Statystyka i Analiza Danych

W1: Statystyka opisowa

dr hab. inż. Katarzyna Filipiak, prof. PP

Instytut Matematyki
Politechnika Poznańska

2023/2024

Organizacja zajęć

Wykład – 32h

Laboratorium – 24h mgr Malwina Mrowińska

R: https://cloud.r-project.org/ oraz RStudio Desktop

Egzamin – uzyskanie minimum 50% punktów na teście z zagadnień przedstawianych na wykładach

Zaliczenie laboratorium – aktywne uczestnictwo w zajęciach oraz uzyskanie minimum 50% punktów z każdego z kolokwium obejmujących materiał przedstawiony na zajęciach

Program

- Statystyka opisowa (interpretacja graficzna danych i miary statystyczne)
- Zmienne losowe i ich rozkłady - przypomnienie
- Statystyka
  - teoria estymacji (dla jednej i dwóch ppopulacji)
  - teoria weryfikacji hipotez (dla jednej i dwóch populacji)
  - analiza wariancji (jendo- i dwu-kierunkowa)
  - analiza regresji (regresja liniowa jednej i wielu zmiennych, regresja wielomianowa)
  - test chi-kwadrat (zgodność i niezależność)
  - testy nieparametryczne

Literatura

- Krysicki, W., J. Bartos, W. Dyczka, K. Królikowska i M. Wasilewski, *Rachunek prawdopodobieństwa i statystyka matematyczna w zadaniach*, cz. II, PWN Warszawa, 1986.
- Bobrowski D. i K. Maćkowiak-Łybacka, *Wybrane metody wnioskowania statystycznego*, Wyd. PP, Poznań 2004.
- Lapin, L.L., *Probability and Statistics for Modern Engineering*, PWS Engineering, Boston, Massachusetts, 1983.
- Ross, S.M., *Introductory Statistics* (3rd ed), Academic Press, 2010.
- Biecek, P., *Przewodnik po pakiecie R*, Oficyna Wyd. GiS, Wrocław 2008.

Populacja i próba statystyczna

Populacja:

- zbiorowość wszystkich elementów stanowiących podmiot badania (populacja przedmiotowa)
- zbiór wszystkich możliwych do zaobserwowania wartości cechy opisującej badane zjawisko (populacja zdarzeniowa)

Próba – podzbiór populacji dostępny badaczowi i stanowiący podstawę jego wnioskowania o populacji statystycznej.

Typy danych (cech)

Dane jakościowe – cechy niemierzalne, np. kolor oczu, kształt liścia, ocena bólu, poziom zarobków

Dane ilościowe – cechy mierzalne:

- dyskretne (skokowe) – gdy zbiór wartości jest skończony lub przeliczalny (pomiaru takich cech dokonujemy na ogół poprzez "zliczanie")
- ciągłe – gdy zbiór wartości jest nieprzeliczalny (pomiaru takich cech dokonujemy na ogół poprzez "mierzenie")

Prezentacja danych dyskretnych

- dane surowe lub szereg pozycyjny:
  - wpisywanie danych: `c(`*dane rozdzielone przecinkami*`)`
  - wczytywanie danych z pliku csv:

standardowo: `read.csv("`*nazwa*`", sep = ";")`
dane kodowane z przecinkiem: `read.csv("`*nazwa*`", sep = ";", dec = ",")`
dane z etykietami: `read.csv("`*nazwa*`", sep = ";", head = TRUE)`

- szereg rozdzielczy punktowy: `table(`*dane*`)`
- histogram odcinkowy (rozkład częstości / liczebności):

  `discrete.histogram(`*dane*`)` `discrete.histogram(`*dane*`,freq = T)`

  UWAGA! Wymagany pakiet "arm"

  `plot(table(`*dane*`)/length(`*dane*`))` `plot(table(`*dane*`))`

- wykres kołowy: `pie(table(`*dane*`))`

Przykład

Przeprowadź eksperyment, w którym studenci zostają zapytani o liczbę rowerów w gospodarstwie domowym (student+rodzice+rodzeństwo lub student+partner+dzieci). Otrzymane dane przedstaw za pomocą histogramu odcinkowego.

wprowadzenie danych w R: *rowery* = `c(1,4,0,...,3)`

instalacja wymaganego pakietu: `install.packages("arm")`

wczytanie pakietu: `library(arm)`

narysowanie histogramu: `discrete.histogram(`*rowery*`)`

Prezentacja danych ciągłych

- dane surowe
- szereg rozdzielczy przedziałowy: `table(cut(`*dane*`,`k`))`
  (k – liczba przedziałów klasowych)
- histogram liczebności:
  `hist(`*dane*`, main=`*tytuł*`, xlab=`*etykieta osi OX*`)`
- histogram częstości:
  `hist(`*dane*`, main=`*tytuł*`, xlab=`*etykieta osi OX*`, freq=FALSE)`
- łamane, wieloboki, krzywe (liczebności, częstości)
- wykresy kołowe liczebności: `pie(table(cut(`*dane*`,`k`)))`

Konstrukcja szeregu rozdzielczego

Zasady ogólne

- klasy obejmują wszystkie jednostki badanej zbiorowości
- klasy są rozłączne
- klasy są niepuste

Liczba klas, k (n - liczba obserwacji):

$$k \approx \sqrt{n}, \quad \frac{\sqrt{n}}{2} \leqslant k \leqslant \sqrt{n}, \quad k \leqslant 5 \log n, \quad k \approx 1 + 3,322 \log n$$

Rozpiętość klas, h:

$$h \approx \frac{x_{(n)} - x_{(1)}}{k} \quad \text{(zaokrąglone w górę)}$$

Przykład

Przedstaw za pomocą szeregu rozdzielczego przedziałowego następujące dane pochodzące z pomiaru koncentracji ozonu w 78 miejscach w badanym regionie. Zilustruj rozkład otrzymanych danych.

3,5	1,7	3,1	4,5	3,0	6,1	6,8	1,1	5,8	4,2	6,0	8,1
2,4	7,5	4,7	5,4	1,4	2,0	6,8	5,8	5,7	6,5	2,8	6,2
5,5	3,4	6,0	7,4	2,5	5,6	6,2	3,1	4,4	5,5	3,7	4,0
5,7	4,4	4,7	5,8	3,3	3,4	9,4	6,6	4,7	1,6	9,4	11,7
6,8	5,4	5,6	1,4	5,3	6,6	6,6	5,6	6,0	5,9	3,5	4,1
1,4	5,3	5,8	3,7	4,1	6,6	2,5	5,1	7,6	4,4	6,7	3,7
3,0	6,2	3,8	4,7	3,9	7,6						

maksimum: `max(dane)` minimum: `min(dane)`

zaokrąglanie w górę: `ceiling(dane)` △

Charakterystyki liczbowe

$x_1, x_2, \ldots, x_n$ – dane surowe

$x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ – dane uporządkowane niemalejąco (szereg pozycyjny)

Miary:

- położenia
- zmienności (rozproszenia, zróżnicownia)
- asymetrii (skośności)
- koncentracji i skupienia
- ...

Miary położenia - średnia

Dla szeregu szczegółowego : $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ `mean(`*dane*`)`

Dla szeregu rozdzielczego punktowego: $\overline{x} = \frac{1}{n}\sum_{i=1}^{k} x_i n_i$

Dla szeregu rozdzielczego przedziałowego:

$\overline{x} = \frac{1}{n}\sum_{i=1}^{k} m_i n_i$ (m_i – środek przedziału klasowego)

Miary położenia - dominanta (moda)

Dla szeregu szczegółowego, pozycyjnego lub rozdzielczego punktowego:

wartość występująca najczęściej

Miary położenia - kwantyle

$100p$-ty **kwantyl (percentyl)**, $0 \leqslant p \leqslant 1$ – ta obserwacja $x_{[p]}$, dla której w szeregu pozycyjnym co najmniej $100p\%$ danych jest niewiększych i co najmniej $100(1-p)\%$ danych jest niemniejszych od $x_{[p]}$:

$$x_{[p]} = \begin{cases} x_{(\lfloor pn \rfloor + 1)} & \text{gdy} \quad pn \notin \mathbb{N} \\ \frac{1}{2}\left(x_{(pn)} + x_{(pn+1)}\right) & \text{gdy} \quad pn \in \mathbb{N} \end{cases}$$

Dla szeregu rozdzielczego punktowego:

- wskazujemy klasę, w której po raz pierwszy liczebność skumulowana osiągnie lub przekroczy pn
- $x_{[p]}$ jest równy wartości cechy we wskazanej klasie

```
quantile(dane, probs = p)
```

Miary położenia - kwartyle

Q_1 – pierwszy kwartyl to 25-kwantyl

$Q_2 = x_{me}$ – drugi kwartyl lub mediana to 50-kwantyl

Q_3 – trzeci kwartyl to 75-kwantyl

`quantile(`*dane*`)`

Miary tendencji centralnej razem:

`summary(`*dane*`)`

Miary rozproszenia (zmienności)

Wariancja:

- dla danych surowych:

$$s^2 = \frac{1}{\ell}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{\ell}\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right)$$

- dla danych pogrupowanych:

$$s^2 = \frac{1}{\ell}\sum_{i=1}^{k}(m_i - \overline{x})^2 n_i = \frac{1}{\ell}\left(\sum_{i=1}^{k} m_i^2 n_i - n\overline{x}^2\right)$$

$\ell = n$ – wariancja populacyjna

$\ell = n - 1$ – wariancja z próby: `var(`*dane*`)`

Odchylenie standardowe: $s = \sqrt{s^2}$ `sd(`*dane*`)`

Miary rozproszenia

Rozstęp: $R = x_{(n)} - x_{(1)}$

Rozstęp ćwiartkowy: $R_Q = Q_3 - Q_1$

Współczynnik zmienności: $v = \frac{s}{\overline{x}} \cdot 100\%$

Interpretacja współczynnika zmienności:

0 – 20%	–	**słabe** zróżnicowanie danych
20 – 40%	–	**umiarkowane** zróżnicowanie danych
40 – 60%	–	**silne** zróżnicowanie danych
powyżej 60%	–	**bardzo silne** zróżnicowanie danych

Przykład

W celu porównania dwóch pięcioosobowych grup studentów ze względu na oceny uzyskane z przedmiotu STATYSTYKA, zebrano następujące dane:

grupa A	3,0	3,0	4,0	4,5	4,5
grupa B	2,0	3,5	4,0	4,5	5,0

Za pomocą poznanych miar położenia i zmienności porównaj wspomniane grupy studentów.

Przykład

Interpretacja graficzna

Wykres pudełkowy (wykres ramka-wąsy, ang. *box-plot*) – pozwala ująć na jednym rysunku miary położenia, rozproszenia i kształtu rozkładu empirycznego badanej cechy.

Wykres tworzymy odkładając na osi poziomej:

$$x_{min},\ Q_1,\ x_{me},\ Q_3,\ x_{max}.$$

Nad osią umieszczony jest prostokąt (pudełko), którego lewy bok jest wyznaczony przez Q_1, a prawy przez Q_3. Szerokość pudełka odpowiada wówczas wartości rozstępu ćwiartkowego R_Q. Wewnątrz prostokąta znajduje się pionowa linia określająca wartość mediany x_{me}. Rysunek pudełka uzupełniany jest po prawej i lewej stronie odcinkami zwanymi wąsami, przy czym końce odcinków wyznaczają odpowiednio x_{min} i x_{max}.

`boxplot(`*dane*`)`

Przykład

W celu porównania dwóch pięcioosobowych grup studentów ze względu na oceny uzyskane z przedmiotu STATYSTYKA, zebrano następujące dane:

grupa A	3,0	3,0	4,0	4,5	4,5
grupa B	2,0	3,5	4,0	4,5	5,0

Porównaj grupy studentów za pomocą wykresów pudełkowych.

boxplot(*grupaA*, *grupaB*)

Reguła Czebyszewa

Dla dowolnego zbioru danych:

- przedział $(\overline{x} - 2s; \overline{x} + 2s)$ zawiera co najmniej 75% $\left(1 - \frac{1}{2^2}\right)$ danych
- przedział $(\overline{x} - 3s; \overline{x} + 3s)$ zawiera co najmniej 89% $\left(1 - \frac{1}{3^2}\right)$ danych
- przedział $(\overline{x} - ks; \overline{x} + ks)$ zawiera co najmniej $1 - \frac{1}{k^2}$ danych $(k > 1)$