Универзитет у Београду
Математички факултет

СТАТИСТИЧКИ СОФТВЕР 1 (3В)

26. август 2025. године

---

1 Основни појмови и функције за рад са подацима

1. У бази `mtcars` налазе се техничке карактеристике аутомобила, при чему су модели аутомобила дати у редовима табеле. Свака променљива представља неку перформансу или техничку спецификацију аутомобила.

   (а) [**1 поена**] Формирати матрицу основних дескриптивних статистика где су колоне матрице карактеристике аутомобила `mpg`, `hp`, `wt` и `qsec`, а редови матрице су средња вредност (`MEAN`), стандардна девијација (`SD`), минимум (`MIN`) и максимум (`MAX`).

   (б) [**1 поен**] Бази `mtcars` додати колону `Index` која представља индекс ефикасности аутомобила (заокружен на две децимале), где се индекс ефикасности дефинише као

   $$\texttt{Index} = \texttt{mpg}^2/\texttt{wt} + 0.5 \cdot \texttt{qsec} - \log(\texttt{hp}).$$

   (в) [**2 поена**] Бази `mtcars` додати колону `Classter` која класификује возила према нивоу ефикасности и масе, на следећи начин:

   - 1: `mpg` и `qsec` су већи од просека, а `Index` мањи од просека,
   - 2: `mpg` је већи, а `qsec` и `Index` су мањи од просека,
   - 3: `mpg`, `qsec` и `Index` су мање од просека,
   - 4: остали аутомобили.

   (г) [**2 поен**] Формирати листу где i-ти елемент листе, за $i = 1, 2, 3, 4$, садржи имена брендова аутомобила који припадају i-том кластеру и просечан индекс ефикасности модела из датог кластера.

2. У бази `airquality` налазе се подаци о квалитету ваздуха у Њујорку током летњих месеци.

   (а) [**1 поен**] Израчунати просек и дисперзију оригиналних и стандардизованих вредности за `Ozone`, `Solar.R`, `Wind` и `Temp`. Вредности заокружити на 5 децимала.

   (б) [**1 поен**] Имплементирати функцију `Iqr_Func()` која за дати нумерички вектор одређује његово интерквартилно растојање.

   (в) [**2 поен**] Имплементирати функцију `HotDays()` која из базе `airquality` брише елементе код којих је температура мања од $6.5 \cdot$ `IQR`, где је `IQR` интерквартилно растојање, а брзина ветра је већа од просека. Функција као повратну вредност исписује нову базу и број обрисаних елемената.

2 Симулације, вероватноћа и статистика

3. Агенција за безбедност саобраћаја истражује просечне вредности брзине вожње током дана и ноћи. Претпоставља се да брзина вожње током дана има нормалну $\mathcal{N}(m_1, \sigma_1^2)$ расподелу, а брзина вожње током ноћи има нормалну $\mathcal{N}(m_2, \sigma_2^2)$ расподелу, где су m_1, m_2, σ_1 и σ_2 непознати. Измерене су брзине у km/h за по 12 случајно изабраних возача у дневним и ноћним условима и добијени су следећи резултати: 58.3, 62.1, 54.7, 59.5, 61.0, 57.8, 63.2, 60.5, 56.9, 59.1, 64.0, 60.2 за дневне вожње, те 65.1, 68.4, 70.2, 66.8, 67.5, 69.3, 71.0, 66.0, 68.1, 67.9, 69.7, 68.3 за ноћне вожње.

   (а) [**2 поен**] Са прагом значајности 0.05 тестирати хипотезу да су средње вредности брзине током дана и брзине током ноћи једнаке против алтернативе да се разликују.

(б) [**2 поен**] Одредити 95% интервал поверења за непознат параметар m_1.

(в) [**2 поен**] Одредити 90% интервал поверења за непознат параметар σ_1^2, као и за непознат параметар σ_2.

4. Електропривреда Србије жели да симулира потрошњу електричне енергије 100 домаћинстава у месецу децембру (претпоставити да је 1. децембар понедељак). Потрошња зависи од температуре, дана у недељи и типичног понашања домаћинства. Претпоставља се да просечан дневна потрошња (у kWh) радним данима има нормалну $\mathcal{N}(28, 25)$ расподелу, а викендом има нормалну $\mathcal{N}(34, 36)$ расподелу. Ако је температура испод $0°C$, потрошња се повећа за 20%. Температура сваког дана је случајна и има нормалну $\mathcal{N}(-2, 16)$ расподелу.

   (а) [**3 поен**] Имплементирати функцију која симулира матрицу потрошње електричне енергије посматраних домаћинстава у месецу децембру.

   (б) [**2 поен**] Графички приказати трајекторију просечне дневне потрошње 100 домаћинстава.

5. [**4 поен**] Александар и Андреј на случајан начин, независно један од другог, бирају по један број из скупа $\{1, 2, \ldots, 25\}$. Нека случајна величина X представља број који је Александар изабрао, а случајна величина Z збир бројева које су изабрали Александар и Андреј. Проценити $E[X|Z = z]$ за све вредности z на основу симулираних 100.000 парова бројева X и Y.

3 Обрада и визуализација података

6. База `satfruit` (пакет `PASWR`) садржи податке о површинама пољопривредних култура у 47 сегмената на три подручја у Шпанији (`R63`, `R67`, `68`). За сваки сегмент дати су подаци о површинама за више врста усева (нпр. пшеница, јечам, кукуруз, маслине, воћке итд).

   (а) [**1 поен**] Користећи функцију `summarise()` израчунати просечану и медијалну површину за усеве луцерке (`AF`), бадема (`AL`) и маслина (`OL`), без обзира на подручје (`SArea`).

   (б) [**2 поен**] Креирати нову променљиву `Fruit.Share` као однос површине засађене воћкама (`FR`) и укупне површине обрадивог земљишта у сегменту (сумарно све пољопривредне културе). Затим израчунати просечну вредност `Fruit.Share` по подручју (`SArea`) и рангирати подручја од најмањег до највећег удела воћака.

   (в) [**1 поен**] Користећи функцију `group_by()` одредити за свако подручје просечну површину под кукурузом (`COR`), медијану површину под сунцокретом (`SF`) и максималну површину под воћкама (`FR`).

7. У бази `flights` (пакет `nycflights13`) налазе се подаци о свим летовима са три њујоршка аеродрома током 2013. године, а у бази `airports` (пакет `nycflights13`) подаци о аеродромима.

   (а) [**3 поен**] Из базе `flights` изабрати у базу `flights_jan` летове који су полетели (`dep_delay` $\neq$ `NA`) и стигли (`arr_delay` $\neq$ `NA`) током јануара. Креирати нову променљиву `air_time_diff` која представља разлику између времена лета (`air_time`) и времена планираног лета (`dep_delay`).

   (б) [**2 поен**] Користећи функцију `left_join()` спојити табелу `flights_jan` и базу `airports`, тако да се уместо кода одредишта (`dest`) прикаже пуно име аеродрома (`name` из `airports`). Назвати нову табелу `joined_data`.

   (в) [**1 поен**] Израчунати просечну вредност `arr_delay` и `air_time_diff` по одредишном аеродрому. Сачувати резултате као табелу `summary_airports`.

   (г) [**2 поен**] За `summary_airports`, креирати нову категоричку променљиву `punctuality_level`:
   - `high`: ако је `arr_delay` мање од 0,
   - `medium`: ако је `arr_delay` између 0 и 10,
   - `low`: ако је `arr_delay` веће од 10.

   (д) [**1 поен**] Израчунати број одредишта по сваком нивоу поузданости (`punctuality_level`).

   (ђ) [**2 поен**] Графички приказати (пакет `ggplot2`) зависност између просечног кашњења при доласку (`arr_delay`) и одступања просечне дужине лета (`air_time_diff`), при чему је боја одређена променљивом `punctuality_level`.