Podstawy statystyki praktycznej

Wykładowca: Tomasz Elsner
Konsultacje (pok. 305): poniedziałek 10^{00}–11^{00} oraz środa 7^{30}–8^{30}
Zasady zaliczenia:

- laboratorium – rozliczone wszystkie sprawozdania + aktywność
- ćwiczenia – dwa kolokwia na wykładzie (20.04 i 8.06) + aktywność
- wykład – egzamin pisemny

Literatura:

- D. Moore, G. McCabe, B. Craig, *Introduction to the Practice of Statistics*
- listy zadań dostępne na stronie: `www.math.uni.wroc.pl/~elsner`

Statystyka

Statystyka

Statystyka to nauka zbierania i interpretowania danych.

Dane

Dane to zebrane informacje o rzeczywistych osobach, przedmiotach lub zdarzeniach.

- Dane to nie są abstrakcyjne liczby, tylko liczby, które mają kontekst.
- Dane charakteryzują się losową zmiennością (chcemy odróżnić sygnał od losowego szumu).

Zbieranie danych

Badanie pewnych parametrów danej populacji

- spis powszechny (dane kompletne)
- próbkowanie (dane częściowe, reprezentatywne)

Badanie zależności między parametrami:

- obserwacja
- eksperyment

Przykłady

Przykład 1

Przeprowadzamy sondaż (i badanie) losowo wybranej grupy osób, by odpowiedzieć na pytanie: Czy aktywność fizyczna powoduje obniżenie poziomu cholesterolu?

Eksperyment musi uwzględniać:

- różny naturalny poziom cholesterolu u badanych ludzi,
- korelację między aktywnością fizyczną a zdrowym trybem życia (dieta, używki itp.),
- wpływ aktywności fizycznej na inne czynniki (np. apetyt),
- problem obiektywnego oznaczenia aktywności fizycznej.

Przykłady

Przykład 2

Eksperyment mikromacierzowy porównuje komórki rakowe i normalne. Czy dwukrotnie wyższy zaobserwowany poziom ekspresji genu dowodzi związku aktywności genu z chorobą?

- Czy wykonano powtórzenie eksperymentu i czy wyniki w kolejnych powtórzeniach są podobne?
- Ile powtórzeń należy wykonać?
- Jak ustalić wartość krytyczną wzrostu poziomu ekspresji genu?

Przykłady

Przykład 3

Pewien gen żyta o dwóch allelach ma trzy genotypy – AA, Aa, aa. W ramach eksperymentu podzielono kłosy żyta na odpowiednie trzy grupy i mierzono przeciętną wydajność każdej grupy.

- Czy różnice w wydajności są wystarczająco duże, aby stwierdzić bliskość genu odpowiadającego za wydajność?

Przykłady

Przykład 4

Dziennikarz przytoczył badania mówiące, że 80% pieszych będących ofiarami nocnych wypadków samochodowych nosiło ciemne ubrania, a 20% – jasne ubrania. Wyciągnął z tego wniosek, że w nocy bezpieczniej jest nosić jasne ubrania.

- Czy przeprowadzone badania upoważniają do takiej konkluzji?

Przykłady

Przykład 5

Reakcja owiec na bakterie wąglika – eksperyment Pasteura.

Reakcja	Szczepione	Nie szczepione
śmierć	0	25
przeżycie	25	0
procent przetrwania	100%	0%

- Brak zmienności – silna konkluzja.

Przykłady

Przykład 6

Rozwiój raka wątroby u myszy.

Wynik	Zakażone bakterią E.coli	Wolne od zarazków
rak wątroby	8	19
zdrowa	5	30
razem	13	49
procent przetrwania	62%	39%

- Duża zmienność – słaba konkluzja.
- Jak duża musi być próba, byśmy w oparciu o nią mogli dowieść wpływu czynnika na wynik eksperymentu?

Proces naukowy/statystyczny

1. Pytanie naukowe
2. Planowanie eksperymentu
3. Eksperyment / zbieranie danych
4. Analiza danych
5. Wnioski statystyczne
6. Wnioski naukowe

Populacja

Populacja (population)

Populacja do zbiór (ludzi, zwierząt, przedmiotów) podlegający badaniu statystycznemu.

Obserwacja (case)

Obserwacja to element populacji, który poddajemy badaniu (typowe oznaczenia: x, y, z lub dla wielokrotnych obserwacji $x_1, x_2, \ldots, x_n$).

Próba (sample)

Próba to zbiór obserwacji. Próba powinna być reprezentatywna dla populacji. Rozmiar próby to liczba obserwacji (typowe oznaczenia: n lub n_1, n_2).

Zmienne

Zmienna (variable)

Zmienna to funkcja, której dziedziną jest zbiór obserwacji, np. wysokość, kolor, wykształcenie (typowe oznaczenia: X, Y, Z).

Etykieta (label)

Etykieta, to specjalna zmienna, która służy jedynie do rozróżniania poszczególnych obserwacji, np. nr indeksu, nazwisko, nr klienta.

Rodzaje zmiennych

Zmienne jakościowe (categorical variables)

Zmienna jakościowa przyporządkowuje każdą obserwację do jednej z grup (kategorii). Zmienne jakościowe dzielimy na:

- porządkowe – kategorie są uporządkowane według nasilenia mierzonej cechy, np. *nigdy–rzadko–czasami–często–zawsze*;
- nie porządkowe, np. płeć, kolor, system operacyjny komputera.

Zmienne ilościowe (quantitative variables)

Zmienna ilościowa każdej obserwacji przypisuje liczbę. Zmienne ilościowe dzielimy na:

- ciągłe – wartość może być dowolną liczbą rzeczywistą z pewnego przedziału,
- dyskretne – zmienna może przyjmować tylko niektóre wartości, np. liczby całkowite.

Zmienne ilościowe

- Zmienne ilościowe pozwalają na wykonywanie operacji arytmetycznych (np. wyliczanie średniej).
- Czasami zmienne jakościowe porządkowe traktujemy jako zmienne ilościowe (np. ankiety ewaluacyjne).
- Zmienne ilościowe muszą mieć podaną jednostkę.

Zbiór danych

Zbiór danych musi również zawierać:

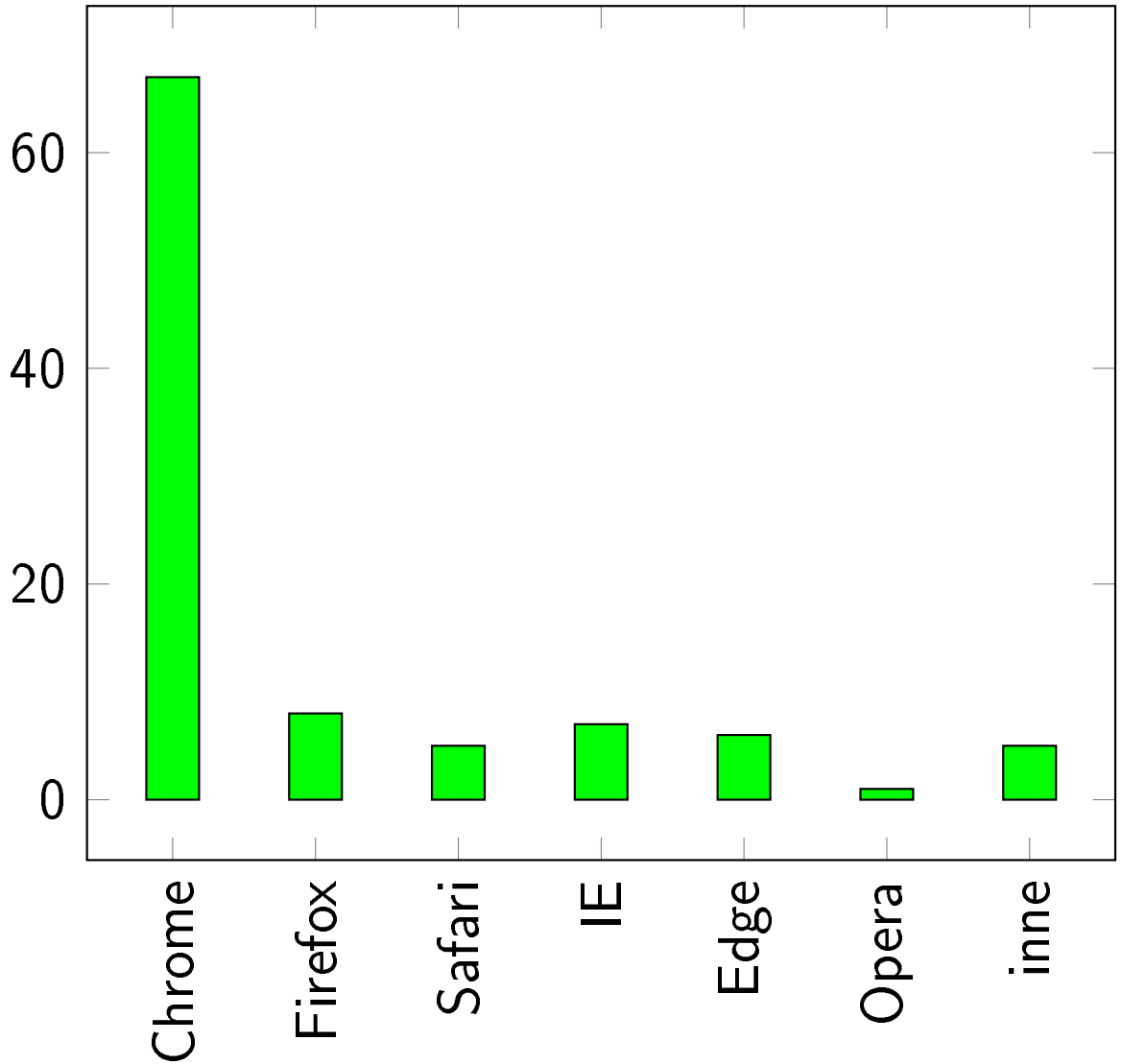
- informacje o całej populacji (np. klienci wybranego banku, studenci pewnego kierunku studiów, mieszkańcy Wrocławia),
- informacje o sposobie wyboru próby z populacji,
- informacje o sposobie zbierania danych,
- precyzyjną definicję zmiennych (np. pytania zadawane w ankiecie),
- definicje kategorii (dla zmiennych jakościowych) lub jednostki (dla zmiennych ilościowych).

Wnioskowanie statystyczne

Wnioskowanie statystyczne (statistical inference)

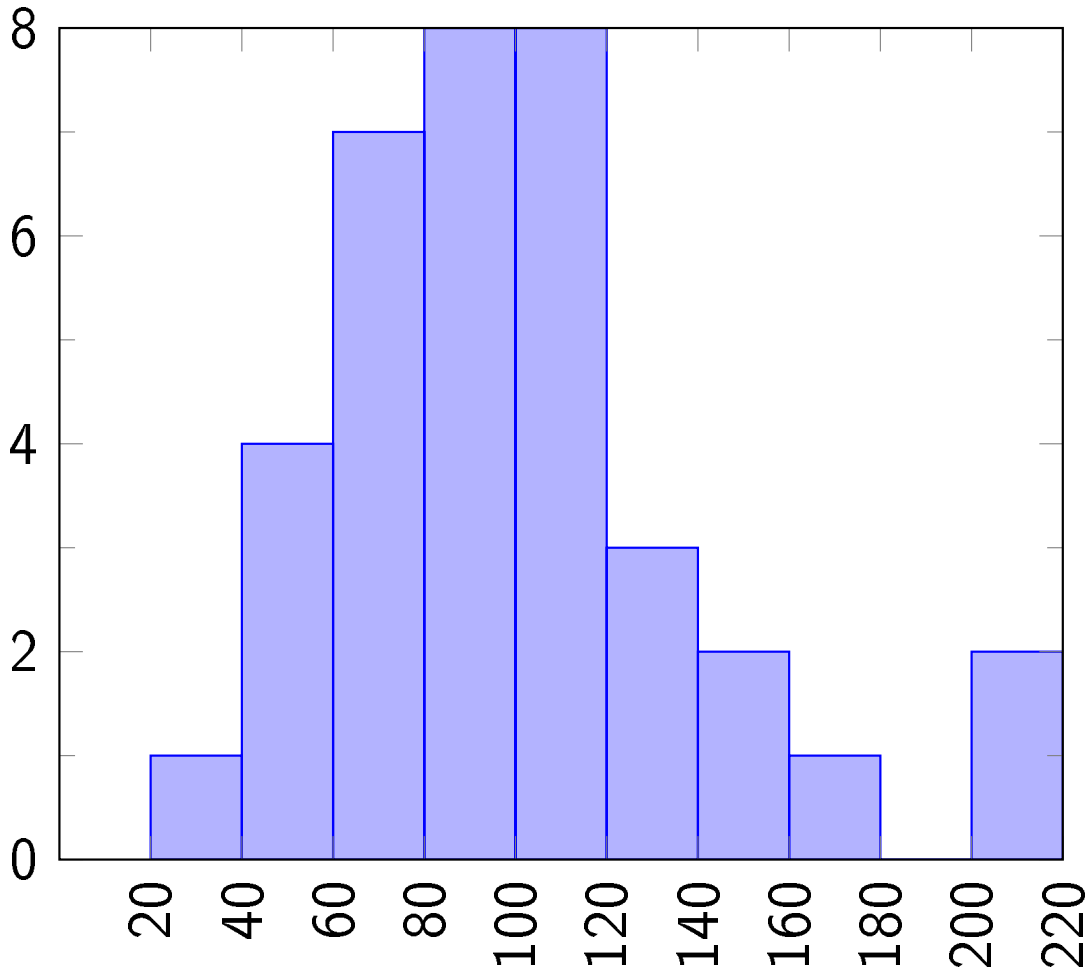
Wnioskowanie statystyczne to wnioskowanie o populacji w oparciu o próbę.

Prezentacja danych jakościowych



Prezentacja danych ilościowych

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



Histogram

Jak wybierać klasy?

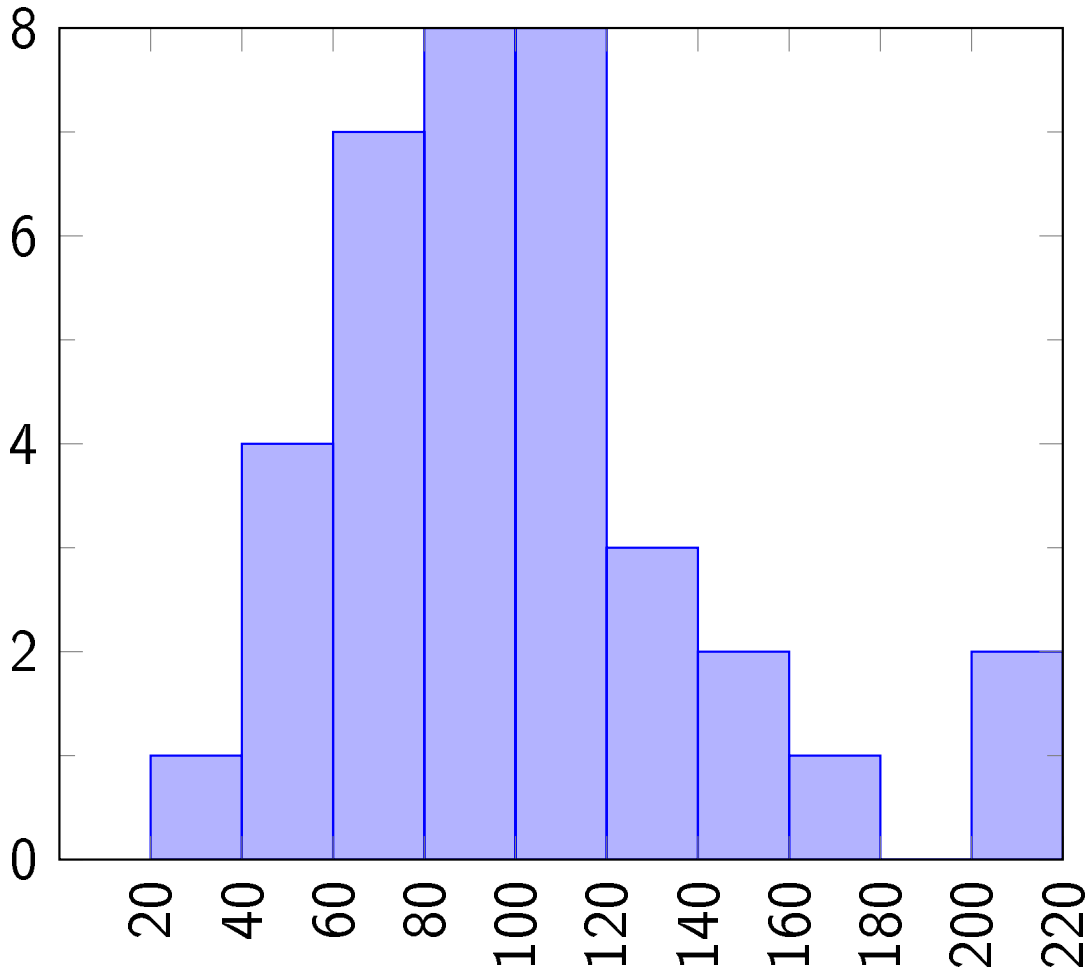
- klasy muszą być rozłączne
- rozmiary (szerokości) klas zwykle są jednakowe
- używamy wygodnych („okrągłych") granic klas
- używamy 5–15 klas dla umiarkowanych zbiorów danych ($n \leq 50$) lub więcej, gdy próba jest duża

Za mała szerokość klas – wykres „postrzępiony", za duża – tracimy informacje

Czasami rysujemy histogramy częstości względnej (częstość/n) – użyteczne, gdy chcemy porównać kilka prób różnych rozmiarów.

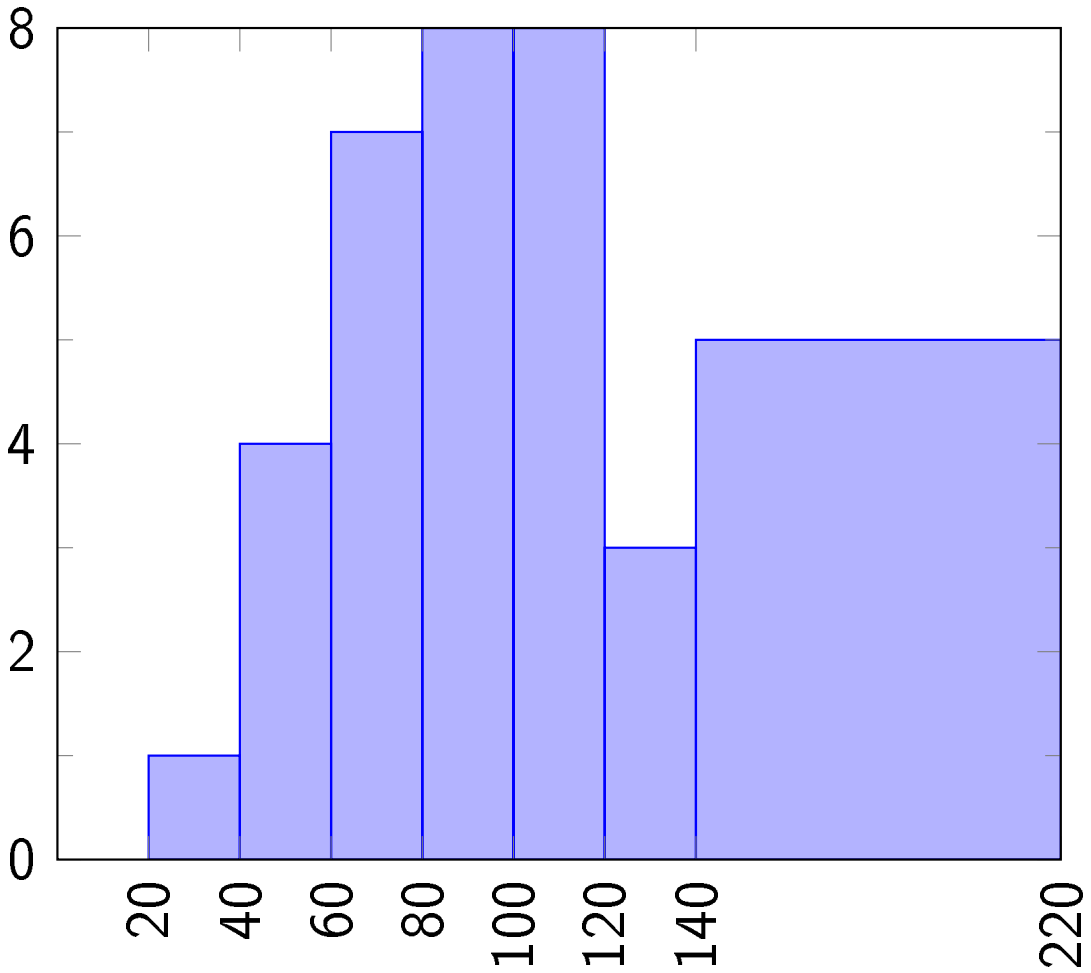
Prezentacja danych ilościowych

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



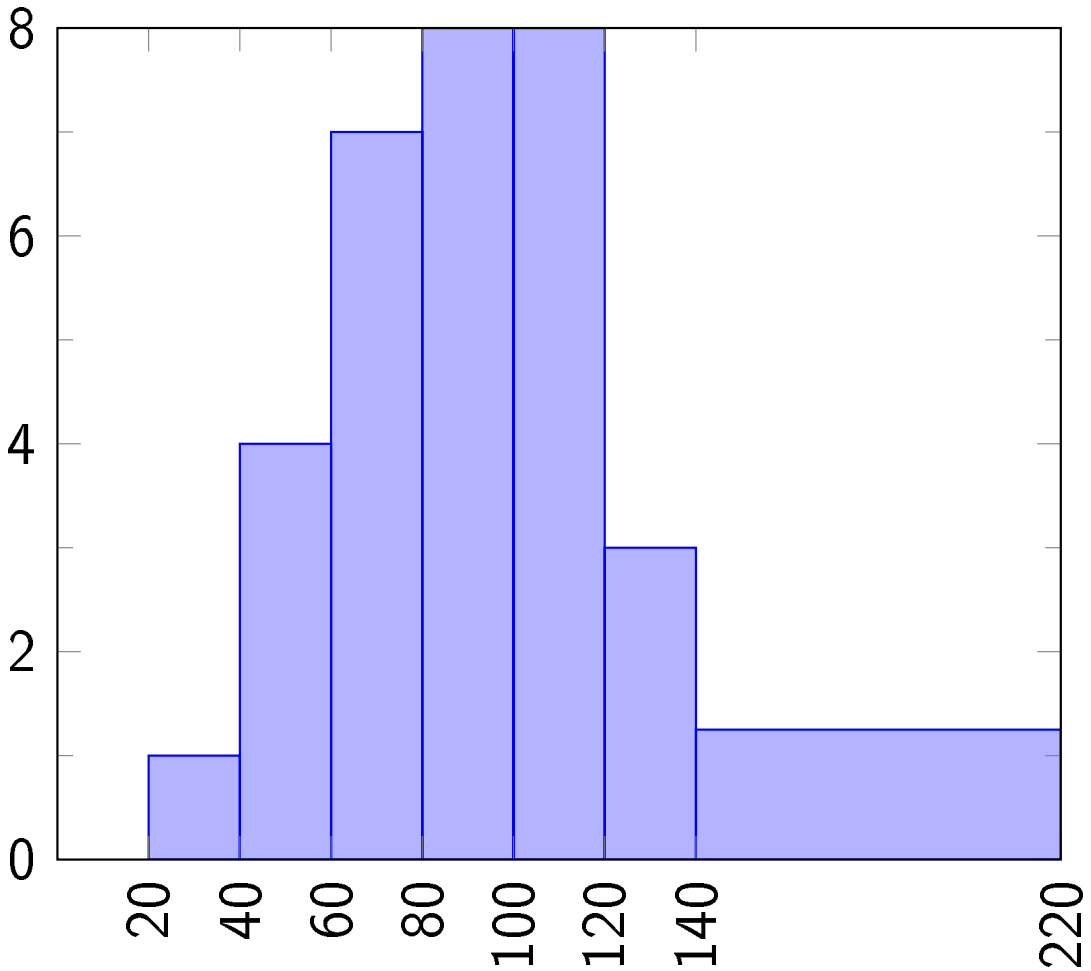
Prezentacja danych ilościowych

Serum CK	Częstość
20–39	1
40–59	4
60–79	7
80–99	8
100–119	8
120–139	3
140–219	5
Suma	36



Prezentacja danych ilościowych

Serum CK	Częstość
20–39	1
40–59	4
60–79	7
80–99	8
100–119	8
120–139	3
140–219	5
Suma	36



Opis histogramu

Kształt

- Histogram symetryczny,
- Histogram asymetryczny, skośny w lewo/prawo.

Środek

- Moda (główny wierzchołek): histogram jedno- / dwumodalny;
- Średnia
- Mediana

Rozrzut

- Rozstęp
- Rozstęp międzykwartylowy
- Odchylenie standardowe / wariancja
- Współczynnik zmienności

Statystyka

Statystyka (statistic)

Statystyka to funkcja próby.

Przykłady: minimum, maximum, średnia, mediana, kwartyle itp.
Oznaczenia: $\bar{Y}$ to funkcja (zmienna), $\bar{y}$ to wartość dla konkretnej próby

Średnia i mediana

Średnia (mean)

Średnią zbioru obserwacji $x_1, \ldots, x_n$ nazywamy liczbę:

$$\bar{x} = \frac{x_1 + \cdots + x_n}{n}$$

Mediana (median)

Medianą uporządkowanego zbioru obserwacji $x_1 \leq \cdots \leq x_n$ nazywamy liczbę:

$$M = \begin{cases} x_{n+1/2}, & \text{gdy } n \text{ nieparzyste} \\ \frac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}), & \text{gdy } n \text{ parzyste} \end{cases}$$

Średnia i mediana

- mediana jest odporna (niewrażliwa na obserwacje odstające), średnia jest nieodporna (obserwacje odstające mają duży wpływ),
- średnia jest łatwiejsza do wyliczenia (do jej wyliczenia wystarczy znajomość sumy zmiennych),
- mediana dzieli pole histogramu na połowę, średnia stanowi środek ciężkości histogramu
- średnia jest częściej wykorzystywana do testowania i estymacji (choć obie miary położenia są jednakowo ważne)
- dla symetrycznego histogramu średnia i mediana są zbliżone; dla histogramu skośnego w prawo średnia jest zwykle większa niż mediana,

Odporność

Przykład 1

- Dane: 6.3 4.9 5.2 9.1 6.7 8.4 4.1 5.9 8.7 3.1
- Średnia $\bar{x} = 6.24$
- Mediana $M = 6.1$

Przykład 2 (błąd w zapisie danych)

- Dane: 6.3 4.9 52 9.1 6.7 8.4 4.1 5.9 8.7 3.1
- Średnia $\bar{x} = 10.92$
- Mediana $M = 6.5$

Kwartyle

Kwartyl (quartile)

Pierwszym kwartylem Q_1 uporządkowanego zbioru obserwacji $x_1 \leq \cdots \leq x_n$ nazywamy medianę obserwacji leżących na lewo od mediany zbioru. Trzecim kwartylem Q_3 nazywamy medianę obserwacji leżących na prawo od mediany całego zbioru. Drugi kwartyl to mediana $Q_2 = M$.

Przykład

3.1 4.1 4.9 5.2 5.9 | 6.3 6.7 8.4 8.7 9.1

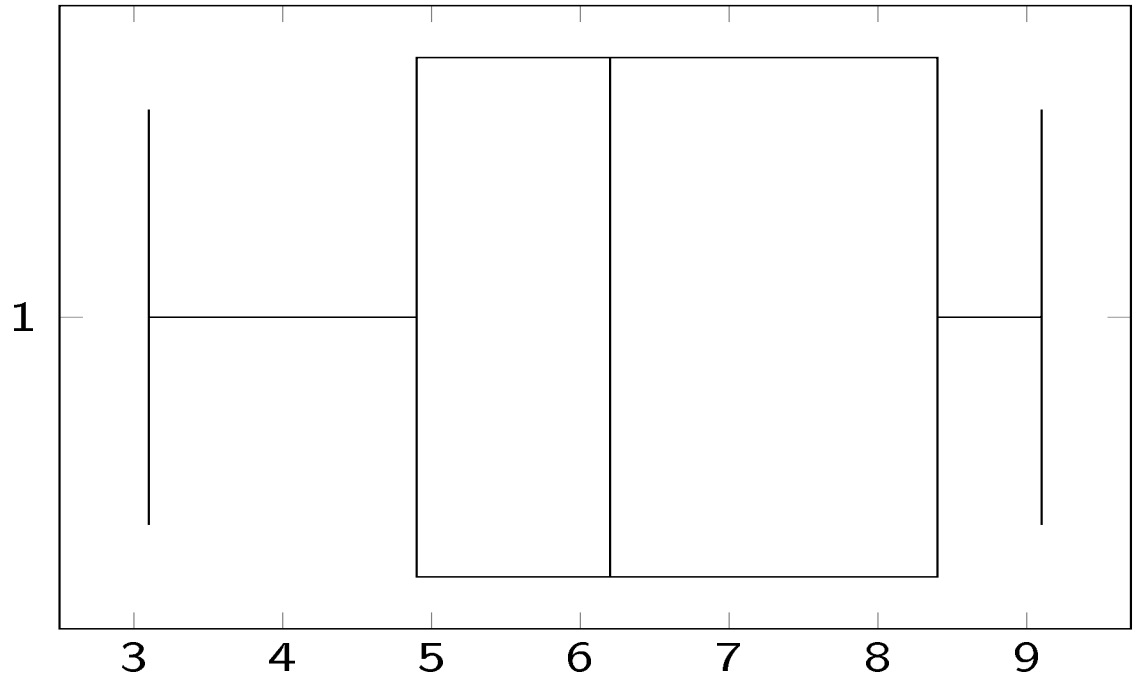
Stąd: $Q_1 = 4.9$ $Q_2 = M = 6.2$ $Q_3 = 8.4$.

Wykres pudełkowy

Poniższe 5 liczb dzieli zbiór obserwacji na 4 równoliczne części:

$$\min \qquad Q_1 \qquad M \qquad Q_3 \qquad \max$$

Przedstawiamy to graficznie w postaci wykresu pudełkowego:



Obserwacje odstające

Obserwacje odstające mogą wynikać z błędu w zapisie danych, błędu maszyny, zmiany warunków eksperymentu. Typowo jako obserwacje odstające traktuje się obserwacje nie mieszczące się w przedziale:

$$[Q_1 - 1.5 \cdot IQR, Q_3 + 1.5 \cdot IQR]$$

Miary rozrzutu

Miary rozrzutu służą do szacowania zmienności danych.

- rozstęp (spread) $=$ max $-$ min (bardzo wrażliwy na obserwacje odstające, nieprzydatny do testowania)
- rozstęp międzykwartylowy (interquartile range IQR) $= Q_3 - Q_1$ (rozstęp środkowych 50% obserwacji)
- odchylenie standardowe (standard deviation SD) / wariancja (variance)
- współczynnik zmienności (coefficient of variation CV) $= \frac{\sigma}{\mu}$

Próbkowe odchylenie standardowe

Odchylenie standardowe (standard deviation)

Próbkowym odchyleniem standardowym nazywamy liczbę:

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

Wariancja (variance)

Próbkową wariancją nazywamy liczbę:

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

Nierówność Czebyszewa

Nierówność Czebyszewa

Mniej niż $\frac{1}{k^2}$ wszystkich obserwacji znajduje się w odległości większej niż $k\sigma$ od średniej.

Wniosek

Przynajmniej 75% obserwacji leży w odległości nie większej niż 2σ od średniej.
Przynajmniej 89% obserwacji leży w odległości nie większej niż 3σ od średniej.