

**Prof. dr Goran Trajković
Doc. dr Zoran Bukumirić**

MEDICINSKA STATISTIKA U R PROGRAMSKOM OKRUŽENJU

**AKADEMSKA MISAO
Beograd, 2020.**

MEDICINSKA STATISTIKA U R PROGRAMSKOM OKRUŽENJU

Drugo izdanje

(Prvo izdanje Akademske misli)

Autori:

Prof. dr Goran Trajković

Doc. dr Zoran Bukumirić

Recenzenti:

Prof. dr Vesna Bjegović-Mikanović, Medicinski fakultet Univerziteta u Beogradu

Prof. dr Biljana Miličić, Stomatološki fakultet Univerziteta u Beogradu

Lektor: Biljana Ristić

Izdavač: Akademska misao, Beograd

Tehnički urednik: Aleksandar Mandić

Dizajn korica: Raša Mladenović

Štampa: Akademska misao, Beograd

Mesto i godina izdanja: Beograd, 2020

Tiraž: 200

ISBN 978-86-7466-817-7

MEDICINSKA STATISTIKA U R PROGRAMSKOM OKRUŽENJU

IZ PREDGOVORA PRVOM IZDANJU

Statistički metod je jedna od najvažnijih komponenata kvantitativnih istraživanja u medicini. Upotreba statističkih softverskih alata učinila je primenu statističkih metoda mnogo efektivnijom. Jedan od najpoznatijih i sve češće primenjivanih statističkih softverskih alata je *R* programski jezik i okruženje. Publikacija „Medicinska statistika u *R* programskom okruženju” prikazuje osnove rada u okruženju programskog jezika *R*, njegovo korišćenje u pripremi, vizuelizaciji i statističkoj analizi podataka. Cilj autora nije bio da prikažu sve mogućnosti *R*-a, već njegovu praktičnu primenu u rešavanju problema iz medicinske statistike.

Publikacija je namenjena, kako polaznicima koji se po prvi put susreću sa okruženjem programskog jezika *R*, kao i sa primenom statističkih metoda, tako i iskusnim korisnicima. Za rad u *R* programskom okruženju i rešavanje primera iz knjige dovoljno je osnovno poznavanje rada na računaru.

Ova knjiga je sadržinski podeljena na 15 poglavlja i obuhvata različite nivoe potreba studenata integrisanih i posle diplomskih studija, istraživača iz oblasti medicinskih nauka, kao i lekara svih specijalnosti za statističkom analizom i interpretacijom podataka i rezultata.

Knjiga je koncipirana tako da uz nju idu baze podataka i *R* skriptovi sa komandama za rešavanje svih primera. Na taj način korisnici mogu ponoviti analizu korak po korak i imati pristup kompletnim izlaznim rezultatima. Nazivi varijabli kucani su bez srpskih dijakritičkih znakova (č, ć, đ, š i ž) zbog potencijalnog neadekvatnog prikaza unutar izlaznih rezultata ili na grafikonima. Zavisno od verzije operativnog sistema i kodne strane računara korisnik može naknadno ukucati nazive varijabli (ćirilicom ili latinicom) u *R* skriptovima pre pokretanja i oni će se prikazati korektno.

Knjiga je rađena u *R* programskom okruženju instaliranom na računaru sa *Windows* 10 operativnim sistemom. *R* skriptovi sa komandama mogu se koristiti i unutar *R*-a instaliranog na *Linux* ili *MacOS* platformama.

Tekst publikacije predstavlja rezultat višegodišnjeg naučnog i stručnog rada autora i sadrži materijale sa više onlajn kurseva koji se koriste u nastavi.

Za preuzimanje većeg dela teksta iz knjige potrebna je dozvola autora. *R* kôd koji se koristi u primerima u knjizi deo je *R* programskog okruženja i može se slobodno kopirati i reprodukovati bez traženja dozvole autora.

Autori se zahvaljuju kolegama koji su dobrovoljno ustupili delove svojih baza podataka za primere u knjizi. Ujedno se zahvaljuju i svima koji su, svako na svoj način, doprineli da knjiga bude bolja.

Beograd, 2019. godine

Autori

PREDGOVOR DRUGOM IZDANJU

Drugo izdanje publikacije „Medicinska statistika u R programskom okruženju” dopunjeno je indeksom termina. Pored toga, uz knjigu se dobija prateći CD na kome se nalaze baze podataka i R skripte sa kompletnim komandama neophodnim za izvršavanje svih primera iz knjige.

Tipografske konvencije korišćene u knjizi:

- *Kurziv* – koristi se za: nazive baza podataka, nazive primenjenih statističkih metoda, naglašavanje pojedinih termina, nove termine i reči na engleskom jeziku.
- **Podebljano** – koristi se za nazive paketa za R i podnaslove.
- ***Kurziv i podebljano*** – koristi se za komande u *Windows* operativnom sistemu.
- *Kurziv i podvučeno* – koristi se za internet adrese.
- Konstantne širine – koristi se za: R komande u tekstu, nazive varijabli, prikaz R outputa, tumačenje delova R outputa u tekstu i shematski prikaz strukture R komandi.
- Konstantne širine markirano – koristi se za prikaz R komandi u primerima.

R output

U R outputu prikazaće se deo izlaznih rezultata koji se interpretira u okviru primera.

NAPOMENA:

Ukazuje na nešto bitno, bilo da je to predlog, savet ili napomena opšteg tipa.

Ovo izdanje izlazi nakon iznenadne smrti Prof. dr Gorana Trajkovića, jednog od autora, koji i na ovaj način nastavlja da nesebično prenosi svoje znanje drugima.

Zahvaljujem se svima koji su podržali i doprineli izdavanju ove publikacije.

Beograd, 2020. godine

Doc. dr Zoran Bukumirić

SADRŽAJ

1. OSNOVNI STATISTIČKI POJMOVI	11
2. R PROGRAMSKO OKRUŽENJE	15
2.1 Zašto koristiti R?	15
2.2 Preuzimanje i instaliranje u <i>Windows</i> operativnom sistemu	16
2.3 Interfejs i struktura R-a	16
2.4 R komande	18
2.5 R skript	19
2.6 Spisak izvršenih komandi	21
2.7 Objekti	21
2.8 Radni prostor	21
2.9 R paketi	22
2.10 R kao kalkulator	23
3. RAZVOJ I DOKUMENTOVANJE BAZE PODATAKA	25
3.1 Formiranje baze podataka u tabelarnim kalkulatorima	25
3.2 Učitavanje baze podataka u R programsko okruženje	27
3.3 Nedostajuće vrednosti u bazi podataka	28
4. DESKRIPTIVNA STATISTIKA	29
4.1 Deskripcija kategorijalnih podataka	33
4.1.1 Učestalosti kategorija jedne varijable	33
4.1.2 Učestalosti u tabeli kontingencije $r \times k$	33
4.2 Deskripcija numeričkih podataka	34
4.2.1 Mere centralne tendencije i mere varijabiliteta jedne varijable	34
4.2.2 Mere centralne tendencije i mere varijabiliteta po grupama	35
5. ANALIZA EMPIRIJSKIH RASPODELA	37
5.1 Provera normalnosti raspodele	37
5.1.1 Testiranje normalnosti raspodele jedne varijable za sve ispitanike	38
5.1.2 Testiranje normalnosti raspodele po grupama	40
5.2 Transformacija podataka	44

6. GRAFIČKO PRIKAZIVANJE PODATAKA I REZULTATA PRIMENJENIH STATISTIČKIH ANALIZA	49
6.1 ggplot2	49
6.2 Izvoz grafikona u odgovarajući grafički format	50
7. ISTRAŽIVANJA U MEDICINI: CILJEVI I DIZAJN	51
8. OCENJIVANJE POPULACIONIH PARAMETARA NA OSNOVU UZORKA	55
8.1 Ocena aritmetičke sredine u populaciji	55
8.2 Ocena proporcije u populaciji	57
9. TESTIRANJE HIPOTEZA	59
10. TESTIRANJE HIPOTEZA O PROPORCIJAMA, SREDNJIM VREDNOSTIMA ILI RASPODELAMA	63
10.1 Testiranje hipoteza za jedan uzorak	63
10.1.1 Hi-kvadrat test slaganja	63
10.1.2 Studentov t -test za jedan uzorak	67
10.1.3 Wilcoxonov test za jedan uzorak	68
10.2 Testiranje hipoteza za nezavisne uzorke	73
10.2.1 Pearsonov hi-kvadrat test	73
10.2.2 Fisherov test tačne verovatnoće	83
10.2.3 Studentov t -test za dva nezavisna uzorka	86
10.2.4 Test sume rangova (Mann-Whitneyev test)	90
10.2.5 Jednofaktorska analiza varijanse	94
10.2.6 Kruskal-Wallisov test	97
10.3 Testiranje hipoteza za zavisne uzorke	100
10.3.1 McNemarov test	100
10.3.2 Studentov t -test za dva zavisna uzorka	102
10.3.3 Test ekvivalentnih parova (Wilcoxonov test)	105
10.3.4 Cochranov Q test	107
10.3.5 Analiza varijanse ponovljenih merenja	111
10.3.6 Friedmanov test	116

11. ISPITIVANJE ODNOSA IZMEĐU VARIJABLI	121
11.1 Ispitivanje povezanosti kategorijalnih varijabli	121
11.1.1 Relativni rizik	121
11.1.2 Odnos šansi	123
11.2 Korelaciona analiza	126
11.2.1 Grafičke metode korelacione analize	126
11.2.1.1 Dijagram rasturanja	126
11.2.1.2 3D Dijagram rasturanja	127
11.2.2 Računske metode korelacione analize – koeficijenti korelacije	128
11.2.2.1 Pearsonov koeficijent linearne korelacije	128
11.2.2.2 Spearmanov koeficijent korelacije rangova	129
11.2.2.3 Korelaciona matrica	132
11.3 Regresiona analiza	136
11.3.1 Linearna regresiona analiza	136
11.3.1.1 Univarijantna linearna regresija	136
11.3.1.2 Multivarijantna linearna regresija	140
11.3.1.3 Poređenje regresionih nagiba	143
11.3.1.4 Regresiona ravan	145
11.3.2 Logistička regresiona analiza	147
11.3.2.1 Univarijantna logistička regresija	148
11.3.2.2 Multivarijantna logistička regresija	150
11.3.2.3 Nomogram – grafičko prikazivanje predikcije modela logističke regresije	154
12. GENERALNI LINEARNI MODELI	157
12.1 Dvofaktorska analiza varijanse	157
12.2 Analiza kovarijanse (ANCOVA)	160
12.3 Multivarijantna analiza varijanse (MANOVA)	163
13. KONTROLA PRIDRUŽENOSTI	169
14. ADEKVATNOST MERENJA	179
14.1 Pouzdanost merenja	180
14.1.1 Interna konzistentnost	180

14.1.2 Test-retest pouzdanost	181
14.1.3 Međuposmatračka pouzdanost	182
14.2 Valjanost merenja	185
14.2.1 Konkurentna valjanost	185
14.2.2 Dijagnostička valjanost	186
14.2.2.1 ROC kriva	190
14.3 Slaganje merenja	192
15. META-ANALIZA	195
LITERATURA	215
DODATAK: Korisničko okruženje i spisak dodatnih paketa	217
INDEKS	219