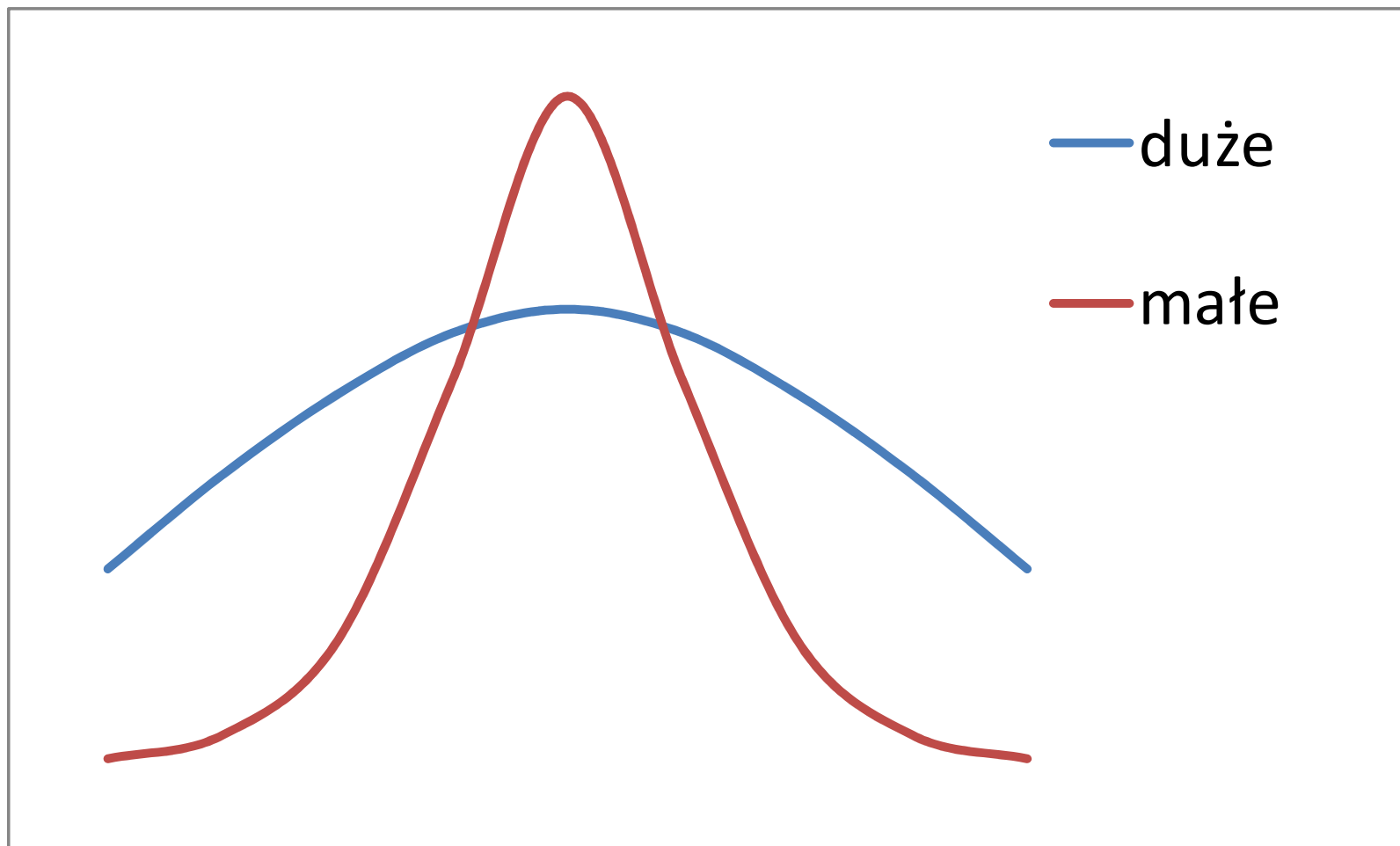
$M_1 = 2$, $M_3 = 4$ a więc

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

Przykład 3 – cd.



Rozproszenie, zmienność, dyspersja



Miary rozproszenia

□ Miary klasyczne

- wariancja, odchylenie standardowe
- odchylenie przeciętne
- współczynnik zmienności (klas.)

□ Miary pozycyjne

- rozstęp
- rozstęp międzykwartyłowy
- *odchylenie ćwiartkowe*
- *współczynnik zmienności (poz.)*



Miary pozycyjne

□ Rozstęp

najprostsza miara, nie bierze pod uwagę żadnych wartości oprócz skrajnych

$$r = X_{n:n} - X_{1:n}$$

□ Rozstęp międzykwartylowy

bardziej odporny na obserwacje nietypowe niż zwykły rozstęp

$$IQR = Q_3 - Q_1$$

długość przedziału, w którym mieści się 50% środkowych obserwacji

na jego podstawie **odchylenie ćwiartkowe** $Q = IQR/2$, oraz pozycyjne współczynniki zmienności $V_Q = Q/Med$ albo $V_{Q_1Q_3} = IQR/(Q_3 + Q_1)$
także typowy przedział zmienności cechy: $[Med - Q, Med + Q]$



Rozstęp, rozstęp międzykwartylowy – przykłady

Przykład 1:

$$r = 5 - 2 = 3,$$

$$IQR = 3,5 - 2 = 1,5$$

Przykład 3:

$$r \cong 120 - 30 = 90$$

(w rzeczywistości $118,9 - 32,45 = 86,45$)

$$IQR \cong 66,67 - 46,09 = 20,58$$



Miary klasyczne

Wariancja

□ dane surowe $\hat{S}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - (\bar{X})^2$

□ szereg rozdzielczy punktowy

$$\hat{S}^2 = \frac{1}{n} \sum_{i=1}^k n_i (X_i - \bar{X})^2 = \frac{1}{n} \sum_{i=1}^k n_i X_i^2 - (\bar{X})^2$$

□ szereg rozdzielczy przedziałowy

$$\hat{S}^2 \cong \frac{1}{n} \sum_{i=1}^k n_i (\bar{c}_i - \bar{X})^2 = \frac{1}{n} \sum_{i=1}^k n_i \bar{c}_i^2 - (\bar{X})^2$$

+ ew. poprawka Shepparda $\bar{S}^2 \cong \hat{S}^2 - \frac{c^2}{12}$

c=długość
przedziału
klasy (jeśli
równe)



lub ogólniej

$$\bar{S}^2 \cong \hat{S}^2 - \frac{1}{12n} \sum_{i=1}^k n_i (c_i - c_{i-1})^2$$

Wariancja – przykłady

[Przykład 1 – cd.](#)

Przykład 1:

$$\hat{S}^2 \approx$$

$$\frac{1}{168} ((2 - 2,842)^2 \cdot 72 + (3 - 2,842)^2 \cdot 42 + (3,5 - 2,842)^2 \cdot 32 + (4 - 2,842)^2 \cdot 11 + (4,5 - 2,842)^2 \cdot 7 + (5 - 2,842)^2 \cdot 4)$$

$$\approx 0,706$$

Przykład 3:

[Przykład 3 – cd.](#)

$$\hat{S}^2 \approx \frac{1}{100} \cdot ((35 - 58,7)^2 \cdot 11 + (45 - 58,7)^2 \cdot 23 + (55 - 58,7)^2 \cdot 33 + (65 - 58,7)^2 \cdot 12 \\ + (75 - 58,7)^2 \cdot 6 + (85 - 58,7)^2 \cdot 8 + (95 - 58,7)^2 \cdot 3 + (105 - 58,7)^2 \cdot 2 + (115 - 58,7)^2 \cdot 2)$$

$$= 331,31$$

$$\bar{S}^2 = 331,31 - \frac{10^2}{12} \approx 322,98$$

w rzeczywistości $\hat{S}^2 = 333,85$

rozkład nie jest normalny, za mała próba na poprawkę Shepparda – większe błędy wynikają z małej próby niż z podziału na klasy



Odchylenie standardowe

W tych samych jednostkach, co wyjściowy szereg

$$\hat{S} = \sqrt{\hat{S}^2}, \quad \bar{S} = \sqrt{\bar{S}^2}$$

Przykład 1:

$$\hat{S} \approx 0,840 \text{ [oceny]}$$

Przykład 3:

$$\hat{S} \approx 18,2 \text{ [m}^2\text{]}$$



Odchylenie przeciętne

średnie odchylenie bezwzględne
obecnie rzadko stosowane, choć łatwiejsze
w obliczeniach, wyrażone w jednostkach
naturalnych
dla danych surowych

$$d = \frac{1}{n} \sum_{i=1}^n |X_i - \bar{X}|$$

itd...

Mamy: $d < S$



WARSAW UNIVERSITY
Faculty of Economic Sciences

Współczynniki zmienności (klasyczne)

Do porównywania tej samej cechy w różnych populacjach lub różnych cech jednej populacji

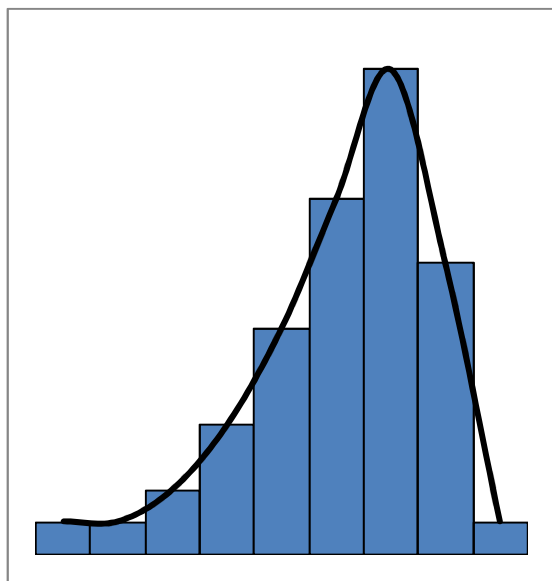
$$V_s = \frac{\hat{S}}{\bar{X}} (\cdot 100\%),$$

$$\text{lub } V_d = \frac{d}{\bar{X}} (\cdot 100\%)$$



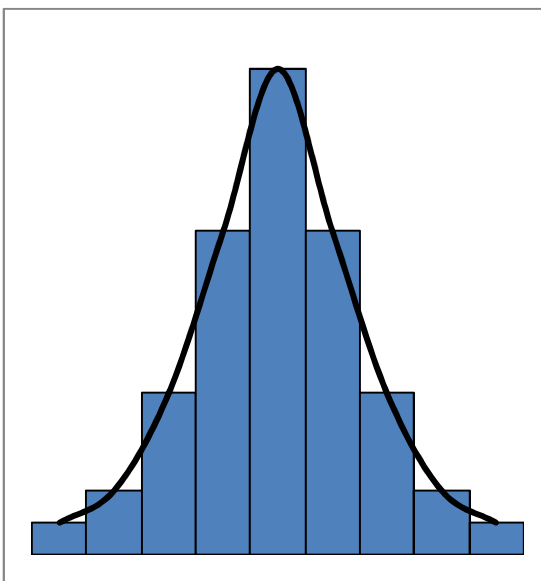
Asymetria

lewostronna
(ujemna)



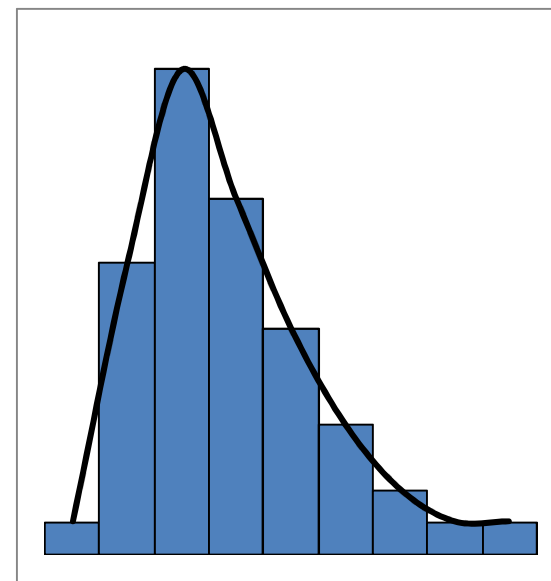
$$\bar{X} < Med < Mo$$

symetria



$$\bar{X} = Med = Mo$$

prawostronna
(dodatnia)



$$\bar{X} > Med > Mo$$

(typowe układy)



WARSAW UNIVERSITY
Faculty of Economic Sciences

Miary asymetrii

□ Współczynnik asymetrii

$$A = \frac{M_3}{\hat{S}^3}$$

gdzie M_3 jest trzecim momentem centralnym

□ Współczynnik skośności

$$A_1 = \frac{\bar{X} - Mo}{\hat{S}} \quad \text{lub} \quad A_1 = \frac{\bar{X} - Med}{\hat{S}}$$

□ Pozycyjny współczynnik asymetrii

$$A_2 = \frac{Q_3 - 2Med + Q_1}{Q_3 - Q_1}$$

mierzy asymetrię tylko dla obserwacji drugiej i trzeciej ćwiartki



Interpretacja

- ☐ Wskaźniki dodatnie = asymetria dodatnia (prawostronna)
- ☐ Wskaźniki ujemne = asymetria ujemna (lewostronna)
- ☐ Dla współczynnika skośności (z medianą) i pozycyjnego wsp. asymetrii ocena siły asymetrii (co do modułu):
 - ☐ 0 - 0,33: słaba
 - ☐ 0,34 - 0,66: średnia
 - ☐ 0,67 – 1: silna



Asymetria – przykłady

Przykład 1: $A = 0,55$

$$A_1 = \frac{2,842 - 3}{0,840} \approx -0,18 \text{ (Med)}$$

$$A_1 = \frac{2,842 - 2}{0,840} \approx 1,00 \text{ (Mo)}$$

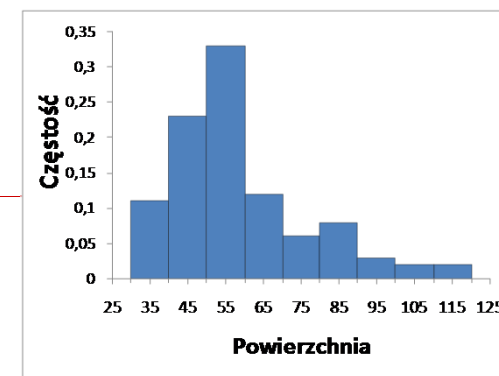
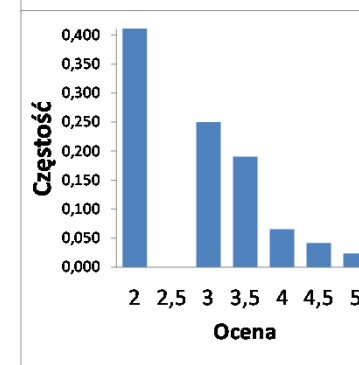
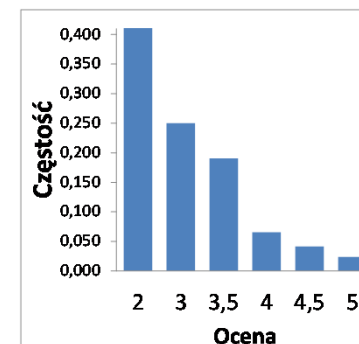
$$A_2 = \frac{3,5 - 2 \cdot 3 + 2}{3,5 - 2} \approx -0,33$$

Przykład 3:

$$A \cong 1,15,$$

$$A_1 \cong \frac{58,7 - 53,23}{18,2} \approx 0,3 \text{ (Mo)} \text{ lub } A_1 = \frac{58,7 - 54,85}{18,2} \approx 0,24 \text{ (Med)}$$

$$A_2 \cong \frac{66,67 - 2 \cdot 54,85 + 46,09}{66,67 - 46,09} \approx 0,15$$



Wykres pudełkowy („pudełko z wąsami”)

- Pozwala porównać graficznie dwie populacje (lub więcej)

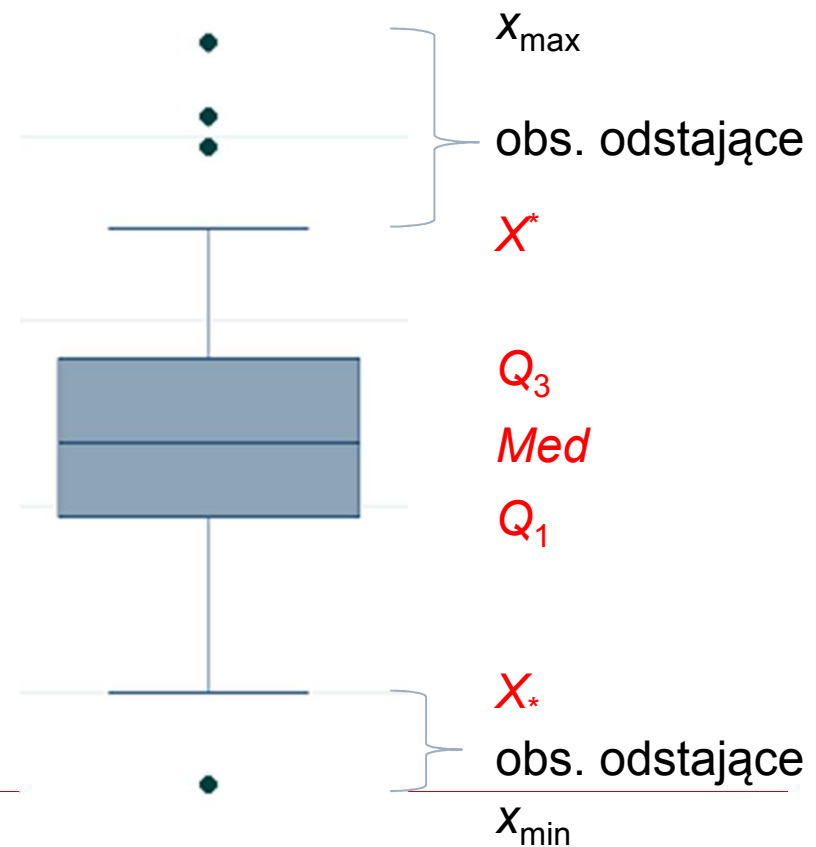
$$X_* = \min\{X_i : X_i \in [Q_1 - \frac{3}{2}IQR, Q_1]\}$$

$$X^* = \max\{X_i : X_i \in [Q_3, Q_3 + \frac{3}{2}IQR]\}$$

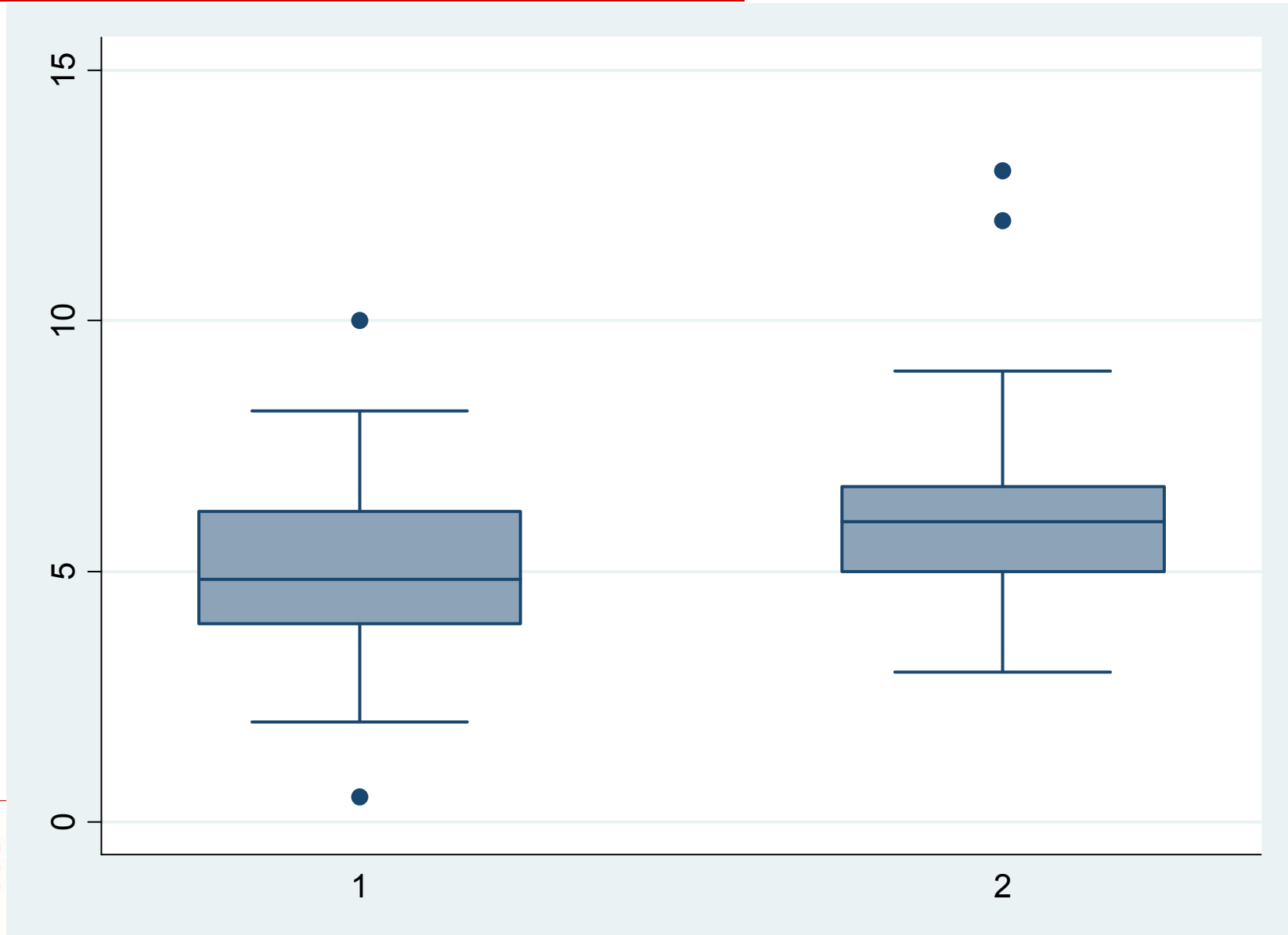
(ewentualne)

obserwacje odstające:

$$x < X_* \text{ lub } x > X^*$$



Wykres pudełkowy – przykład porównania



Przykłady zestawień statystycznych (1)

TABL. 16. CHARAKTERYSTYKI OPISOWE CEN DETALICZNYCH NIEKTÓRYCH TOWARÓW ŻYWNOŚCIOWYCH
W CZERWCU 2009 R.

DESCRIPTIVE CHARACTERISTICS OF RETAIL PRICES OF SELECTED FOODSTUFFS IN JUNE 2009

Towary Goods	Średnia cena w złotych Average price in zł				Współczynnik zmienności w % Variability coefficient in %	
	najniższa lowest	najwyższa highest	krajowa national		ogółem total	między- woje- wódzkiej interview- vodship
	województwo voivodship		arytme- tyczna arithmetic	mediana median		
Chleb pszenno-żytni - za 0,5 kg Wheat-rye bread - per 0.5 kg	1,59 świętokrzyskie	2,19 śląskie	1,88	1,86	19,6	7,6
Mąka pszenna - za 1 kg Wheat flour - per kg	1,74 lubuskie	2,19 świętokrzyskie	1,93	1,99	17,5	4,9
Kasza jęczmienna - za 0,5 kg Pearl-barley - per 0.5 kg	1,54 podkarpackie	2,13 mazowieckie	1,82	1,80	32,3	10,0
Mięso wołowe z kością (rostbef) - za 1 kg Beef meat, bone-in (roast beef) - per kg	17,93 łódzkie	23,51 kujawsko-pomorskie	19,84	19,50	15,9	6,8
Mięso wołowe bez kości (z udźca) - za 1 kg Beef meat, boneless (gammon) - per kg	21,89 małopolskie	29,63 zachodniopomorskie	26,15	25,80	15,4	9,5



Przykłady zestawień statystycznych (2)

Tabl. 3.2. Średnie zużycie, wydatki i ceny energii elektrycznej w gospodarstwach domowych

Miary statystyczne	Ilość				Wartość	Cena	
	razem		w tym gospodarstwa prowadzące działalność rolniczą				
	w kWh	w GJ	w kWh	w GJ		w zł	w zł/kWh
Średnia arytmetyczna	2303	8,3	2983	10,7	1245	0,53	147,2
Mediana.....	2039	7,3	2558	9,2	1105	0,52	144,4
Zakres decylowy	956 – 4080	3,4 – 14,7	1240 – 4800	4,5 – 17,3	520 – 2160	0,48 – 0,60	133,3 – 166,7

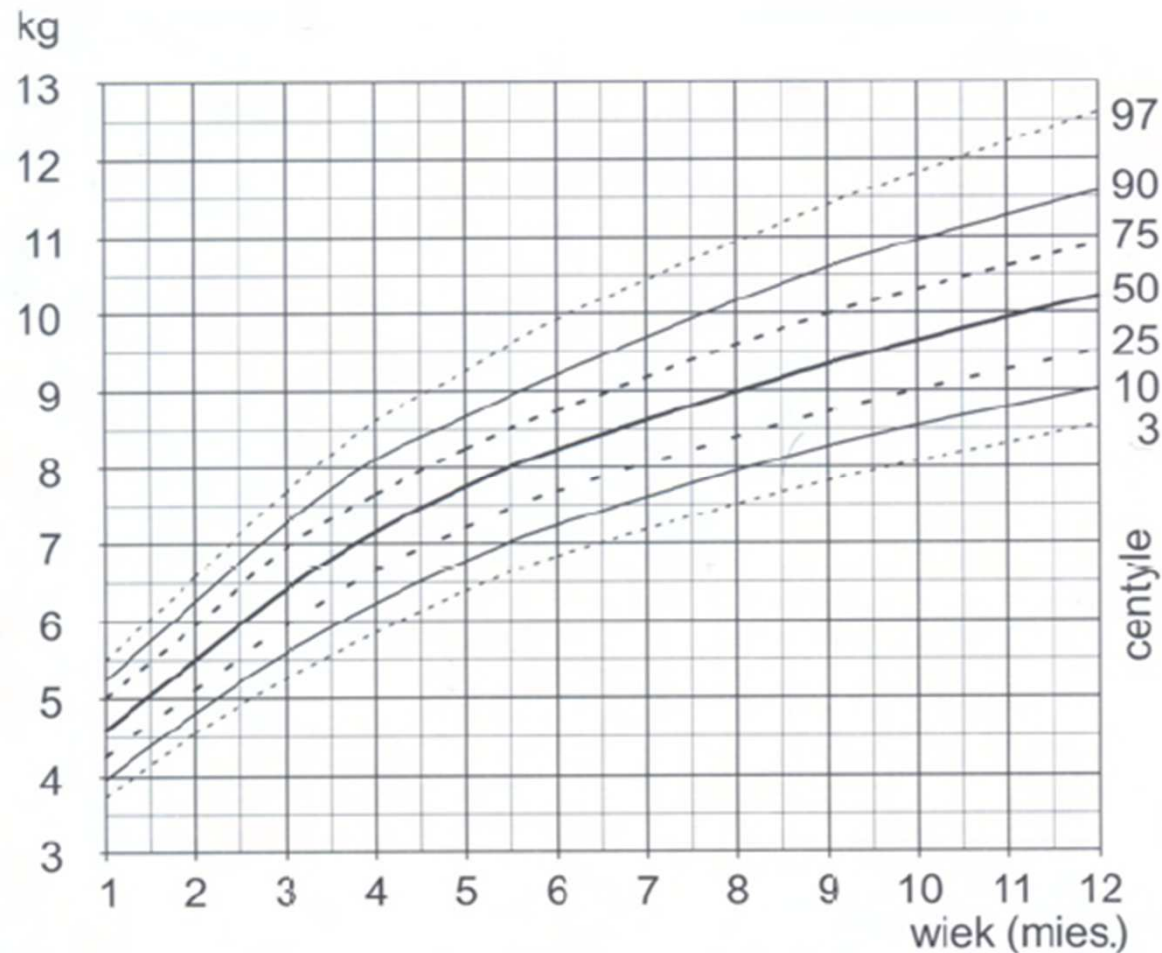
A. Miary wybranych cech mieszkań

Cecha mieszkania	Jednostka miary	Średnia arytmetyczna	Pierwszy decyl	Pierwszy kwartyl	Mediana	Trzeci kwartyl	Dziewiąty decyl
Ogólna powierzchnia użytkowa mieszkania.....	m ²	74,9	38,0	48,0	63,0	100,0	140,0
Powierzchnia ogrzewana.....	m ²	73,1	37,0	48,0	61,0	96,0	135,0
Udział powierzchni ogrzewanej w całkowitej	%	97,5	97,1	100,0	100,0	100,0	100,0
Wysokość mieszkania	m	2,6	2,4	2,5	2,5	2,6	2,8



Przykłady zestawień statystycznych (3)

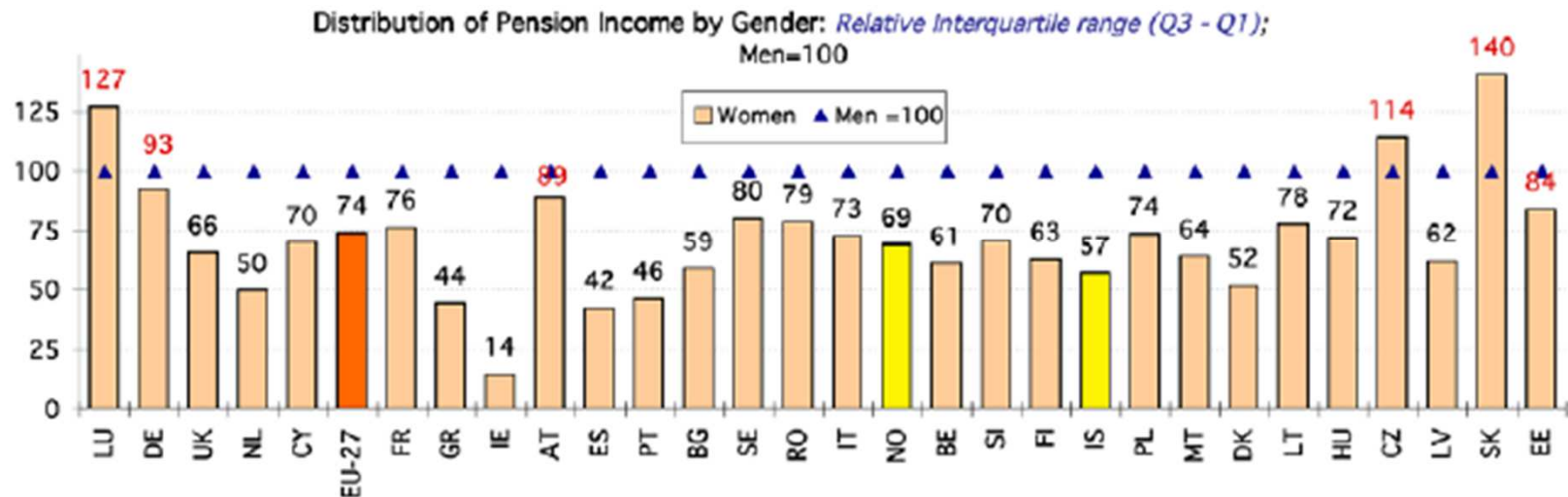
siatka centylowa masy chłopców w W-wie



Przykłady zestawień statystycznych (4)

Względne rozstępy międzykwartylowe

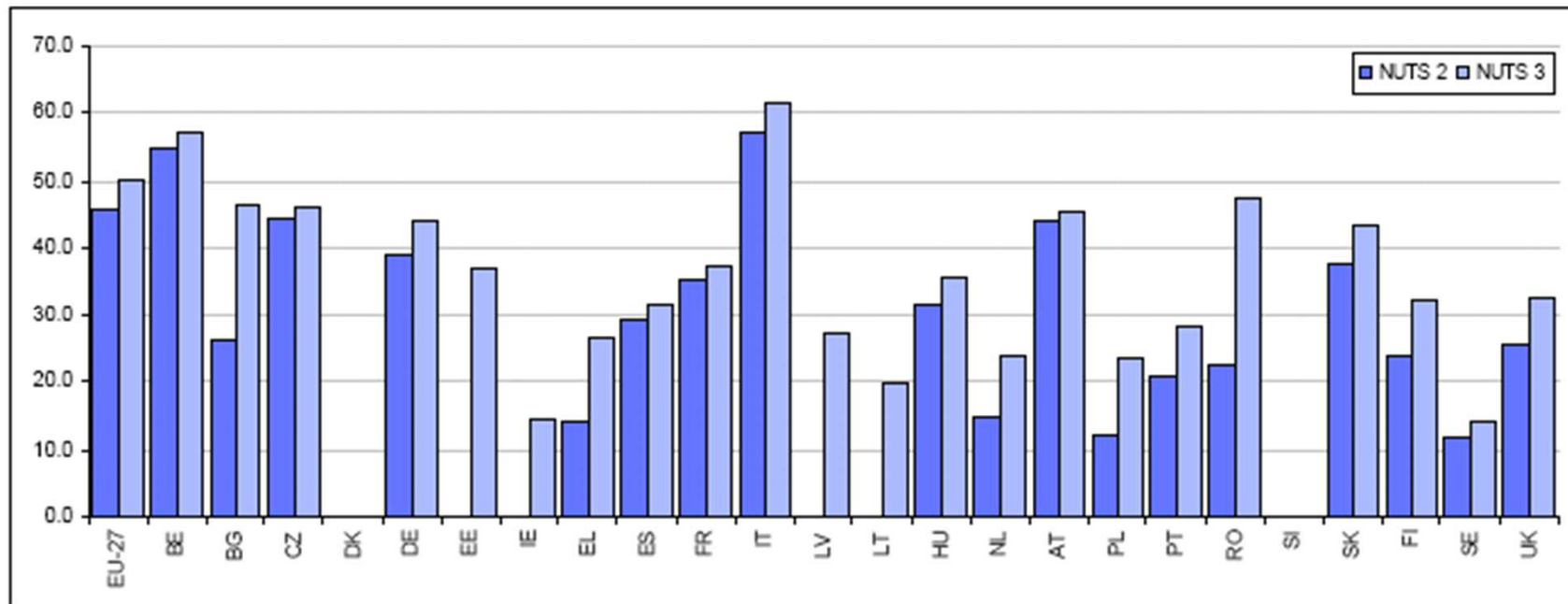
Rozstęp międzykwartylowy poziomu emerytury
wg płci



Przykłady zestawień statystycznych (5)

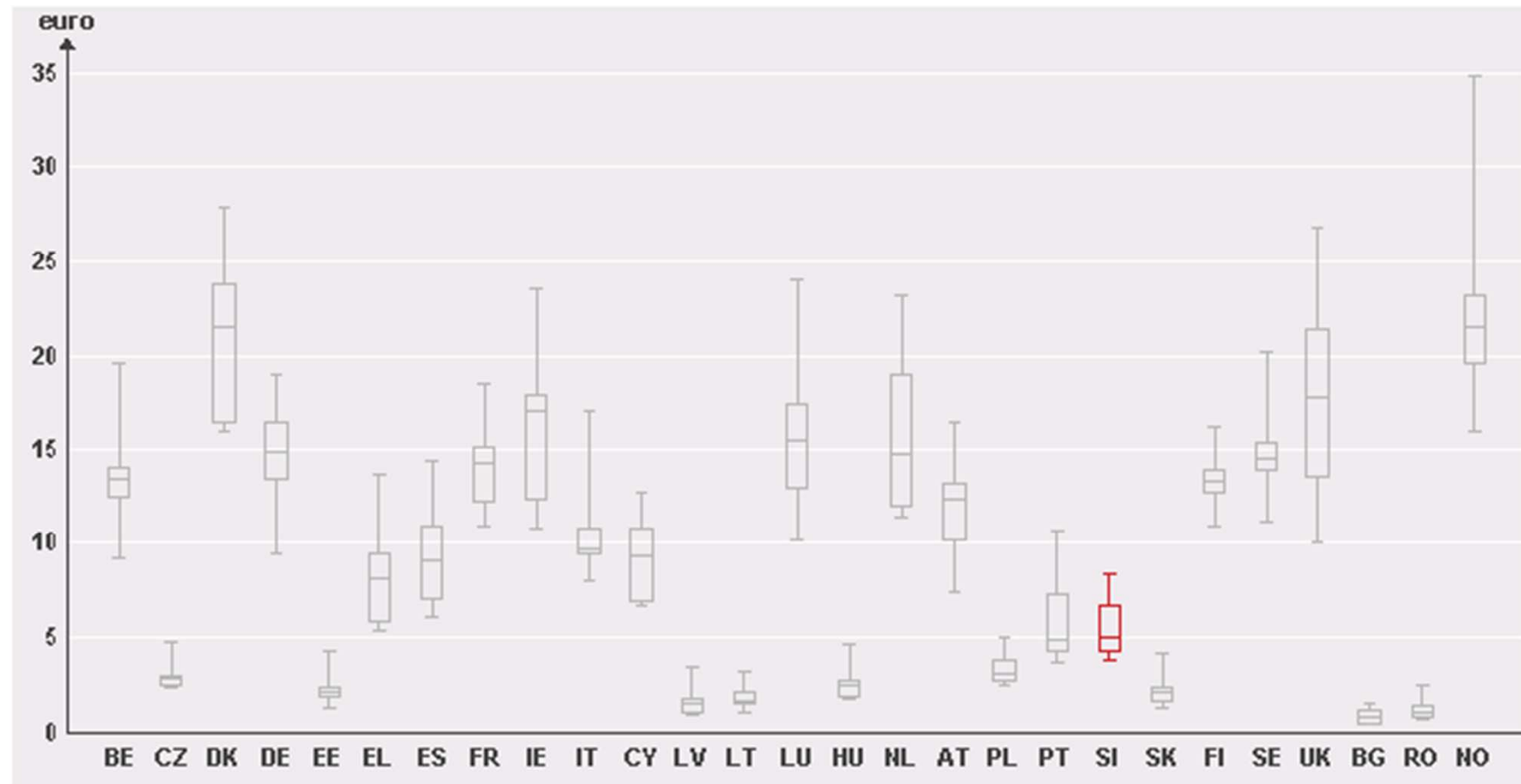
Zróżnicowanie poziomów bezrobocia

Zróżnicowanie poziomów bezrobocia w ujęciu regionalnym
(mierzone współczynnikiem zmienności), 2006



Przykłady zestawień statystycznych (6)

Godzinowe wynagrodzenia brutto, 2002



Źródło: komisja Europejska 2005



WARSAW UNIVERSITY
Faculty of Economic Sciences

STATYSTYKA MATEMATYCZNA



WARSAW UNIVERSITY
Faculty of Economic Sciences

Założenia statystyki matematycznej

Dane doświadczalne są wynikiem działania pewnego mechanizmu losowego.

A zatem: mamy do czynienia ze zmiennymi losowymi określonymi na pewnej przestrzeni probabilistycznej, których realizacjami (wartościami) są zebrane dane. Problem: nie znamy (dokładnego) rozkładu tych zmiennych losowych...



Różnica w podejściach RP i SM:

1. RP, przykład:

- Sformułowanie: w procesie produkcyjnym każdy konkretny wyrób może być wadliwy. Dzieje się tak z prawdopodobieństwem 10%. Wady poszczególnych sztuk są niezależne.
- Problemy: Jaka jest szansa, że w partii 50 sztuk dokładnie 6 będzie wadliwych? Ile średnio sztuk będzie wadliwych? Jaka jest najbardziej prawdopodobna liczba sztuk wadliwych?
- Rozwiązanie: budujemy model probabilistyczny, tu: Schemat Bernoulliego dla $n=50$, $p=0,1$

Ewentualnie, jeśli interesują nas też inne pytania (np. jaka jest szansa, że pierwsze 5 sztuk wadliwych),

model dla ciągów



Różnica w podejściach RP i SM – cd.

2. SM, przykład:

- Sformułowanie: Kontroler przebadął partię 50 sztuk towaru. Wyniki są następujące (1 – towar wadliwy, 0 – bez wad):

0 1 0 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1

- Problemy: jakie jest prawdopodobieństwo, że produkt jest wadliwy (oszacowanie)? Czy prawdą może być deklaracja producenta, że wadliwość to 10%?
 - Rozwiązanie: budujemy model statystyczny, czyli model probabilistyczny z nieznanym(i) parametrem(ami) rozkładu
-



Model Statystyczny

w RP było:

Model statystyczny: (X, F_X, P)

(Ω, F, P)

gdzie:

X – przestrzeń wartości obserwowanej
zmiennej losowej X (często n -wymiarowa,
jeśli mamy n -wymiarową próbkę
zmiennych $X_1, \dots, X_n$)

F_X – σ -ciało na X

P – rodzina rozkładów prawdopodobieństw
 P_θ , indeksowana parametrem $\theta \in \Theta$



Model statystyczny – przykład

$X = \{0, 1\}^n$ – przestrzeń próbkowa

Łączny rozkład prawdopodobieństwa:

$$\begin{aligned} P_{\theta}(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) &= \prod_{i=1}^n \theta^{x_i} (1 - \theta)^{1-x_i} \\ &= \theta^{\sum x_i} (1 - \theta)^{n - \sum x_i} \end{aligned}$$

dla $\theta \in [0, 1]$

(u nas $n=50$ oraz $X_2 = X_{10} = X_{15} = X_{32} = X_{42} = X_{50} = 1$, pozostałe $X_i = 0$)



Model statystyczny – przykład cd.

Alternatywne sformułowanie (jeśli notujemy tylko *liczbę* wadliwych elementów w próbie):

$X = \{0, 1, 2, \dots, n\}$ – przestrzeń próbkowa

Łączny rozkład prawdopodobieństwa:

$$P_{\theta}(X = x) = \binom{n}{x} \theta^x (1 - \theta)^{n-x}$$

dla $\theta \in [0, 1]$

(u nas $n=50$ oraz $X=6$)



Model statystyczny – przykład cd. (2): pytania

Mamy konkretne dane (próbkę):

☐ Jaka jest wartość parametru θ ?

- interesuje nas konkretna wartość
- interesuje nas przedział (*ufności*)

→ *zagadnienie estymacji*

☐ Weryfikacja hipotezy, że $\theta = 0,1$

→ *testowanie hipotez statystycznych*

☐ → *ew. predykcje*

Statystyki

Estymację parametrów (punktową, przedziałową) czy testowanie hipotez statystycznych przeprowadza się na podstawie tzw. *statystyk*:

Statystyka = dowolna funkcja obserwacji, czyli zmienna losowa postaci

$$T = T(X_1, X_2, \dots, X_n)$$

Rozkład statystyki T zależy od rozkładu zmiennej X , ale *statystyka jako taka nie*

może zależeć od parametru θ , np. ~~$X_1 + X_2 - \theta$~~

Statystyki – przykład

$$T_1 = \sum_{i=1}^n X_i, \quad T_2 = \frac{1}{n} \sum_{i=1}^n X_i, \quad T_3 = \frac{1}{n} \sum_{i=1}^n X_i - 0,1$$

są statystykami dla pierwszego sformułowania;

$$T_1 = X, \quad T_2 = \frac{X}{n}, \quad T_3 = \frac{X}{n} - 0,1$$

są statystykami dla drugiego sformułowania

Wybór statystyki zależy od pytania, na które mamy odpowiedzieć.



Model Statystyczny: Przykład 2

Wzrosty na giełdzie. Analityk bada długość *okresów wzrostowych* na giełdzie. Interesuje go czas wzrostu kursu (do pierwszego spadku), w dniach. Załóżmy, że czasy wzrostu $X_1, X_2, \dots, X_n$ są próbką z rozkładu wykładniczego $\text{Exp}(\lambda)$.

λ – nieznaną parametr

$X = (0, \infty)^n$ – przestrzeń próbkowa

Łączny rozkład prawdopodobieństwa:

$$P_\lambda(X_1 \leq x_1, X_2 \leq x_2, \dots, X_n \leq x_n) = \prod_{i=1}^n (1 - e^{-\lambda x_i})$$

$$f_\lambda(x_1, x_2, \dots, x_n) = \lambda^n e^{-\lambda \sum x_i}$$

dla $\lambda > 0$



Model Statystyczny: Przykład 3

Pomiar z błędem losowym: powtarzamy pomiar wielkości μ , wyniki poszczególnych pomiarów są niezależnymi zmiennymi los. $X_1, X_2, \dots, X_n$, bo maszyna do pomiaru niedoskonała. Każdy z pomiarów ma jednakowy rozkład normalny $N(\mu, \sigma^2)$.

μ, σ^2 – nieznane parametry (a więc $\theta = (\mu, \sigma)$)

$X = \mathbb{R}^n$ – przestrzeń próbkowa

Łączny rozkład prawdopodobieństwa:

$$P_{\mu, \sigma}(X_1 \leq x_1, X_2 \leq x_2, \dots, X_n \leq x_n) = \prod_{i=1}^n \Phi\left(\frac{x_i - \mu}{\sigma}\right) \text{ lub}$$

$$f_{\mu, \sigma}(x_1, x_2, \dots, x_n) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2\right)$$

dla $\mu \in \mathbb{R}, \sigma > 0$





WARSAW UNIVERSITY
Faculty of Economic Sciences
