Statystyka i analiza danych - laboratorium 1
Wprowadzenie, grupowanie i histogramy

Paweł Misiorek

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

28 lutego/3 marca 2023

Podstawowe informacje (1/2)

- Strona przedmiotu wspomagająca eKurs:
  http://pawel.misiorek.pracownik.put.poznan.pl/siad.html
- prowadzący laboratorium: Paweł Misiorek
  - Instytut Informatyki
  - Zakład Technologii Przetwarzania Danych
  - dyżury: wt. 13:00-14:15, pt. 9:35-10:20
  - pokój 318 (BM)
  - pawel.misiorek@put.poznan.pl
    (pisząc maile zaczynać tytuł wiadomości od [siad])

Podstawowe informacje (2/2)

- Statystyka i analiza danych
- Wykłady prowadzi prof. Jerzy Stefanowski (środa 11:45)
  - materiały wykładowe są dostępne na platformie eKursy
- 13 grup laboratoryjnych - wspólne zasady

Plan na dziś

- Krótkie zapoznanie z przedmiotem
- Zasady zaliczenia
- Harmonogram
- Rejestracja (kwestionariusz/Moodle/Slack/DataCamp)
- Pierwsze ćwiczenia z grupowania i histogramów
- Wskazanie materiałów do samodzielnego wprowadzenia do R

Zasady zaliczenia (1/3)

- na ocenę z laboratorium składają się:
  - 70% - kartkówki (9 lub 10 kartkówek obejmujących materiał z poprzednich zajęć) oraz 2 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)
- kartkówki i tutoriale: planuje się 9 lub 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 (obejmują dwa pierwsze zajęcia)

Zasady zaliczenia (3/3)

- kartkówki: krótkie, 1-3 zadania sprawdzające wiedzę z ostatnich zajęć (pierwsza kartkówka odbędzie się za 2 tygodnie - czyli 14. i 17. marca obejmując zajęcia poprzednie z 7. i 10. marca)
- 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 na stronie przedmiotu wraz z podaniem terminu oddania
  - każdy tydzień zwłoki skutkować będzie odjęciem 10% od oceny z zadania domowego (w przypadku tutoriali każdy dzień spóźnienia również oznacza odjęcie 10%)

Wstępny harmonogram

- Wstępny harmonogram naszkicowano na stronie przedmiotu
- Uwaga: przedmiot zakłada dokładnie 12 spotkań laboratoryjnych (w maju z powodu wyjazdu prowadzącego przez 2 tygodnie zajęcia nie będą się odbywać).

Podstawowe narzędzia

- Arkusz kalkulacyjny: MS Excell, LibreOffice
- Juputer Notebook z wtyczką R, Google Colab
- Platforma DataCamp

Rejestracja

- Kwestionariusz Google (link na stronie przedmiotu)
- eKurs
  - kurs Statystyka i analiza danych
    - grupa L10 [PM] wtorek 8:0 ma kod dostępu: sad-L10
    - grupa L4 [PM] wtorek 9:45 ma kod dostępu: sad-L4
    - grupa L1 [PM] piątek 11:45 ma kod dostępu: sad-L1
  - wspólny kurs dla wykładu i laboratorium, zawiera też dedykowana sekcja dla grup prowadzonych przeze każdego prowadzącego.
- Platforma DataCamp - darmowy dostęp dla studentów

DataCamp

- darmowy dostęp (poziom Premium) dla studentów po zapisaniu się na kurs *statistics-and-data-analysis-23*
- pełen dostęp do kursów dotyczących zaawansowanego przetwarzania i analizy danych z użyciem języka R, Python oraz SQL
- dostęp do 27.08.2023 (6 miesięcy)
- mail z zaproszeniem został/zostanie rozesłany dziś (po dokonaniu przez studenta rejestracji z użyciem formularza (link na stronie przedmiotu))
- dwa tutoriale jako dzisiejsze zadanie domowe (pierwszy - mniejszy - na za tydzień oraz drugi - większy - na za dwa tygodnie)

Podstawowe pojęcia

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

Grupowanie i histogramy (1/2)

- histogram - wykres słupkowy liczności w poszczególnych kolejnych przedziałach
- wybór liczby przedziałów k (n - 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)

Kolejne kroki działania:

- ustalenie liczby przedziałów
- ustalenie szerokości przedziału
- zdefiniowanie początku pierwszego przedziału
- zliczenie obserwacji w utworzonych przedziałach

Zadania

Dziś: Rozpoczęcie ćwiczeń w arkuszu kalkulacyjnym:

- ćwiczenie 1
- ćwiczenie 2
- ćwiczenie 3
- ćwiczenie 4

oraz wskazanie materiałów wprowadzających do języka R (w tym budowanie histogramów w R)

Dziękuję za uwagę

Proszę o pytania