Rachunek prawdopodobieństwa i statystyka

Statystyka opisowa

Prof. UEK dr hab. Paweł Ulman

Program - zarys

1. Wprowadzenie w problematykę rachunku prawdopodobieństwa i statystyki
2. Parametry rozkładu zmiennej losowej oraz rozkładu empirycznego
3. Podstawowe rozkłady zmiennej losowej, zmienna losowa dwuwymiarowa i wielowymiarowa
4. Podstawowe pojęcia statystyki matematycznej
5. Estymacja
6. Weryfikacja hipotez – testy statystyczne
7. Statystyczne metody analizy dynamiki zjawisk

Historia statystyki i rachunku prawdopodobieństwa

Termin „statystyka” rozumieć można jako:

- Pozyskiwanie danych (informacji); zbiór (zestawienie) danych,
- Naukę,
- Jako pewną wielkość, charakteryzującą się znanym rozkładem.

Pierwsze badania natury statystycznej – spisy (powszechne)

Sumeria (3-4 tys. lat przed Chrystusem),

Egipt - c*o najmniej 3 tys. lat przed Ch.* razem ze spisem majątków

Izrael - Biblia, księga liczb (Numerii) - 2 spisy mężczyzn: pierwszy -mężczyźni powyżej 20 lat (603550), drugi - mężczyźni od 20 lat życia wzwyż (601730) plus Lewici (23000) w wieku od 1 miesiąca wzwyż.

Historia statystyki i rachunku prawdopodobieństwa

Chiny (co najmniej od 2238 przed Ch.) - w latach 1368 - 1644 po Ch. cesarzowie z dynastii Ming wprowadzili system spisów co 10 lat.

Rzym (od VI w. przed Ch.) - Cesarz August nakazał 3 razy przeprowadzenie spisu, z czego drugi przeszedł do historii dzięki św. Łukaszowi

Francja (w 786 r. Karol Wielki zarządził spis poddanych powyżej 12 roku życia*)*

Anglia (Wilhelm Zdobywca zarządza w 1085 r. szczegółowy opis Anglii - Domesday Book)

Historia statystyki i rachunku prawdopodobieństwa

Spisy ludności w nowożytnym świecie

Szwecja - 1749

Polska - 1777 - spis ludności miast

- 1789 - Sejm Czteroletni zarządza I powszechny spis ludności

- pozostałe spisy 1921, 1931, 1950, 1960, 1970, 1978, 1988, 2002, 2011

- mikrospisy (metodą reprezentacyjną: 1974, 1984, 1995).

Stany Zjednoczone Am. Płn. - 1790 (co 10 lat)

Francja (1801, co 5 lat, po wojnie co 7 lat),

Wielka Brytania (1801, co 10 lat)

Norwegia (1815), Holandia (1829), Dania (1840), Belgia (1846) - Adolphe Quetelet - metodolog spisu, który stał się wzorcem dla innych krajów,

Niemcy (1871)

Rosja (1896/1897), ZSRR (1926, 1937, 1939, 1959, 1970, 1979, 1989)

Historia statystyki i rachunku prawdopodobieństwa

Statystyka jako nauka

1. Badania arytmetyków politycznych:
    - J. Graunt (1620-1674),
    - W. Petty (1623-1687),

wprowadzili rozumienie statystyki jako metody wnioskowania na podstawie danych liczbowych umożliwiającej wykrywanie prawidłowości wśród pozornie chaotycznych zjawisk masowych.

2. badania państwowoznawcze:

nazwa statystyka wywodzi się z łaciny (od słowa *status* oznaczającego stan, miejsce),

G.Achenwall (1719-1772), pierwszy raz użył terminu „statystyka” w piśmie, w znaczeniu zbioru szeroko ujmowanych wiadomości o stanie państwa,

Natural and *Political*

OBSERVATIONS

Mentioned in a following INDEX,
and made upon the

Bills of Mortality.

BY

Capt. *JOHN GRAUNT*,
Fellow of the *Royal Society*.

With reference to the *Government*, *Religion*, *Trade*, *Growth*, *Air*, *Diseases*, and the several Changes of the said CITY.

——*Non, me ut miretur Turba, laboro,*
Contentus paucis Lectoribus.——

The Fifth Edition, much Enlarged.

LONDON,
Printed by *John Martyn*, Printer to the *Royal Society*, at the Sign of the *Bell* in St. *Paul's* Church-yard. MDCLXXVI.

Źródło: Wikipedia

Historia statystyki i rachunku prawdopodobieństwa

Metoda jaką [...] *stosuje nie jest jeszcze często w użyciu. Zamiast używać tylko słów porównujących i opisujących oraz argumentów intelektualnych, wybieram kierunek (jako specjalista od arytmetyki politycznej czyniłem to od dawna) na wyrażanie się za pomocą liczb, wag i miar, korzystając tylko z argumentów rozumu lub rozważając tylko te przyczyny, które mają swe podstawy w przyrodzie.*

William Petty

źródło: Wikipedia

Historia statystyki i rachunku prawdopodobieństwa

Statystyka jako nauka

W ramach nurtu państwowoznawczego wykształcił się tabularyzm jako metoda pojmowania danych liczbowych w formie tabel,

J.K. Kirgiłow – pierwszy opis tabelaryczny Rosji,

J.P. Anchersen – pierwszy opis tabelaryczny Danii.

3. dalszy rozwój statystyki jest związany z powstaniem matematycznej teorii rachunku prawdopodobieństwa:

B. Pascal (1623-1662),

P. Fermat (1601-1665).

Definicje statystyki i rachunku prawdopodobieństwa

STATYSTYKA

OPISOWA

1) rozwinięcie nurtu badań państwowo-znawczych,
2) cel: statystyczny opis struktury (budowa, elementy składowe), współzależności (związek między elementami składowymi) i dynamiki (zmienność w czasie),
3) sposób: deterministyczne metody opisowe,

MATEMATYCZNA (INDUKCYJNA)

1) rozwinięcie nurtu badań arytmetyków,
2) cel: uogólnianie wyników dotyczących struktury, współzależności i dynamiki zjawisk masowych otrzymywanych z próby losowej dla całej populacji,
3) sposób: wnioskowanie statystyczne (metody statystyczne), podstawa teoretyczna (rachunek prawdopodobieństwa),

uwaga: nie ma wnioskowania bez opisu,

Definicje statystyki i rachunku prawdopodobieństwa

Statystyka – nauka traktująca o metodach ilościowych wykorzystywanych w celu poszukiwania prawidłowości w pozornie chaotycznych zjawiskach masowych.

Rachunek prawdopodobieństwa – dział matematyki zajmujący się zdarzeniami losowymi. Rachunek prawdopodobieństwa zajmuje się badaniem abstrakcyjnych pojęć matematycznych stworzonych do opisu zjawisk, które nie są deterministyczne: zmiennych losowych w przypadku pojedynczych zdarzeń oraz procesów stochastycznych w przypadku zdarzeń powtarzających się (w czasie).

Podstawowe pojęcia statystyki

1. zbiorowość statystyczna (populacja/masa statystyczna) – zbiór dowolnych elementów objęty badaniem statystycznym
2. jednostka statystyczna (jednostka badania lub obserwacji) – element składowy badanej zbiorowość
3. cecha statystyczna – właściwość jednostki statystycznej:

- cecha stała i zmienna;
- cecha jakościowa i ilościowa (mierzalna);
- cecha skokowa (dyskretna) i ciągła.

Podstawowe pojęcia statystyki

4. Badanie statystyczne – zespół czynności zmierzających do określenia prawidłowości w badanej zbiorowości.

Rodzaje:

- ciągłe, cykliczne (okresowe), doraźne;
- pełne, częściowe, szacowanie.

Spis powszechny, rejestracja bieżąca, badanie ankietowe, badanie monograficzne, badanie reprezentacyjne, interpolacja, ekstrapolacja

Podstawowe pojęcia statystyki

Etapy badania statystycznego:

1. Przygotowanie: cel, jednostka, przedmiot i metoda badania;
2. Obserwacja statystyczna: materiał statystyczny – pierwotny i wtórny;
3. Kontrola materiału statystycznego: formalna, merytoryczna (logiczna, arytmetyczna). Błędy losowe i systematyczne;
4. Przetwarzanie i prezentacja materiału statystycznego: tabele, wykresy, szeregi;
5. Analiza statystyczna.

Przetwarzanie i prezentacja materiału statystycznego

Objaśnienia znaków umownych

Kreska	(—)	— zjawisko nie wystąpiło.
Zero	(0,0)	— zjawisko istniało, jednakże w ilościach mniejszych od liczb, które mogły być wyrażone uwidocznionymi w tablicy znakami cyfrowymi.
Kropka	(.)	— zupełny brak informacji albo brak informacji wiarygodnych.
Znak	x	— wypełnienie pozycji, ze względu na układ tablicy, jest niemożliwe lub niecelowe.
„W tym"		— oznacza, że nie podaje się wszystkich składników sumy.
Znak	#	— oznacza, że dane nie mogą być opublikowane ze względu na konieczność zachowania tajemnicy statystycznej w rozumieniu ustawy o statystyce publicznej.

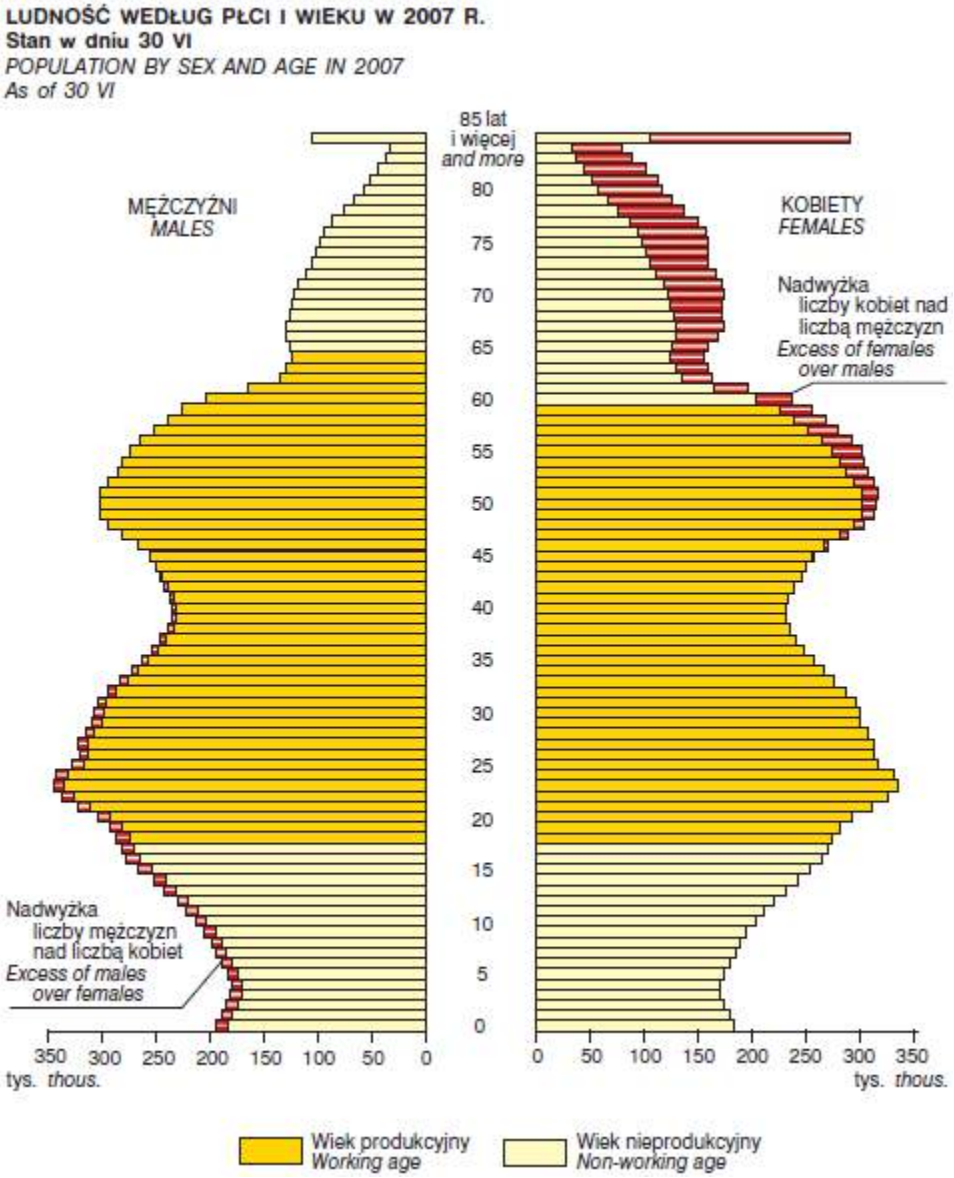
Przetwarzanie i prezentacja materiału statystycznego

LUDNOŚĆ — POPULATION

TABL. I (60). **LUDNOŚĆ NA PODSTAWIE SPISÓW**
POPULATION BASED ON CENSUS DATA

Data spisu *Census date*	Ogółem *Total*	Mężczyźni *Males*	Kobiety *Females*	Z liczby ogółem — w % — ludność *Of total number — in % — population*		Ludność na 1 km² *Population per 1 km²*
	w tys. *in thous.*			miejska *urban areas*	wiejska *rural areas*	
W GRANICACH Z DNIA 31 III 1938 R. (powierzchnia 389 tys. km²) *WITHIN BORDERS AS OF 31 III 1938 (area 389 thous. km²)*						
30 IX 1921	27177	13133	14044	24,6	75,4	70
9 XII 1931	32107	15619	16488	27,4	72,6	83
1938 (stan w dniu 31 XII) *(as of 31 XII)*	**34849**	**17000**	**17849**	**30,0**	**70,0**	**90**
W GRANICACH OBECNYCH (powierzchnia ogólna 313 tys. km²) *WITHIN PRESENT BORDERS (total area 313 thous. km²)*						
14 II 1946	23930	10954	12976	31,8	68,2	77
3 XII 1950	25008	11928	13080	39,0	61,0	80
6 XII 1960	29776	14404	15372	48,3	51,7	95
8 XII 1970	32642	15854	16788	52,3	47,7	104
7 XII 1978	35061	17079	17982	57,5	42,5	112
7 XII 1988	37879	18465	19414	61,2	38,8	121
20 V 2002	**38230**	**18516**	**19714**	**61,8**	**38,2**	**122**

Przetwarzanie i prezentacja materiału statystycznego



Przetwarzanie i prezentacja materiału statystycznego

TABL. 7 (66). **LUDNOŚĆ[a] WEDŁUG GŁÓWNEGO ŹRÓDŁA UTRZYMANIA**
POPULATION[a] BY MAIN SOURCE OF MAINTENANCE

Główne źródło utrzymania *Main source of maintenance*	1988	2002				
		ogółem *total*			miasta *urban areas*	wieś *rural areas*
		ogółem *total*	mężczyźni *males*	kobiety *females*		
	w tys. *in thous.*		w odsetkach *in percent*			
OGÓŁEM ***TOTAL***	**37879**	**38230**	**100,0**	**100,0**	**100,0**	**100,0**	**100,0**
Praca *Work*	17218	12355	32,3	36,3	28,5	33,1	31,1
poza rolnictwem *outside agriculture*	13178	10710	28,0	31,2	25,0	32,6	20,7
w tym w sektorze prywatnym *of which in private sector*	1132	6325	16,5	20,4	12,9	18,9	12,8
w rolnictwie *in agriculture*	4040	1645	4,3	5,1	3,5	0,5	10,4
w tym w sektorze prywatnym *of which in private sector*	3116	1607	4,2	5,0	3,4	0,5	10,2
Niezarobkowe źródło *Non-earned source*	6807	10692	28,0	24,4	31,3	27,9	28,0
w tym: *of which:*							
emerytura *retirement pay*	3457	5323	13,9	11,8	15,9	13,9	14,0
renta[b] *pension[b]*	2771	3516	9,2	7,8	10,5	8,8	9,8
zasiłek dla bezrobotnych *unemployment benefit*	—	611	1,6	2,0	1,2	1,6	1,6
zasiłek pomocy społecznej *social assistance benefit*	59	242	0,6	0,6	0,7	0,7	0,6
Dochody z własności *Incomes from property*	—	27	0,1	0,1	0,1	0,1	0,0
Na utrzymaniu *Dependents*	13854	14547	38,1	37,6	38,5	36,8	40,2
Nieustalone źródło *Unknown source*	—	609	1,6	1,6	1,6	2,1	0,7

a Dane spisów powszechnych. *b* Łącznie ze świadczeniem rehabilitacyjnym.
a Data of national censuses. *b* Including rehabilitation benefit.

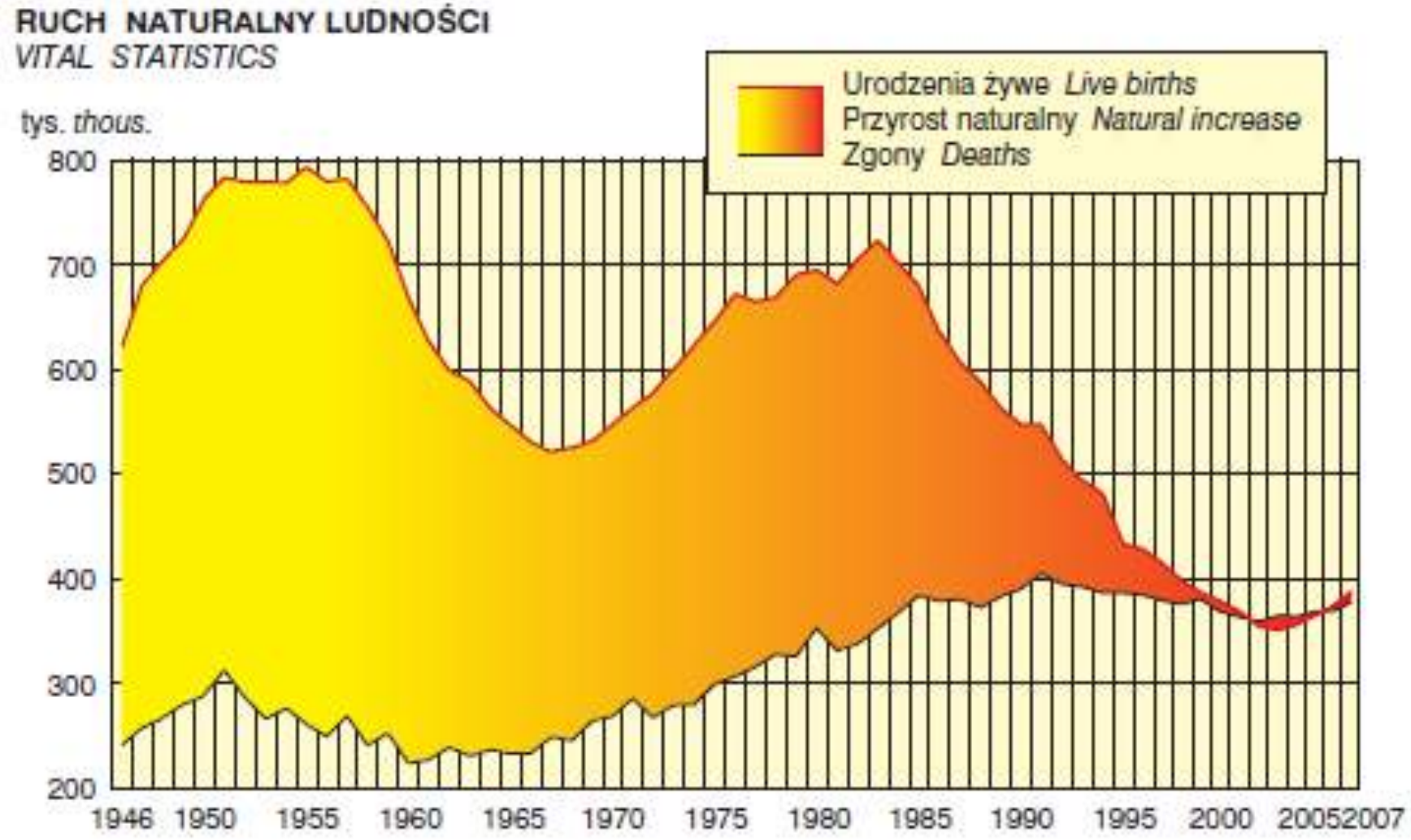
Przetwarzanie i prezentacja materiału statystycznego

TABL. 9 (68). **RODZINY**[a]
FAMILIES[a]

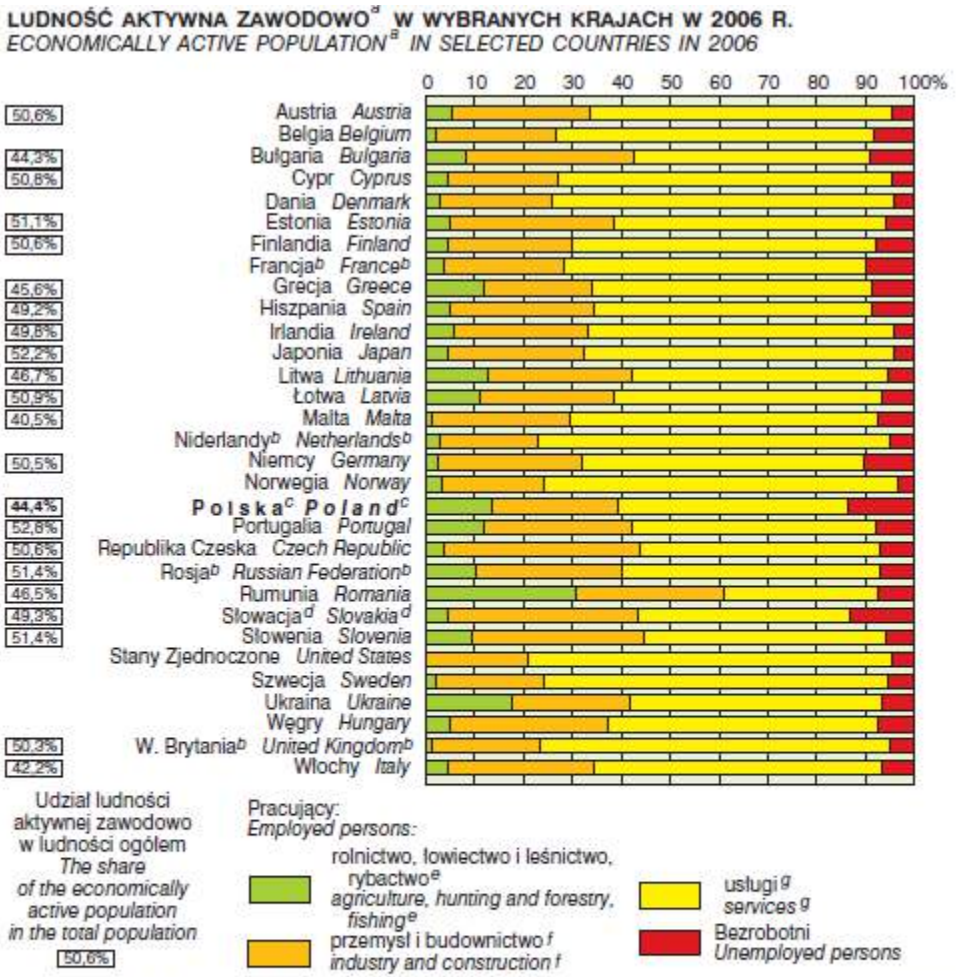
Wyszczególnienie	1988 ogółem *total*	1988 miasta *urban areas*	1988 wieś *rural areas*	2002 ogółem *total*	2002 miasta *urban areas*	2002 wieś *rural areas*	*Specification*
	w tys. *in thous.*						
OGÓŁEM	**10226**	**6364**	**3862**	**10458**	**6597**	**3861**	***TOTAL***
w tym z dziećmi do 24 lat pozostającymi na utrzymaniu	6210	4012	2198	6079	3794	2285	*of which dependent children up to age 24*
Małżeństwa bez dzieci[b]	2329	1418	911	2370	1543	827	*Marriages without children*[b]
Małżeństwa z dziećmi[b]	6323	3874	2449	5860	3511	2349	*Marriages with children*[b]
Partnerzy bez dzieci	.	.	.	87	70	17	*Cohabiting couples without children*
Partnerzy z dziećmi	.	.	.	111	78	33	*Cohabiting couples with children*
Samotne matki z dziećmi	1396	958	438	1798	1241	557	*Lone mothers with children*
Samotni ojcowie z dziećmi	178	114	64	232	154	78	*Lone fathers with children*

a Dane spisów powszechnych. *b* W 1988 r. łącznie ze związkami partnerskimi.
a *Data of national censuses.* *b* *In 1988 including cohabiting couples.*

Przetwarzanie i prezentacja materiału statystycznego



Przetwarzanie i prezentacja materiału statystycznego

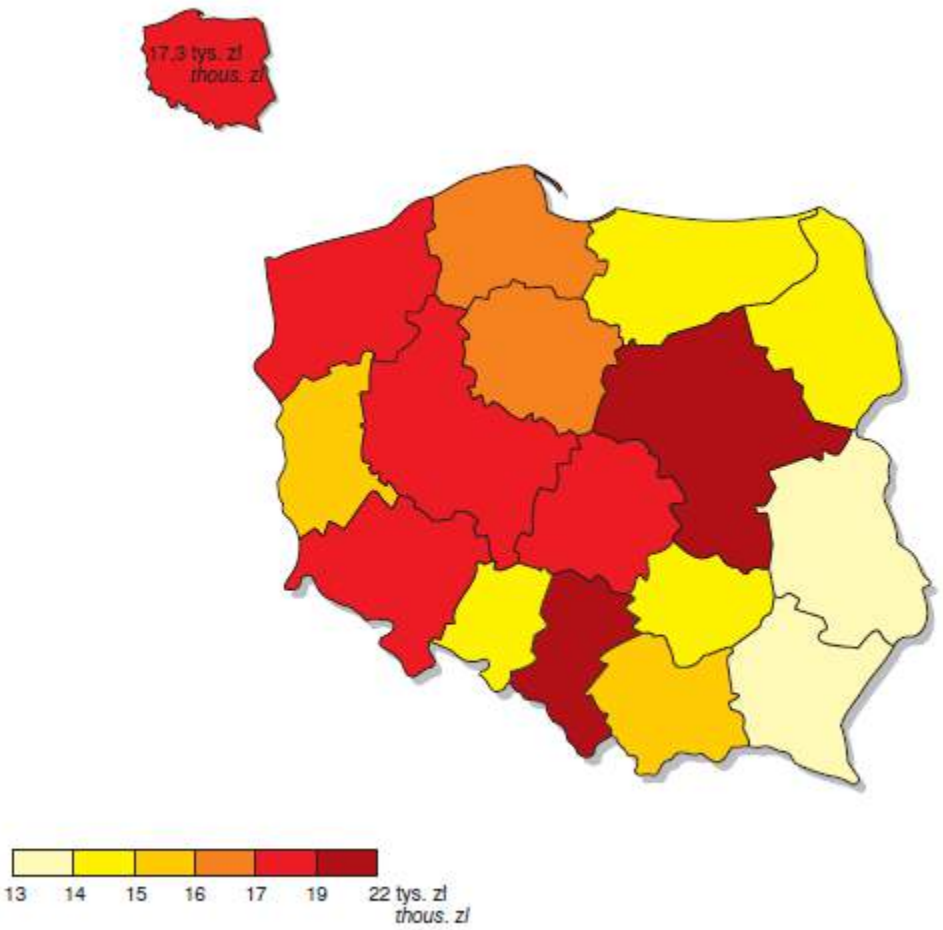


a Dane na podstawie Badania Aktywności Ekonomicznej Ludności (BAEL) opracowane według metodologii Międzynarodowej Organizacji Pracy (MOP). b 2005 r. c Przeciętne w roku. d Bez osób na urlopach wychowawczych. e Dla Danii, Malty, Rumunii i Słowenii bez rybactwa. f Dla Danii bez górnictwa. g Dotyczy pozostałych sekcji PKD (NACE).

a Data based on the Labour Force Surveys (LFS) compiled according to the methodology of International Labour Organization (ILO). b 2005. c Annual averages. d Excluding persons on child-care leave. e For Denmark, Malta, Romania and Slovenia excluding fishing. f For Denmark excluding mining and quarrying. g Concerns remaining NACE sections.

Przetwarzanie i prezentacja materiału statystycznego

NOMINALNE DOCHODY DO DYSPOZYCJI BRUTTO W SEKTORZE GOSPODARSTW DOMOWYCH[a] NA 1 MIESZKAŃCA WEDŁUG WOJEWÓDZTW W 2005 R.
GROSS NOMINAL DISPOSABLE INCOME OF HOUSEHOLDS SECTOR[a] PER CAPITA BY VOIVODSHIP IN 2005



a W podziale według województw nie uwzględniono zmian podanych w nocie na str. 160.
a *In division according to voivodships without taking into consideration changes given in note on page 161.*

Przetwarzanie i prezentacja materiału statystycznego

I. WAŻNIEJSZE DANE O SYTUACJI SPOŁECZNO-GOSPODARCZEJ KRAJU (cd.)
MAJOR DATA REGARDING THE SOCIO-ECONOMIC SITUATION OF THE COUNTRY (cont.)

Lp.	Wyszczególnienie	1950	1960	1970	1980	1990	1995
							EDUKACJA
	Uczniowie w szkołach[d] (stan na początku roku szkolnego) w tys.:						
17	podstawowych[e]	3303	4875	5342	4265	5287	5104
18	gimnazjach	x	x	x	x	x	x
19	zasadniczych zawodowych[f]	318	303	838	732	815	722
20	liceach ogólnokształcących[g]	195	261	402	346	445	683
21	liceach profilowanych	x	x	x	x	x	x
22	technikach[gh]	223	224	475	601	637	828
23	artystycznych ogólnokształcących[i]	.	.	.	.	.	10,5
24	policealnych	x	37,7	67,1	135	108	161

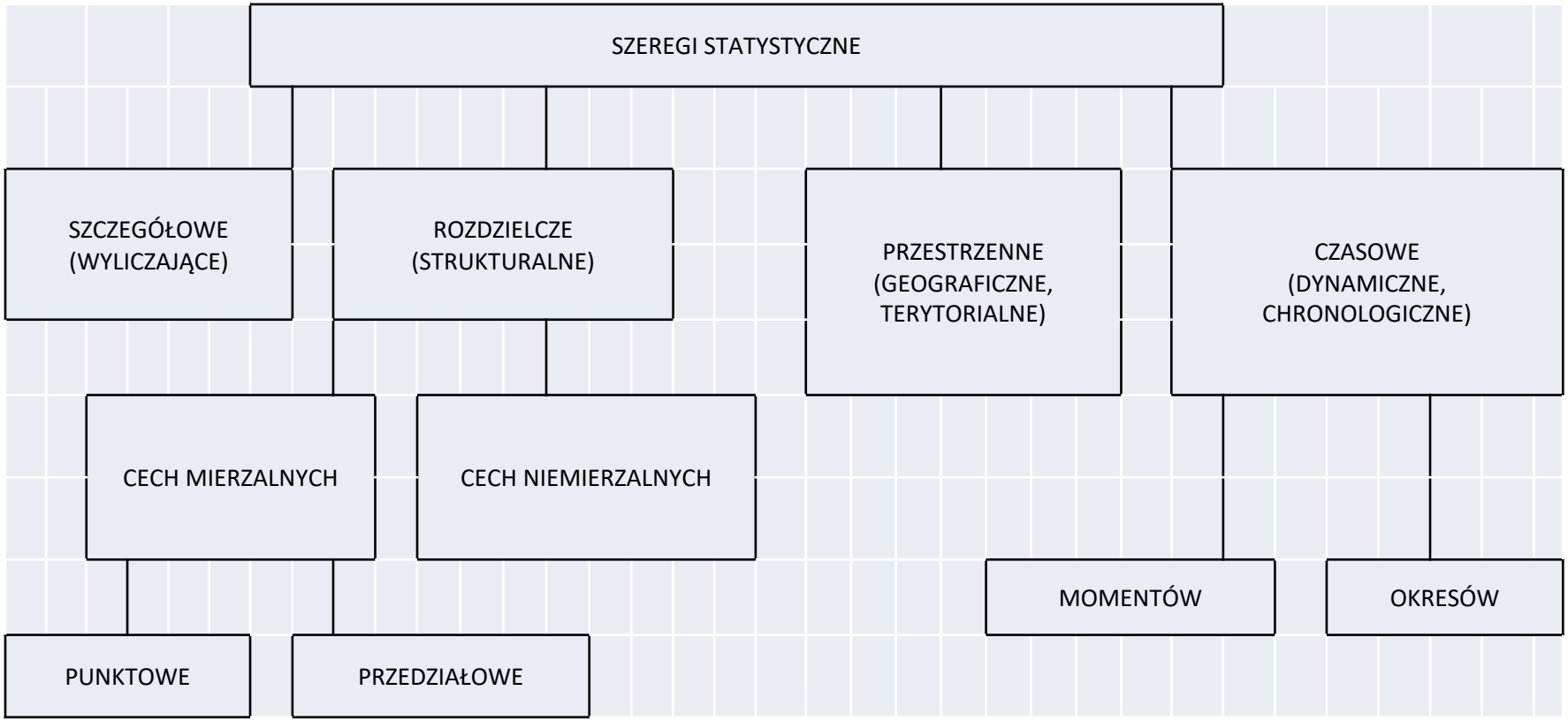
Przetwarzanie i prezentacja materiału statystycznego

TABL. 26 (213). **STAN ZDROWOTNY LASÓW**[a] **(dok.)**
FOREST CONDITION[a] *(cont.)*

Wyszczególnienie	Drzewa — w odsetkach — o stopniu uszkodzenia *Damaged trees in percent*				*Specification*
	0 (bez uszkodzeń) *(none)*	1 (uszkodzenia słabe) *(slight)*	2 (uszkodzenia średnie) *(moderate)*	3[b] (uszkodzenia silne) *(severe)*	
Drzewa iglaste	24,1	54,7	20,4	0,8	*Coniferous trees*
sosna	21,8	57,3	20,2	0,7	*pine*
świerk	38,0	40,0	20,4	1,6	*spruce*
jodła	33,6	43,2	22,4	0,8	*fir*
pozostałe	48,5	28,6	22,6	0,0	*others*
Drzewa liściaste	33,2	48,6	17,6	0,6	*Broadleaved trees*
dąb	16,5	53,7	29,6	0,2	*oak*
buk	53,9	36,1	10,0	0,0	*beech*
brzoza	27,6	55,0	16,9	0,5	*birch*
pozostałe	41,7	42,7	14,6	1,0	*others*

Przetwarzanie i prezentacja materiału statystycznego

Szereg statystyczny - zbiór wyników obserwacji jednostek pod względem pewnej cechy.



Przetwarzanie i prezentacja materiału statystycznego

Szereg szczegółowy:

23, 25, 21, 23, 27, 21, 20, 24

20, 21, 21, 23, 23, 24, 25, 27

0, 1, 5, 0, 0, 1, 1, 2, 2, 1, 2, 3, 3, 4, 3, 4, 1, 3, 2, 2

Liczba dzieci	0	1	2	3	4	5
Liczba rodzin	3	5	5	4	2	1

Przetwarzanie i prezentacja materiału statystycznego

Szereg strukturalny:

Oceny x_i	Liczba studentów studiów zaocznych n_i	Liczba studentów studiów dziennych n_i
2	600	100
3	1200	300
4	900	400
5	300	200
Ogółem	3000	1000

Oceny x_i	Liczba studentów studiów zaocznych n_i	Liczba studentów studiów dziennych n_i
niedostateczny	600	100
dostateczny	1200	300
dobry	900	400
bardzo dobry	300	200
Ogółem	3000	1000

Zużycie energii	2 – 4	4 – 6	6 – 8	8 – 10	10 – 12	12 – 14
Liczba rodzin	6	10	30	40	10	4

Przetwarzanie i prezentacja materiału statystycznego

Szereg geograficzny:

Kraj	Przeciętna miesięczna pensja brutto (w USD)
Słowenia	935
Chorwacja	620
Polska	380
Czechy	370
Węgry	322
Estonia	305
Litwa	265
Słowacja	250
Łotwa	240
Rumunia	115
Bułgaria	110
Rosja	60

Szereg czasowy:

1.01. danego roku	1950	1951	1952
liczba jednostek chorych	100	120	200

lata	<1950-1955)	<1955-1960)	<1960-1965)
liczba nowych zachorowań	80	40	60

Skąd się biorą liczby?

Pomiar – proces empiryczny, w którym przyporządkowuje się liczby poszczególnym kategoriom cechy w taki sposób, aby relacje między liczbami odzwierciedlały relacje między kategoriami cechy.

Skale pomiarowe

1. Nominalna
2. Porządkowa (rangowa)
3. Przedziałowa
4. Ilorazowa
5. Absolutna

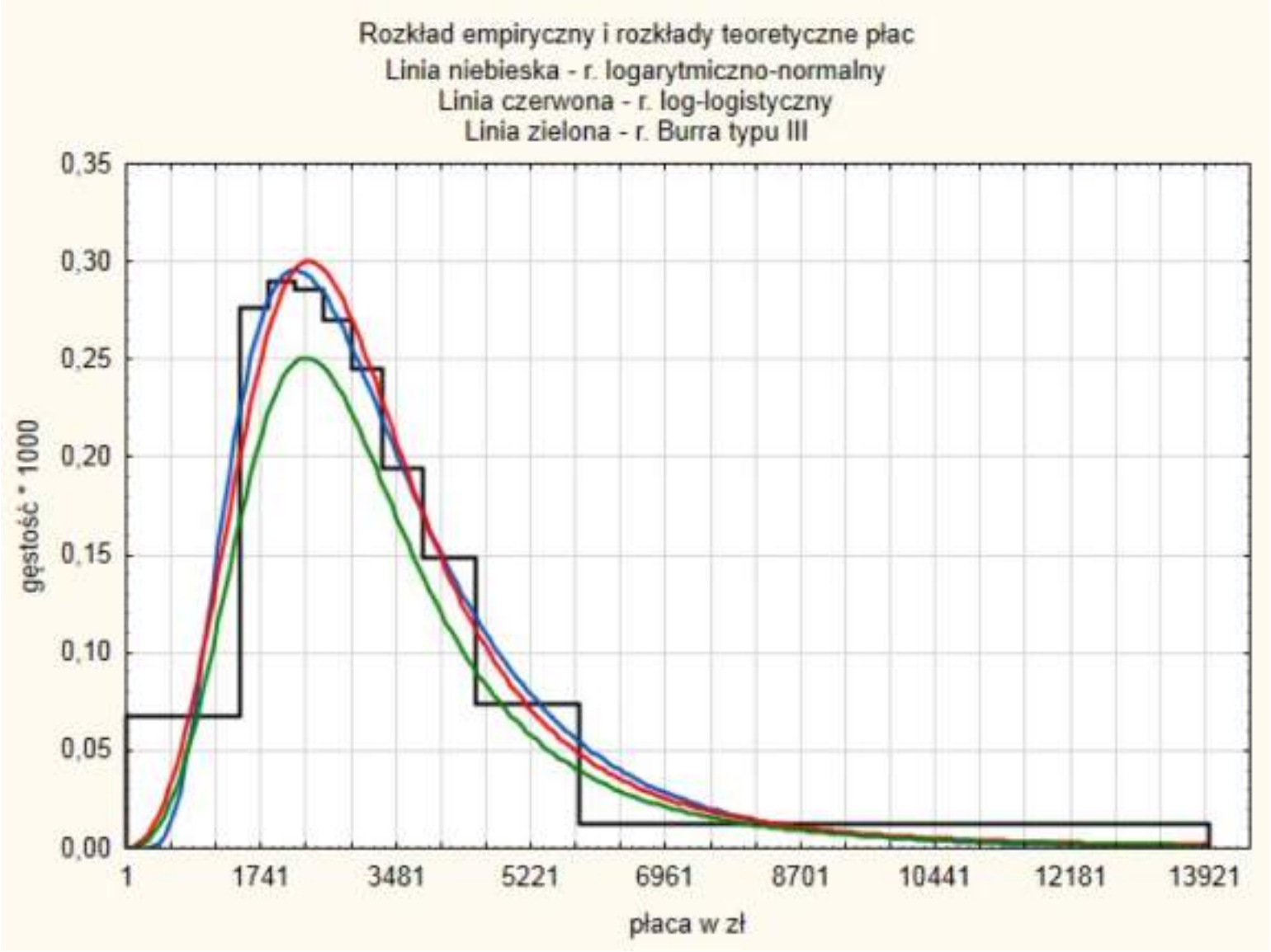
Analiza statystyczna

1. Badanie struktury zjawisk i procesów
2. Badanie zależności zjawisk i procesów
3. Badanie dynamiki zjawisk.

Charakterystyki rozkładu jednej cechy

Rozkład empiryczny (cechy) zmiennej – przyporządkowanie kolejnym wartościom lub wariantom (cechy) zmiennej x_i odpowiadających im liczb lub częstości w_i jednostek posiadających daną wartość lub wariant x_i

Analiza statystyczna – rozkład empiryczny



Analiza statystyczna – rozkład empiryczny

Charakterystyki rozkładu:

1. Miary położenia rozkładu
2. Miary zmienności
3. Miary asymetrii
4. Miary koncentracji

Analiza statystyczna – rozkład empiryczny

Miary położenia rozkładu:

Przeciętne:

- średnie:
  - arytmetyczna, harmoniczna,
  - geometryczna, potęgowa
- przeciętne pozycyjne:
  - modalna, mediana

Kwantyle:

- kwartyle, kwintyle, decyle, centyle

Analiza statystyczna – rozkład empiryczny

Średnia arytmetyczna:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^{k} f_i\, x_i \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^{k} f_i\, x_i'$$

Średnia geometryczna:

$$\bar{i} = \sqrt[n-1]{i_{2/1} \cdot i_{3/2} \cdot i_{4/3} \cdots i_{n/n-1}} = \sqrt[n-1]{\frac{y_2}{y_1} \cdot \frac{y_3}{y_2} \cdots} = \sqrt[n-1]{\frac{y_n}{y_1}}$$

Analiza statystyczna – rozkład empiryczny

Mediana

$$M_e = x_{\frac{n+1}{2}}$$

$$M_e = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2};$$

$$M_e = x_{me} + \frac{h_{me}}{f_{me}}\left[\frac{n}{2} - \sum_{i=1}^{me-1} f_i\right]$$

Modalna

$$M_o = x_m + \frac{f_m - f_{m-1}}{\left(f_m - f_{m-1}\right) + \left(f_m - f_{m+1}\right)} h$$

Analiza statystyczna – rozkład empiryczny

Kwantyl rzędu p

$$q_p = x_p + \frac{h_p}{f_p}\left[pn - \sum_{i=1}^{q-1} f_i \right]$$

Kwartyl - 3
Kwintyl - 4
Decyl - 9
Centyl - 99

Analiza statystyczna – rozkład empiryczny

Siatka centylowa

Analiza statystyczna – rozkład empiryczny

Miary zmienności: bezwzględne i względne

Bezwzględne:

Rozstęp: $\mathrm{R} = \max_{\mathrm{i}}\{\mathrm{x_i}\} - \min_{\mathrm{i}}\{\mathrm{x_i}\}$ $\qquad R_Q = Q_3 - Q_1$

Odchylenie przeciętne: $D = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - \bar{x}\right|$

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

Odchylenie standardowe: $s = \sqrt{s^2}$

Analiza statystyczna – rozkład empiryczny

Miary zmienności: bezwzględne i względne

Względne:

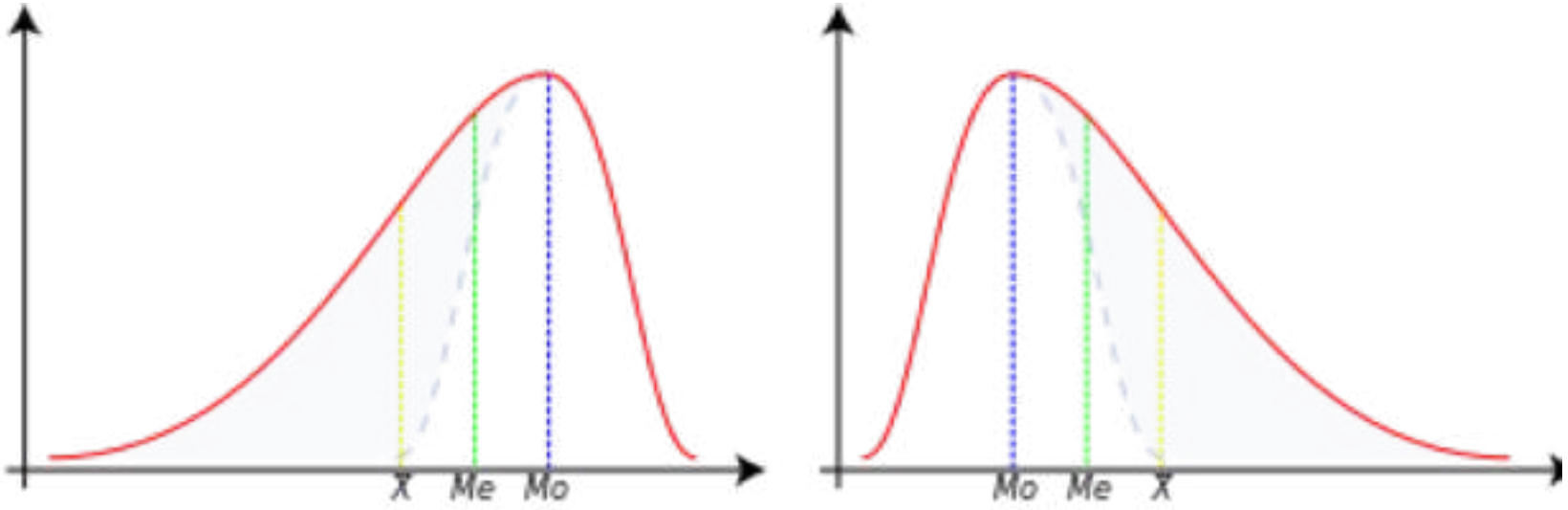
Współczynnik zmienności:

$$V_S = \frac{s}{|\bar{x}|} \cdot 100\% \quad V_D = \frac{D}{|\bar{x}|} \cdot 100\% \qquad V_Q = \frac{Q}{Me}$$

gdzie: $Q = \frac{Q_3 - Q_1}{2}$; Uwaga na moduł w mianowniku.

Analiza statystyczna – rozkład empiryczny

Asymetria rozkładu



Analiza statystyczna – rozkład empiryczny

Miary asymetrii:

$$A_S = \frac{\bar{x} - M_o}{S}; \quad A_S = \frac{3\left(\bar{x} - M_e\right)}{S};$$

$$A_S = \frac{\left(Q_3 - M_e\right) - \left(M_e - Q_1\right)}{\left(Q_3 - Q_1\right)}$$

Analiza statystyczna – rozkład empiryczny

Miary koncentracji:

Kurtoza/Eksces: $K = \frac{M_4}{s^4} - 3$

gdzie: $s^4 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4$

Współczynnik koncentracji Lorenza – współczynnik Giniego

Analiza statystyczna – badanie zależności

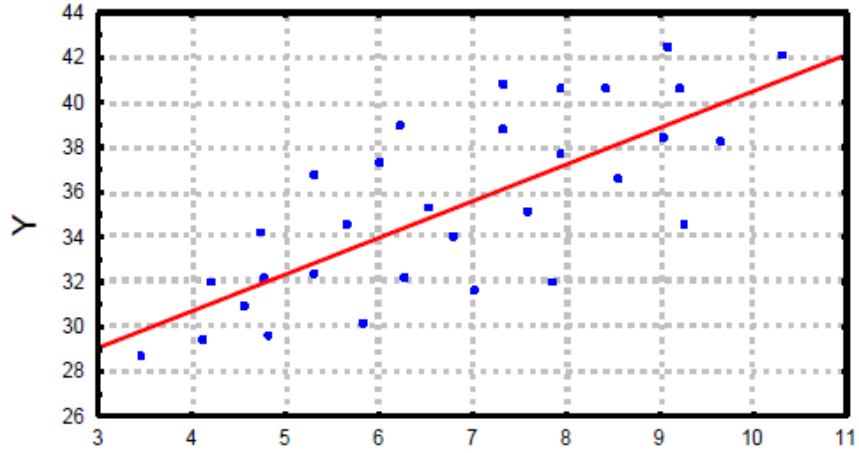
Zależność statystyczna występuje wówczas, gdy istnieje logiczny związek między dwiema lub więcej cechami w badanej zbiorowości potwierdzony danymi statystycznymi.

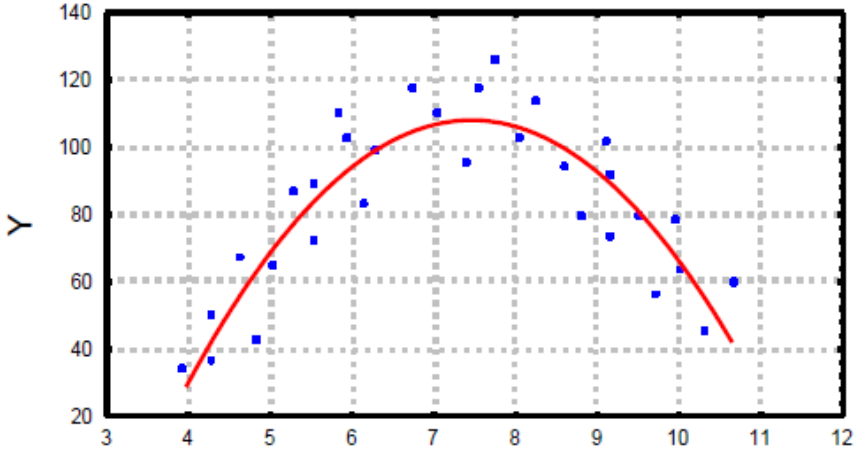
Zależność funkcyjna

Zależność stochastyczna

Zależność korelacyjna

Analiza statystyczna – badanie zależności





Analiza statystyczna – badanie zależności

Pomiar zależności dla zmiennych mierzonych na słabych skalach – **skala nominalna**

Tablica kontyngencji

Zmienna X		**Zmienna Y**				**suma**
		y_1	y_2	…	y_s	
Kategorie zmiennej X	x_1	n_{11}	n_{12}	…	n_{1s}	$n_{1.}$
	x_2	n_{21}	n_{22}	…	n_{2s}	$n_{2.}$
	…	…	…	…	…	…
	x_k	n_{k1}	n_{k2}	…	n_{ks}	$n_{k.}$
Suma		$n_{.1}$	$n_{.2}$	…	$n_{.s}$	n

Analiza statystyczna – badanie zależności

$$\chi^2 = \frac{n(ad - bc)^2}{(a+b)(a+c)(b+d)(c+d)}$$

$$\varphi = \sqrt{\frac{\chi^2}{n}}$$

$$\boldsymbol{\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{s} \frac{\left(n_{ij} - \hat{n}_{ij}\right)^2}{\hat{n}_{ij}}}$$

$$V = \sqrt{\frac{\chi^2}{n \cdot \min(r-1; k-1)}}$$

Analiza statystyczna – badanie zależności

Przykład

Wykształcenie (X)	Rodzaj programu (Y)				Ogółem
	film	teatr	programy rozrywkowe	programy publicystyczne	
Podstawowe	105	10	75	10	200
Średnie	120	60	80	40	300
Wyższe	35	30	15	20	100
Ogółem	260	100	170	70	600

$\chi^2 = 62{,}04$ $\quad$ $V = 0{,}2274$

$$\varphi = 0{,}3216$$

Analiza statystyczna – badanie zależności

Skala porządkowa

Współczynnik Spearmana

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

Współczynnik tau Kendalla

$$K_{tq} = \frac{2R}{1/2 * n(n-1)} - 1$$

gdzie:

R – suma not +1
n – liczba ocenianych obiektów

Analiza statystyczna – badanie zależności

Przykład

Zbadano 5 uczelni ekonomicznych w Polsce ze względu na orientację na studenta oraz selektywność. Wyniki w punktach przedstawia tabela:

Uczelnia	Liczba punktów	
	Orientacja na studenta (X)	Selektywność (Y)
SGH	84	67
UEP	66	62
UEK	76	58
UEW	61	61
UEKat	55	48

Wyznaczyć wartość współczynnika korelacji Spearmana i tau Kendalla

Analiza statystyczna – badanie zależności

r Spearmana:

W pierwszym kroku nadajemy rangi uczelniom osobno ze względu na jedną i drugą cechę

X: 1; 3; 2; 4; 5

Y: 1; 2; 4; 3; 5

Później obliczamy różnice między rangami d_i i podnosimy je do kwadratu, następnie sumujemy i podstawiamy do wzoru:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} = 1 - \frac{6 * 6}{5(25-1)} = 1 - \frac{36}{120} = 0,7$$

Analiza statystyczna – badanie zależności

tau-Kendalla

W pierwszym kroku nadajemy rangi uczelniom osobno ze względu na jedną i drugą cechę

X: 1; 3; 2; 4; 5

Y: 1; 2; 4; 3; 5

Następnie porządkujemy rangi ze względu na jedną z cech

X: 1; 2; 3; 4; 5

Y: 1; 4; 2; 3; 5

Po tym dla każdej rang cechy Y tworzymy pary z następującymi po niej rangami

1 (1;4), (1;2), (1;3), (1;5)

4 (4;2), (4;3), (4;5)

2 (2;3), (2;5)

3 (3;5)

Analiza statystyczna – badanie zależności

Jeśli poprzednik jest mniejszy od następnika to nadajemy notę 1 (pary oznaczone na zielono) w przeciwnym przypadku notę -1 (dla rang powiązanych mogłoby zaistnieć zero). Zliczamy jedynki; jest ich 8

$$K_{tq} = \frac{2*8}{1/2*5*(5-1)} - 1 = 0{,}6$$

Analiza statystyczna – badanie zależności

Skala co najmniej przedziałowa

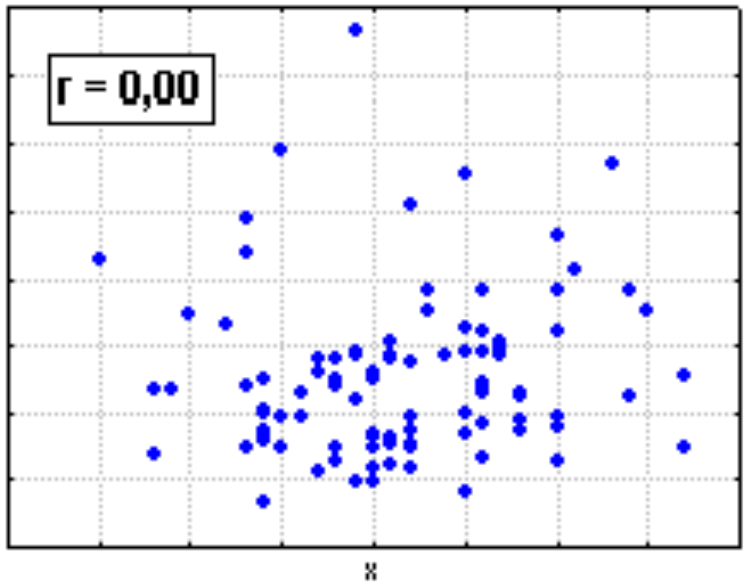
Kowariancja

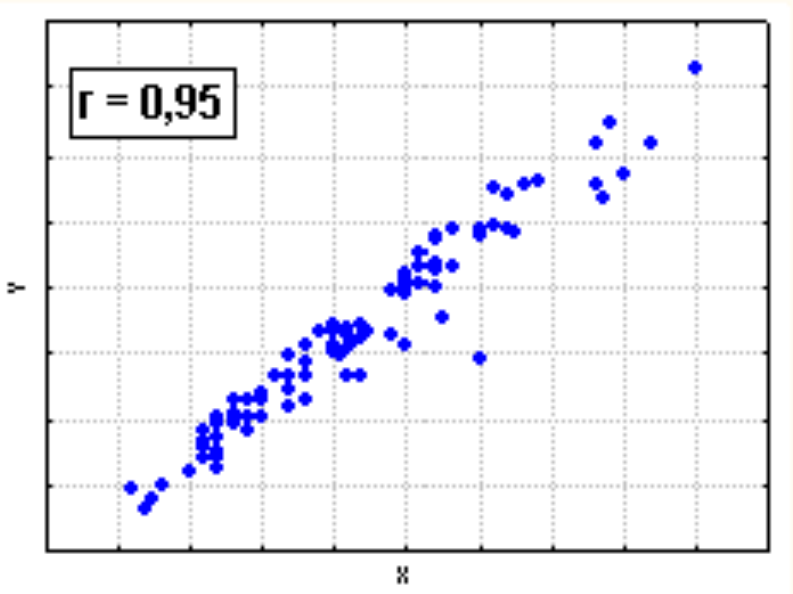
$$cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

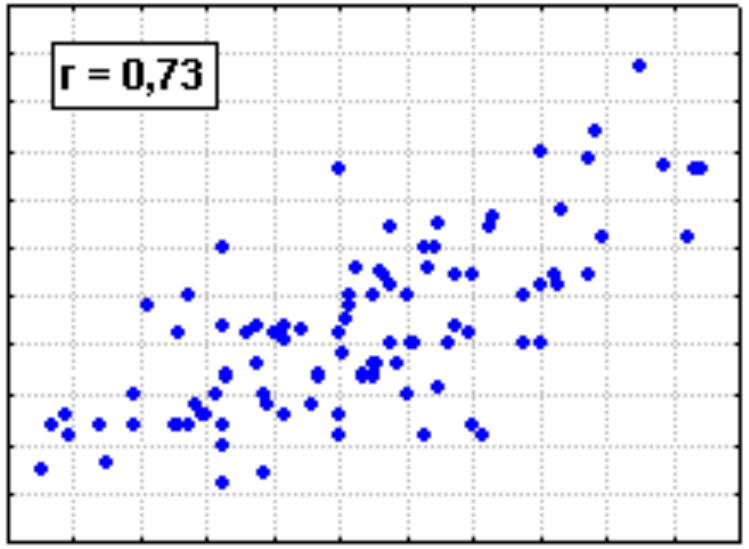
Współczynnik korelacji liniowej Pearsona

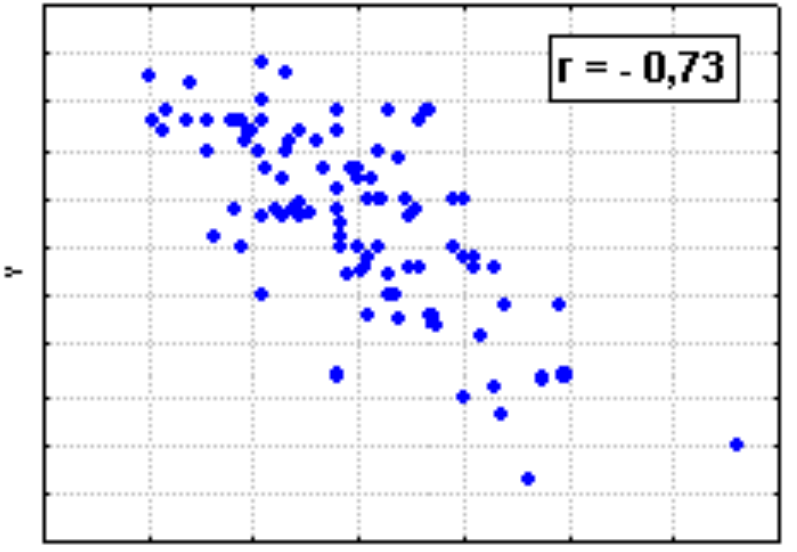
$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Analiza statystyczna – badanie zależności









Analiza statystyczna – badanie zależności

Funkcje regresji

Funkcja regresji – analityczny wyraz przyporządkowania średnich wartości zmiennej objaśnianej (zależnej) konkretnym wartościom zmiennych objaśniających (niezależnych).

I rodzaju – to funkcja, która realizacjom zmiennych objaśniających przypisuje średnie warunkowe zmiennej objaśnianej.

Dla jednej zmiennej objaśniającej:

$$E(Y|X = x_i) = g(x_1)$$

Analiza statystyczna – badanie zależności

Funkcja regresji II rodzaju –

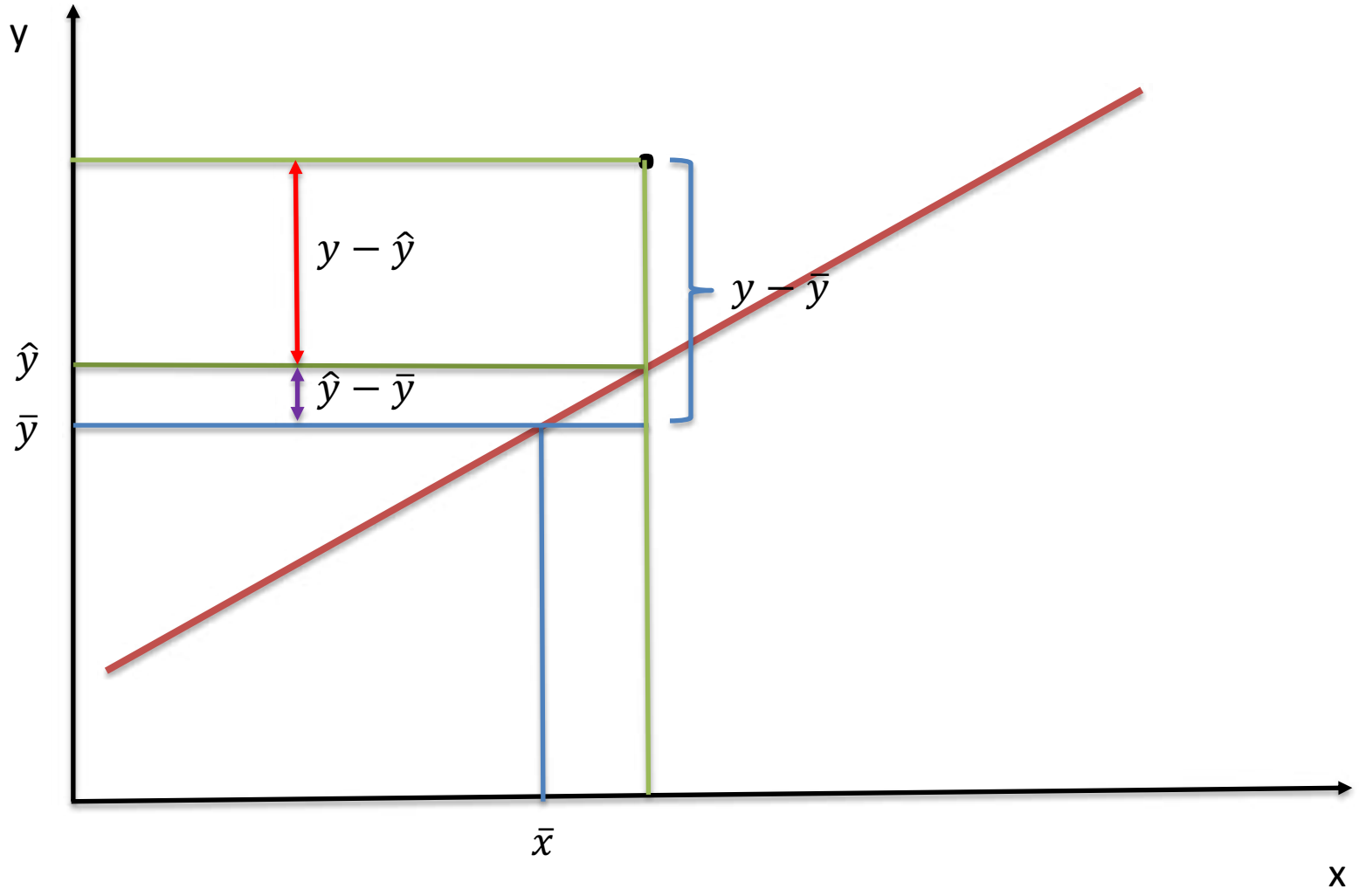
$$\hat{y} = a_0 + a_1 x + \varepsilon$$

MNK – Metoda Najmniejszych Kwadratów

$$\begin{cases} na_0 + a_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ a_0 \sum_{i=1}^{n} x_i + a_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \end{cases} \rightarrow \begin{cases} a_1 = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \\ a_0 = \bar{y} - a_1 \bar{x} \end{cases} \rightarrow \begin{cases} a_1 = r \dfrac{s_y}{s_x} \\ a_0 = \bar{y} - a_1 \bar{x} \end{cases}$$

Analiza statystyczna – badanie zależności

Funkcja regresji II rodzaju



Analiza statystyczna – badanie zależności

Funkcja regresji II rodzaju

$$\varphi^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$\varphi^2 + R^2 = 1$$

$$s_\varepsilon = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Analiza statystyczna – badanie zależności

Zależność wielu cech

$$R = \begin{bmatrix} 1 & r_{xy} & r_{xz} \\ r_{yx} & 1 & r_{yz} \\ r_{zx} & r_{zy} & 1 \end{bmatrix}$$

$$r_{xy.z} = \frac{r_{xy} - r_{xz} r_{yz}}{\sqrt{\left(1 - r_{xz}^2\right)\left(1 - r_{yz}^2\right)}}$$

$$r_{yz.x} = \frac{r_{yz} - r_{yx} r_{zx}}{\sqrt{\left(1 - r_{yx}^2\right)\left(1 - r_{zx}^2\right)}}$$

$$r_{xz.y} = \frac{r_{xz} - r_{xy} r_{zy}}{\sqrt{\left(1 - r_{xy}^2\right)\left(1 - r_{zy}^2\right)}}$$

$$R_{y.xz} = \sqrt{\frac{r_{xy}^2 + r_{yz}^2 - 2 r_{xy} r_{xz} r_{yz}}{1 - r_{xz}^2}}$$

Analiza statystyczna – badanie zależności

Zależność wielu cech

$$\hat{y} = a_{y0} + a_{yx}x + a_{yz}z$$

$$\begin{cases} a_{yx} = \dfrac{r_{yx} - r_{yz}r_{xz}}{1 - r_{xz}^2} \cdot \dfrac{s_y}{s_x} \\ a_{yz} = \dfrac{r_{yz} - r_{yx}r_{zx}}{1 - r_{zx}^2} \cdot \dfrac{s_y}{s_z} \\ a_{yo} = \bar{y} - a_{yx}\bar{x} - a_{yz}\bar{z} \end{cases}$$