Statystyka danych

Elżbieta Pociask

Literatura

- W. Krysicki i in. *Rachunek prawdopodobieństwa i statystyka matematyczna w zadaniach* część 1i 2
- Bąk Iwona i in. Statystyka w zadaniach - Część 1 Statystyka opisowa, PWN 2024
- Bąk Iwona i in. Statystyka w zadaniach - Część 2 Statystyka matematyczna, PWN 2018
- Tadeusiewicz R., Izworski A., Majewski J.: *Biometria*, Wydawnictwa AGH, Kraków, 1993
- Górecki T., *Podstawy statystyki z przykładami w R*, Wydawnictwo BTC 2011
- Ćwik J.,Mielniczuk J., *Statystyczne systemy uczące się – ćwiczenia w oparciu o pakiet R*, Oficyna Wydawnicza Politechniki Warszawskiej 2009
- Gągolewski M., *Programowanie w języku R – Analiza danych, obliczenia, symulacje*, Wydawnictwo PWN 2014
- Biecek P., *Przewodnik po pakiecie R,* Oficyna Wydawnicza GiS 2008 i wyżej
- Grzegorzewski P. i in. *Wnioskowanie statystyczne z wykorzystaniem środowiska R,* Warszawa 2014

Harmonogram

Data	Temat
6.10.2025	Organizacyjne + wstep do statystyki opisowej
13.10.2025	Zdarzenia, prawdopodobieństwo
20.10.2025	Zmienna losowa dyskretna
27.10.2025	Zmienna losowa ciągła
03.11.2025	Estymacja przedziałowa
10.11.2027	Testy parametryczne
17.11.2025	Powtórka
24.11.2025	**Kolokwium I**
01.12.2025	Analiza danych: wczytywanie danych, walidacja zbioru danych, brakujące dane, wartości odstające, statystyka opisowa
08.12.2025	Analiza danych: wizualizacja danych, badanie korelacji, zależności
15.12.2025	Analiza danych: weryfikowanie hipotez
12.01.2026	Analiza danych: regresja liniowa, regresja logistyczna, regresja Cox'a, szukanie predyktorów
19.01.2026	**Kolokwium II**
26.01.2026	Odrabianie zaległości, wystawianie ocen

Organizacja pracy

- Zajęcia stacjonarne (zajęcia tablicowe oraz komputerowe)
  - W sytuacjach wyjątkowych oraz w przypadku zagrożenia - zajęcia zdalne przy wykorzystaniu narzędzi MS Teams
- Praca w środowisku R
- Obecność – dopuszczalna jedna nieusprawiedliwiona
- Przygotowanie do zajęć:
  - zakres materiału z wykładów oraz dodatkowej literatury
  - zestawy zadań
- Sprawdzanie wiadomości
  - praca na zajęciach – rozwiazywanie zadań przy tablicy / komputerach
  - kolokwia
- Platforma MS Teams
  - Kurs: Statystyka danych pomiarowych 2025/26
  - Kod: 74f0a5n

Warunki i sposób zaliczenia

- **Obecność obowiązkowa (jedna dopuszczalna nieobecność)**
- **Aktywność na zajęciach (praca na zajęciach)**
- **Sposób obliczania oceny końcowej**
  - Suma punktów z kolokwium I i II
- **Zaległości**
  - Dodatkowe zadania/projekt, w terminie ustalonym indywidualnie z prowadzącą lub na zajęciach z inną grupą (przy zachowaniu maksymalnej liczby osób w grupie zgodnie z Regulaminem Studiów AGH – 15)

Kolokwium

- Obejmuje zakres materiału z wykładów oraz ćwiczeń
- Zaliczenie : minimum 50% punktów z dwóch kolokwiów
- Brak zaliczenia w I terminie:
  - Kolokwium poprawkowe z całego semestru:
    - jeśli zaliczone (>=50%) – zaliczenie w II terminie - z max. ocena 4.5;
    - jeśli nie zaliczone (<50%) – możliwość uzyskania zaliczenia w ostatnim III terminie, z max. oceną 3.0;
- Możliwość poprawy oceny w II terminie

Co to jest R?

- R to darmowy i potężny **język programowania** przeznaczony do **obliczeń statystycznych** i **wizualizacji danych** .
- Platforma programistyczna wyposażona w interpreter tego języka oraz nazwa projektu, w ramach którego rozwijany jest zarówno język jak i środowisko.
- **Języka R** można używać do:
  - obliczeń, jako kalkulatora
  - obliczania szerokiej gamy klasycznych testów statystycznych
  - przeprowadzania analiz klasyfikacyjnych, predykcyjnych, ML
  - rysowania wiele rodzajów wykresów
  - Automatycznego generowania raportów

Dlaczego R?

- **R** jest oprogramowaniem typu open source , więc jest darmowe.
- **R** jest kompatybilny z wieloma platformami , więc można go zainstalować w systemach Windows, MAC OSX i Linux
- **Język R** udostępnia szeroką gamę technik statystycznych i możliwości graficznych .
- **R** umożliwia przeprowadzanie powtarzalnych badań poprzez osadzanie skryptów i wyników w jednym pliku.
- **R** ma ogromną społeczność zarówno w środowisku akademickim, jak i biznesowym
- **Język R** jest niezwykle rozszerzalny i posiada tysiące dobrze udokumentowanych rozszerzeń (nazywanych pakietami R) dla bardzo szerokiego zakresu zastosowań w sektorze finansowym, opiece zdrowotnej,...
- Łatwo jest tworzyć pakiety R do rozwiązywania konkretnych problemów

Instalowanie i uruchamianie R

- https://pbiecek.gitbooks.io/przewodnik/content/Programowanie/podstawy/jak_zainstalowac_R.html
- Najnowszą wersję R najlepiej zainstalować ze strony https://cran.r-project.org/.
- Najnowszą wersję RStudio Desktop można pobrać ze strony http://www.rstudio.com/products/rstudio/download/.

Definicje

- Statystyka i statystyki (statistics , statistic)

Dyscyplina naukowa

- zajmująca się opisywaniem i analizą zjawisk masowych przy użyciu metod rachunku prawdopodobieństwa (zbieranie, porządkowanie, analiza i interpretacja danych)

Wielkość liczbowa

- miary położenia,
- miary rozrzutu,
- miary kształtu rozkładu.

Celem obliczeń nie są same liczby, lecz ich zrozumienie.

Wstępna analiza danych – na czym polega?

- **Ile danych?:**
  - ile zmiennych (cech: płeć, wykształcenie, staż, zarobki) ?
  - ile przypadków ?
- **Jakie typy?:**
  - dane jakościowe
    - porządkowe (zdefiniowany ma porządek kategoriiwykształcenie : podstawowe, zawodowe, średnie, wyższe)
    - nominalna (nie ma określonego porządku, nie można powiedzieć, która jest wyżej w rankingu np. płeć, kolor oczu)
  - dane ilościowe (staż pracy, płaca)
- **Ile braków, jakie, jak je zastąpić ?**

Cechy, którymi wyróżniają się jednostki wchodzące w skład zbiorowości, nazywa się **cechami statystycznymi.**

Wyniki obserwacji – jak wyrazić?

Wyniki obserwacji i pomiarów mogą być wyrażone w postaci:

- Tekstu (cechy jakościowe)
- Liczb całkowitych
- Przedziałów liczbowych

Dane źródłowe zawierają się w:

- zbiorze,
- zbiorze uporządkowanym, **zwanym szeregiem szczegółowym lub szeregiem czasowym**
- zbiorze podzielonym na klasy, zwanym **szeregiem rozdzielczym**

Szeregi statystyczne

Celem tych działań jest przejście od danych indywidualnych do danych zbiorowych.

- Materiał źródłowy należy odpowiednio posegregować i policzyć, w wyniku otrzymuje się tzw. tablice robocze.

Klasyfikacja danych musi być przeprowadzona:

- w sposób rozłączny, jednostki o określonych cechach muszą być jednoznacznie przydzielone do poszczególnych klas
- w sposób zupełny, tzn. klasy muszą objąć wszystkie występujące cechy danej zbiorowość

Technika zestawiania zależy od rodzaju skali pomiarowej

Szereg szczegółowy

Uporządkowany ciąg wartości badanej cechy statystycznej, wartości cechy porządkujemy rosnąco lub malejąco.

- Badana cecha przyjmuje niewielką liczbę jednostek (mała grupa):

{x1, …. ,xn}

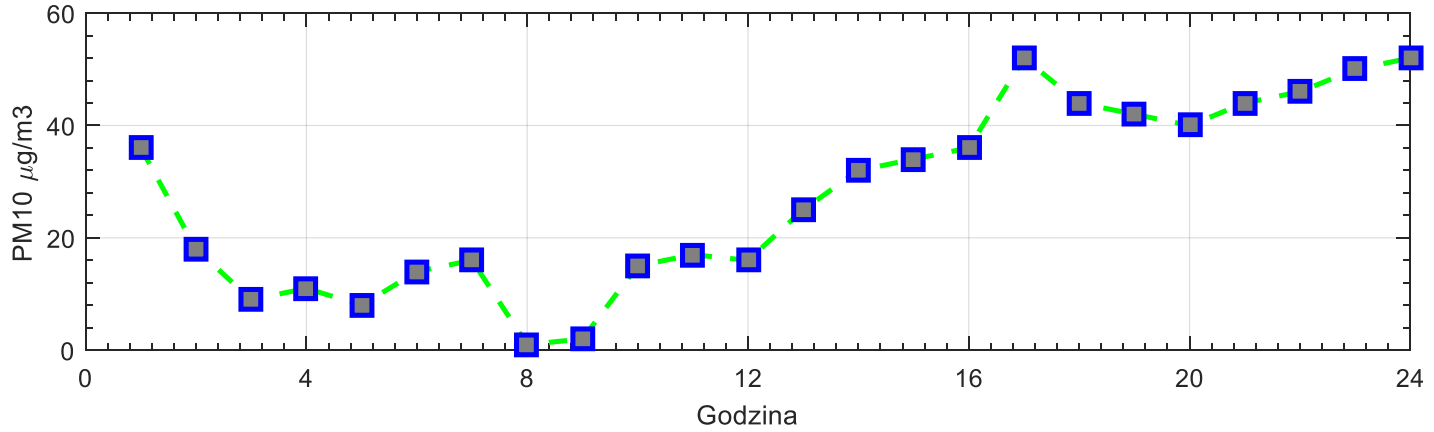
- Wartości porządkuje się:
  - Rosnąco x1< …. <xn

lub

  - Malejąco x1> …. >xn

Szereg czasowy

- Otrzymuje się w wyniku grupowania
  - typologicznego (wyodrębniającego różne jakościowo cechy)
  - wariancyjnego (porządkującego zbiorowość przez łączenie w klasy jednostek mających odpowiednie wartości cech) gdy podstawą grupowania jest zmiana badanego zjawiska w czasie
- Co można powiedzieć o stężeniu pyłu PM10 w danym punkcie pomiarowym na podstawie ***szeregu czasowego***?



Szereg rozdzielczy

- zagregowany sposób dostarczania danych, stosuje się w przypadku dużej liczby obserwacji.
- grupuje się wówczas obserwacje w kilka do kilkunastu klas oraz podaje się jedyni granice przedziałów klasowych i liczby obserwacji w poszczególnych klasach.
- dane takie często przedstawia się graficznie w postaci histogramów.

i	Dolna granica przedziału x_{1i}	Środek przedziału $\dot{x}$	Górna granica przedziału x_{2i}	Liczebność przedziału n_i	Liczebność skumulowana $n(x_{2i})$	Częstość w_i	Częstość skumulowana $F(x_{2i})$
1.							
2.							

Etapy budowy przedziałów w szeregach rozdzielczych przedziałowych

- Ustalenie wartości cechy maksymalnej i minimalnej
- Wyznaczenie rozstępu z próby (R = max – min)
- Ustalenie liczby klas k , gdzie k – liczba klas, N – liczba obserwacji
- Ustalanie rozpiętości przedziałów klasowych , przedziały [...,...)

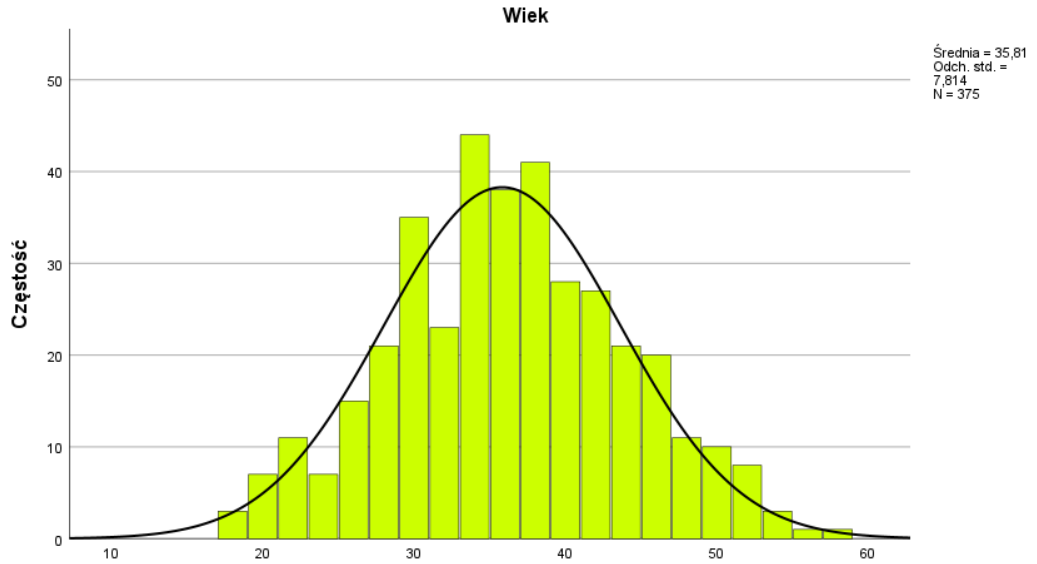
h, h ≈ R/K jest to przybliżenie z nadmiarem, a więc h >= R/K

- Wyznaczanie lewego końca pierwszego przedziału klasowego

a = xmin − α/ 2 , gdzie α jest dokładnością pomiaru

Wizualizacja struktury zbioru danych – Histogram

- Histogram to jeden z graficznych sposobów przedstawiania rozkładu cechy.
  - Składa się z szeregu prostokątów umieszczonych na osi współrzędnych.
  - Prostokąty te są wyznaczone przez
    - przedziały klasowe wartości cechy; szerokość przedziału; krok
    - natomiast ich wysokość jest określona przez liczebności lub częstości elementów należących do określonego przedziału klasowego



Parametry statystyczne

- Istnieje często potrzeba wyrażenia serii wartości (np. pomiarów) w postacie jednej liczby odzwierciedlającej ogólny poziom zjawiska**, jego przeciętną tendencję**.
  - miara skupienia, położenia, miara tendencji centralnej.
- Poza wskazaniem liczby, wokół której koncentrują się wyniki próby niezbędne jest określenie **stopnia rozproszenia** wyników próby wokół miary skupienia
  - miara zmienności, rozproszenia
- **Parametrami statystycznymi** (statystykami) nazywamy liczby umożliwiające sumaryczny opis zbiorowości.
- Parametry te tak dokładnie charakteryzują zbiorowość, że mogą być wykorzystane do porównywania różnych zbiorowości.

Wyliczanie parametrów statystyki opisowej dla szeregu szczegółowego:

- ***Miary położenia:***
- Średnia arytmetyczna: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- Mediana: $Me = \begin{cases} x_{(n+1)/2}, & gdy\ n\ -\ nieparzyste \\ \frac{1}{2}(x_{n/2} + x_{1+n/2}), & gdy\ n -\ parzyste \end{cases}$
- Pierwszy kwartyl: $Q_1 = x_{0,25(n-1)+1}$
- Trzeci kwartyl: $Q_3 = x_{0,75(n-1)+1}$

Wyliczanie parametrów statystyki opisowej dla szeregu szczegółowego:

- ***Miary zmienności:***
- Wariancja: $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- Odchylenie standardowe: $s = \sqrt{s^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$
- Odchylenie przeciętne: $d = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$
- Rozstęp: $R = x_{max} - x_{min}$
- Rozstęp międzykwartylowy: $IQR = Q_3 - Q_1$
- Współczynnik zmienności: $V_s = {}^{s}/_{\bar{x}}$

gdzie: n – liczebność próby; x_i – kolejne wartości cechy z szeregu szczegółowego;

Wyliczanie mediany i kwartyli dla szeregu rozdzielczego:

- $Me = x_{1Me} + (\frac{1}{2}n - \sum_{i=1}^{m} n_i) \cdot \frac{c}{n_{Me}}$
- $Q_1 = x_{1Q_1} + (\frac{1}{4}n - \sum_{i=1}^{m} n_i) \cdot \frac{c}{n_{Q_1}}$
- $Q_3 = x_{1Q_3} + (\frac{3}{4}n - \sum_{i=1}^{m} n_i) \cdot \frac{c}{n_{Q_3}}$

gdzie:

x_{1Me} – dolna granica przedziału zawierającego medianę (lub kwartyl);

m – liczba przedziałów poprzedzających przedział z medianą (kwartylem);

c – długość przedziału, w którym znajduje się mediana (kwartyl);

n_{Me} – liczebność przedziału, w którym znajduje się mediana (kwartyl);

Przykład

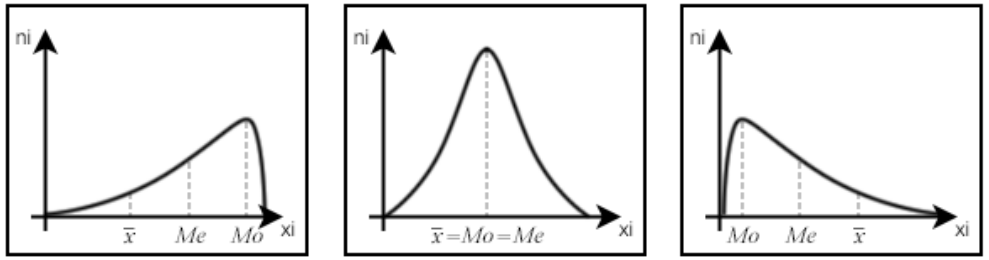
Jednym z parametrów, które najczęściej używa się w spirometrii jest FEV1.

FEV1 to nasilona pierwszosekundowa objętość wydechowa (ang. *forced expiratory volume in one second*). Jest to objętość powietrza wydmuchnięta z płuc w czasie pierwszej sekundy maksymalnie natężonego wydechu. W miarę starzenia się organizmu, czynność płuc stale się zmniejsza. Palenie tytoniu powoduje m.in. znaczne przyspieszenie tempa spadku FEV1. Parametr ten zależy -od wielkości płuc (czyli od tzw. pojemności życiowej) oraz od drożności dróg oddechowych. Zwężenie oskrzeli u chorego z napadem astmy albo z astmą słabo kontrolowaną ogranicza przepływ powietrza w czasie wydechu i powoduje zmniejszenie FEV1. [źródło: pulmonologia.mp.pl]

Osoba	A	B	C
$\bar{x}$	4,12	3,13	3,92
s	0,15	0,41	0,08
Me	4,10	3,00	3,87
Q_1	4,10	2,82	3,84
Q_3	4,22	3,35	3,94
IQR	0,12	0,53	0,11

Miary asymetrii i koncentracji

- **Skośność** - jest statystyką umożliwiającą porównanie rozkładu analizowanej zmiennej z hipotetycznym rozkładem normalnym. Wskazuje na rozbieżności pomiędzy wartością średnią, a centrum danego rozkładu.



Asymetria lewostronna

$\bar{x}$ < Me < Mo

Rozkład symetryczny

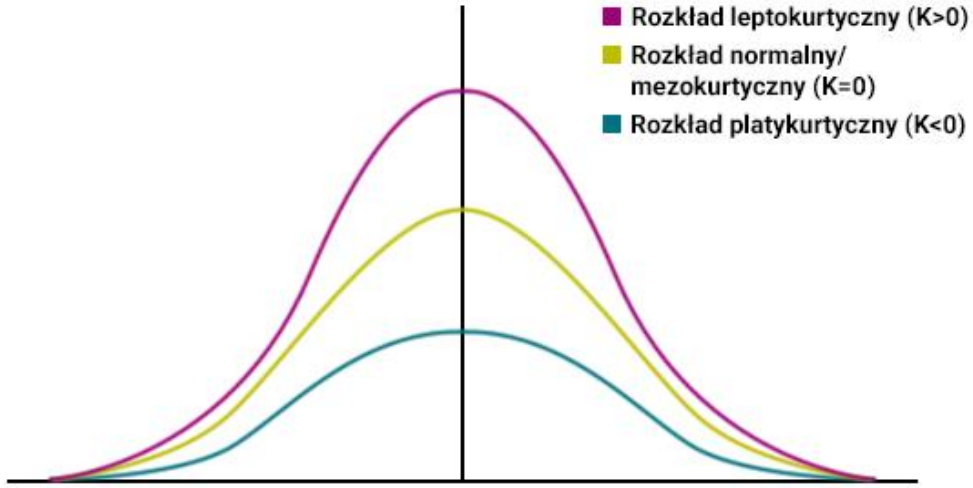
$\bar{x}$ = Me = Mo

Asymetria prawostronna

$\bar{x}$ > Me > Mo

Miary asymetrii i koncentracji

- **Kurtoza** – jest miarą występowania wartości odstających.
- W zależności od wartości kurtozy wykreślony rozkład może mieć **„grubsze” lub „węższe ogony”,** na co wpływ ma intensywność wartości skrajnych.
  - ***leptokurtyczny (K>0)*** – rozkład ma tzw „grube ogony”, czyli intensywność wartości skrajnych jest większa niż w rozkładzie normalnym.
  - ***mezokurtyczny (K=0)*** - rozkład jest zbliżony do normalnego.
  - ***platykurtyczny (K<0)*** – rozkład ma „węższe ogony” niż rozkład normalny, intensywność wartości ekstremalnych jest mniejsza niż w przypadku rozkładu normalnego.



Wykres pudełkowy

- wykresy skrzynkowe wskazują na to czy w bazie danych występują obserwacje odstające czy nie.
- **wynik nietypowy**, odstający od reszty, outlier czy nawet dewiant to nazwa obserwacji (najczęściej wyniku badanej osoby lub innego podmiotu badań), której rezultat może negatywnie wpłynąć na wyniki przeprowadzanych testów statystycznych.