Wprowadzenie do statystyki

cz. 1

Podstawowe pojęcia i opis statystyczny

1::

Statystyka

- zbiór przetworzonych i zsyntetyzowanych danych liczbowych,
- nauka o ilościowych metodach badania zjawisk masowych,
- zmienna losowa będąca funkcją próby.

Podstawowe pojęcia:

- populacja (zbiorowość statystyczna),
- jednostka statystyczna,
- próba.

Cechy:

- ilościowe (mierzalne),
  - skokowe (dyskretne),
  - quasi ciągłe,
  - ciągłe,
- jakościowe (niemierzalne).

2::

Jednostka statystyczna: jednostki statystyczne w danej populacji różnią się od innych jednostek spoza danej populacji poprzez swoje własności wspólne (cechy stałe), jednocześnie różnią się między sobą cechami (cechy zmienne)

Cechy statystyczne – właściwości jednostek statystycznych Cechy stałe – jednakowe dla wszystkich jednostek badania: rzeczowa (co? kto? jest badane/y) przestrzenna (gdzie?) czasowa (kiedy?)

Cecha ilościowa (quantitative), np. numer buta, wiek; Cecha jakościowa (qualitative), np. płeć

Skale

- słabe, niemetryczne, jakościowe:
  - nominalna (kategoryjna, wariantowa):
    - dwudzielna (dychotomiczna): np. kobieta/mężczyzna, tak/nie, 0/1,
    - wielodzielna (politomiczna): np. kolor, marka, gatunek,
  - porządkowa (rangowa),
    - np. oceny, preferencje, nie/raczej nie/nie mam zdania/raczej tak/tak
- silne, metryczne, ilościowe:
  - przedziałowa (interwałowa),
    - np. temperatura w skali Celsjusza,
  - ilorazowa (stosunkowa),
    - np. temperatura w skali Kelvina,
- zamknięte/otwarte,
- Zmienna ze skali mocniejszej może być rozpatrywana w skali słabszej, ale nie odwrotnie.

3::
Four measurement scales: nominal (nominalna), ordinal (porządkowa), interval (interwałowa) and ratio (ilorazowa)
Skale interwałowe nie nie mają **prawdziwego zera** (brak cechy), co powoduje subtelny problem przy mnożeniu. Przykładowo temperatura w skali Celsjusza jest skalą interwałową: $2 \times 0 = 0C$ (a powinno być 2 razy cieplej niż przy temperaturze 0C; W skali Kelvina tak jest $2 \times 273{,}15 = 546{,}3$ ale dla zera bezwzględego $2 \times 0 = 0K$ się zgadza z tym czego należało oczekiwać, tj. że mnożenie przez 0 daje 0/brak cechy)

Badanie statystyczne

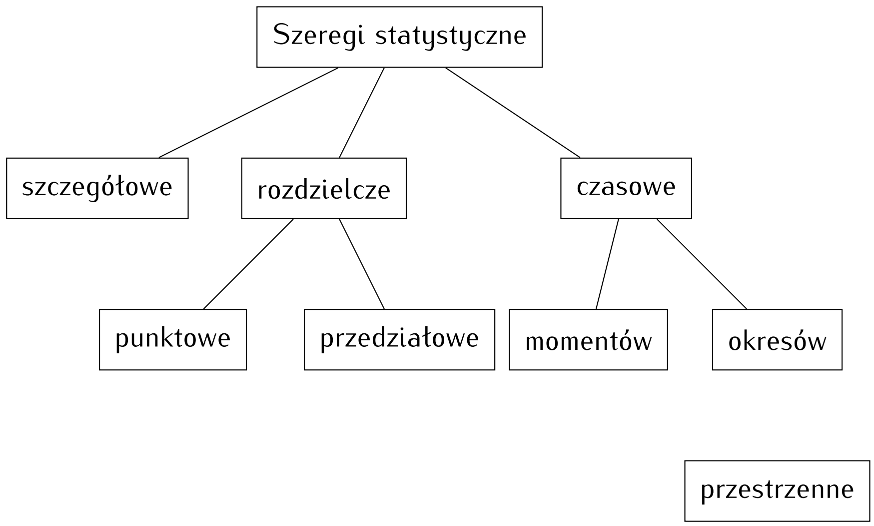
- pełne, częściowe,
- ciągłe, okresowe, doraźne,
- badania ankietowe, monograficzne, próbkowe (metoda reprezentacyjna).
- szacunki: interpolacja, ekstrapolacja,

Etapy badania statystycznego:

- przygotowanie badania (cel, populacja, jednostka, metoda),
- obserwacja statystyczna,
- opracowanie i prezentacja materiału statystycznego,
- opis lub wnioskowanie statystyczne.

4::
Główny Urząd Statystyczny (GUS) – podległy Prezesowi RM urząd zajmujący się zbieraniem/udostępnianiem informacji statystycznych na temat różnych dziedzin życia publicznego/prywatnego.
W 2016 wydatki GUS wyniosły 409,7 mln PLN. **Średnie zatrudnienie** (co to?) w przeliczeniu na pełne etaty wyniosło 5834 osoby.
EUROSTAT – urząd zajmujący się sporządzaniem prognoz/analiz statystycznych dot. obszaru UE/EFTA, i koordynowaniem/monitorowaniem prac narodowych urzędów statystycznych (unifikacja metod badań/klasyfikacji)

Szeregi statystyczne



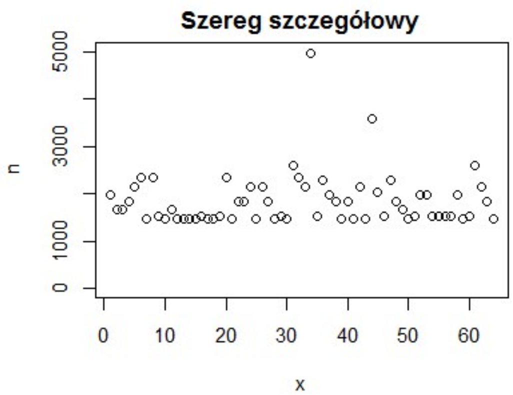
5::

Dane statystyczne możemy w ogólności podzielić na **dane przekrojowe** (cross-sectional data) – wiele jednostek obserwowanych w jednym momencie/okresie czasu; **szeregi czasowe** (time-series) – jedna jednostka obserwowana w wielu momentach/okresach czasu; **dane panelowe** (panel data, cross-sectional time-series data) – wiele jednostek obserwowanych w wielu m/o czasu.

Mówiąc inaczej: **Szeregi czasowe**: dane oznaczone stemplem czasu; **Dane przestrzenne** (spatial): dane oznaczone pozycją na powierzchni ziemi.

Szereg szczegółowy (wyliczający, dane indywidualne)

3590, 1520, 2340, 1460, 1990, 1830, 1830, 1520, 1460, 1990, 2612, 1520, 2340, 2145, 1460, 1830, 1520, 2299, 1460, 1460, 1520, 2145, 1990, 1830, 1990, 1830, 1460, 1460, 1660, 1660, 1830, 1990, 1460, 1520, 1830, 1830, 1460, 1460, 1460, 1460, 1660, 1520, 2340, 1460, 2045, 1520, 2145, 2145, 2299, 1660, 1520, 2340, 1520, 1520, 1460, 2145, 2145, 1460, 1460, 1520, 1460, 1460, 4960, 2612



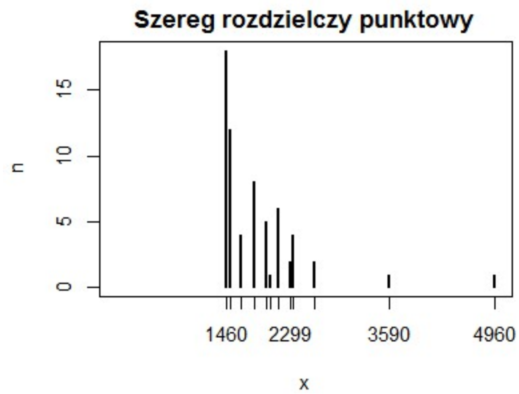
6::

Individual/Discrete/Continuous (Data) Series;

W innym aspekcie używa się **individual** vs **organizational** data, co oznacza dane na poziomie indywidualnym albo na poziomie organizacji (przedsiębiorstwa)

Szereg rozdzielczy punktowy

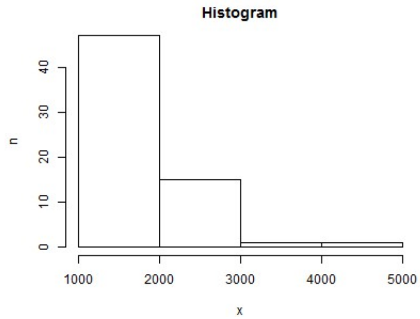
i	x_i	n_i
1	1460	18
2	1520	12
3	1660	4
4	1830	8
5	1990	5
6	2045	1
7	2145	6
8	2299	2
9	2340	4
10	2612	2
11	3590	1
12	4960	1
Razem	×	64



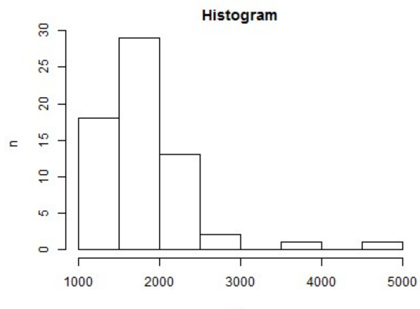
7::
Discrete (Data) Series
Rozkład częstości (rozkład empiryczny zmiennej, szereg rozdzielczy): przyporządkowanie kolejnym wartościom zmiennej (xi) odpowiadających im liczebności (ni) lub udziałów. Przedstawia **strukturę zbiorowości** dla określonej cechy (stąd analiza struktury); **Frequency table/distribution** vs **Relative frequency table/distribution**
Tablica statystyczna: Część liczbowa + część opisowa: tytuł; boczek (nazwy wierszy); główka (n. kolumn); źródło danych; ewentualne uwagi/objaśnienia.

Szereg rozdzielczy przedziałowy

x_i	n_i
1000 – 1999	47
2000 – 2999	15
3000 – 3999	1
4000 – 4999	1



x_i	n_i
mniej niż 1500	18
1500 – 2000	29
2000 – 2500	13
2500 – 3000	2
3000 i więcej	2



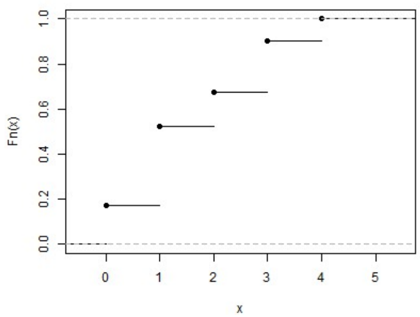
8::
Continuous (Data) Series
Zasady grupowania danych: 1) Równe rozpiętości przedziałów; 2) Niezerowe liczebności wszystkich przedziałów; 3) Zdefiniowane wszystkie końce przedziałów; 4) Niedominująca liczebność przedziału; 5) „Dobrze wyglądające" końce przedziałów (kończące się na zero/pięć/liczbą całkowitą)
Liczba przedziałów: określona konwencjami w domenie zastosowań/celem badania. Raczej nie mniej niż 6–8. (In case of doubt copy good reference.)

Liczebność skumulowana, dystrybuanta empiryczna

- Liczebność skumulowana: liczba obserwacji nie większa od danej wartości cechy:
$$n(x) = \sum_{i:x_i \leqslant x} n_i$$
- Dystrybuanta empiryczna: frakcja (część) obserwacji nie większa od danej wartości cechy:
$$F_n(x) = \sum_{i:x_i \leqslant x} f_i = \frac{n(x)}{n}.$$

9::
cumulative frequency (table)
empirical **distribution function**

Liczba zadań	Liczebność	Częstość	Skumulowana liczebność	Dystrybuanta empiryczna
x_i	n_i	f_i	$n(x_i)$	$F_n(x_i)$
0	7	0.175	7	0.175
1	14	0.35	21	0.525
2	6	0.15	27	0.675
3	9	0.225	36	0.900
4	4	0.1	40	1
5	0	0	40	1
Suma	40	1	×	×



Miary opisu struktury

- poziom przeciętny (położenie, średni poziom wartości): średnia, mediana, dominanta (moda),
- zróżnicowanie (rozproszenie, dyspersja, zmienność): wariancja, odchylenie standardowe, odchylenie przeciętne, odchylenie ćwiartkowe, rozstęp,
- asymetria (skośność): skośność, współczynnik Yule'a-Kendalla,
- koncentracja (spłaszczenie): kurtoza, współczynnik Giniego, entropia,

11::
Mean (or Average); Median; Mode
Dispersion: variance, standard deviation, range (rozstęp)
Skewness (positive/negative)
Concentration

Średnia arytmetyczna

- dla danych indywidualnych:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + \cdots + x_n}{n},$$

- dla szeregów rozdzielczych punktowych:

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

- dla szeregów rozdzielczych przedziałowych:

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

- $x_{\min} \leqslant \bar{x} \leqslant x_{\max}$, $\sum_{i=1}^{n} x_i = n\bar{x}$, $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$ (lub $\sum_{i=1}^{k}(x_i - \bar{x})n_i = 0$ lub $\sum_{i=1}^{k}(\dot{x}_i - \bar{x})n_i = 0$),

x_i	n_i	$\dot{x}_i$	$\dot{x}_i n_i$
$(20, 25]$	11	22.5	247.5
$(25, 30]$	23	27.5	632.5
$(30, 35]$	16	32.5	520.0
Σ	50	$\times$	1400

$$\bar{x} = \frac{1400}{50} = 28.$$

Mediana – wartość środkowa, kwantyle

- Medianą z próby Me nazywamy taką wartość, że co najmniej połowa obserwacji ma wartość nie większą niż Me i równocześnie co najmniej połowa obserwacji ma wartość nie mniejszą niż Me.
- Inaczej: jest to najmniejsza wartość, dla której

$$F_n(Me) \geqslant \frac{1}{2} \qquad \text{lub rownoważnie} \qquad n(Me) \geqslant \frac{n}{2}.$$

- dla szeregów szczegółowych:

$$Me = \begin{cases} x_{\frac{n+1}{2}} & \text{gdy } n \text{ jest nieparzyste,} \\ \dfrac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2} & \text{gdy } n \text{ jest parzyste.} \end{cases}$$

- kwantylem empirycznym rzędu p, gdzie $0 < p < 1$, nazywamy najmniejszą wartość q_p cechy, dla której zachodzi:

$$F_n(q_p) \geqslant p.$$

Kwantyle, kwartyle

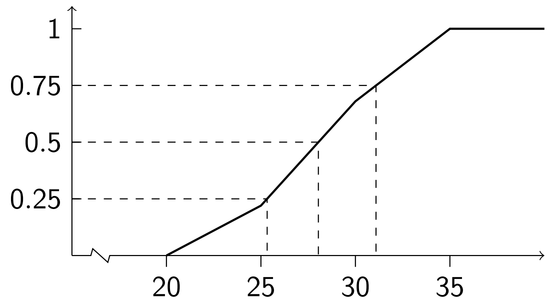
- dla szeregów przedziałowych kwantyle aproksymujemy wzorem

$$q_p \approx x_{0p} + [pn - n(x_{0p})] \cdot \frac{h_p}{n_p} = x_{0p} + [p - F_n(x_{0p})] \cdot \frac{n \cdot h_p}{n_p}$$

  - p – rząd kwantyla,
  - x_{0p} – dolna granica przedziału kwantyla: $F_n(x_{0p}) \leqslant p < F_n(x_{1p})$,
  - n_p – liczebność przedziału kwantyla,
  - h_p – szerokość przedziału kwantyla,
  - $n(x_{0p})$ – liczebność skumulowana w przedziale poprzedzającym przedział kwantyla,
  - $F_n(x_{0p})$ – wartość dystrybuanty empirycznej na końcu przedziału poprzedzającego przedział kwantyla,
- kwartyle: $Q_1 = q_{0.25}$, $Q_2 = Me = q_{0.5}$, $Q_3 = q_{0.75}$,
- w szczególności dla $p = \frac{1}{2}$ otrzymujemy wzór dla mediany:

$$Me \approx x_{0M} + \left[\frac{n}{2} - n(x_{0M})\right] \cdot \frac{h_M}{n_M} = x_{0M} + \left[\frac{1}{2} - F_n(x_{0M})\right] \cdot \frac{n \cdot h_M}{n_M}$$

x_i	n_i	$F_n(x_i)$
$(20, 25]$	11	0.22
$(25, 30]$	23	0.68
$(30, 35]$	16	1



$$Q_1 = 25 + [0.25 - 0.22] \cdot \frac{50 \cdot 5}{23} \approx 25.33,$$
$$Me = 25 + [0.5 - 0.22] \cdot \frac{50 \cdot 5}{23} \approx 28.04,$$
$$Q_3 = 30 + [0.75 - 0.68] \cdot \frac{50 \cdot 5}{16} \approx 31.09.$$

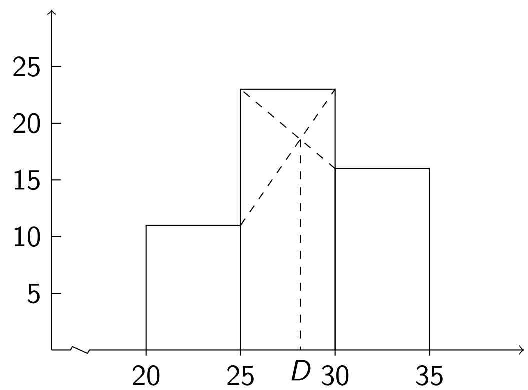
Dominanta

- Dominantą (modą, modalną) nazywamy wartość zmiennej, która występuje najczęściej,
- można wyznaczać tylko w rozkładach jednomodalnych,
- w szeregach szczegółowych i punktowych jest to wartość cechy odpowiadająca największej liczebności,
- w szeregach przedziałowych aproksymujemy ją wzorem:

$$D \approx x_{0D} + \frac{n_D - n_{D-1}}{(n_D - n_{D-1}) + (n_D - n_{D+1})} h_D,$$

  - x_{0D} – dolna granica przedziału dominanty (o największej liczebności),
  - n_D, n_{D-1}, n_{D+1} – odpowiednio liczebność przedziału dominanty, przedziału poprzedniego i następnego,
  - h_D – rozpiętość przedziału dominanty.
- „Wzór Pearsona”: $Me \approx \frac{1}{3}D + \frac{2}{3}\bar{x}$.

x_i	n_i
$(20, 25]$	11
$(25, 30]$	23
$(30, 35]$	16



$$D = 25 + \frac{23-11}{(23-11)+(23-16)} \cdot 5 \approx 28.16.$$

Uwaga: w przypadku przedziałów o różnej szerokości liczebności n_i zastępujemy gęstościami: $g_i = n_i / h_i$.

Wariancja

- dla danych indywidualnych:

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

- dla szeregów rozdzielczych punktowych:

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

- dla szeregów rozdzielczych przedziałowych:

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

poprawka Shepparda: $\bar{S}^2 = S^2 - \frac{h^2}{12}$,

Variance

- odchylenie standardowe: $S = \sqrt{S^2}$,
- współczynnik zmienności:

$$V = \frac{S}{\bar{x}},$$

- odchylenie przeciętne:

$$d = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}| \qquad \left(d = \frac{1}{n}\sum_{i=1}^{k}|x_i - \bar{x}| \cdot n_i\right),$$

- odchylenie ćwiartkowe:

$$Q = \frac{Q_3 - Q_1}{2},$$

- pozycyjny współczynnik zmienności: $V = \dfrac{Q}{Me}$,
- rozstęp: $R = x_{\max} - x_{\min}$,
- rozstęp ćwiartkowy (międzykwartylowy): $IQR = Q_3 - Q_1$,

20::

Odchylenie standardowe (*standard deviation*); **Odchylenie przeciętne** (*average absolute deviation*); Rozstęp ćwiartkowy, rozstęp międzykwartylowy (interquartile range (IQR), midspread); Odchylenie ćwiartkowe Odchylenie ćwiartkowe (*Quartile coefficient of dispersion*)

$$\bar{x} = \frac{1400}{50} = 28,$$

x_i	n_i	$\dot{x}_i$	$\dot{x}_i n_i$	$(\dot{x}_i - \bar{x})^2$	$(\dot{x}_i - \bar{x})^2 n_i$
$(20, 25]$	11	22.5	247.5	30.25	332.75
$(25, 30]$	23	27.5	632.5	0.25	5.75
$(30, 35]$	16	32.5	520.0	20.25	324.00
Σ	50	$\times$	1400	$\times$	662.50

$$S^2 = \frac{662.5}{50} = 13.25,$$
$$S = \sqrt{13.25} \approx 3.64,$$
$$Q \approx \frac{31.09 - 25.33}{2} = 2.88,$$

Równość wariancyjna

- mamy informacje o k grupach: ich liczebności n_i, średnie $\bar{x}_i$ oraz wariancje (wewnątrzgrupowe) S_i^2,
- średnia ogólna, to średnia ważona liczebnościami:

$$\bar{x} = \frac{\sum_{i=1}^{k} \bar{x}_i \cdot n_i}{\sum_{i=1}^{k} n_i}.$$

- liczbę

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

  nazywamy wariancją międzygrupową,
- wariancja ogólna wyraża się wzorem:

$$S^2 = \overline{S_i^2} + S^2(\bar{x}_i) = \frac{\sum_{i=1}^{k} S_i^2 \cdot n_i}{\sum_{i=1}^{k} n_i} + \frac{\sum_{i=1}^{k} (\bar{x} - \bar{x}_i)^2 \cdot n_i}{\sum_{i=1}^{k} n_i}.$$

- Moment zwykły rzędu r:

$$m_r = \frac{1}{n}\sum_{i=1}^{n} x_i^r, \qquad \left(m_k = \frac{1}{n}\sum_{i=1}^{k} x_i^r \cdot n_i\right)$$

- Moment centralny rzędu r:

$$M_r = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^r, \qquad \left(M_k = \frac{1}{n}\sum_{i=1}^{k}(x_i - \bar{x})^r \cdot n_i\right)$$

Asymetria

- klasyczny współczynnik asymetrii:

$$\gamma_3 = \frac{M_3}{S^3},$$

- współczynnik Yule'a-Kendalla:

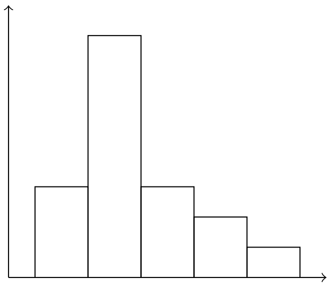
$$A_Q = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_2) + (Q_2 - Q_1)},$$

- współczynnik skośności Pearsona:

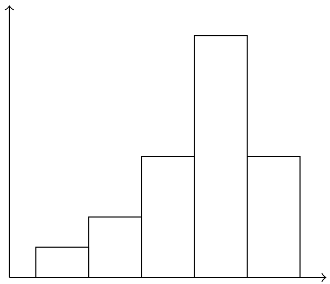
$$A_S = \frac{\bar{x} - D}{S},$$

Skewness

Asymetria

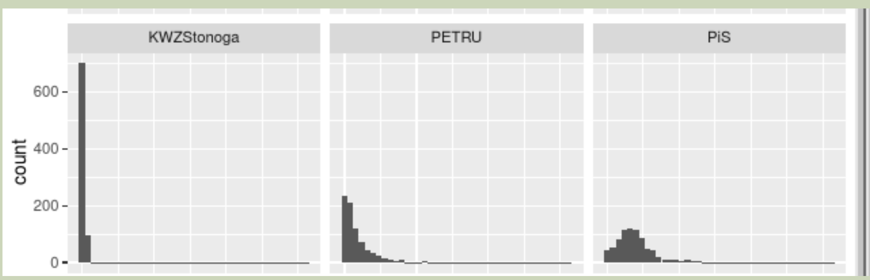


wartości dodatnie
asymetria prawostronna



wartości ujemne
asymetria lewostronna

25::



Krzywa koncentracji Lorenza, współczynnik Giniego

- linia łamana powstała z połączenia punktów o współrzędnych:

$$(x_0, y_0) = (0, 0), \qquad (x_j, y_j) = \left(\frac{j}{n}, \frac{\sum_{i=1}^{j} z_i}{\sum_{i=1}^{n} z_i} \right), \quad j = 1, \ldots, n.$$

- dla szeregu rozdzielczego: $(x_0, y_0) = (0, 0)$,

$$(x_j, y_j) = \left(\frac{\sum_{i=1}^{j} n_i}{n}, \frac{\sum_{i=1}^{j} z_i \cdot n_i}{\sum_{i=1}^{k} z_i \cdot n_i} \right), \quad j = 1, \ldots, k.$$

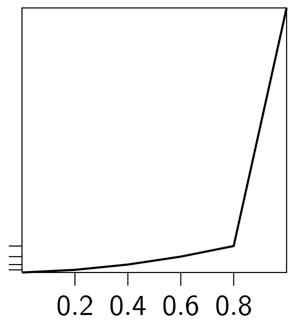
- podwojone pole obszaru między krzywą Lorenza a przekątną kwadratu jednostkowego nazywamy współczynnikiem koncentracji Giniego:

$$G = \frac{\sum_{j=1}^{n}(2j - n - 1)z_j}{n^2 \cdot \bar{z}}.$$

- współczynnik Giniego przyjmuje wartości z przedziału $[0, 1]$, gdzie 0 oznacza rozkład równomierny, a wartość 1 – rozkład skupiony w pojedynczej wartości,

26::
Herfindahl-Hirschman Index (HHI): commonly accepted measure of market concentration. It is calculated by squaring the market share of each firm competing in a market, and then summing the resulting numbers, and can range from close to zero to 10,000. The U.S. Department of Justice uses the HHI for evaluating potential mergers issues. (cf https://www.investopedia.com/terms/h/hhi.asp)
Przykład: na rynku udziały 20 podmiotów są następujące:
$F_1 = 40\%, F_2 = 30\%, F_3 = 14\%, F_4$–$F_20 = 1\%$ każda;
HHI $= 40^2 + 30^2 + \ldots 1^2 =$ = 1,600 + 900 + 196 + 20 = 2,716

j	z_j	$\sum_{i=1}^{j} z_i$	x_j	y_j	$2j - n - 1$	$(2j - n - 1)z_j$
1	1	1	0.2	0.01	–4	–4
2	2	3	0.4	0.03	–2	-6
3	3	6	0.6	0.06	0	0
4	4	10	0.8	0.10	2	20
5	90	100	1	1	4	400



$$G = \frac{-4 - 6 + 20 + 400}{5^2 \cdot 20} = \frac{410}{500} = 0.82.$$

Analiza dynamiki

- Szereg czasowy:
$$y_1, y_2, \ldots, y_{n-1}, y_n.$$
- y_t – poziom (wartość) badanego zjawiska w okresie lub chwili t.
- Szeregi czasowe momentów dotyczą zasobów.
- Szeregi czasowe okresów dotyczą strumieni.
- Strumienie możemy agregować: na przykład dane miesięczne do kwartalnych lub rocznych.
- Zasobów nie możemy agregować w ten sposób.
- Szereg czasowy powinien zawierać wielkości jednorodne i porównywalne.

Składniki szeregu czasowego

Dekompozycja – wyodrębnianie składowych szeregu czasowego:

- tendencja rozwojowa (trend) T,
- wahania okresowe (sezonowe, koniunkturalne) S,
- wahania przypadkowe P.

Składowe mogą się łączyć poprzez:

- dodawanie – szereg addytywny:

$$Y = T + S + P,$$

- mnożenie – szereg multiplikatywny,

$$Y = T \cdot S \cdot P.$$

Średnia ruchoma arytmetyczna (SMA)

- krocząca, o długości $2q+1$:

$$\bar{y}_t = \frac{1}{2q+1}\sum_{r=-q}^{q} y_{t+r}, \qquad t = q+1, q+2, \dots, n-q,$$

- scentrowana (chronologiczna), o długości $2q$:

$$\bar{y}_t = \frac{1}{2q}\left[\frac{1}{2}y_{t-q} + \sum_{r=-q+1}^{q-1} y_{t+r} + \frac{1}{2}y_{t+q}\right], \ t = q+1, q+2, \dots, n-q,$$

Przyrosty względne i absolutne

- przyrosty absolutne (bezwzględne) o podstawie stałej:

$$y_1 - y_{t_0},\ y_2 - y_{t_0},\ y_3 - y_{t_0},\ \ldots,\ y_{n-1} - y_{t_0},\ y_n - y_{t_0},$$

  gdzie podstawą jest okres t_0,
- przyrosty absolutne (bezwzględne) łańcuchowe:

$$y_2 - y_1,\ y_3 - y_2,\ y_4 - y_3,\ \ldots,\ y_{n-1} - y_{n-2},\ y_n - y_{n-1},$$

- przyrosty względne o podstawie stałej w okresie t_0:

$$\frac{y_1 - y_{t_0}}{y_{t_0}},\ \frac{y_2 - y_{t_0}}{y_{t_0}},\ \frac{y_3 - y_{t_0}}{y_{t_0}},\ \ldots,\ \frac{y_{n-1} - y_{t_0}}{y_{t_0}},\ \frac{y_n - y_{t_0}}{y_{t_0}},$$

- przyrosty względne łańcuchowe:

$$\frac{y_2 - y_1}{y_1},\ \frac{y_3 - y_2}{y_2},\ \frac{y_4 - y_3}{y_3},\ \ldots,\ \frac{y_{n-1} - y_{n-2}}{y_{n-2}},\ \frac{y_n - y_{n-1}}{y_{n-1}},$$

Wskaźniki dynamiki (indeksy)

- stosunek wielkości badanego zjawiska w danym okresie (momencie) – **badanym, sprawozdawczym** – do wielkości tego samego zjawiska w innym okresie (momencie) – **bazowym, podstawowym** – przyjętym za podstawę porównań,
- indeksy jednopodstawowe o podstawie stałej w okresie t_0:

$$\frac{y_1}{y_{t_0}}, \frac{y_2}{y_{t_0}}, \ldots, \frac{y_{n-1}}{y_{t_0}}, \frac{y_n}{y_{t_0}},$$

- indeksy łańcuchowe:

$$\frac{y_2}{y_1}, \frac{y_3}{y_2}, \ldots, \frac{y_{n-1}}{y_{n-2}}, \frac{y_n}{y_{n-1}},$$

- na indeksach wygodniej wykonuje się operacje algebraiczne,

Zamiany indeksów

- przyrosty względne na indeksy:

$$\text{łańcuchowe:} \qquad \frac{y_i}{y_{i-1}} = \frac{y_i - y_{i-1}}{y_{i-1}} + 1,$$

$$\text{jednopodstawowe:} \qquad \frac{y_i}{y_{t_0}} = \frac{y_i - y_{t_0}}{y_{t_0}} + 1,$$

- indeksy na przyrosty względne:

$$\text{łańcuchowe:} \qquad \frac{y_i - y_{i-1}}{y_{i-1}} = \frac{y_i}{y_{i-1}} - 1,$$

$$\text{jednopodstawowe:} \qquad \frac{y_i - y_{t_0}}{y_{t_0}} = \frac{y_i}{y_{t_0}} - 1,$$

Zamiana podstawy indeksu jednopodstawowego

- należy wszystkie indeksy podzielić przez wskaźnik wyrażający zmianę zjawiska między okresem starej (s) a nowej podstawy (n):

$$\frac{y_i}{y_n} = \frac{y_i}{y_s} \cdot \frac{y_s}{y_n} = \frac{y_i}{y_s} : \frac{y_n}{y_s}.$$

- zmiana indeksu jednopodstawowego na łańcuchowy:

$$\frac{y_i}{y_{t_0}} : \frac{y_{i-1}}{y_{t_0}} = \frac{y_i}{y_{i-1}},$$

Zamiana indeksów łańcuchowych na jednopodstawowe

- w okresie bazowym indeks jednopodstawowy wynosi: $\frac{y_{t_0}}{y_{t_0}} = 1$,
- w okresie następującym bezpośrednio po okresie bazowym indeks jednopodstawowy jest równy łańcuchowemu: $\frac{y_{t_0+1}}{y_{t_0}}$,
- kolejne indeksy po okresie bazowym otrzymujemy mnożąc poprzedni indeks jednopodstawowy przez bieżący indeks łańcuchowy:

$$\frac{y_i}{y_{t_0}} = \frac{y_{i-1}}{y_{t_0}} \cdot \frac{y_i}{y_{i-1}},$$

- indeksy przed okresem bazowym obliczamy dzieląc poprzedni indeks jednopodstawowy przez bieżący indeks łańcuchowy:

$$\frac{y_{t_0-1}}{y_{t_0}} = 1 : \frac{y_{t_0}}{y_{t_0-1}}, \frac{y_{t_0-2}}{y_{t_0}} = \frac{y_{t_0-1}}{y_{t_0}} : \frac{y_{t_0-1}}{y_{t_0-2}}, \ldots, \frac{y_i}{y_{t_0}} = \frac{y_{i+1}}{y_{t_0}} : \frac{y_{i+1}}{y_i},$$

  inaczej, jest to odwrotność iloczynu indeksów między okresem bazowym a badanym:

$$\frac{y_i}{y_{t_0}} = 1 / \left[\frac{y_{i+1}}{y_i} \cdot \frac{y_{i+2}}{y_{i+1}} \cdots \frac{y_{t_0-1}}{y_{t_0-2}} \cdot \frac{y_{t_0}}{y_{t_0-1}} \right].$$

Średnie tempo zmian

- interesuje nas średnie tempo zmian zjawiska w okresie od chwili 1 do chwili n (za $n-1$ okresów),
- jest to tempo $\bar{r}$, które będąc stałe w całym rozważanym okresie, przyniosłoby taką samą zmianę całkowitą,
- odpowiada mu taki średni indeks $\bar{g} = 1 + \bar{r}$, że

$$y_n = (\bar{g})^{n-1} y_1 = (1+\bar{r})^{n-1} y_1.$$

- zatem średni indeks możemy obliczamy jako:
  - pierwiastek $(n-1)$-tego stopnia z ilorazu badanej wielkości na końcu i początku badanego okresu:

$$\bar{g} = \sqrt[n-1]{\frac{y_n}{y_1}}.$$

  - pierwiastek $(n-1)$-tego stopnia z ilorazu ostatniego i pierwszego indeksu jednopodstawowego:

$$\bar{g} = \sqrt[n-1]{\frac{y_n}{y_{t_0}} : \frac{y_1}{y_{t_0}}}.$$

Średnie tempo zmian c.d.

- najczęściej średni indeks obliczamy jako średnią geometryczną indeksów łańcuchowych, które są indeksami ilustrującymi dynamikę zmian w kolejnych okresach

$$\bar{g} = \sqrt[n-1]{\frac{y_2}{y_1} \cdot \frac{y_3}{y_2} \cdots \frac{y_{n-1}}{y_{n-2}} \cdot \frac{y_n}{y_{n-1}}} = \sqrt[n-1]{\prod_{i=2}^{n} \frac{y_i}{y_{i-1}}}.$$

- średnie tempo zmian obliczamy wówczas jako $\bar{r} = \bar{g} - 1$,
- średnie tempo zmian możemy wykorzystać do sporządzania prognozy (naiwnej):

$$y^*_{n+1} = y_n \cdot (1 + \bar{r}).$$

- Uwaga: porównaj z wzorem na oprocentowanie przeciętne przy kapitalizacji złożonej:

$$\bar{r} = \sqrt[n]{(1 + r_1)(1 + r_2) \cdots (1 + r_n)} - 1.$$

Indeksy indywidualne i agregatowe

- p – cena, q – ilość, $w = p \cdot q$ – wartość,
- 0 – okres bazowy, 1 – okres badany,
- indywidualny indeks cen
$$i_p = \frac{p_1}{p_0}$$
- indywidualny indeks ilości
$$i_q = \frac{q_1}{q_0}$$
- indywidualny indeks wartości
$$i_w = \frac{w_1}{w_0} = \frac{p_1 \cdot q_1}{p_0 \cdot q_0} = i_p \cdot i_q$$
- agregatowy indeks wartości:
$$I_w = \frac{\sum w_1}{\sum w_0} = \frac{\sum p_1 q_1}{\sum p_0 q_0}$$

Indeksy agregatowe

- agregatowy indeks ilości Laspeyresa:

$$I_q^L = \frac{\sum q_1 p_0}{\sum q_0 p_0}$$

- agregatowy indeks ilości Paaschego:

$$I_q^P = \frac{\sum q_1 p_1}{\sum q_0 p_1}$$

- agregatowy indeks cen Laspeyresa:

$$I_p^L = \frac{\sum p_1 q_0}{\sum p_0 q_0}$$

- agregatowy indeks cen Paaschego:

$$I_p^P = \frac{\sum p_1 q_1}{\sum p_0 q_1}$$

Indeksy agregatowe, c.d.

	Laspeyresa	Paaschego
ilości	$I_q^L = \frac{\sum q_1 p_0}{\sum q_0 p_0}$	$I_q^P = \frac{\sum q_1 p_1}{\sum q_0 p_1}$
cen	$I_p^L = \frac{\sum p_1 q_0}{\sum p_0 q_0}$	$I_p^P = \frac{\sum p_1 q_1}{\sum p_0 q_1}$

- agregatowy indeks ilości Fishera:

$$I_q^F = \sqrt{I_q^L \cdot I_q^P}$$

- agregatowy indeks cen Fishera:

$$I_p^F = \sqrt{I_p^L \cdot I_p^P}$$

- zachodzą związki:

$$I_w = I_p^L \cdot I_q^P = I_p^P \cdot I_q^L = I_p^F \cdot I_q^F$$

Indeksy agregatowe, c.d.

Gdy nie dysponujemy szczegółowymi danymi, możemy zauważyć, że:

$$I_w = \frac{\sum w_0 \cdot i_w}{\sum w_0} = \frac{\sum w_1}{\sum \frac{w_1}{i_w}}$$

$$I_q^L = \frac{\sum w_0 \cdot i_q}{\sum w_0} \qquad I_q^P = \frac{\sum w_1}{\sum \frac{w_1}{i_q}}$$

$$I_p^L = \frac{\sum w_0 \cdot i_p}{\sum w_0} \qquad I_p^P = \frac{\sum w_1}{\sum \frac{w_1}{i_p}}$$

Indeksy dla wielkości stosunkowych

- indywidualnie:

$$x = \frac{a}{b} \quad \Longleftrightarrow \quad a = bx \quad \Longleftrightarrow \quad b = \frac{a}{x}$$

- zespołowo:

$$X = \frac{A}{B} = \frac{\sum a}{\sum b} = \frac{\sum xb}{\sum b} = \frac{\sum a}{\sum \frac{a}{x}}$$

- indeksy agregatowy wszechstronny (o zmiennej strukturze):

$$I_X = \frac{X_1}{X_0} = \frac{\sum a_1}{\sum b_1} : \frac{\sum a_0}{\sum b_0} = \frac{\sum x_1 b_1}{\sum b_1} : \frac{\sum x_0 b_0}{\sum b_0} = \frac{\sum a_1}{\sum \frac{a_1}{x_1}} : \frac{\sum a_0}{\sum \frac{a_0}{x_0}}$$

Indeksy dla wielkości stosunkowych c.d.

- indeksy o stałej strukturze:

$$I_{x/a_0} = \frac{\sum a_0}{\sum \frac{a_0}{x_1}} : \frac{\sum a_0}{\sum \frac{a_0}{x_0}} \qquad I_{x/b_0} = \frac{\sum x_1 b_0}{\sum b_0} : \frac{\sum x_0 b_0}{\sum b_0}$$

$$I_{x/a_1} = \frac{\sum a_1}{\sum \frac{a_1}{x_1}} : \frac{\sum a_1}{\sum \frac{a_1}{x_0}} \qquad I_{x/b_1} = \frac{\sum x_1 b_1}{\sum b_1} : \frac{\sum x_0 b_1}{\sum b_1}$$

- indeksy zmian strukturalnych:

$${}_bI_{x/x_0} = \frac{\sum x_0 b_1}{\sum b_1} : \frac{\sum x_0 b_0}{\sum b_0} \qquad {}_aI_{x/x_0} = \frac{\sum a_1}{\sum \frac{a_1}{x_0}} : \frac{\sum a_0}{\sum \frac{a_0}{x_0}}$$

$${}_bI_{x/x_1} = \frac{\sum x_1 b_1}{\sum b_1} : \frac{\sum x_1 b_0}{\sum b_0} \qquad {}_aI_{x/x_1} = \frac{\sum a_1}{\sum \frac{a_1}{x_1}} : \frac{\sum a_0}{\sum \frac{a_0}{x_1}}$$

$$I_x = I_{x/a_1} \cdot {}_aI_{x/x_0} = I_{x/a_0} \cdot {}_aI_{x/x_1} = I_{x/b_1} \cdot {}_bI_{x/x_0} = I_{x/b_0} \cdot {}_bI_{x/x_1}$$

Wykresy

Dane jakościowe: struktura (kołowy, słupkowy/barplot)
Wysokości słupków są równe odpowiednim liczebnościom (lub częstościom); szerokość słupków jest jednakowa, zalecane jest uporządkowanie
Wykres kołowy: Mało czytelne, gdy występuje więcej kategorii; Porównanie dwóch wykresów jest trudniejsze niż dla wykresów słupkowych.

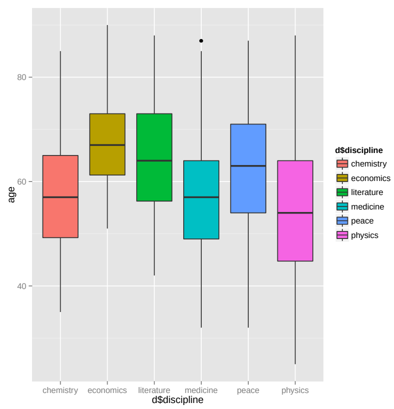
Dane przekrojowe: rozkład (słupkowy zwany także diagramem liczebości/częstości dla szeregu rozdzielczego jednostopniowego, histogram dla sz.r. wielostopniowego)

Szeregi czasowe: dynamika (słupkowy, liniowy)

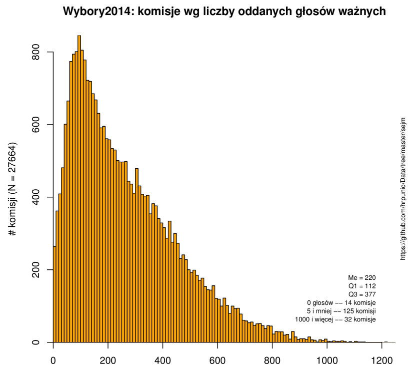
Porównanie: struktury, rozkładów, dynamiki (słupkowy, pudełkowy dla danych przekrojowych)

Wykres pudełkowy (boxplot): 5 podstawowych wskaźników sumarycznych na jednym wykresie: Q_1 dolna krawędź, Q_3 górna krawędź, Me środek pudełka; wąsy to maksimum/minimum (lub $\pm$1,5 $\times$ IQR)
Dwie zmienne: wykresy rozrzutu (scatterplots)

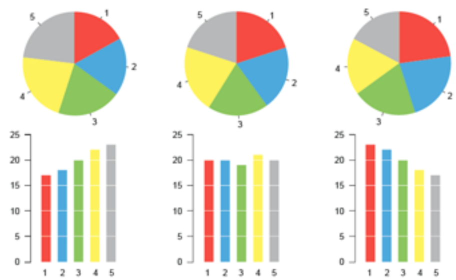
Wykresy

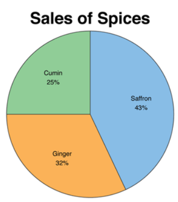


age
d$discipline
chemistry
economics
literature
medicine
peace
physics

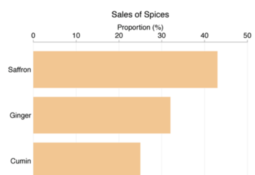


Wybory2014: komisje wg liczby oddanych głosów ważnych
# komisji (N = 27664)
Me = 220
Q1 = 112
Q3 = 377
0 głosów --- 14 komisje
5 i mniej --- 125 komisji
1000 i więcej --- 32 komisje
https://github.com/hrpunio/Data/tree/master/sejm





Sales of Spices
Cumin
25%
Saffron
43%
Ginger
32%



Sales of Spices
Proportion (%)
Saffron
Ginger
Cumin