w1

Statystyczna analiza danych

27.02.2023r.

Dawno, dawno temu….

» Pierwotnie termin „**statystyka**" był używany na oznaczenie **wiedzy o państwie**, poczynając od najdawniejszych spisów w Sumerii, między 5000 a 2000 p. n. e.

» STATYSTYKA z łac. *status* – państwo

„Do ok. połowy XIX wieku termin *statystyka* oznaczał *podany w tabelarycznej formie zbiór danych na temat stanu państwa*"

» 1794 r. (j. niemiecki) oznaczenie nauki o „ gromadzeniu, przetwarzaniu i wykorzystaniu danych przez państwo"

» 1770 r. (j. angielski) „nauka, która poucza nas, jakie są porządki polityczne we wszystkich współczesnych państwach w znanym świecie

» 1809 r. (j. polski) „O statystyce Polski. Krótki rzut wiadomości potrzebnych tym, którzy ten kraj chcą oswobodzić i tym którzy chcą w nim rządzić"

Teraz….

Industry 4.0 — wizja przyszłości przemysłu

czwarta rewolucja przemysłowa, **Przemysł 4.0**

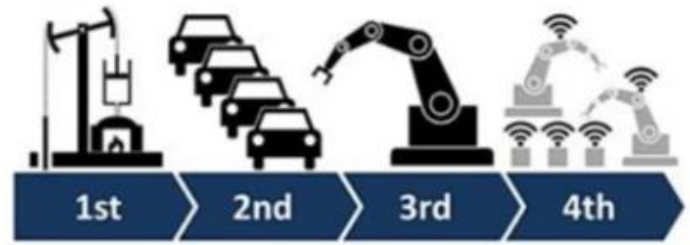
— systemy cyberfizyczne (CPS): integracja maszyn (autonomicznych) z warstwą systemów umożliwiających: wizualizację, monitoring, sterowanie i optymalizację procesów produkcyjnych

»1.0: mechanika - mechaniczne krosno tkackie

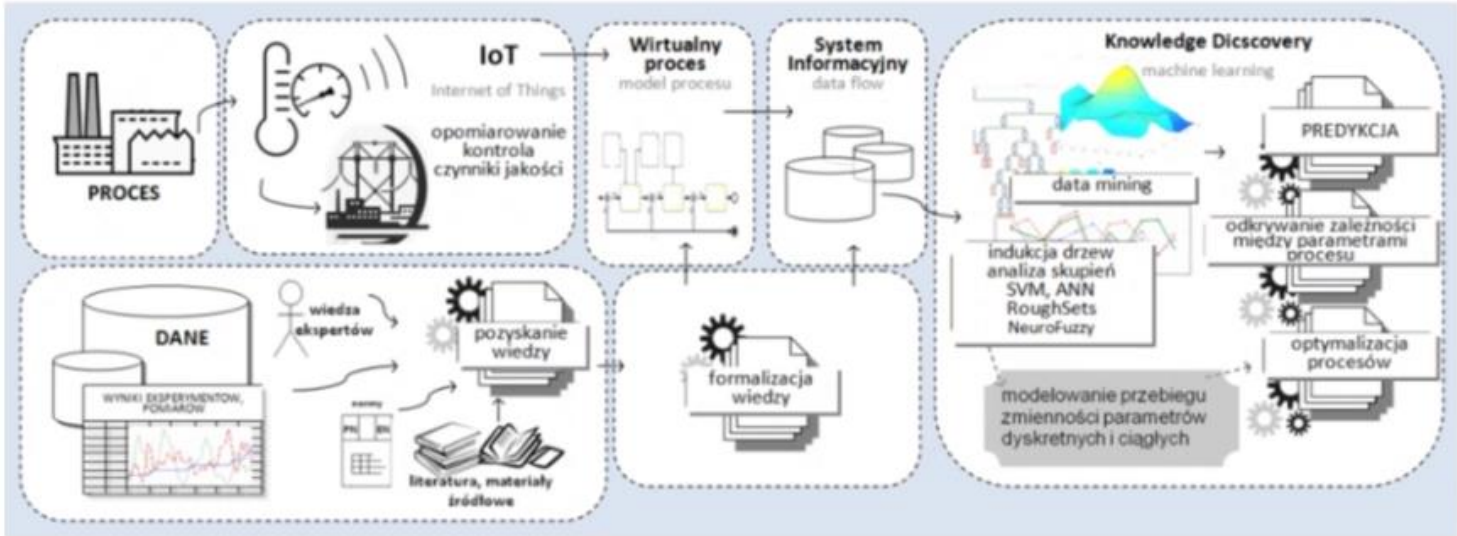
»2.0: produkcja seryjna - masowa produkcja z wykorzystaniem elektryczności

»3.0: sterowanie i automatyzacja – programowalny układ logiczny

»4.0: koncepcja cyberfizycznych systemów produkcyjnych – internet – ludzi/usług/rzeczy/danych



Systemy cyberfizyczne

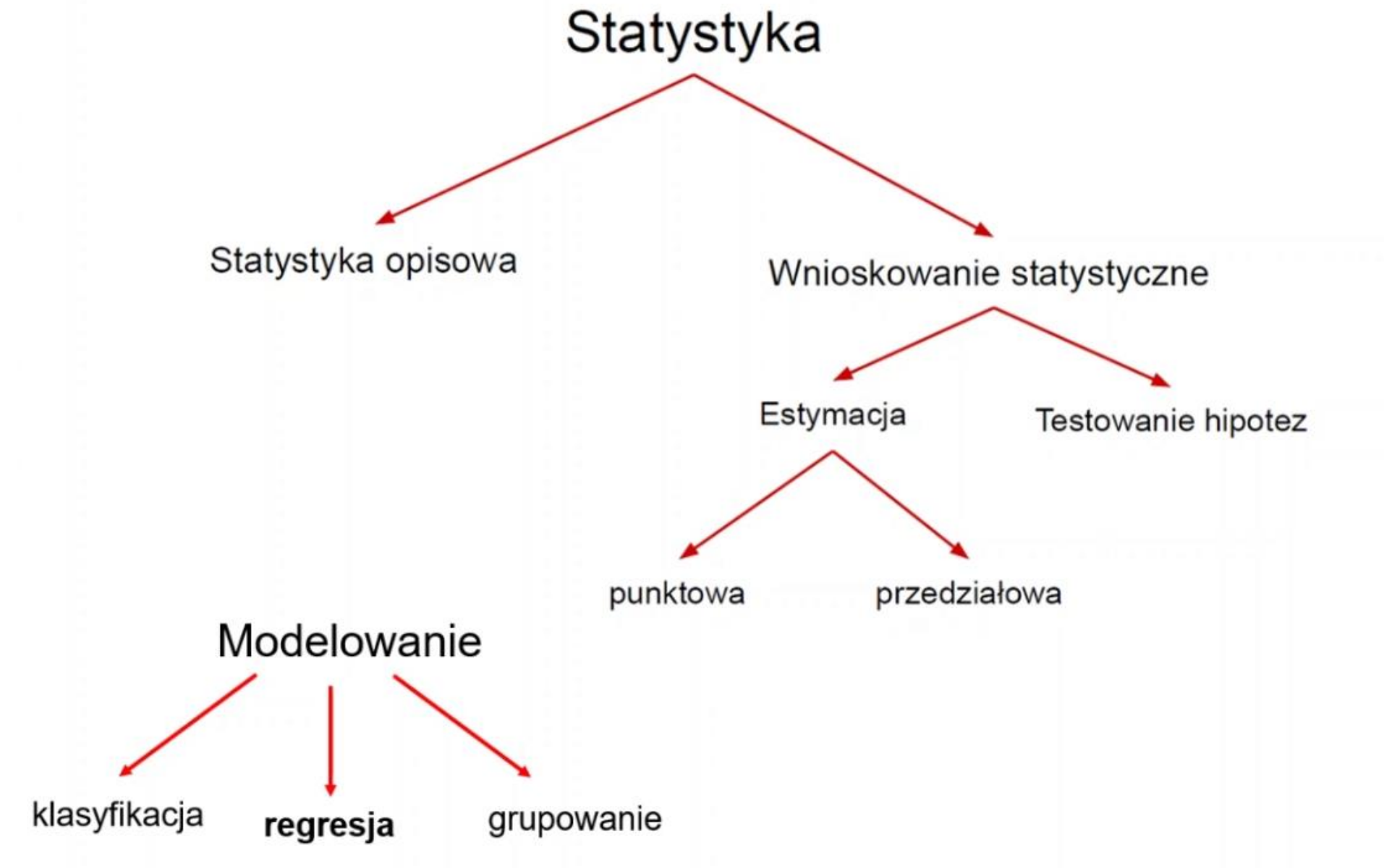


Industry 4.0 to ciągłe przewidywanie rezultatów bieżących operacji. Tworzenie modeli predykcyjnych:

1.poszukiwanie zależności pomiędzy parametrami procesu

2.zastosowanie algorytmu, który analizując te zależności będzie w stanie przewidzieć nieznaną wartość parametrów istotnych dla efektywności procesu (zmiennej zależnej)

Statystyka


Statystyka
Statystyka opisowa
Wnioskowanie statystyczne
Estymacja
Testowanie hipotez
punktowa
przedziałowa
Modelowanie
klasyfikacja
regresja
grupowanie

SAD – cel ☺

- **Znaczenie i rola** statystycznej analizy danych we współczesnych badaniach inżynierskich i biznesowych.
- Podstawowe **pojęcia** w statystyce.
- Nabycie **umiejętności** stosowania narzędzi statystycznych form analizy danych.
- Zdobycie **wiedzy** o metodach i technikach statystyki matematycznej i **poprawnym ich stosowaniu** w badaniach naukowych (opracowanie wyników pomiarów i obserwacji) oraz w gospodarce (analiza rynku, kontrola jakości).

SAD – treści

- **Badania statystyczne**; Podstawowe pojęcia. Statystyka opisowa miary położenia, miary zmienności, asymetrii i koncentracji, reprezentacja graficzna danych. Szeregi.
- Elementy **rachunku prawdopodobieństwa**: interpretacja zdarzeń, prawdopodobieństwo – podstawowe twierdzenia. Zmienne losowe, ich rozkłady i parametry rozkładu.
- Techniki **wnioskowania statystycznego**: estymacja i estymatory, weryfikacja hipotez statystycznych, testy statystyczne parametryczne i nieparametryczne.
- Analiza **struktury zbiorów danych**. Dopasowanie rozkładu empirycznego do teoretycznego. Analiza wariancji.
- **Szukanie i badanie zależności**. Podstawy korelacji i regresji: pojęcia podstawowe, korelacje cząstkowe, korelacje nieparametryczne, funkcje regresji. Ocena dopasowania funkcji do danych.

SAD znaczenie i rola statystyki

- Statystyka pozwala wydobyć wiedzę z chaosu (z danych szczegółowych)
- Stale posługujemy się statystyką, np. uogólniając sądy
  - Zarabiamy mniej niż w innych krajach UE,
  - Dłużej żyjemy
  - Częściej chorujemy
- Stosujemy pojęcia statystyczne w języku potocznym:
  - Przeciętny konsument
  - Podwyżka energii pociąga za sobą wzrost cen żywności
- Skąd wynika zła opinia o statystyce
  - Hermetyczna i trudna terminologia,
  - Brak wiedzy na temat metod statystycznej analizy, które chronią przed skutkami niepewności wynikającej z przypadkowości, ze współdziałania wielu czynników i umożliwiają podejmowanie najlepszych decyzji w warunkach niepewności
  - Niepoprawne (świadome) stosowanie statystyki dla osiągania ściśle określonych celów np. politycznych, komercyjnych itp.

ETAPY BADANIA STATYSTYCZNE

Badanie statystyczne to złożony proces składający się z kilku etapów:

Przygotowanie badania

1. Ustalenie celu badania statystycznego.
2. Określenie przedmiotu badania (zbiorowości i jednostki statystycznej).
3. Właściwe określenie jednostki sprawozdawczej (źródeł danych).
4. Decyzja co do metody badania (pełne czy częściowe).

Obserwacja statystyczna

1. Ustalenie wartości cech ilościowych lub odmian cech jakościowych u wszystkich jednostek badanej zbiorowości (generalnej bądź próbnej).
2. Kontrola formalna i merytoryczna zebranych danych

Opracowanie i prezentacja materiału statystycznego

1. Grupowanie lub klasyfikacja.
2. Zliczanie danych.
3. Tabelaryczna prezentacja materiału statystycznego.
4. Graficzna prezentacja materiału statystycznego.

Analiza statystyczna

1. Opis statystyczny.
2. Wnioskowanie statystyczne (badanie częściowe – próba losowa)

Badania statystyczne – próby losowe

- Losowy dobór próby polega na tym, że o fakcie znalezienia się poszczególnych elementów populacji w próbie **decyduje przypadek** (randomization).
- Jest to taki sposób wyboru przy którym spełnione są następujące dwa warunki;
  - każda jednostka populacji ma dodatnie, znane prawdopodobieństwo znalezienia się w próbie
  - istnieje możliwość ustalenia prawdopodobieństwa znalezienia się w próbie dla każdego zespołu elementów populacji

Żargon statysty

AGH

Poszczególne elementy zbiorowości to **obserwacje** (przypadki, rekordy, instancje, case'y).
Właściwości obserwacji to **cechy** (zmienne, atrybuty, kolumny).

Cechy dzielimy na:

- Stałe - wspólne dla wszystkich obserwacji, nie podlegają badaniu, decydują o przydziale do populacji
  - rzeczowe – charakteryzują określony zbiór osób, rzeczy, zjawisk, itp.
  - czasowe – charakteryzują okres jaki obejmuje badanie, lub w jakim momencie się ono odbywa
  - przestrzenne – gdzie badamy
- Zmienne

ZMIENNA – dowolna cecha, która przyjmuje **różne wartości** dla **poszczególnych obserwacji z próby**

Nr gospodarstwa	Klasa	Wydatki wyżywienie	Wydatki rachunki	Wydatki leczenie	...
Gospodarstwo1	małe	936,34 zł	217,44 zł	59,42 zł	
Gospodarstwo2	małe	1006,54 zł	230,54 zł	130,44 zł	
Gospodarstwo3	duże	1821,33 zł	476,34 zł	320,36 zł	
Gospodarstwo4	średnie	1136,54 zł	318,11 zł	150,57 zł	
Gospodarstwo5	duże	1723,44 zł	426,31 zł	540,20 zł	
...					

Populację stanowią gospodarstwa domowe w Krakowa. Badanie zostało przeprowadzone 20.01.2009.

Żargon statystyczny

AGH

ZMIENNE

ilościowe/liczbowe

Właściwości, które można zmierzyć i porównać

dyskretne

Elementy, które można **policzyć** (**przeliczyć**). Lista możliwych wartości może być **ustalona** (**skończona**) lub dążyć do **nieskończoności** (być **przeliczalna**). Np. liczba orłów w 100 rzutach monetą przyjmuje wartości od 0 do 100 (wartość skończona), ale liczba rzutów aby wypadło 100 orłów, przybiera wartości od 100 (najszybszy scenariusz) do nieskończoności.

ciągłe

To wyniki pomiaru. Ich możliwych wartości **nie można policzyć**, mogą one zostać opisane za pomocą przedziałów na osi liczbowej. Np. czas działania baterii AAA może teoretycznie wynosić od 0 do nieskończoności**, z wszystkimi możliwymi wartościami pośrednimi.**

jakościowe/niemierzalne

Oznaczają **cechy** (płeć, stan cywilny, rodzaj filmu, ...). Mogą przyjmować wartości liczbowe (1- kobieta, 2-mężczyzna), ale nie mają charakteru znaczącego (nie można ich poddać operacjom matematycznym)

porządkowe

Szeregują natężenie badanej właściwości przedstawionej w sposób opisowy, np. gdy oceniamy restaurację w skali od 0 do 4 gwiazdek, oceny studentów, bdb, db, ...

Skale pomiaru cechy (types of variables)

Zmienne ilościowe

- Skala równomierna (przedziałowa). Stosowana do pomiaru cech ilościowych, zakłada że zbiór wartości cechy składa się z liczb rzeczywistych określona przez wskazanie stałej jednostki miary i relacji **przyporządkowującej liczbę** każdemu wynikowi obserwacji (czas kalendarzowy, temperatura oC)
- Skala ilorazowa. Posiada wszystkie właściwości skali przedziałowej ale pomiary wg tej skali charakteryzują się **stałymi stosunkami** i **bezwzględnym zerem**, ma zastosowanie w fizyce, technice, np. długość czy czas

Zmienne kategoryczne

- Skala nominalna –dotyczy cech jakościowych, operacją pomiarową jest identyfikacja **kategorii** do której należy zaliczyć wynik, prowadzi do podziału zbioru na zbiory rozłączne (np. samochody wg kolorów).
- Skala porządkowa – stosowana jest do badania cech których **natężenie** jest określane przez przymiotniki, pociąga za sobą porządkowanie lub uszeregowanie badanej zmiennej (np. poniżej normy, w normie, powyżej normy, albo za mały, mały, średni, duży...)

Skala ilorazowa

- na **skali ilorazowej** wolno dokonywać operacji matematycznych,

tzn. bezpiecznie można stwierdzić, że np. dwa kilogramy cukru są 2x cięższe od jednego kilograma, a trzymetrowa deska jest trzy razy dłuższa niż deska o długości jednego metra

- wynika to z obecności absolutnego zera (gdyby cukru było 0 kg to znaczy, że nie byłoby go wcale)

- Przy użyciu skali stosunkowej (ilorazowej) możliwe jest podanie rozkładu częstości zmiennej, obliczenie m.in. dominanty, mediany, średniej, odchylenia standardowego i wariancji.

Skala nominalna

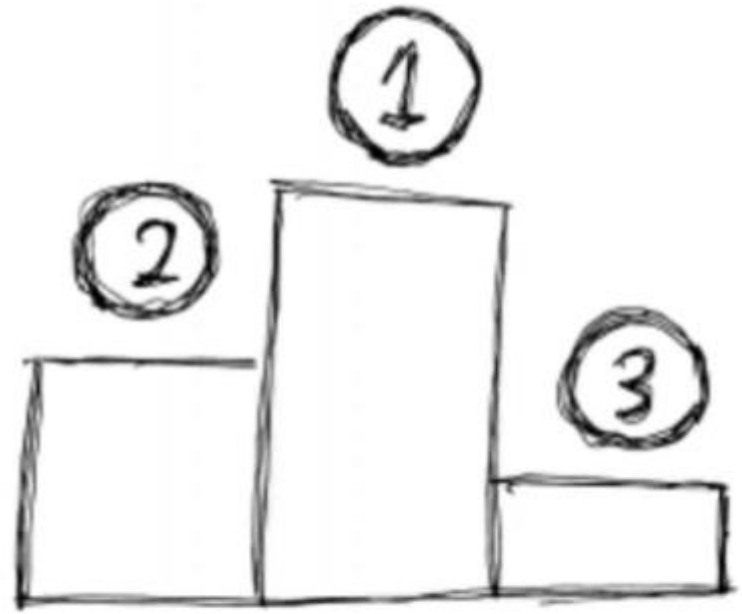
Brak fizycznej interpretacji dla kolejności wartości

Ale...

można je zakodować za pomocą liczb

Skala porządkowa

Miejsce na podium to z kolei przykład skali porządkowej.



Jak ocenić ile razy złoty medal jest więcej wart niż srebrny?

Skala pomiarowa a statystyka opisowa

Rodzaj skali pomiarowej	„Dozwolone" parametry statystyk opisowych
Nominalna	N, n kategorii, ω kategorii, Mo
Porządkowa	N, n kategorii, ω kategorii, Mo Kwantyle: Min, Q_1, Me, Q_3, Max,
Przedziałowa	N, Średnia , SD, V_c , Mo Kwantyle: Min, Q_1, Me, Q_3, Max,
Ilorazowa	N, Średnia , SD, V_c , Mo Kwantyle: Min, Q_1, Me, Q_3, Max,

Analiza zbioru danych

Populację stanowią mieszkańcy ul. Statystycznej. Badanie zostało przeprowadzone 20.01.2009.

Dane, zwykle prezentowane są w postaci **tabeli**.
Tablica może zawierać jedną lub kilka cech.

Tabela 1. Rejestr mieszkańców

Inicjały	Płeć	Kolor oczu	Waga, kg	Wzrost	Wiek	Długość włosów
IO	K	Niebieski	60	160	30	Średniej długości
BM	K	Brązowy	55	355	40	Krótkie
MW	K	Niebieski	35	130	10	Długie
PW	M	Piwny		180	40	Krótkie
AO	M	Brązowy	90	185	45	Krótkie

Analiza zbioru danych

Duży zbiór danych - Jak wychwycić błędy?

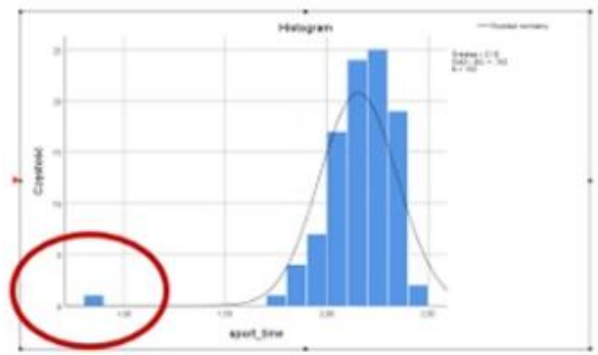
Statystyka opisowa

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$s = \sqrt{\frac{\sum_{i=1}^{n}\left(w_i\left(x_i - \frac{\sum_{i=1}^{n} x_i w_i}{n}\right)^2\right)}{n-1}},$$

....

Wykresy



Analiza zbioru danych

Statystyki opisowe to **liczby**, które ułatwiają odbiór informacji zawartych w danych źródłowych, prezentując je w skondensowanej, a przy tym bardziej zrozumiałej formie.

Do statystyk opisowych możemy zaliczyć:

- miary położenia/ miary tendencji centralnej
- wskaźniki rozproszenia
- miary asymetrii i koncentracji

Wskaźniki położenia/miary tendencji centralnej

ŚREDNIA ARYTMETYCZNA

suma wszystkich liczb
podzielona przez ich ilość

$$\bar{x} = \frac{\sum_{i=0}^{n} x_i}{n}$$

9,3,1,8,3,6

9+3+1+8+3+6=30

30 / 6 = 5

Średnia arytmetyczna wynosi 5

Średnia to przydatna i precyzyjna miara położenia, ale **ma swoje wady!**

Bardzo mocno zależy od wartości odstających!

MEDIANA

środkowa wartość z uporządkowanego zbioru liczb

1. Szeregujemy wartości od najmniejszych do największych

2. $$M_e = \begin{cases} x_{\frac{n+1}{2}} & dla\ n\ nieparzystego \\ \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2} & dla\ n\ parzystego \end{cases}$$

9,3,1,8,3,6

1,3,3,6,8,9

Mediana wynosi 4,5

MODA/DOMINANATA

najczęściej występująca liczba w zbiorze liczb

9,3,1,8,3,6

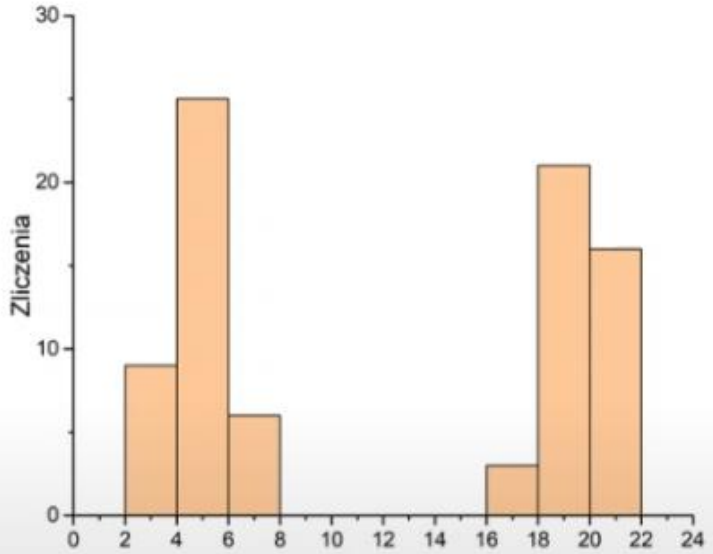
Modą jest liczba 3

Wskaźniki rozproszenia – rozstęp

Rozstęp

Różnica między **największą** wartością a **najmniejszą**

$$R = x_{max} - x_{min}$$



9,3,1,8,3,6

9 - 1 = 8

Wskaźniki rozproszenia – wariancja/ odchylenie standardowe

Wariancja

Odchylenie standardowe

Informują o tym, jak duże jest standardowe zróżnicowanie wyników w danym zbiorze danych.

$$s^2 = \frac{\sum_{i=0}^{n}(x_i - \bar{x})^2}{n-1}$$

Wada: Występuje w jednostkach kwadratowych, np. m^2

$$s = \sqrt{\frac{\sum_{i=0}^{n}(x_i - \bar{x})^2}{n-1}}$$

To standardowa (lub **typowa**) **wartość** odchylenia (lub **odległości**) od **przeciętnej** (średniej).

- Nie może być liczbą ujemną.
- Najmniejsza wartość 0. Wyłącznie w aranżowanych sytuacjach – w zbiorze danych są dokładnie takie same liczby.
- Na odchylenie/wariancję mają wpływ wartości odstające.

Wskaźniki rozproszenia – wariancja/ odchylenie standardowe

AGH

Kwantyl/ Percentyl

Wyznacza **względną pozycję** w zbiorze danych. Ważne jest to gdzie się znajdujemy (nie w odniesieniu do średniej), ale do wszystkich innych jednostek w zbiorze.

Interpretacja:
Znalezienie się na poziomie 95 percentyla oznacza to samo bez względu na to czy chodzi o wyniki egzaminu (👍), czy masę ciała (👎). 95 percentyl oznacza, że 95% pozostałych wartości leży poniżej, a 5% powyżej.

Kwartyl

Kwartyle (inaczej – wartości ćwiartkowe) to wartości, które dzielą zebrane obserwacje na **cztery równe, co do ilości** elementów, grupy.
Pierwszy kwartyl (Q1) – 25% obserwacji położonych jest poniżej pewnej wartości, a 75% powyżej tej wartości
Drugi kwartyl (Q2) - 50% , inaczej mediana dzieli zbiór obserwacji na dwie równe części (**Mediana**)
Trzeci kwartyl (Q3) – 75% obserwacji położona jest poniżej pewnej wartości, a 25% powyżej tej wartości

Miary asymetrii i koncentracji

Skośność

Określa kierunek i siłę asymetrii

$As > 0$ – asymetria prawostronna (rozkład ma dłuższy prawy „ogon")

$As = 0$ – symetria rozkładu

$As < 0$ – asymetria lewostronna (rozkład ma dłuższy lewy „ogon")

$$A_d = \frac{\mu - d}{s}$$

$$A_m = 3\frac{\mu - m}{s}$$

$$A_Q = \frac{Q_1 + Q_3 - 2m}{2Q} = \frac{Q_1 + Q_3 - 2m}{Q_3 - Q_1}$$

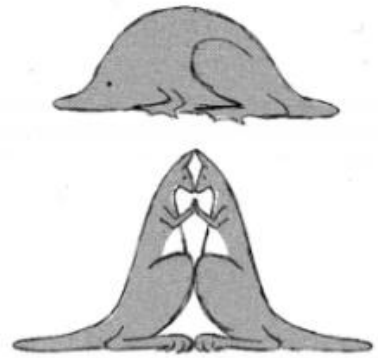
μ – średnia arytmetyczna,
m – mediana,
d – dominanta (moda),
s – odchylenie standardowe,
Q_1, Q_3 – pierwszy i trzeci kwartyl,
Q – odchylenie ćwiartkowe.

Kurtoza

Koncentracja wokół średniej

$K < 0$ - rozkład jest bardziej spłaszczony od normalnego

$K > 0$ - rozkład jest bardziej wysmukły od normalnego



$$\text{Kurt} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^4}{\sigma^4} - 3,$$

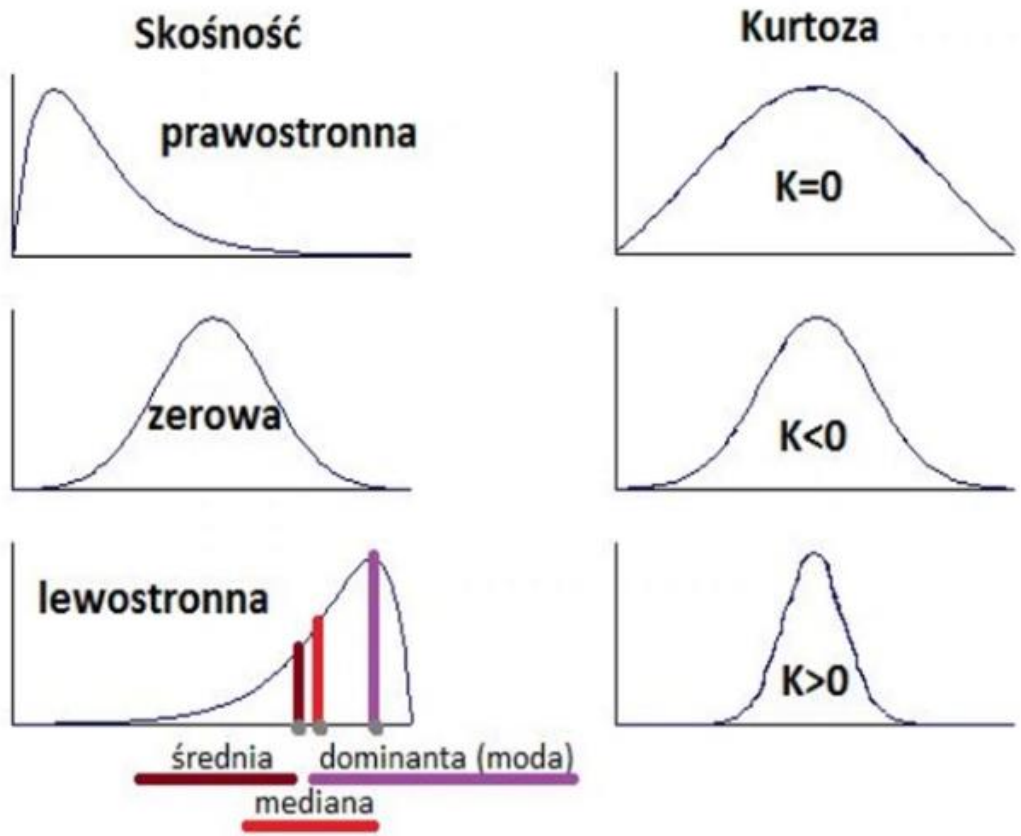
gdzie:

x_i – i-ta wartość cechy,
μ – wartość oczekiwana w populacji,
σ – odchylenie standardowe w populacji,
n – liczebność próby.

Miary asymetrii i koncentracji

Skośność

Kurtoza



Średnia – własności

Średnia spełnia 8 podstawowych własności!

1. Spełnia relację $x_{min} < \bar{x} < x_{max}$

2. Spełnia zależność: $\sum_{i=1}^{n} x_i = \bar{x}n$

3. Suma odchyleń poszczególnych wartości zmiennej X od średniej arytmetycznej jest równa 0

4. Suma kwadratów odchyleń poszczególnych wartości zmiennej X od średniej jest wartością minimalną

5. Średnia artmetyczna sumy (różnicy) zmiennych równa jest sumie (różnicy) ich średnich artmetycznych

6. Średniej nie można obliczać w przypadkach występowania licznych zakresów klasowych otwartych

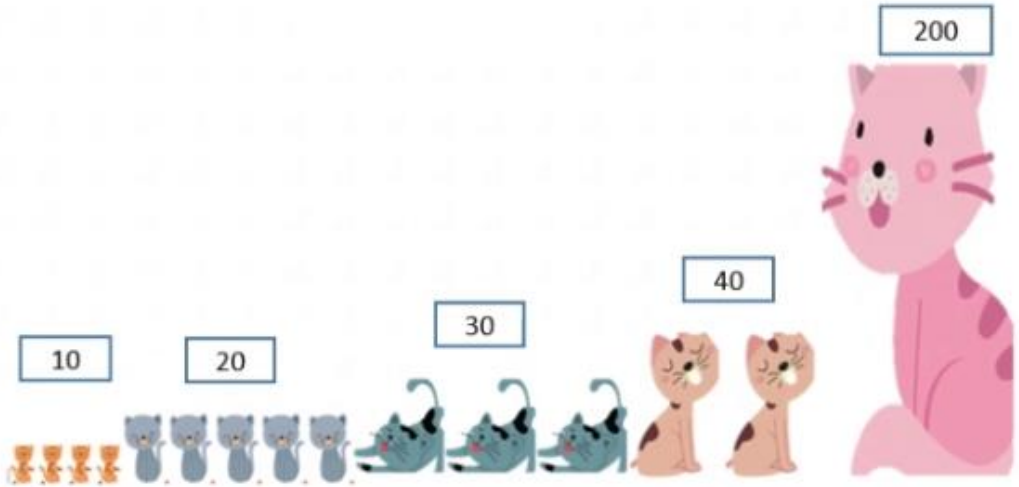
7. Średnia artmetyczna na podstawie próby losowej jest dobrym oszacowaniem wartości oczekiwanej w populacji generalnej

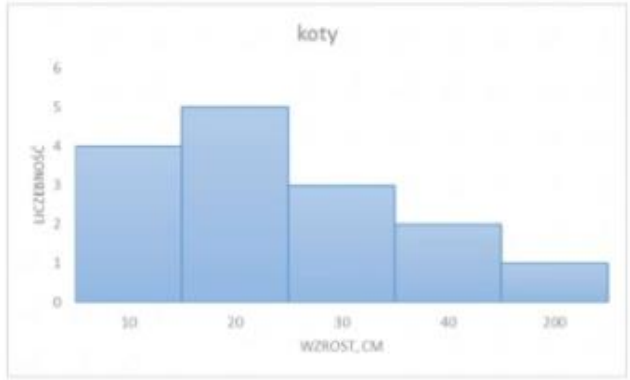
8. Średnia wrażliwa na wartości skrajne.

Średnia – własności

Średnia to przydatna i precyzyjna miara położenia, ale **ma swoje wady**!

Bardzo mocno zależy od wartości odstających!





$$\bar{x} = \frac{4 \cdot 10 + 5 \cdot 20 + 3 \cdot 30 + 2 \cdot 40 + 200}{15} = 34$$

$$M_{e\left(\frac{15+1}{2}=8\right)} = 20$$

$$M_o = 20$$