Podstawy statystyki praktycznej

- **Wykładowca** : dr. hab, prof. UWr Małgorzata Bogdan
- **Biuro**: 513
- **Godziny konsultacji**:
- Czwartki 11:00-12:00

Oceny

- Zaliczenie laboratorium
- A) Rozliczone wszystkie sprawozdania
- B) Dwa kolokwia na wykładzie: 5 kwietnia, 7 czerwca
- Zaliczenie wykladu – egzamin w sesji
- Ocena bardzo dobra z laboratorium zwalnia z egzaminu

Książki

- *Statistics for the Life Sciences*, Myra L. Samuels i Jeffrey A. Witmer
- *Introduction to the Practice of Statistics*, David S. Moore, George P. McCabe, Bruce A. Craig
- Listy zadań dostępne w internecie

Dane

Danych uzywamy aby odpowiedzieć na różne pytania naukowe

- Na ogół dane charakteryzują się losową zmiennością
- Oceniamy informację zawartą w danych
- Chcemy odróżnić sygnał od losowego szumu

Co to jest statystyka?

- Nauka dotycząca zrozumienia danych i podejmowania decyzji w obliczu losowości
- Zbiór metod do planowania eksperymentu i analizy danych służących do uzyskania maksimum informacji i ilościowej oceny ich wiarygodności

Przykład 1

- **Badania dotyczące wpływu aktywności fizycznej na poziom cholesterolu. Pytanie - Czy poziom cholesterolu jest niższy u osób, które ćwiczą ? Przeprowadzono kwestionariusz na losowo wybranej grupie osob.**
  - **Ludzie mają naturalnie różne poziomy cholesterolu**
  - **Różny stopnień zaangażowania w realizację planu ćwiczeń**
  - **Wpływ diety**
  - **Ćwiczenia mogą wpływać na inne czynniki (np. apetyt)**

Przykład 2

- **Eksperyment mikromacierzowy porównuje komórki rakowe i normalne. Czy dwukrotnie wyższy zaobserwowany poziom ekspresji genu dowodzi zwiazku aktywnosci genu z choroba ?**
- Czy mamy powtórzenia eksperymentu ? Czy w kolejnych powtórzeniach wyniki są podobne ?
- Jak ustalić właściwą wartość krytyczną ?

Przykład 3 (Lokalizacja genów)

- Gen o dwóch allelach – trzy genotypy AA, Aa, aa
- Dzielimy kłosy żyta odpowiednio na trzy grupy
- Czy różnice w przeciętnej wydajności między tymi trzema grupami są wystarczająco duże aby stwierdzić bliskość genu odpowiadającego za wydajność ?

Przykład 4

- **W artykule wyczytaliśmy, że stwierdzono, że 80 % pieszych będących ofiarami nocnych wypadków samochodowych nosiło ciemne ubrania a 20 % jasne ubrania. Wyciągnięto wniosek, że w nocy bezpiecznie jest nosić jasne ubrania.**
  - **Czy przeprowadzone badania upoważniają do takiej konkluzji ?**

Przykład 5
Reakcja owiec na bakterie wąglika – eksperyment Pasteura

Reakcja	Szczepione	Nie szczepione
Śmierć	0	25
Przeżycie	25	0
Procent przetrwania	100 %	0 %

Przykład 6
Rozwój raka wątroby u myszy

wynik	E.coli	Wolne od zarazków
Rak wątroby	8	19
Zdrowa	5	30
Suma	13	49
Procent myszy z rakiem wątroby	62 %	39 %

- Przykład 5 – brak zmienności – mocna konkluzja
- Przykład 6 – duża zmienność – słaba konkluzja
- Jak duża musi być próba abyśmy w oparciu o nią mogli dowieść wpływu czynnika na wynik eksperymentu ?

Proces naukowy/statystyczny

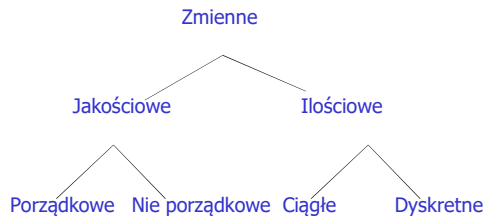
- **Pytanie naukowe**
- **Planowanie eksperymentu**
- **Eksperyment / zbieranie danych**
- **Analiza danych**
- **Wnioski statystyczne**
- **Wnioski naukowe**

Próba, Zmienna

- Próba
  - Obserwacje lub wyniki eksperymentu
  - Reprezentuje kolejne realizacje eksperymentu
- Przykłady
  - Wysokości 10 kłosów żyta (10 obserwacji)
  - Poziom hemoglobiny u 35 dawców
  - Kolor i kształt 556 fasolek w drugiej generacji (żółte/zielone, gładkie/pomarszczone)

- Rozmiar próby
  - "n"
  - n=10,n=35,n=556
- Zmienna
  - To co mierzymy
  - Wysokość, poziom hemoglobiny, kolor/kształt

Rodzaje zmiennych



Rodzaje zmiennych

- Jakościowe – kwalifikujące do kategorii
  - Porządkowe : wybory w ankiecie ; nigdy, rzadko, czasami, często, zawsze
  - Nie porządkowe : faktura, kolor; gładkie & żółte, gładkie & zielone, pomarszczone & żółte, pomarszczone & zielone

- Ilościowe – wynik jest liczbą
  - Ciągłe : wzrost, waga, stężenie
  - Dyskretne : liczba wadliwych elementów, liczba gładkich i żółtych fasolek

Oznaczenia

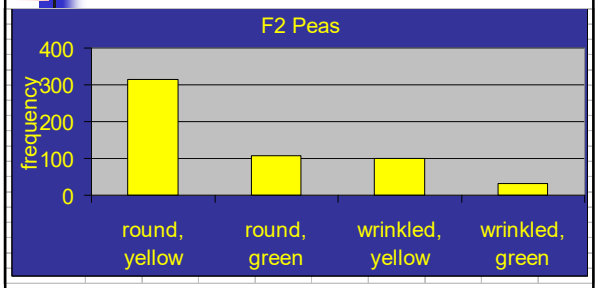
- Rozmiar próby = n, czasami n_1, n_2
- zmienne : X,Y,Z; np. Y=wzrost
- obserwacje (wyniki) : x,y,z
- Wielokrotne obserwacje $y_1, y_2, \ldots, y_n$

Reprezentacja danych jakościowych: Tabela częstości

Fasolki:gładkie/pomarszczone, zielone/żółte

Klasy	Liczba
Gładkie, żółte	315
Gładkie, zielone	108
Pomarszczone, żółte	101
Pomarszczone, zielone	32

Wykres słupkowy



Dane ilościowe dyskretne

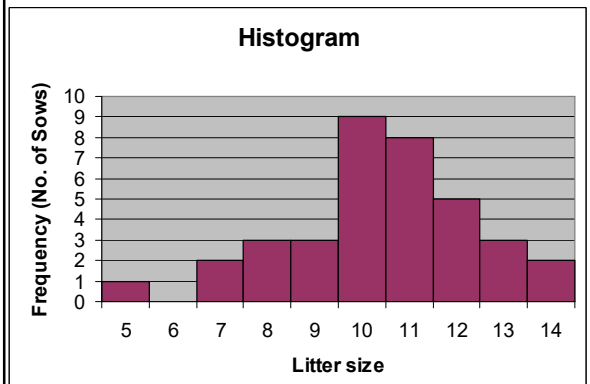
- Liczba potomków u n=36 macior. Liczba potomków jest liczbą całkowitą (zmienna dyskretna).

Dane

10	12	10	7	14	11
14	11	10	13	10	10
8	11	7	13	12	13
10	8	5	11	11	12
11	11	9	8	12	10
9	11	10	12	10	9

Rozkład częstości

Liczba potomków	Liczba macior
5	1
6	0
7	2
8	3
9	3
10	9
11	8
12	5
13	3
14	2



Histogram

- Zwykle jest pomocne grupowanie podobnych obserwacji
- Tak na ogół postępujemy z danymi ciągłymi
- Definiujemy "klasy" obserwacji i zliczamy liczbę obserwacji w każdej klasie

Jak wybierać klasy

- Każda obserwacja musi wpadać do dokładnie jednej klasy (klasy są rozłączne)
- Rozmiar (szerokość) wszystkich klas jest zwykle taki sam
- Używamy wygodnych granic, np. 20-29 a nie 19.82 – 29.26
- Używamy 5 do 15 klas dla umiarkowanych zbiorów danych ($n \leq 50$); więcej gdy próba jest duża

Przykład

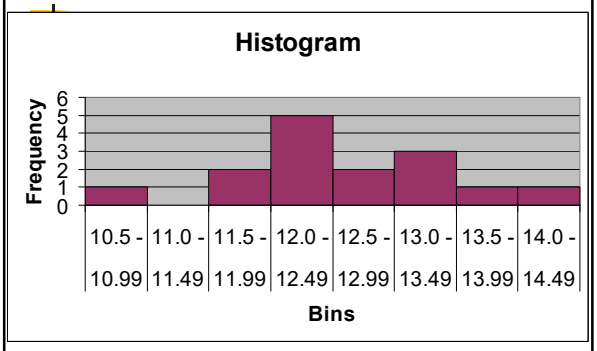
- Dane : długość łodygi papryki (n=15)

12.4	12.2	13.4	10.9	12.2
12.1	11.8	13.5	12	14.1
12.7	13.2	12.6	11.9	13.1

- Min=10.9, max=14.1, zakres=max-min=3.2
- Wybieramy szerokość klasy, np. 0.5 i punkt początkowy 10.5 aby pokryć przedział 10.5 – 14.5.
- Liczymy rozkład częstości i rysujemy histogram.
- Zmieniamy szerokość klas aby uzyskać pożądany kształt
- Za mała szerokość klas = ``postrzępiony", za duża = tracimy informację

Tabela częstości

Klasa			Częstość
10.5	-	10.99	1
11.0	-	11.49	0
11.5	-	11.99	2
12.0	-	12.49	5
12.5	-	12.99	2
13.0	-	13.49	3
13.5	-	13.99	1
14.0	-	14.49	1



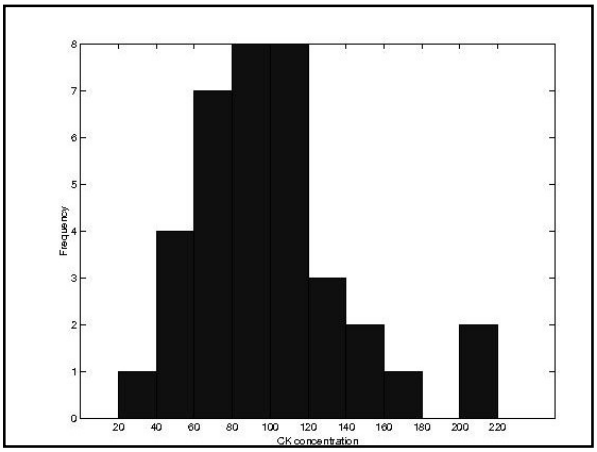
- Czasami rysujemy histogramy częstości względnej = częstość / n
- Użyteczne gdy chcemy porównać kilka zbiorów o różnych rozmiarach

Przykład Serum CK

121	82	100	151	68	58
95	145	64	201	101	163
84	57	139	60	78	94
119	104	110	113	118	203
62	83	67	93	92	110
25	123	70	48	95	42

- Min=25, max=203
- Rozstęp =178
- Szerokość klasy =20
- Punkt początkowy=20

Serum CK	Częstość
20 - 39	1
40 - 59	4
60 - 79	7
80 - 99	8
100 - 119	8
120 - 139	3
140 - 159	2
160 - 179	1
180 - 199	0
200 - 219	2
Suma	36



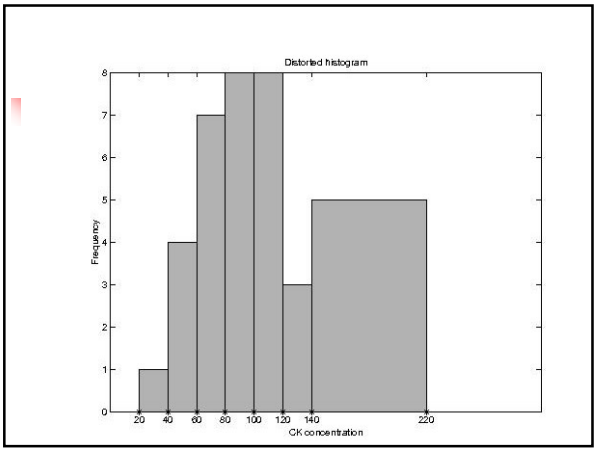
- Centralny szczyt (moda) w okolicach 100 U/Li
- Zasadnicza masa rozkładu między 40 a 140 U/Li
- Niesymetryczny – skośny na prawo

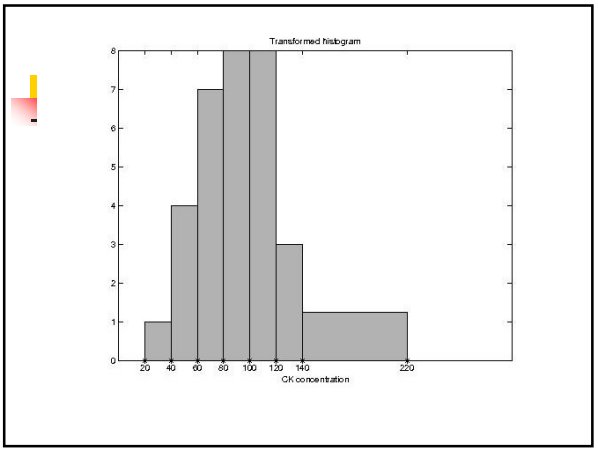
Całkowanie powierzchni pod histogramem (równa szerokość klas)

- Odcinek 60 -100 U/Li
- 42 % całkowitej powierzchni pod histogramem
- 42 % (16 out of 36) wartości CK

Nierówna szerokość klas

- Powierzchnia pod histogramem nie jest proporcjonalna do częstości
- W tak ``spaczonym" histogramie (patrz następna strona) powierzchnia między 140 – 220 stanowi 39 % całkowitej powierzchni (tylko 14 % obserwacji)
- Rozwiązanie – Podzielić odpowiednią częstość przez liczbę zgrupowanych klas
- Oś Y na przekształconym histogramie – średnia częstość w zgrupowanych klasach





Opis histogramu (rozkładu)

- Symetryczny / asymetryczny
- Skośny na prawo lub lewo
- Jednomodalny (jeden główny wierzchołek)
- Dwumodalny (dwa główne wierzchołki)
- Rozrzut (duży lub mały)

Statystyka

- Statystyka – funkcja próby
- Przykłady statystyk

próba: y_1=24, y_2=35, y_3=26 ,y_4=36

min=24, max=36, t= y_1+y_2=59

Miary położenia rozkładu

- Średnia z próby
- symbol $\bar{y}$ oznacza liczbę (arytmetyczną średnią z obserwacji)
- Symbol $\bar{Y}$ oznacza pojęcie średniej z próby
- Średnia jest ``środkiem ciężkości'' zbioru danych

Przykład: Przyrost wagi owiec

- Dane : 11, 13, 19, 2, 10,1
- y_1=11, y_2=13,..., y_6=1

$$\sum_{i=1}^{6} y_i = y_1 + y_2 + \ldots + y_6 = 11+13+\ldots+1 = 56$$

$$\bar{y} = 56/6 = 9.33$$

Odchylenia

$$dev_i = y_i - \bar{y}$$

$$dev_1 = y_1 - \bar{y} = 11 - 9.33 = 1.67$$

Σ dev_i= ?

Mediana próbkowa

- Ustawiamy obserwacje w porządku rosnącym
- Środkowa obserwacja jeżeli n jest nieparzyste
- Średnia z dwóch środkowych wartości gdy n jest parzyste

Przykłady

- **Przykład 1 (n = 5)**
  - Dane: 6.3 5.9 7.0 6.9 5.9
  - Średnia z próby $\bar{y}$ = 32/5 = 6.4
  - Mediana =
- **Przykład 2 (n = 6)**
  - Dane: 366 327 274 292 274 230
  - Średnia z próby $\bar{y}$ = 293.8
  - Mediana =

Średnia a mediana

- **Przykład 1 (n = 5)**
  - Dane: 6.3 5.9 7.0 6.9 5.9
  - Średnia $\bar{y}$ = 32/5 = 6.4
  - Mediana = 6.3
- **Błąd w zapisie danych**
  - Data: 6.3 5.9 **70** 6.9 5.9
  - Średnia $\bar{y}$ = 19
  - Mediana = 6.3

Średnia a mediana

- Mediana dzieli powierzchnię histogramu na połowę
- Jest **odporna –** nie mają na nią wpływu obserwacje ``odstające"
- Średnia to ``środek ciężkości" histogramu
- Obserwacje odstające mają duży wpływ na średnią – średnia nie jest **odporna**

Średnia a Mediana

- Jeżeli histogram jest w przybliżeniu symetryczny to średnia i mediana są zbliżone.
- Jeżeli histogram jest skośny na prawo to średnia jest zwykle większa niż mediana.
- Obie miary położenia są jednakowo ważne
- Średnia jest częściej wykorzystywana do testowania i estymacji (czego nauczymy się wkrótce).

Kwartyle

- Kwartyle dzielą zbiór danych na cztery grupy.
- Drugi kwartyl (Q2) to mediana.
- Pierwszy kwartyl (Q1) to mediana połowy obserwacji leżących poniżej Q2.
- Trzeci kwartyl (Q3) to mediana połowy obserwacji leżących powyżej Q2.

Przykład

- Dane: 3 5 6 2 1 7 4

Przykład (n=15)

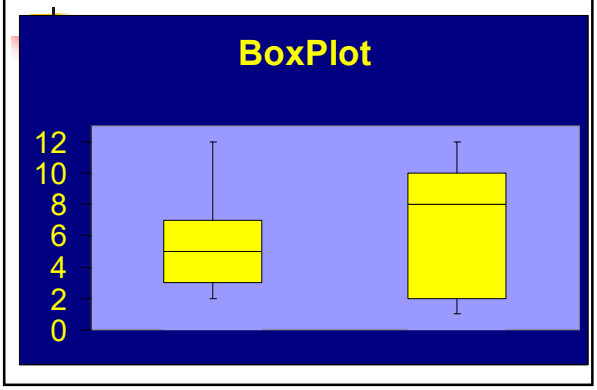
7	12	8	2	4	3	5	5
4	3	4	5	6	9	3	

Rozstęp międzykwartylowy

- IQR=Q3-Q1

Wykres pudełkowy (Boxplot)

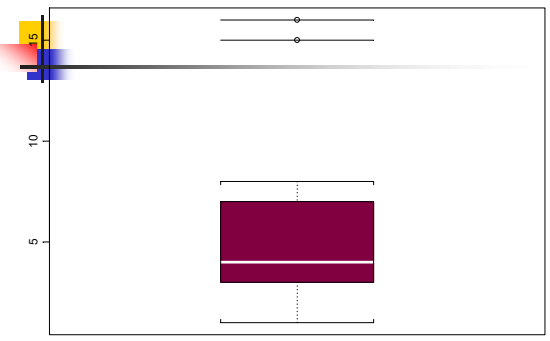
- Boxplot – graficzna reprezentacja mediany, kwartyli, maximum i minimum z danych.
  - ``Pudełko" powstaje z obrysowania kwartyli
  - Linie ciągą się do wartości najmniejszej i największej.



Zmodyfikowany Boxplot

- Obserwacja odstająca
  - Np. błąd w zapisie danych, błąd maszyny, zmiana warunków eksperymentu
- Które obserwacje są odstające ?
- Typowa propozycja:
  - Dolna granica = Q1 - 1.5*IQR
  - Górna granica = Q3 + 1.5*IQR

- Dane : 1 2 2 3 3 4 4 4 5 6 6 7 8 15 16



Miary rorzutu

- Opis danych : kształt, centrum, rorzut
- Miary rozrzutu
  - Rozstęp (max – min) – bardzo wrażliwy na obserwacje odstające, nieprzydatny do testowania
  - Rozstęp między-kwartylowy (IQR=Q3-Q1) – rozstęp środkowych 50% obserwacji
  - Standardowe odchylenie/ Wariancja
  - Współczynnik zmienności (CV)

Próbkowe odchylenie standardowe (SD, symbol s)

- Wyrażone w jednostkach pomiarowych
- Mówi jak przeciętnie obserwacje są odległe od średniej.

$$s = \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2 / (n-1)} \quad \text{(definition)}$$

$$= \sqrt{(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2)/(n-1)} \quad \text{(calculations)}$$

$$s = \sqrt{\frac{SS}{n-1}} \quad \text{,where}$$

$$SS = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$$

Próbkowa wariancja

- Przeciętny kwadrat odległości od średniej próbkowej – s^2
- Mierzona w jednostkach będących kwadratem jednostek, w których wyrażone są dane.

Dlaczego n-1 ?

- s^2 jest nieobciążonym estymatorem wariancji w populacji
- Σ dev_i=0

$$dev_n = -\sum_{i=1}^{n-1} dev_i$$

n-1 jednostek informacji = n-1 stopni swobody

Miary rozrzutu

- Współczynnik zmienności (CV)

$$CV = s / \bar{y}$$

- Przykład

  Dane : 35.1,30.6,36.9,29.8 (n=4)

  - Rozstęp =

- Suma obserwacji: Σy = 35.1 + 30.6 + 36.9 + 29.8 = 132.4
- średnia: $\bar{y}$
- SD z definicji:

  SS =

wariancja: s^2 =

- Współczynnik zmienności: CV=
- **Uwaga:** Proszę zachować dużo cyfr znaczących przy rachunkach. Zaokrąglamy dopiero na koniec.

Standardowe odchylenie (cd)

Duże SD = Duży rozrzut. Małe SD = mały rozrzut.

- Ogólne zasady
- Jeżeli rozkład jest dzwonowy (bliski normalnemu) wtedy zwykle
- 68% obserwacji jest w odległości $\pm$ 1 SD od średniej
- 95% obserwacji jest w odległości $\pm$ 2 SD od śedniej
- > 99% obserwacji jest w odległości $\pm$ 3 SD od średniej

Nierówność Czebyszewa

- Nawet gdy rozkład nie jest normalny to
- Co najmniej 75% obserwacji jest w odległości $\pm$ 2 SD od średniej
- Co najmniej 89% obserwacji jest w odległości $\pm$ 3 SD od średniej.

- Przykład

13	14	12	14	13
12	17	14	13	19
14	11	10	14	15
13	20	20	18	12

Przykład cd

- Średnia $\overline{y}$ = 14.4 i odchylenie standardowe s = 2.9.

Porównanie miar rozrzutu i położenia

- Miary rozrzutu służą do oszacowania zmienności w danych.
- Odporność
- Załóżmy, że mamy dość skupiony ``dzwonowy" (normalny) zbiór danych.
- Co się stanie gdy jedną dużą obserwację zastąpimy **bardzo dużą** wartością.

- Mediana
- Rozstęp
- Średnia
- Kwartyle i rozstęp międzykwartylowy
- Standardowe odchylenie