Podatci i uzorak

Tvrtko Tadić[1]

Sažetak

Ovo je prvi od niza članaka u kojemu pokušavamo ilustrirati neke pojmove i ideje iz statistike na primjerima stvarnih podataka. U ovom članku objašnjavamo važnost podataka u današnjem svijetu i na primjeru ilustriramo potrebu da informacije sažmemo u formi tablice i histograma. U drugom dijelu članka objašnjavamo što je to uzorak i simulacijama tumačimo zašto zaključke često donosimo upravo na temelju uzorka. Učenici sve mogu samostalno provjeriti koristeći *Excel* i programski jezik *Python*.

Uvod – važnost podataka danas

U umreženom i digitaliziranom svijetu akumulirale su se ogromne količine podataka. Brojne tvrtke znaju puno o svojim korisnicima i koriste te podatke kako bi poboljšale usluge koje pružaju.

Evo nekih primjera:

- *Amazon*, najveća svjetska *online* trgovina, skuplja podatke o tome što njegovi korisnici širom svijeta naručuju kako bi na skladištu imao potrebne proizvode. Također rade na tome da smanje troškove i vrijeme isporuke.
- Tražilice *Google* i *Bing* skupljaju podatke o napravljenim pretragama kako bi bolje ocijenili koje bi stranice na internetu mogle biti zanimljive njihovim korisnicima.
- Banke skupljaju podatke o svojim korisnicima u svrhu pouzdanijeg donošenja odluka o davanju kredita. U mnogim zemljama postoje specijalizirani kreditni uredi koji skupljaju podatke za banke i tvrtke za kreditno bodovanje koje procjenjuju rizike.
- Osiguravajuća društva također prikupljaju podatke s ciljem boljeg upravljanja povjerenog im novca.
- Prijevoznici, primjerice velike zrakoplovne tvrtke, prate broj putnika na svojim linijama kako bi bolje planirale broj potrebnih mjesta, upotrebu voznog parka i cijenu karata.

[1]Tvrtko Tadić, PMF-MO, Zagreb / Microsoft Corporation, Redmond / University of Washington, Seattle

- Investicijske kompanije prikupljaju razne podatke kako bi osigurale što bolje upravljanje svojom imovinom i napravile bolja ulaganja. Danas postoji cijeli niz kompanija koje imaju potpuno automatiziran sustav ulaganja i gdje odluke, primjerice o prodaji dionica, donose računala.
- ...

Sve ovo stavilo je svijet pred niz novih izazova:

- gdje držati sve te podatke;
- kako ih iskoristiti.

Razvijen je čitav niz novih tehnoloških rješenja, kao i posve nova područja znanstvenog interesa.

Poučavanje statistike

Statistika je izrasla u važnu disciplinu čije je poznavanje postalo iznimno bitno za donošenje pravih odluka. Odmah na početku treba istaknuti jednu stvar: statistika je iznimno bliska matematici, no ona se ne smatra dijelom matematike nego zasebnom **matematičkom znanošću**. Statistika ima svoju **primjenu** (u brojnim društvenim, prirodnim i tehničkim znanostima), **izazove računalne implementacije**, kao i svoju (matematičku) **teoriju**.

Poučavanje statistike predstavlja velik izazov iz više razloga:

- statistiku treba širok niz stručnjaka raznih profila,
- potrebno je dosta znanja matematike da bi se ona razumjela,
- teorija je iznimno zahtjevna,
- u različitim područjima primjene metode se mogu bitno razlikovati. To je vrlo često uzrokovano manjom ili većom dostupnošću podataka. Primjerice, u medicini ćemo za neku rijetku bolest imati bitno manje podataka nego o kreditima u banci koja ima milijune klijenata.

U Hrvatskoj se već dulji niz godina pokušavaju uvesti elementi statistike u osnovne i srednje škole. Imajući u vidu navedeno, smatramo da za dobro razumijevanje statistike to nije moguće napraviti samo kroz nastavu matematike, te da bi bilo dobro da se dio toga tereta preraspodijeli i na druge predmete (poput informatike ili, primjerice, zemljopisa/geografije).

Kako bi pojasnili svoje tvrdnje u ovom i idućih nekoliko članaka, iznijet ćemo neke metodičke primjere iz statistike koji bi trebali omogućiti nastavnicima da lakše prenesu učenicima što je to statistika.

U svrhu obrade podataka koristit ćemo *Microsoft Excel* i programski jezik *Python*.

Statistika

Što je statistika? Počet ćemo s klasičnom definicijom pojma *statistika*.

> ***Statistika*** *je disciplina koja se bavi prikupljanjem, analizom, tumačenjem i prikazivanjem podataka.*

Vidimo da je definicija iznimno općenita te da praktično svatko tko prati neke podatke može biti statističar. Tijekom stoljeća ljudi su skupljali podatke i susretali se s raznim izazovima:

- Kako sažeti informacije o prikupljenim podacima?
- Kako protumačiti podatke?
- Mogu li se napraviti predviđanja na temelju prikupljenih podataka?
- Što ako nije moguće prikupiti sve podatke?

Razvijena je cijela teorija i razne praktične tehnike obrade podataka. Mi ćemo se kroz primjere stvarnih podataka upoznati s važnom terminologijom i nekim tehnikama obrade podataka. Pritom ćemo se osvrnuti na matematičku teoriju koja stoji u pozadini, ali nećemo ulaziti u detalje.

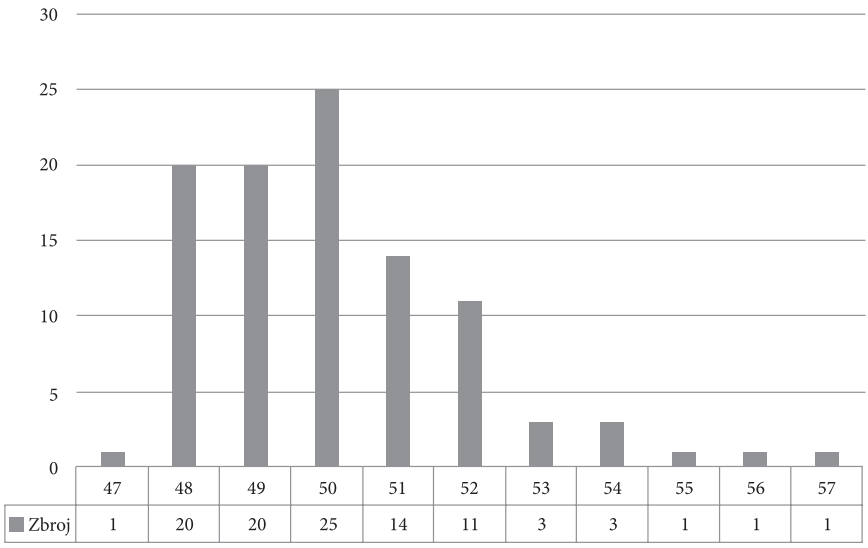
Otkucaji srca

Sljedeći niz podataka predstavlja 100 mjerenja prosječnog broja otkucaja autorova srca u minuti tijekom sna (prikupljenih od 2. 1. 2016. do 18. 4. 2016. pomoću uređaja *Microsoft Band* i povučenih iz baze *Microsoft Health*):

51	49	52	53	52	51	55	52	52	57	50	54
50	54	51	50	52	52	48	50	49	49	49	49
50	50	51	49	50	50	48	47	49	48	50	48
49	51	52	50	53	53	50	52	50	50	49	52
48	48	52	51	50	48	56	48	51	49	48	50
50	50	49	49	50	48	51	49	48	48	54	50
50	48	48	50	49	49	49	51	51	48	50	50
51	50	48	48	48	49	51	49	48	51	49	49
48	50	51	52								

Tablica frekvencija i histogram

Ovi brojevi ovako poslagani slabo nam govore o otkucajima srca. Jedna od ideja obrade podataka je pretvoriti podatke u pregledniju formu. U ovom slučaju bilo bi korisno kad bismo znali **frekvenciju**, tj. koliko je puta zabilježen pojedini broj. Također ćemo nacrtati **histogram** – stupčasti dijagram koji će nam vizualno dočarati frekvencije pojedinih brojeva. Ovo je lako napraviti u *Excelu*.



	47	48	49	50	51	52	53	54	55	56	57
Zbroj	1	20	20	25	14	11	3	3	1	1	1

Slika 1. Histogram podataka o prosječnom broju otkucaja srca

Što smo dobili? Iz histograma i tablice frekvencija preglednije vidimo neke informacije. Bez dubljeg ulaženja u analizu sa slike možemo odmah reći da autor tijekom sna ima prosječno od 47 do 57 otkucaja srca u minuti. Možemo očitati određene odnose. Primjerice:

- autor rijetko u prosjeku ima više od 54 otkucaja i manje od 48 po minuti;
- prosjek broja otkucaja srca u 90 % slučajeva nalazi se u intervalu [48, 52];
- autor je u najviše slučajeva imao prosjek od 50 otkucaja u minuti.

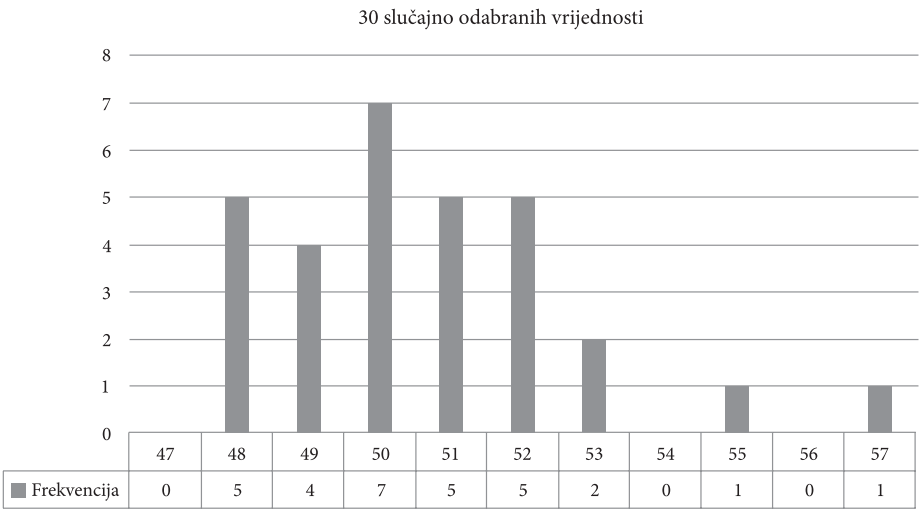
U ovom primjeru imamo samo 100 podataka koji poprimaju tek 11 vrijednosti pa nije jednostavno napraviti ni tablicu ni histogram bez pomoći računala i specijaliziranih programa.

Uzorak

Vrlo često neće nam biti dostupni svi podatci, nego ćemo baratati tek manjim dijelom njih – **uzorkom**. No i to će nam često biti **dovoljno da bi smo došli do određenih zaključaka o svim podatcima**.

Promatrajući podatke možemo vidjeti da će, ako uzmemo manji uzorak, ponašanje podataka biti ***slično*** kao da smo uzeli sve podatke.

To ćemo ilustrirati na podatcima o prosječnom broju otkucaja srca. Slučajno odaberimo 30 vrijednosti. Jedna od mogućih realizacija ovakvog odabira podataka dana je histogramom i tablicom frekvencija prikazanom na slici 2.



Slika 2. Histogram uzorka

Uočimo sličnosti s originalnim podatcima i prikazom podataka na slici 1:

- brojka 50 pojavljuje se najviše puta;
- 86.66 % podataka i dalje se nalazi u intervalu [48, 52];
- vrijednosti iznad 54 i ispod 48 su rijetke.

Uočimo kako smo slične zaključke imali kod svih podataka. Brojna svojstva podataka prenijet će se na uzorak s velikom vjerojatnošću. To ćemo pokazati eksperimentalno u sljedećem odjeljku, nakon što objasnimo uzimanje uzorka.

Tipovi uzorka

Svaki uzorak sadržavat će određen broj podataka. Taj broj zovemo **duljina uzorka**. Postoje dva načina generiranja uzorka:

- **slučajni uzorak** gdje slučajno biramo podatak i postupak ponavljamo na svim podatcima sve dok ne dobijemo uzorak željene duljine (isti podatak može biti izvučen više puta);
- **reprezentativni uzorak** gdje slučajno biramo podatak i postupak ponavljamo na podatcima koji još nisu bili izabrani sve dok ne dobijemo uzorak željene duljine.

Ova dva načina uzimanja uzorka postoje iz raznih praktičnih i teorijskih razloga u koje sada nećemo ulaziti. Većina računalnih jezika više razine (poput Pythona) ima ugrađene funkcije za generiranje ovih uzoraka.

Napravit ćemo idući eksperiment (pomoću računala, za detalje vidi dodatak na kraju članka):

- 1 000 000 (milijun) puta uzet ćemo uzorak duljine 30.
- Za svaki uzorak provjerit ćemo:
  - je li 50 broj koji se najviše puta pojavio u uzorku (uključujući mogućnost da se pojavio jednako mnogo puta kao neki drugi broj);
  - nalazi li se 85-95 % brojeva u uzorku u intervalu [48, 52];
  - ima li manje od 5 % brojeva vrijednost iznad 54 ili manje od 48.

Rezultati eksperimenta su sljedeći:

provjera / tip uzorka	slučajni	reprezentativni
broj 50 najčešće se pojavljuje u uzorku	u 53.8 % slučajeva	u 59.7 % slučajeva
85-95 % brojeva u uzorku nalazi u intervalu [48, 52]	u 64.2 % slučajeva	u 72.6 % slučajeva
manje od 5 % brojeva ima vrijednost iznad 54 ili manje od 48	u 66.1 % slučajeva	u 65.2 % slučajeva

Rezultati ilustriraju činjenicu da **uzorak često preuzima brojna svojstva koja ima cijeli skup podataka**. To će još češće biti slučaj za velike uzorke još većeg skupa podataka. Primjer toga su, primjerice, izlazne ankete na izborima (vidi članak (Tadić, 2011)) koje jako dobro uspiju procijeniti pobjednika izbora unatoč tome što su anketirale tek nekoliko desetaka tisuća ljudi od milijun glasača. U sljedećim člancima vidjet ćemo kako to funkcionira u drugim slučajevima.

Zbog navedenog, **brojna svojstva cijelog skupa podataka pokušavamo procijeniti na uzorku**.

Sve izrečeno može se matematički precizno formulirati koristeći napredne alate matematičke analize i terminologiju teorije vjerojatnosti.

Čemu služe ovi podatci u praksi?

Ovi podatci govore o uobičajenim vrijednostima prosječnog broja otkucaja autorova srca. Ukoliko bi te vrijednosti počele izlaziti izvan okvira uobičajenih vrijednosti, npr. ako bi autor zabilježio tri dana za redom prosječan broj otkucaja veći od 55, to bi bio znak da je došlo do određene promjene koja može biti uzrokovana:

- zdravstvenim stanjem,
- pokvarenim uređajem za mjerenje,

- lošim snom,
- nečim sasvim drugim.

Obično se prati više takvih signala na temelju kojih se može donijeti zaključak (u skladu s prijašnjim iskustvima) koje su promjene vjerojatno nastupile.

Veliki problem kod donošenja ovakvih zaključaka je **nedostupnost podataka**. Primjerice, podatci o broju otkucaja srca su privatni, podatci o trenutnom zdravstvenom stanju se ne bilježe. Sve to predstavlja izazove s kojima se susreću oni koji u primjenama pokušavaju donijeti zaključke.

O odnosu prakse i teorije

U statistici i općenito praktičnim primjenama matematike koriste se razne matematičke metode. Često se u praksi nađu razni postupci koje je vrlo teško teorijski objasniti. Kako (posebice ako se radi o kompanijama) rijetko tko može čekati 5 godina da se neki postupak koji heuristički ima smisla i radi u praksi objasni, takvi postupci najčešće se opravdavaju simulacijama (slično kao što su neke stvari izložene u ovome članku). U mnogo slučajeva, kad se napokon nađe teorijsko objašnjenje, ono zna uključivati nerealne pretpostavke.

Zaključak

Cilj ovoga članka bio je na jedan drukčiji način pokušati objasniti neke postupke u statistici. Ideja je da se kroz stvarne podatke i računalne simulacije opravdaju neki statistički postupci, a da se pritom izbjegne zahtjevna matematička teorija. U poučavanju statistike ovakav pristup ima jednaku važnost kao otkrivanje geometrijskih činjenica crtanjem. Kao što ćemo u idućim člancima vidjeti, za obradu podataka i izvođenje zaključaka upotreba računala bit će neizbježna. Kako bi učenici stekli ideju o tome što je statistika, nužno je da se upoznaju i eksperimentiraju sa stvarnim podatcima koristeći pomoć proračunskih tablica i programskih jezika. U ovom članku pokušali smo objasniti što je uzorak i ilustrirati zašto ga uzimamo. Osnovni pregled mnogih pojmova dan je u člancima (Varošanec, 2013.) i (Varošanec, 2014). Za više razrede srednje škole knjiga (Sarapa, 1996.) temeljito daje uvod u statistiku.

Dodatak – podatci i kod

Podatci i kod korišteni u izradi ovoga članka mogu se preuzeti s internetske stranice: https://web.math.pmf.unizg.hr/~tvrtko/metodikaStatistike/clanak1

Dodatak – *Excel*

Excel je standardni program za pripremu proračunskih tablica i crtanje grafikona koji je dio *Microsoft Office* ponude. Postoji i besplatna (nešto jednostavnija)

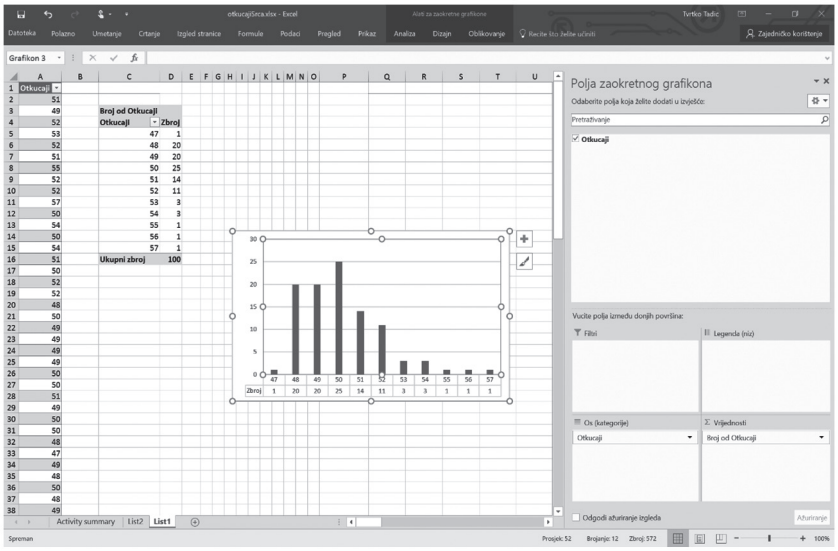
online verzija. Ovdje ćemo kratko prikazati kako nacrtati histogram i kako napraviti reprezentativni odnosno slučajni uzorak. Sve ovdje navedeno napravljeno je u verziji *Excel* 2016.

Crtanje histograma

Otvorimo novi *Excel* dokument i odaberimo stupac u koji ćemo smjestiti podatke. U prvi red stupca upišimo naziv – *Otkucaji*, a nakon toga smjestimo podatke.

Odaberimo podatke i napravimo tablicu (Polazno -> Oblikuj kao tablicu).

U idućem koraku odaberemo cijelu tu tablicu i napravimo zaokrenuti dijagram (Umetanje -> Zaokretni grafikon). Postavljanjem da *Os (kategorije)* budu *Otkucaji*, a *Vrijednosti* stavimo da je *Broj od Otkucaji* dobivamo i tablicu frekvencija i histogram.



Slika 3. Pripremanje histograma

Uzimanje slučajnog i reprezentativnog uzorka

Prvo podatke stavimo u neki stupac. Primjera radi, kao na slici 3, pretpostavimo da vrijedi sljedeće:

- podatci se nalaze u stupcu A i zauzimaju redove 2 do 101;
- želimo simulirati uzorak duljine 30.

Slučajni uzorak u *Excelu* generira se na sljedeći način:

- odaberemo neki drugi slobodni stupac i u prvih 30 redova toga stupca upišemo kombinaciju naredbi:

=INDEX($A:$A,RANDBETWEEN(2,101),1)

Ova naredba iz stupca A na slučajan način odabire polje u redu između 2 i 101 i ispisuje njegov sadržaj.

Reprezentativni uzorak generira se na nešto drukčiji način:

- u stupcu B u redovima 2 do 101 upišemo naredbu:

=RAND()

Ona generira slučajan broj iz intervala [0,1].

- Odaberemo stupce A i B od 1 do 101 reda. Napravimo sortiranje (uzlazno ili silazno) po vrijednostima u stupcu B (Podatci -> Sortiranje).
- Prvih 30 vrijednosti predstavlja reprezentativni uzorak.

Dodatak – *Python*

Excel je alat prikladan za brzi pregled podataka i izradu grafova. Za zahtjevniju obradu podataka često koristimo programske jezike. U praksi se kao alati za obradu podataka koriste *R*, *Python*, *Matlab* i neki drugi popularni alati. Kako se *Python* koristi u nastavi informatike (vidi primjerice (L. Budin, 2012.)), prikazat ćemo kao napraviti 1 000 000 simulacija reprezentativnog i slučajnog uzorka te provjeriti jesu li pojedina svojstva izvornih podataka ostala sačuvana u uzorku.

Idejno, algoritam izgleda ovako:

unos: podatci, duljine uzorka, broj eksperimenata

za j = 1 **do** broj eksperimenata:

uzmi uzorak iz danih podataka

provjeri svojstva na uzorku

za svako svojstvo:

ispiši $\frac{\textit{koliko je puta svojstvo bilo zadovoljeno}}{\textit{broj eksperimenata}}$

Implementacija u *Pythonu* izgleda ovako:

```
#unos paketa za generianje slucajnih brojeva
import random;

duljinaUzorka = 30;
brojEksperimenata = 100000;
```

```
#brojaci
najviseBroj50 = 0;
punoPutaUIntervalu4852 = 0;
maloPutaIzvanIntervala4854 = 0;

#ucitavanje podataka
with open('otkucaji.txt') as podaciIzvor:
  podaci = [int(x) for x in podaciIzvor.readlines()]

duljinaPodataka = len(podaci);

for k in range(0,brojEksperimenata):
  #reprezenativni uzorak
  uzorak = random.sample(podaci, duljinaUzorka);
  #slucajni uzorak
  #uzorak = [podaci[random.randrange(duljinaPodataka)] for j
  in range(0,duljinaUzorka)]

  vrijednostiUzorka = set(uzorak);
  vrijednostiPodatakaBrojPojavljivanjaUUzroku = dict(
      (vrijednost,uzorak.count(vrijednost)) for vrijednost in
vrijednostiUzorka);

  #provjeri da li je broj 50 najcesci u uzorku
  if(uzorak.count(50) == max(vrijednostiPodatakaBrojPojavlji-
  vanjaUUzroku.values())):
      najviseBroj50 = najviseBroj50 + 1;

  #provjeri da li je 85%-95% vrijednosti uzorka u intervalu [48,52]
  relativnaFrekvencijaIntrevala4852 = sum(
     [vrijednostiPodatakaBrojPojavljivanjaUUzroku[vrijednost]
for vrijednost in
      vrijednostiUzorka if (vrijednost <= 52 and vrijednost >= 48)]
      )/duljinaUzorka;
  if(relativnaFrekvencijaIntrevala4852 <= 0.95 and 0.85 <= re-
  lativnaFrekvencijaIntrevala4852):
      punoPutaUIntervalu4852 = punoPutaUIntervalu4852 + 1;

  #provjeri da li manje od 5% vrijednosti manje od 48 i vece od 54
  relativnaFrekvencijaBrojevaIzvanIntervala4854 = sum(
     [vrijednostiPodatakaBrojPojavljivanjaUUzroku[vrijednost]
for vrijednost in
      vrijednostiUzorka if (vrijednost > 54 or vrijednost < 48)]
      )/duljinaUzorka;
```

```
   if(relativnaFrekvencijaBrojevaIzvanIntervala4854 <= 0.05):
      maloPutaIzvanIntervala4854 = maloPutaIzvanIntervala4854 + 1;

#ispisi rezultat
print([najviseBroj50/brojEksperimenata,
     punoPutaUIntervalu4852/brojEksperimenata,
     maloPutaIzvanIntervala4854/brojEksperimenata]);
```

Literatura:

1. Budin, L. i drugi (2012.). *Rješavanje problema programiranjem u Pythonu : za 2. i 3. razred gimnazije.* Zagreb: Element.
2. Sarapa, N. (1996.). *Vjerojatnost i statistika 2. dio: Osnove statistike - slučajne varijable.* Zagreb: Školska knjiga.
3. Tadić, T. (2011.). Matematika iza anketa – primjer izbora. *Poučak br. 43.*
4. Varošanec, S. (2013.). Grupirani podaci I. *Matematika i škola br. 72.*
5. Varošanec, S. (2014.). Grupirani podaci II. *Matematika i škola br. 74.*
6. Wasserman, L. (2005.). *All of Statistics: A Concise Course in Statistical Inference.* New York: Springer.