Илустрација истраживања о неверству у браку

Павле Вилотијевић, 78/2021

Напомена: за сва тестирања наведене су одговарајуће p-вредности. Приликом доношења закључака узиман је ниво значајности 0.05.

1. Опис базе података

Подаци су преузети из анкете спроведене 1969. од стране Psychology today. База садржи 601 обзервацију са 9 променљивих (2 категоричке, 7 нумеричких):

affairs

нумеричка променљива. Колико често су испитаници били у ванбрачним сексуалним односима током прошле године. 0 = ни једном, 1 = једном, 2 = два пута, 3 = 3 пута, 7 = 4–10 пута, 12 = једном месечно или чешће

gender

фактор који указује на пол.

age

нумеричка променљива којом је изражена старост у годинама: 17,5 = испод 20, 22 = 20–24, 27 = 25–29, 32 = 30–34, 37 = 35–39, 42 = 40–44, 47 = 45–49, 52 = 50–54, 57 = 55 или више.

yearsmarried

нумеричка променљива којом је изражен број година у браку: 0,125 = 3 месеца или мање, 0,417 = 4–6 месеци, 0,75 = 6 месеци–1 година, 1,5 = 1–2 године, 4 = 3–5 година, 7 = 6–8 година , 10 = 9–11 година, 15 = 12 или више година.

children

фактор који указује на то да ли има деце у браку

religiousness

нумеричка променљива којом је представљена религиозност: 1 = против, 2 = нимало, 3 = мало, 4 = донекле, 5 = веома.

education

Нумеричка променљива којом је изражен ниво образовања: 9 = основна школа, 12 = дипломирани средња школа, 14 = неки факултет, 16 = дипломирани факултет, 17 = неки дипломски рад, 18 = магистарска диплома, 20 = докторат, доктор наука или друго напредни степен.

occupation

нумеричка променљива којом је изражено занимање према Холингсхед класификацији (слика десно).

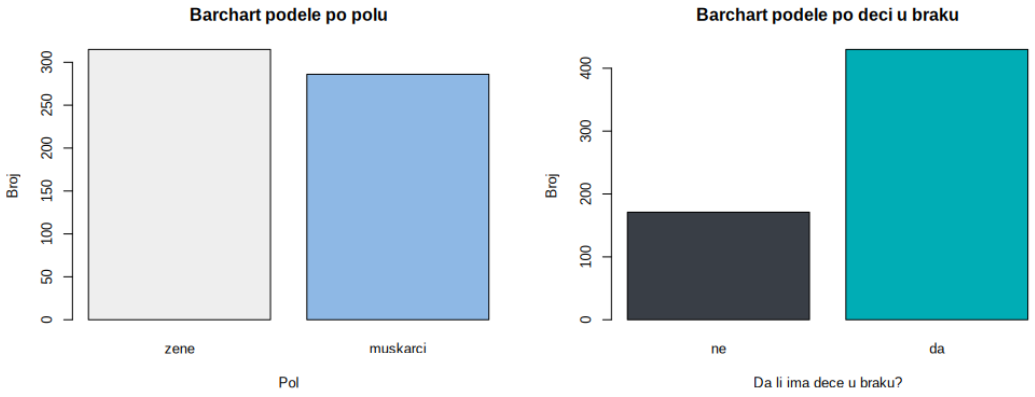
Code	Occupation
1	Higher executive, major professional, etc.
2	Business manager, etc.
3	Administrative personnel, etc.
4	Clerical and sales, technician, etc.
5	Skilled manual
6	Machine operators, semi-skilled
7	Unskilled
8	Never employed

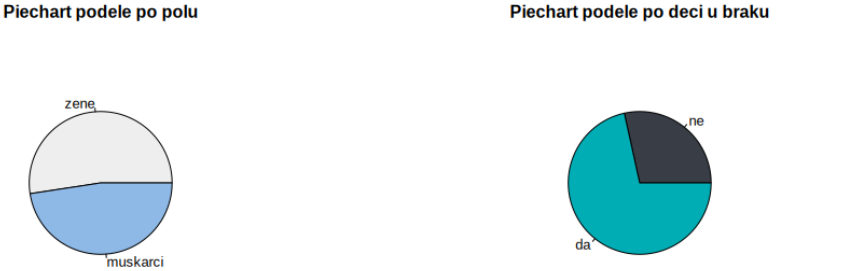
rating

нумеричка променљива којом је изражена сопствена процена брака: 1 = веома несрећан, 2 = донекле несрећан, 3 = просечан, 4 = срећнији од просека, 5 = веома срећан.

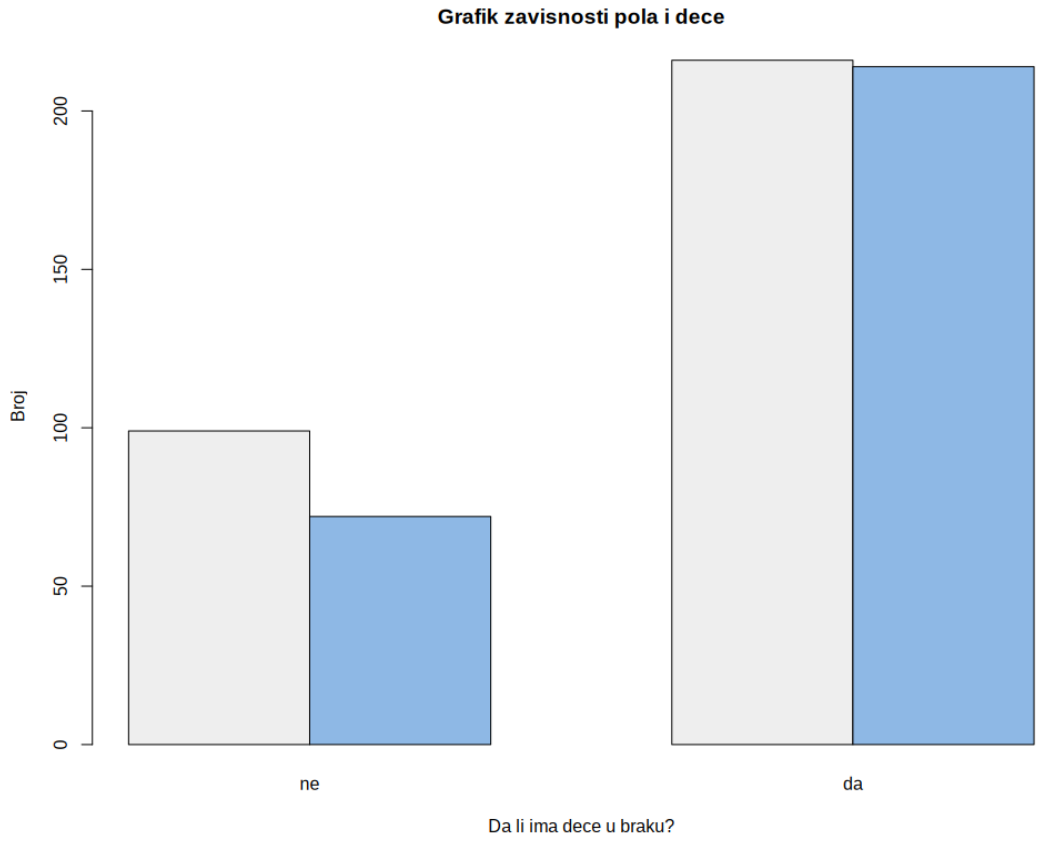
2. Категоричке променљиве

База података садржи 2 категоричке променљиве, gender и children. Ово су њихови графички прикази:





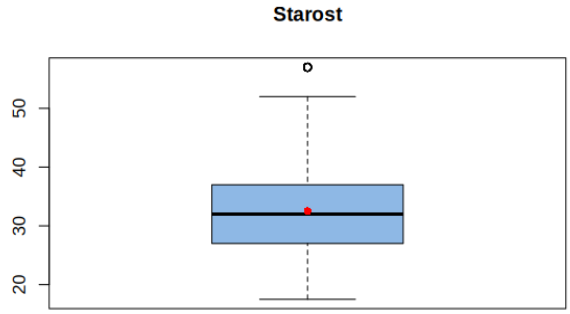
Као што са графика можемо видети, нешто је више жена него мушкараца (око 52.4%), док је удео особа које имају децу значајно већи него који немају (око 71.5%). Испитајмо сада зависност ове две променљиве:

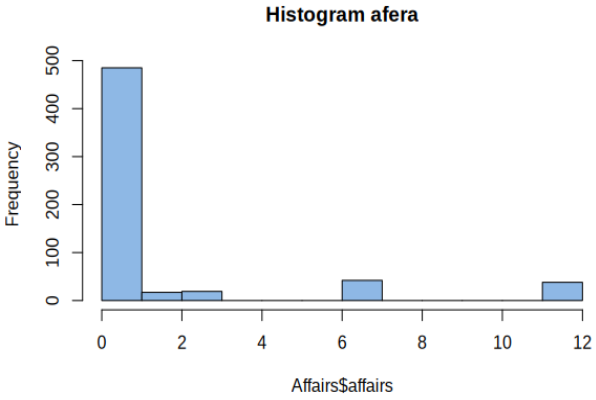


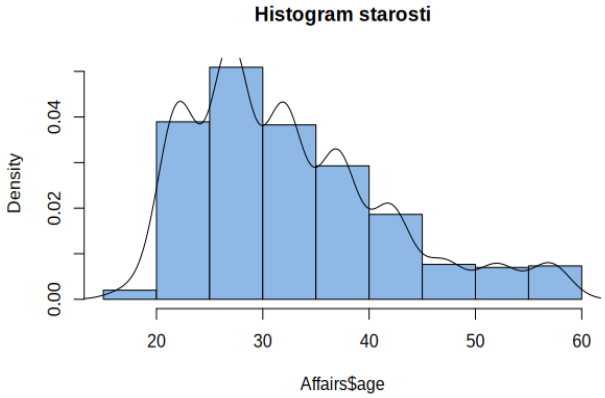
Осматрањем графика, видимо да су међу људима који имају децу подједнако заступљени мушкарци и жене, док међу људима који немају децу има више жена. Дакле, природно је поставити питање - да ли су ове две променљиве незасвисне? Провером помоћу хи квадрат теста (H_0 – променљиве су независне, H_1- променљиве нису зависне), добија се p-вредност 0.1082, што је релативно велико (може се аргументовати да је вредност близу 0.1, што се понекад узима као ниво значајности, па би у зависности од контекста истраживања могли да се донесу другачији закључци. Ипак, наставићемо са 0.05, што је стандардно за друштвене науке, у ком случају је добијена вредност очигледно велика). Ово нас упућује да не одбацимо нулту хипотезу, па је закључак да су ове две расподеле независне.

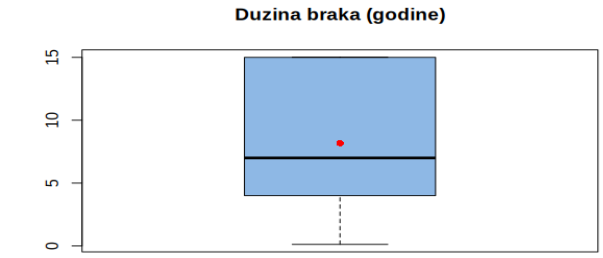
3. Нумеричке променљиве

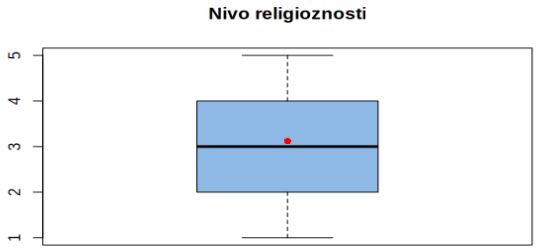
База садржи 7 нумеричких променљивих: affairs, age, yearsmarried, religiousness, education, occupation и rating. Погледајмо сада њихове графичке приказе и основне карактеристике.

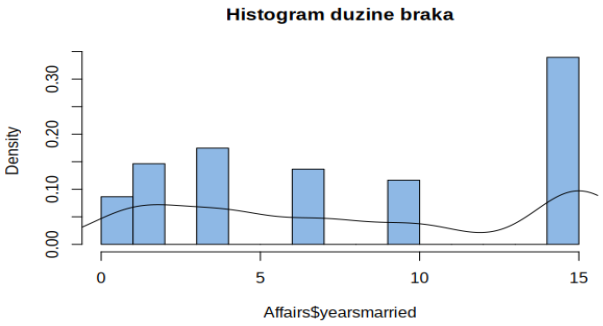


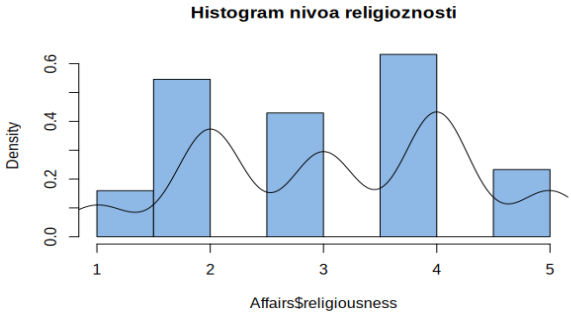


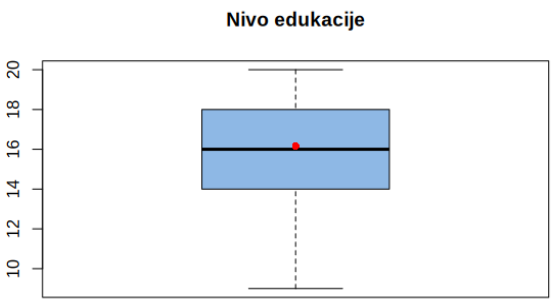


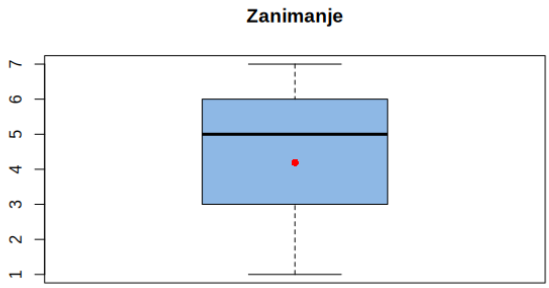


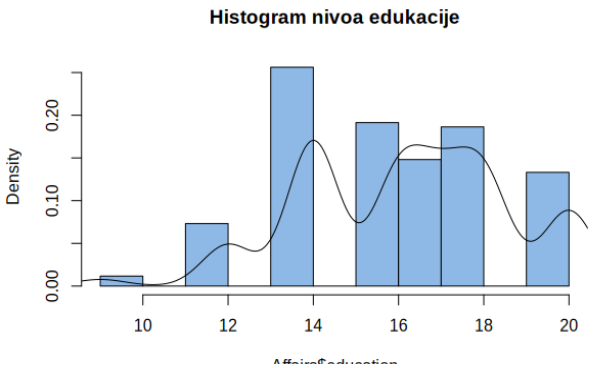


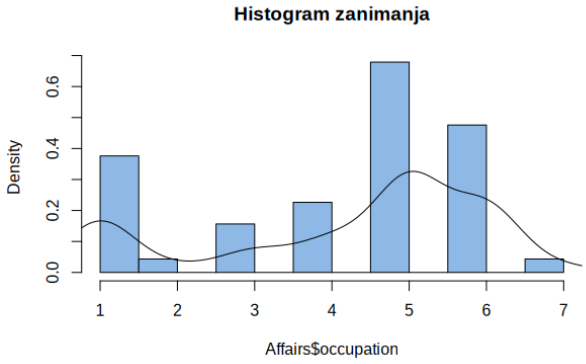


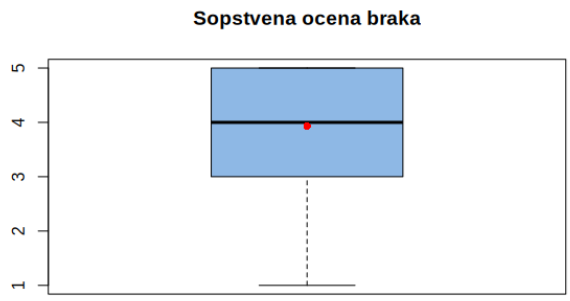




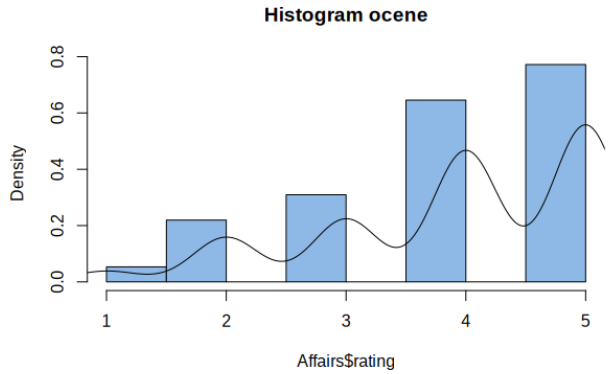








1.455907	3.298758
32.48752	9.288762
8.177696	5.571303
3.116473	1.167509
16.16639	2.402555
4.194676	1.819443
3.93178	1.103179



Табела са средњим вредностима и стандардним одступањима нумеричких променљивих

На основу хистограма, ни једна од расподела не подсећа претерано на неку познату. Такође, ни једна од променљивих није апсолутно непрекидна, па КС тест припадности није адекватан. Ипак, демонстрације ради, тестирајмо припадност нивоа религиозности нормалној расподели. Добијамо p-вредност <2.2e-16, па би закључак био да променљива нема нормалну расподелу.

Постоје 2 поделе података на основу фактора: мушкарци и жене, особе са децом и особе без деце. За обе поделе, урадићемо тест једнакости средње вредности нивоа религиозности:

Када урадимо тест са поделом на мушкарце и жене, добијамо да су им средње вредности религиозности редом 3.125874 и 3.107937. p-вредност теста је 0.8514, па како је она велика, не одбацујемо нулту хипотезу. Закључак је да су средње вредности нивоа религиозности међу мушкарцима и женама једнаке.

Када урадимо тест са поделом на људе са децом и без деце, добијамо да су им средње вредности религиозности редом 3.211628 и 2.877193. p-вредност теста је 0.001513, па како је она мала, одбацујемо нулту хипотезу у корист алтернативне - постоји разлика у нивоу религиозности људи са децом и без деце. Ако поновимо тест, али изменимо алтернативну хипотезу да буде: ниво религиозности међу људима без деце мањи је од нивоа религиозности људи са децом, добијамо p-вредност 0.0007567. Ово нас упућује на одбацивање нулте хипотезе, па закључујемо да је ниво религиозности код људи са децом већи од нивоа религиозности људи без деце.

Поновимо овај поступак за број афера током прошле године (битно је напоменути да је ова расопдела очигледно није ни близу нормалне, док је ниво религиозности макар делимично подсећао на њу- хистограм је био симетричан, висине на крајевима су биле мале, а при средини веће):

При подели на мушкарце и жене, добијамо p-вредност 0.774, па закључујемо да је средња вредост броја афера током прошле године у овом случају једнака.

При подели на људе са децом и без деце, p-вредност је 0.004726, па закључујемо да се у овом случају средња вредност броја афера у току прошле године разликује. Оно што је можда изненађујуће можемо видети када посматрамо средње вредности: за људе без деце, средња вредност је 0.9122807, док је за људе са децом она 1.6720930. Ово нас упућује на нову алтернативну хипотезу: средња вредност броја афера људи са децом већа је од средње вредности броја афера људи без деце (нулта је и даље да су једнаке). Тестирањем добијамо p-вредност 0.002363. Дакле, одбацујемо нулту хипотезу у корист алтернативне - средња вредност броја афера међу људима са децом већа је него међу људима без деце.

4. Тест знакова

Тест знакова одабран је као тест који није рађен на часовима. Разлог зашто је он одабран је то што је непараметарски - не постоји никаква претпоставка о расподели обележја, па како у овој бази података ни једна нумеричка променљива не личи претерано на неку познату расподелу, овај тест је користан јер то није потребно да би се применио.

Опис:

Тестира се нулта хипотеза $H_0 : m_e = m_{e0}$, где је m_e медијана расподеле коју има посматрано обележје. Алтернаривне могу бити $m_e \neq m_{e0}$, $m_e < m_{e0}$, $m_e > m_{e0}$. Тест статистика је

$T_n = \sum_{i=1}^{n} I\{X_i > m_{e0}\}$, а може се користити и $T_n' = \sum_{i=1}^{n} I\{X_i < m_{e0}\}$. При нултој хипотези, $T_н$ има биномну расоделу са параметрима n, 1/2.

У R-у не постоји директно имплементиран тест знакова. Уместо тога, користи се функција binom.test, са позивом облика: binom.test(sum(X > m_{e0}), length(X), p=0.5). Из тог разлога, биће описана функција binom.test. Позив:

binom.test(x, n, p = 0.5, alternative = c("two.sided", "less", "greater"), conf.level = 0.95)

Аргументи:

x - број успеха или вектор дужине 2 у коме је редом записан број успеха и неуспеха

n - број покушаја. Игнорисано ако је x дужине 2

p - вероватноћа успеха из алтернативне хипотезе

alternative - указује на облик алтернативне хипотезе, мора да буде једна од наведених вредности. Може се ставити и само прво слово

conf.level - ниво поверења за враћени интервал поверења (то јест 1 - ниво значајности)

Излаз: листа са класом htest која садржи следеће компоненте:

statistic - број успеха

parameter - број покушаја

p.value - p-вредност теста

conf.int - интервал поверења за вероватноћу успеха

estimate - процењена вероватноћа успеха

null.value - вероватноћа успеха при нултој хипотези

alternative - стринг који описује алтернативну хипотезу

method - стринг "Exact binomial test"

data.name - стринг који садржи имена података

Тестирајмо нулту хипотезу да је медијана нивоа религиозности једнака 3 против алтернативне да је различита. Коришћењем binom.test функције у R-у као што је изнад описано, добијамо p-

вредност 0.001084, што нас упућује на одбацивање нулте хипотезе у корист алтернативне, дакле медијана је различита од 3. "Пешке" имплементацијом тог теста добија се p-вредност 0.001084398, што је једнако вредности коју смо добили преко функције из R-а, па је тест добро имплементиран.

Напомена: овај закључак није тачан, медијана заправо јесте 3, али проблем је у самој природи теста. Разлог за то је мали број различитих вредности које променљива може да узме (1, 2, 3, 4, 5). Цела вредност 3 биће искључена приликом израчунавања тест статистике, па ће то утицати на крајњи резултат, а уколико бисмо ставили ≥ онда ће цела бити укључена, што ће опет имати утицај на крајњи резултат. У ретроспективи, ово вероватно није била идеална тест статистика с обзиром на податке.

5. Закључак

Видели смо како изгледају категоричке и нумеричке променљиве на различитим врстама дијаграма. Установили смо да су две категоричке променљиве, пол и постојање деце у браку, међусобно независне. Поделили смо узорак према њима, и посматрали разлике средњих вредности две нумеричке променљиве - нивоа религиозности и броја афера током прошле године. При подели узорка на мушкарце и жене, нисмо приметили разлике у средњим вредностима ових променљивих. За разлику од тога, при подели на особе са децом и особе без деце, установили смо да, на основу узорка, средња вредност нивоа религиозности међу особама без деце мања је од средње вредности нивоа религиозности међу особама са децом. Такође, закључили смо да је средња вредност броја афера међу људима са децом већа од средње вредности броја афера међу особама без деце (сва тестирања рађена су са нивоом значајности 0.05). На крају, на примеру: "Да ли је медијана нивоа религиозности једнака 3?", илустрован је тест знакова - његова имплементација у R-у преко binom.test фунцкије, као и "пешке" имплементација. У оба случаја добијена је иста p-вредност, и (нетачан) закључак да медијана није 3.