CC4 (4B) - Задаци за семинарски

1. Одабрати базу[1] података и у форми мини истраживања илустровати рад са пакетима и функцијама наученим у току овог курса. Од саме базе и вас зависи шта ћете испитивати и на шта ће се истраживање фокусирати. Важно је да то што радите има неког смисла, тј. да можете да доносите закључке и интерпретирате резултате. Потребно је:

а) Илустровати рад са основним функцијама из пакета *dplyr*. (8)

б) Променљиве од интереса приказати различитим графицима из пакета *ggplot*2. Избор графика је препуштен вама и зависи од базе, али је пожељно направити неколико графика, средити их, приказати неке од занимљивих и специфичних графика, итд. (8)

ц) Илустровати основне функције за сређивање база података. (8)

д) Одабрати још једну базу која је у вези са том (ако не можете да пронађете одговарајућу, направите је сами!) па приказати основне функције за рад са релацијама између њих. (8)

2. Одабрати текст по избору и на њему илустровати основне функције за рад са стринговима. Требало би да подаци буду интерпретабилни и занимљиви колико је то могуће (потрудите се да формирате неке мало сложеније регуларне изразе). (8)

3. Нека су X и Y случајне величине са заједничком расподелом:

$X \setminus Y$	0	1
0	0.5	0.1
1	0.3	0.1

Израчунати условне расподеле, а затим, користећи *Gibbs*-ово узорковање, извадити узорак обима 10000 из заједничке расподеле.
(НАПОМЕНА: Узорачка расподела треба да буде приближна правој расподели датој у табели!) (12)

4. Стандардна претпоставка при моделовању генотипова са двоструким алелима је да се укрштање врши на случајан начин.
Према томе, за популацију где је p вероватноћа алела A, генотипови AA, Aa и aa имају вероватноће $p^2, 2p(1-p)$ и $(1-p)^2$.
Претпоставимо да p има $\mathcal{U}[0,1]$ априорну расподелу.
Претпоставимо да имамо узорак од n јединки: n_{AA} са генотипом AA, n_{Aa} са генотипом Aa и n_{aa} са генотипом aa.
Направити функцију $MCMCsampler(nAA, nAa, naa, iter, start\_value, prop\_sd)$ која ће коришћењем Metropolis алгоритма вратити апроксимативно узорке из апостериорне расподеле за p. Предлог креирати додавањем шума из $\mathcal{N}(0, prop\_sd)$ расподеле.
Ако је $n_{AA} = 50, n_{Aa} = 21$ и $n_{aa} = 29$, покренути алгоритам за 10000 итерација, са почетном вредношћу 0.5 и ширином расподеле предлога 0.01.
Нацртати хистограм, а на њега доцртати густину праве апостериорне расподеле.
Покренути алгоритам за другу почетну вредност, и мањи број итерација, нпр. 0.1 и 1000, редом. Нацртати график ланца (временске серије) и на основу њега проценити колико би почетних вредности требало одбацити као burn-in. (12)

5. Претпоставимо да за низ случајних величина $Y_1, Y_2, ..., Y_n$ важи:

$$Y_i \overset{iid}{\sim} \mathcal{P}(\lambda_1) \quad \text{за } i = 1, ..., m$$

$$Y_i \overset{iid}{\sim} \mathcal{P}(\lambda_2) \quad \text{за } i = m+1, ..., n$$

Априорне расподеле за непознате параметре λ_1, λ_2 и m су дате са:

$$\pi(\lambda_1) \sim \gamma(a_1, b_1)$$

$$\pi(\lambda_2) \sim \gamma(a_2, b_2)$$

$$\pi(m) \sim \frac{1}{n}$$

[1]Сваки студент мора имати различиту базу. Када одаберете базу, пријавите је мејлом, и почните са радом овог задатка тек када добијете потврду.

Извести условне апостериорне расподеле $\pi(\lambda_1|\mathbf{Y}, m)$, $\pi(\lambda_2|\mathbf{Y}, m)$ и $\pi(m|\mathbf{Y}, \lambda_1, \lambda_2)$, а затим, кроз 10000 итерација узорковати:

$$\theta_1^{(k)} \sim \pi(\theta_1|\mathbf{Y}, m^{(k-1)})$$
$$\theta_2^{(k)} \sim \pi(\theta_2|\mathbf{Y}, m^{(k-1)})$$

па

$$m^{(k)} \sim \pi(m|\mathbf{Y}, \lambda_1^{(k)}, \lambda_2^{(k)})$$

Дати су подаци $\mathbf{Y} = (4, 5, 4, 1, 0, 4, 3, 4, 0, 6, 3, 3, 4, 0, 2, 6, 3, 3, 5, 4, 5, 3, 2, 4, 4, 1, 5, 5, 3, 4, 2, 5, 2, 2, 3, 4, 2, 1, 3, 2, 2, 1, 1, 1, 1, 3, 0, 0, 1, 0, 1, 1, 0, 0, 3, 1, 0, 3, 2, 2, 0, 1, 1, 1, 0, 1, 0, 1, 0, 0, 0, 2, 1, 0, 0, 0, 1, 1, 0, 2, 3, 3, 1, 1, 2, 1, 1, 1, 1, 2, 4, 2, 0, 0, 0, 1, 4, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1)$, $a_1 = 3$, $a_2 = 1$, $b_1 = 0.5$ и $b_2 = 0.5$.

НАПОМЕНА: Као условне апостериорне расподеле за λ_1 и λ_2 добијају се познате расподеле, док се за m добија израз који не указује ни на једну познату расподелу. Међутим, лако можемо добити апостериорну условну расподелу за m, $\pi(m|\mathbf{Y}, \lambda_1^{(k)}, \lambda_2^{(k)})$, тако што прођемо добијеним изразом кроз све вредности које може да узме случајна величина m, сумирамо, и нормализујемо (поделимо са добијеном сумом, да добијемо праву расподелу, тј. да у збиру буде 1), а онда узоркујемо $m^{(k)} \sim \pi(m|\mathbf{Y}, \lambda_1^{(k)}, \lambda_2^{(k)})$, као најмањи број из скупа допустивих вредности за m, за који функција расподеле (израчунате у претходном кораку) прелази случајно изабрани број из (0,1) (помоћ: за добијање функције расподеле може се користити функција *cumsum*).

Нацртати хистограме појединачних компоненти и оценити вредности одговарајућих параметара. (12)

БОНУС: Из стринга ("06435.213", "aswww", "2112*121", "011", "232424321", "1232", "23423aaa21321", "0.1", "3424", "123*131", "232 232", "2.3", "4543.45") издвојити исправно написане бројеве (прослеђивањем одговарајућег регуларног израза функцији *str_view*()). (4)