Projektna naloga pri Statistiki

Jimmy Zakeršnik

17.7.2022

Povzetek

V tej projektni nalogi pri predmetu Statistika, bom obravnaval tri naloge po navodilih. Vsaka naloga bo obravnavana v svojem lastnem poglavju. Da bo se lažje sklicevati na njih, bo prva naloga poimenovana *Kibergrad*, druga *Slučajni sprehod* in tretja *Temperature*.

1 Kibergrad

Priložena datoteka *Kibergrad.csv*, ki vsebuje podatke o dani populaciji (prebivalci mesta Kibergrad), je bila odprta v programu LibreOffice Calc. S pomočjo vgrajenega orodja so nato, bili zbrani vzorci velikosti 500 po postopku enostavnega vzorčenja. V priloženi datoteki *Kibergradwork.ods* so na prvi strani izpisani vsi podatki ter vseh pet pridobljenih vzorcev, na drugi strani je posebej obravnavan prvi vzorec v smislu kvartilov in extremnih vrednosti, na tretji strani pa se na enak način obravnavajo vsi vzorci glede na dohodek družin tipa 1. Obe primerjavi sta dodatno podprti s pomočjo škatel z brki in na koncu sta izračunani še s tipi pojasnjena varianca in nepojasnjena varianca dohodkov. Pri tem je v veliko pomoč programski jezik R.

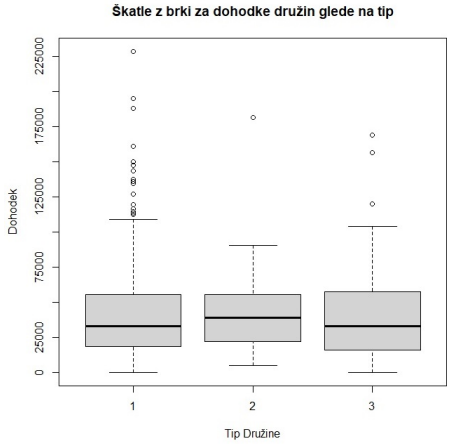
1.1 Primerjava dohodkov med tipi družin prvega vzorca

Poglejmo si najprej prvi vzorec in primerjajmo dohodke družin glede na tip družine. V delovnem okolju jezika **R** odpremo in poženemo skripto *Kibergrad_a.R*. Ob pogonu se v konzoli izpišejo vrednosti o dohodku, ki so navedene v spodnji tabeli.

	Tip 1	Tip 2	Tip 3
Max	228727	181696	168926
Q3	55360	54859	57631
Med	32975	38883	33310
Q1	18700	22011	16071
Min	0	5184	0

Tabela 1: *Tabela vrednosti, ki so potrebne za risanje škatel z brki za vsak tip družine*

Istočasno skripta izriše vzporedne škatle z brki, kot lahko vidimo na spodnji sliki, s pomočjo katerih lahko grafično primerjamo dohodke družin različnih tipov.



Slika 1: *Vzporedno narisane škatle z brki tipov družin* 1, 2 *in* 3

Škatle z brki nam ponudijo nekaj zanimivih ugotovitev. V minimalnih dohodkih ni velikih odstopanj, razen pri tipu 2, ki ima za razliko od ostalih pozitiven minimalen dohodek. Če ignoriramo ostale vzorce in tem rezultatom naivno verjamemo, so enostarševske družine z očetom (torej družine tipa 2) relativno bolj premožne od ostalih tipov. To domnevo podpira tudi opazka, da je povprečna vrednost tipa 2 višja kot povprečni vrednosti ostalih dveh tipov, kar lahko preberemo iz izpisa na konzoli.
Vrednosti so hkrati tudi dostopne v spodnji tabeli.

Tip	1	2	3
Povprečje	41655	46301	39931

Tabela 2: *Povprečne vrednosti dohodkov po tipih*

Če ne bi imeli že izračunanih povprečij, bi lahko še vedno sklepali o njihovih velikostih s pomoćjo škatel z brki. Opazimo namreč, da se prvi kvartili nahajajo na približno enaki višini z maksimalno razliko v okolici 6000 v prid družinam tipa 2, kar velja tudi za mediane. Šele pri tretjem kvartilu ostala dva tipa premagata tip 2, a tudi tu je največja razlika v rangu 4000. V tem primeru tipu 2 pomaga to, da ima izmed vseh tipov družin najmanjšo razdaljo med tretjim kvartilom in mediano, kar pomeni, da so vrednosti, ki pripadajo temu intervalu, bolj gosto porazdeljene.
Višje povprečje dohodka družin tipa 2 ni edino, kar izstopa pri škatlah z brki. Družine tipa 1 izstopajo v tem, da imajo, v primerjavi z ostalimi tipi, veliko število osamelcev (torej tistih vrednosti, ki so na sliki označene s krogi izven škatel).
Družine tipa 3 se odlikujejo po tem, da imajo najširši interkvartilni razmik. Za lažji pregled in primerjavo, so vsi interkvartilni razmiki navedeni v konzoli (po pogonu skripte) ter v tabeli spodaj:

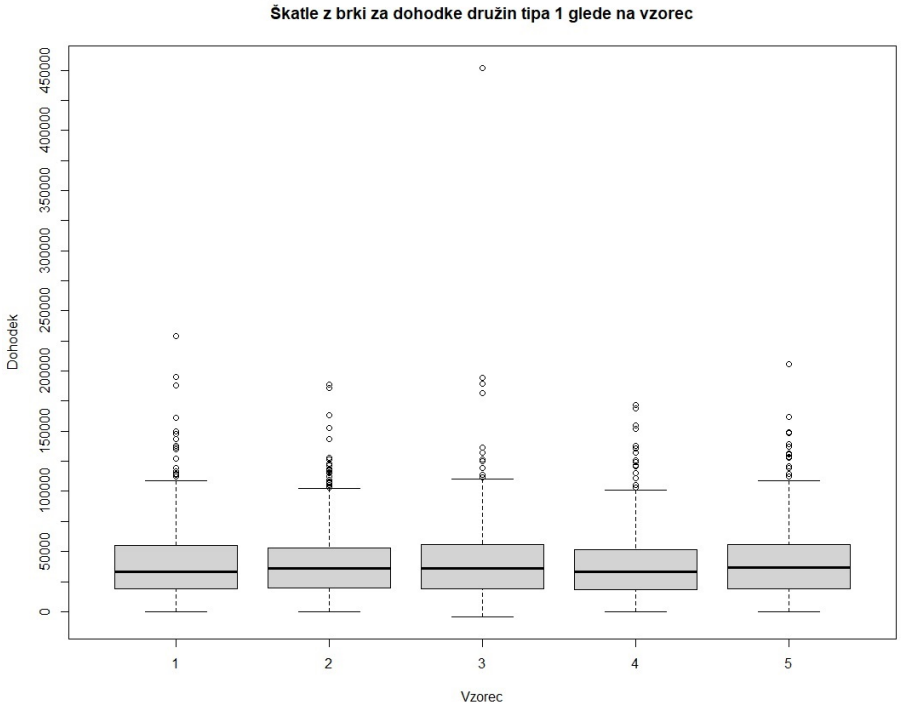
Tip	1	2	3
IQR	36660	32848	41560

Tabela 3: *Interkvartilni razmiki dohodkov po tipih*

Družine tipa 2 so torej v povprečju bolj premožne od družin ostalih tipov in družine tipa 3 so v povprečju najmanj premožne. Vrednosti družin tipa 3 so hkrati tudi najbolj razpršene v »srednji polovici«, kar nam pove velikost interkvartilnega razmika. Družine tipa 1 se v obeh primerih nahajajo v sredini med družinami tipa 2 in 3. Hkrati imajo tudi razmeroma več osamelcev od ostalih tipov. V vsakem primeru nam rezultati namigujejo, da obstaja povezava med tipom družine in njenim dohodkom.

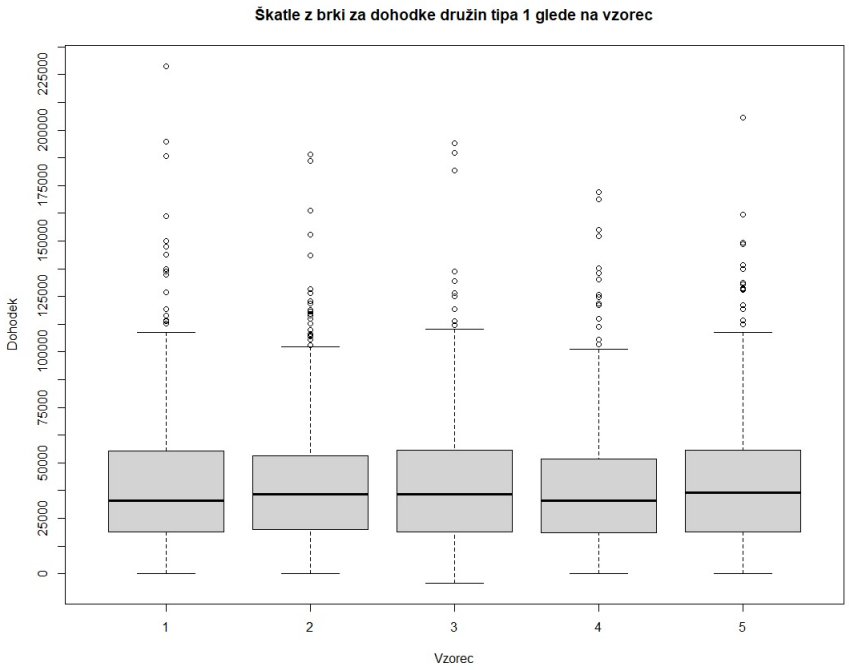
1.2 Primerjava dohodkov družin tipa 1 v petih vzorcih

Da nadaljujemo analizo podatkov si oglejmo porazdelitev dohodkov družin nekega tipa preko 5 neodvisno izbranih vzorcev. Pri tem za prvega vzamemo kar vzorec, ki smo ga obravnavali v prejšnjem podpoglavju, ostale štiri pa pridobimo s pomočjo orodij v LibreOffice Calc. Vsi vzorci so posebej shranjeni v lastni datoteki tipa *.csv* z imeni tipa *KibergradVzorec#.csv*.
Da izrišemo škatle z brki, s pomočjo katerih bomo primerjali dohodke preko vzorcev, poženemo skripto *Kibergrad_b.R*. V njej najprej naložimo vzorce, nato vsakemu vzorcu dodamo stolpec vrednosti, ki nam pove kateremu vzorcu pripada dan podatek. Torej vzorcu 1 dodamo stolpec samih enk, vzorcu 2 stolpec samih dvojk itd. Te tabele nato združimo v eno samo tabelo, iz nje prefiltriramo družine vseh tipov razen 1 in nato s pomočjo te nove tabele po enakem postopku kot v prejšnjem podpoglavju primerjamo dohodke družin glede na vzorec.
Dobljene škatle z brki so prikazane spodaj.



Slika 2: *Vzporedno narisane škatle z brki družin tipa* 1 *po vzorcih*

Takoj opazimo, da je prikazan graf razpotegnjen, v glavnem na račun enega osamelca iz tretjega vzorca. Da dobimo bolj pregleden graf, odstranimo vse vrednosti, ki so večje od 250000 (v resnici je taka zgolj ena). Škatle z brki, ki jih dobimo po tem popravku in so prikazane spodaj, skripta samostojno izriše.



Slika 3: *Popravljene vzporedno narisane škatle z brki družin tipa* 1 *po vzorcih*

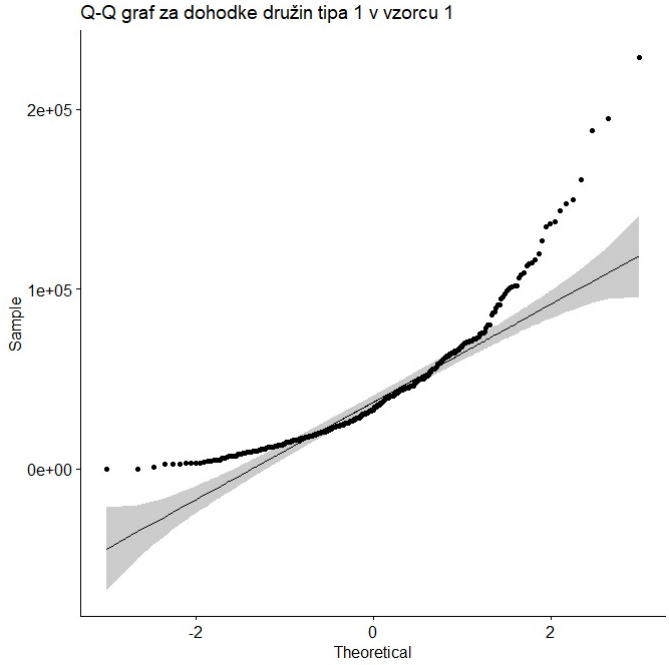
Ker smo pri »popravku« zanemarili zgolj eno vrednost, nam to bistveno ne pokvari primerjave. Izoliran osamelec bi lahko na našo obravnavo vplival kvečjemu negativno. Zato bomo pri izračunu povprečja za vsak vzorec uporabili tabelo, ki tega osamelca ne vsebuje.
Vrednosti (kvartili, povprečja, maksimalna in minimalna vrednost, IQR), ki jih skripta izpiše v konzolo, so prikazane v spodnji tabeli:

	Vzorec 1	Vzorec 2	Vzorec 3	Vzorec 4	Vzorec 5
Max	228727	188899	194230	171999	205712
Q3	55360	53121	55700	51700	55863
Med	32975	36000	35850	33006	36527
Q1	18700	20008	18800	18327	18668
Min	0	0	−4198	0	0
Povprečje	41655	41822	41337	39344	41697
SD	32974,23	31037,97	30517,25	29867,4	31099,04
IQR	36660	33113	36900	33373	37195

Tabela 4: *Ekstremi, kvartili, povprečja, standardni odkloni in interkvartilni razmiki dohodkov družin tipa* 1 *po vzorcih*

Sedaj, ko imamo narisane škatle z brki in zraven napisano tabelo, lahko komentiramo rezultate. V prvi vrsti opazimo, da so si povprečja dokaj blizu. Vsa povprečja razen povprečje četrtega vzorca se nahajajo v okolici 41500 ± 500, povprečje vzorca 4 pa se od 41500 razlikuje za manj kot 2500. Če bi si izbrali še več vzorcev, bi se po vsej verjetnosti njihova povprečja tudi nahajala v neki

bližnji okolici 41500. Podobno obnašanje standardnih odklonov, ki se nabirajo v okolici 31000 ± 2000 nas privede do nepresenetljivega sklepa, da je dohodek družin tipa 1 porazdeljen po vzorcih, torej porazdelitev dohodka ni odvisna od vzorca. To potrjuje tudi relativna bližina kvartilov v tabeli (npr. tretji kvartili se zbirajo v okolici 53000 ± 3000).
Na tej točki bi želeli preveriti, ali je porazdelitev slučajno normalna. Tako test s primerjalnim kvartilnim grafikonom kot Shapiro-Wilkov test na vzorcu 1 nam poveta, da to ne drži. V primeru Shapiro-Wilkovega testa, dobimo vrednost $p < 2.2e{-}16 < 0.05$, torej porazdelitev ni normalna. To vidimo tudi na primerjanlen kvartilnem grafikonu spodaj. Enak sklep seveda velja tudi za ostale vzorce.



Slika 4: *Primerjalni kvartilni grafikon dohodkov družin tipa* 1 *v vzorcu 1*

Četudi dohodki niso porazdeljeni normalno, so še vedno razmeroma konsistentni preko obravnavanih vzorcev. To je v kontrastu z razlikami in odstopanji, ki smo jih opazili, ko smo v prvem vzorcu primerjali dohodke glede na tip družine. Ta kontrast dodatno potrjuje sklep, da tip družine netrivialno vpliva na dohodek družine.

1.3 S tipom pojasnjena varianca populacije

Na koncu prejšnjega podpoglavja smo prišli do sklepa, da ima tip družine netrivialen vpliv na dohodek družine. Če to drži ali ne, lahko preverimo z izračunom s tipom družine pojesnjene variance. Čim smo poračunali to, se lahko skličemo na zvezo med varianco in pojasnjeno ter nepojasnjeno varianco $(Var(X) = Pojasnjena_varianca + Nepojasnjena_varianca)$in poračunamo še slednjo.
Vsi računi in primerjave se prikažejo ob pogonu skripte *Kibergrad_c.R.* Pred-

postavimo, da so dohodki tipov družin X_i medseboj neodvisni. Predpostavko upravičimo z argumentom, da v splošnem dohodek soseda ne vpliva na naš dohodek. Z n_i označimo število družin tipa i ter z N velikost naše populacije. Z X označimo dohodke družin, z Y pa slučajno spremenljivko tipov družin, ki ima porazdelitev $P(Y = i) = n_i/N$. Z $\bar{X}_i$ še označimo pričakovano vrednost dohodka v družini tipa i. S tipom pojasnjeno varianco potem izračunamo po formuli $Var(E[X|Y]) = 1/(N-1) * \sum_{i=1}^{3} (\bar{X}_i - \bar{X})^2 * n_i$.