Statystyka i analiza danych
Laboratorium 1
Wprowadzenie do laboratorium, Mini-wprowadzenie do R, Grupowanie i histogramy

Andrzej Szwabe

Instytut Informatyki
Politechnika Poznańska (PP)
Piotrowo 3, 60-965 Poznan, Poland
Email: Andrzej.Szwabe@put.poznan.pl

25 lutego 2023

Podstawowe informacje (1/2)

- Materiały laboratorium znajdują się na eKursie - dostępnym na stronie *http://andrzej.szwabe.pracownik.put.poznan.pl/laboratorium-statystka-i-analiza-danych/*
- Prowadzący laboratorium to dr inż. Andrzej Szwabe
  - Instytut Informatyki
  - Zakład Inteligentnych Systemów Wspomagania Decyzji
  - Dyżur: Do odwołania dyżury odbywają się w pokoju 318 w budynku A1 (z zegarem) (możliwy udział zdalny studenta):
    - we wtorki pomiędzy 13:00 a 15:00,
    - Proszę o wcześniejsze ustalenie godziny i/lub linku via email/Slack.
  - andrzej.szwabe@put.poznan.pl
    (Pisząc emaile proszę zaczynać tytuł wiadomości od i identyfikatora grupy laboratoryjnej (np. *[sad-l3]*).)

Podstawowe informacje (2/2)

- Profil przedmiotu Statystyka i analiza danych jest przedstawiany na pierwszym wykładzie.
- Wykłady prowadzi prof. Jerzy Stefanowski (w środę od 11:45 w s. L053 BT)
  - Główne materiały (w tym karta ECTS/sylabus) znajdują się na stronie eKursu.
  - Lista podręczników znajduje się w stosownej sekcji karty ECTS.
- Zasady prowadzeniach zajęć i zaliczania w 13 grupach laboratoryjnych są z założenia wspólne.

Plan na dziś

- Zapoznanie z laboratorium (przede wszystkim z oprogramowaniem, w tym ze środowiskiem języka R)
- Krótkie zapoznanie z przedmiotem
- Zasady zaliczenia
- Harmonogram
- Rejestracja (kwestionariusz/Moodle/DataCamp)
- Ćwiczenia z grupowania i histogramów

Zasady zaliczenia (1/3)

- Na ocenę z laboratorium składają się:
  - 70% - kartkówki (na początku zajęć obejmujące materiał z poprzednich zajęć) oraz tutoriale (na platformie DataCamp),
  - 30% - zadanie domowe.
- By zaliczyć laboratorium należy łącznie uzbierać co najmniej 51% punktów. Część przedziałowego szereg rozdzielczego skali ocen dla ocen pozytywnych (;-)) jest następujący:
  - 51-60% -> 3.0
  - 61-70% -> 3.5
  - 71-80% -> 4.0
  - 81-90% -> 4.5
  - 91-100% -> 5.0

Zasady zaliczenia (2/3)

- Dopuszcza się maksymalnie 2 nieusprawiedliwione nieobecności.
- Nieobecności należy usprawiedliwiać w ciągu 2 tygodni.
- Planuje się 10 kartkówek oraz 2 tutoriale.
- Tutoriale wykonywane będą na platformie DataCamp - dotyczyć będą programowania w R, realizowane będą jako zadanie domowe z tygodniowym czasem na wykonanie.

Zasady zaliczenia (3/3)

- Kartkówki będą miały postać quizu na eKursie z około 2 zadaniami sprawdzającymi wiedzę z poprzednich zajęć.
- Pierwsza kartkówka odbędzie się za 3 tygodnie i dotyczyć będzie 'materiału' z zajęć z poprzedzającego tygodnia (tj. z zajęć, które będą za 2 tygodnie).
- Ocena za kartkówki i tutoriale będzie liczona jako średnia z ocen jednostkowych z wyłączeniem dwóch najgorszych ocen (wśród których mogą być zera wynikające z nieobecności).
- Zadanie domowe:
  - zdefiniowane zostanie w trakcie semestru (prawdopodobnie pod koniec kwietnia) wraz z podaniem terminu oddania
  - każdy dzień zwłoki skutkować będzie odjęciem 5% od oceny z zadania domowego (to samo dotyczy tutoriali)

Wstępny harmonogram

- 28.02 (wtorek), 2.03 (czwartek) – laboratorium 1: Grupowanie i histogramy
- 7.03 (wtorek), 9.03 (czwartek) – praca własna / konsultacje: Wprowadzenie do R
- 14.03 (wtorek), 16.03 (czwartek) – laboratorium 2: Statystyki opisowe
- 21.03 (wtorek), 23.03 (czwartek) – laboratorium 3: Rozkłady prawdopodobieństwa
- 28.03 (wtorek), 30.03 (czwartek) – laboratorium 4: Estymacja punktowa i przedziałowa
- 4.04 (wtorek), 6.04 (czwartek) – laboratorium 5: Testy frakcji
- 18.04 (wtorek), 13.04 (czwartek) – laboratorium 6: Testy t i Z
- 25.04 (wtorek), 20.04 (czwartek) – laboratorium 7: Testy dwóch populacji
- 9.05 (wtorek), 27.04 (czwartek) – laboratorium 8: Korelacja i regresja cz. 1
- 16.05 (wtorek), 4.05 (czwartek) – laboratorium 9: Korelacja i regresja cz. 2 i ogłoszenie zadania domowego
- 23.05 (wtorek), 11.05 (czwartek) – laboratorium 10: Test chi-kwadrat
- 30.05 (wtorek), 18.05 (czwartek) – laboratorium 11: Testy nieparametryczne / konsultacje zadania domowego
- 6.06 (wtorek), 5.05 (czwartek) – laboratorium 12: Konsultacje
- 13.06 (wtorek), 1.06 (czwartek) – praca własna / konsultacje
- 15.06 (czwartek) – praca własna / konsultacje

[https://andrzej.szwabe.pracownik.put.poznan.pl/laboratorium-statystyka-i-analiza-danych/]

Podstawowe narzędzia

- Arkusz kalkulacyjny: MS Excell, LibreOffice
- Google Colab z interpreterem R [colab.to/r] lub (nie rekomendowane!) Juputer Notebook z wtyczką R
  - Linki do instrukcji instalacji na stronie http://andrzej.szwabe.pracownik.put.poznan.pl/laboratorium-statystka-i-analiza-danych/
  - Niezalecana opcja dla trochę bardziej 'zaawansowanych technicznie': możliwość instalacji na serwerze Pionier/PCSS (instrukcja na stronie http://andrzej.szwabe.pracownik.put.poznan.pl/laboratorium-statystka-i-analiza-danych/r-i-jupyter-notebook-na-serwerze-ubuntu/)
- Platforma DataCamp

Rejestracja

- Wypełnienie kwestionariusza Google (adres na stronie http://andrzej.szwabe.pracownik.put.poznan.pl/laboratorium-statystyka-i-analiza-danych/)
- Zalogowanie się (przynajmniej jednokrotne) do eKursu PP
  - kurs 'Statystyka i analiza danych (laboratorium)' - wykorzystywany m.in. do przeprowadzania kartkówek-wejściówek
- Platforma DataCamp - darmowy dostęp dla studentów

Wykorzystanie DataCamp w ramach lab. SiAD (1/2)

- Po zapisaniu się na kurs Statistics-and-data-analysis z użyciem adresu @student.put.poznan.pl student uzyskuje:
  - darmowy dostęp (klasy "Premium") działający do września 2023 r. (6 miesięcy),
  - pełen dostęp do kursów dotyczących zaawansowanego przetwarzania i analizy danych z użyciem języka R, Python oraz SQL.
- Email z zaproszeniem zostanie rozesłany w tym tygodniu.

Wykorzystanie DataCamp w ramach lab. SiAD (2/2)

- Zadania domowe z wykorzystaniem DataCamp:
  - Pierwszy tutorial jest najbliższym zadaniem domowym: Tutorial R na platformie DataCamp: należy wykonać pierwsze trzy rozdziały (Intro to basics, Vectors, Matrices) z “Introduction to R”. Termin: do końca dnia (23:59) poprzedzającego zajęcia w tygodniu rozpoczynającym się od 6.03
  - Drugi tutorial jest drugim zadaniem domowym (zadanym w drugim tygodniu zajęć): Tutorial R na platformie DataCamp: należy wykonać pierwsze trzy rozdziały (Conditionals and Control Flow, Loops, Functions) z “Intermediate R”. Termin: do końca dnia (23:59) poprzedzającego zajęcia w tygodniu rozpoczynającym się od 13.03.

Mini-wstęp: podstawowe pojęcia

- data science
- badanie statystyczne
- obserwacja, eksperyment
- populacja, próba
- dobór próby (na przykładach metod podziału danych historycznych na tzw. zbiór uczący, walidacyjny i testowy stosowanych w eksperymentalnej ocenie algorytmów systemów uczących się: uwzględniających chronologię i nie uwzględniających)
- skale pomiarowe
  - numerical - ilościowe (dyskretne/ciągłe)
  - categorical - jakościowe (nominalne/porządkowe)

Grupowanie i histogramy (1/2)

- Histogram (przedziałowy): wykres słupkowy liczności w poszczególnych kolejnych przedziałach
  - Czym jest histogram punktowy? ;-)
- Popularne heurystyki wyboru liczby przedziałów k (gdzie n to liczba próbek):
  - $k = \sqrt{n}$
  - $k = 1 + 3,322 \log n$
  - $k < 5 \log n$
  - $h = 2,64 \times IQR \times n^{-1/3}$
    IQR - rozstęp międzykwartylowy = zakres 50% "środkowych" wartości w próbce
  - $h \approx \frac{x_{max} - x_{min}}{k}$,
    gdzie
    - h -szerokość przedziału
    - x_{min}, x_{max} -wartości najmniejsza i największa

Grupowanie i histogramy (2/2)

Krokami na drodze do utworzenia szeregu rozdzielczego są:

1. ustalenie liczby przedziałów (warto znać popularne heurystyki),
2. ustalenie szerokości przedziału,
3. zdefiniowanie początku pierwszego przedziału,
4. zliczenie obserwacji w utworzonych przedziałach.

Dzisiejsze zadania

- Zapoznanie się z ćwiczeniami w arkuszu kalkulacyjnym ze wsparciem w postaci prezentacji ’alternatywnej implementacji’ w R:
  - ćwiczenie 1
  - ćwiczenie 2
  - ćwiczenie 3
  - ćwiczenie 4
- Próba budowania histogramów w Google Colab z interpreterem R (w ramach wstępnego, ’zwinnego’ wprowadzenia do R")

Dziękuję za uwagę...

...i proszę o pytania.