Statystyka i Analiza Danych

Laboratorium 1

Wprowadzenie do laboratorium, Mini-wprowadzenie do R, Grupowanie i histogramy

Konrad Miazga

na podstawie slajdów Andrzeja Szwabe

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

03 marca 2023

Podstawowe informacje (1/3)

- Materiały laboratorium znajdują się na:
  - eKursach – `https://ekursy.put.poznan.pl/course/view.php?id=3679`
  - stronie prowadzącego – `https://www.cs.put.poznan.pl/kmiazga`

Podstawowe informacje (2/3)

- Prowadzący laboratorium to mgr inż. Konrad Miazga
  - Instytut Informatyki
  - Zakład Inteligentnych Systemów Wspomagania Decyzji
  - Dyżury:
    - pokój 2.7.13/6 w budynku Biblioteki Technicznej,
    - we piątki pomiędzy 15:10 a 16:40,
    - aby uzyskać dostep do pokoju, należy zadzwonić pod numer 61 665 3058,
    - proszę o wcześniejsze potwierdzenie obecności poprzez wiadomość email,
    - możliwe są też konsultacje online.
  - konrad.miazga@put.poznan.pl
    - pisząc emaile proszę zaczynać tytuł wiadomości od tagu [SiAD] i identyfikatora grupy laboratoryjnej (np. *[SiAD][L7]*)

Podstawowe informacje (3/3)

- 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 (eKursy/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:
  - 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 eKursach 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

- **(03.03) Laboratorium 1: Grupowanie i histogramy**
- (10.03) Praca własna / Konsultacje: Wprowadzenie do R
- **(17.03) Laboratorium 2: Statystyki opisowe**
- **(24.03) Laboratorium 3: Rozkłady prawdopodobieństwa**
- **(31.03) Laboratorium 4: Estymacja punktowa i przedziałowa**
- (07.04) –
- **(14.04) Laboratorium 5: Testy frakcji**
- **(21.04) Laboratorium 6: Testy t i Z**
- **(28.04) Laboratorium 7: Testy dwóch populacji**
- **(05.05) Laboratorium 8: Korelacja i regresja cz. 1**
- **(12.05) Laboratorium 9: Korelacja i regresja cz. 2 + ogłoszenie zadania domowego**
- **(19.05) Laboratorium 10: Test chi-kwadrat**
- **(26.05) Laboratorium 11: Testy nieparametryczne + konsultacje zadania domowego**
- **(02.06) Laboratorium 12: Konsultacje**
- (09.06) –
- (16.06) Praca własna / Konsultacje

Podstawowe narzędzia

- Arkusz kalkulacyjny: MS Excel / LibreOffice / Google Sheets
- Google Colab z interpreterem R (dostępny pod adresem `colab.to/r`) lub (nie rekomendowane!) Jupyter 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

- 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 przeciągu tygodnia.

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 09.03.23 (23:59).
  - 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 16.03.23 (23:59).

Mini-wstęp: podstawowe pojęcia

- data science
- badanie statystyczne
- obserwacja, eksperyment
- populacja, próba
- dobór próby
- skale pomiarowe
  - categorical - jakościowe (nominalne/porządkowe)
    - nominalne
    - porządkowe
  - numerical - ilościowe (dyskretne/ciągłe)
    - dyskretne
    - ciągłe

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.