УНИВЕРЗИТЕТ У БЕОГРАДУ
МАТЕМАТИЧКИ ФАКУЛТЕТ

Бојана Милошевић

ОСНОВИ СТАТИСТИКЕ

— Уџбеник —

Београд, 2021.

доц. др Бојана Милошевић
Математички факултет

ОСНОВИ СТАТИСТИКЕ

Прво издање

Издавач: Математички факултет, Студентски трг 16, Београд

За издавача: проф. др Зоран Ракић

Издавачки одбор: проф. др Зоран Петровић, председник
проф. др Олга Атанацковић
проф. др Мирјана Ђорић
проф. др Предраг Јаничић
проф. др Милош Арсеновић

Рецензенти: доц. др Марко Обрадовић
доц. др Јелена Станојевић

Штампа и повез: Службени гласник, Београд

Тираж: 200 примерака

© 2021. Математички факултет у Београду. © 2021. Аутор.
Сва права задржана. Ниједан део ове публикације не може бити репродукован нити смештен у систем за претраживање или трансмитовање у било ком облику (електронски, механички, фотокопирањем, снимањем или на други начин), без претходне писане дозволе издавача и аутора.

Предговор

Уџбеник *Основи статистике* написан је првенствено да прати програм предмета *Увод у статистику* и *Статистика*, али је осим студентима Математичког факултета, намењен свима којима је циљ да се упознају са основним статистичким методама и начину на који се примењују у пракси. Свако поглавље садржи и практичне примере у програмском језику R. За њихово разумевање неопходно је само познавање основне синтаксе програмсог језика. У многим ситуацијама је приказан и псеудокод чиме је читаоцима омогућено да самостално имплементирају методологију у програмском језику који користе. За праћење садржаја уџбеника није потребно напредно знање из теорије вероватноће већ познавање основних концепата и појмова.

Уџбеник се састоји из неколико целина. У првом делу је приказан *пут од података до једноставног статистичког модела* и састоји се од уводног поглавља у коме је акценат на графичком приказу и дескриптивним статистикама, поглавља у коме је приказано тачкасто и интервално оцењивање параметара расподела, и поглавља о тестирању статистичких хипотеза. Поред тестова о вредностима параметара неких расподела или параметара популације, приказано је и неколико тестова независности, као и неколико тестова сагласности са расподелом. Значајан део прве целине је посвећен, кроз приказ Монте Карло метода, емпиријском закључивању о особинама оцена као и особинама представљених тестова. У другом делу су представљени основни регресиони модели и приказано како се могу користити за предикцију и класификацију. Фокус трећег дела је на уводу у Бајесово закључивање са циљем да се читаоцима прикаже овај начин статистичког закључивања и дочара њен значај у статистици 21. века.

Свако поглавље завршава се задацима којима се може проверити да ли је градиво текућег поглавља савладано. Један део задатака је практичног типа и може представљати основу самосталних истраживачких пројеката.

На крају уџбеника налазе се два додатка, један у коме су приказани подаци који су коришћени у претходним поглављима и други који садржи приказ најчешће коришћених расподела у статистици. За сваку од расподелу графички је приказан закон расподеле, или функција густине, наведене основне нумеричке карактеристике, и наведена њихова примена. Такође, за сваку од расподела приказане су функције у R-у које се могу користити за одређивање функције расподеле, закона расподеле (густине расподеле), квантила расподеле и за генерисање псеудослучајних бројева из те расподеле.

Захваљујем се Марку, Благоју и Данијелу. Многи примери су резултат наших свакодневних дискусија како да студентима што боље приближимо значај статистике као науке и помогнемо при развоју статистичког начина мишљења. Хвала Пеђи што ми је указао на важност постојања адекватне писане литературе као и Тамари и Марији и свим студентима који су својим коментарима допринели квалитету садржаја исте. На крају, топло се захваљујем Јелени и Марку на пажљиво прочитаном рукопису и корисним сугестијама.

Београд, март 2021. Аутор

Садржај

1 Основни статистички појмови **1**

1.1 Осмишљавање експеримента и узорковање 4

1.2 Прелиминарна анализа 5

1.3 Идентификација аутлајера 18

1.4 Особине узорачке средине и узорачке дисперзије . . 22

1.5 Емпиријска функција расподеле 27

2 Оцењивање непознатих параметара расподела **31**

2.1 Тачкасте оцене параметара 31

2.1.1 Метод момената 32

2.1.2 Метод максималне веродостојности 36

2.2 Особине оцена 41

2.3 Интервалне оцене параметара 53

2.3.1 Закључивање у моделу са нормалном расподелом 54

2.3.2 Закључивање у моделу са биномном $\mathcal{B}(1,p)$ расподелом 66

2.3.3 Закључивање у моделу са Пуасоновом $\mathcal{P}(\lambda)$ расподелом 71

2.3.4 Интервал поверења за средњу вредност . . 73

3 Тестирање статистичких хипотеза **75**

3.1 Параметарски тестови 86

3.1.1 Тестови у нормалном моделу 86

3.1.2 Тестови у биномном моделу 99

3.1.3 Тестови у Пуасоновом моделу 104

3.2 Непараметарски тестови 108

3.2.1 Тест знакова 110
3.2.2 Вилкоксонов тест заснован на ранговима и знаковима 115
3.2.3 Тестови сагласности са расподелом 126
3.2.4 Тестови о једнакој расподељености два узорка 142
3.2.5 Тестови независности 145

4 Регресиони модели 154
4.1 Линеарна регресија 156
4.2 Логистичка регресија 197

5 Увод у Бајесову статистику 231
5.1 Бајесово оцењивање 232
5.2 Бајесово тестирање статистичких хипотеза 249
5.3 Неколико потенцијалних проблема 256

6 Додатак - важне расподеле и њихове особине 258
6.1 Дискретне расподеле 258
6.2 Апсолутно непрекидне расподеле 261
6.3 Важне расподеле у R-у 266

7 Додатак – Подаци 277

Литература 286

Списак појмова 289

1

Основни статистички појмови

Статистика је наука о подацима. Бави се њиховим прикупљањем и анализом, презентовањем и закључивањем, као и доношењем одлука. Зато можемо слободно да кажемо да је статистика основни алат модерне цивилизације. Томе је свакако значајно допринео развој рачунара. Поред ове дефиниције, *статистика* има и друго значење са којим ћемо се убрзо упознати.

Основни задатак статистичара је да предложи одговарајући *математички* модел којим би се подаци адекватно описали, након чега је могуће вршити даље анализе и предвиђања. Ваш задатак је да се упустите у статистичку авантуру заједно са мном и да овладате што већим бројем познатих модела, да научите када их треба примењивати и на који начин. Самостално предлагање нових модела, на основу стеченог знања, свакако је један од ваших задатака, а ја се искрено надам да ће вам овај уџбеник помоћи у томе. Како то обично на почетку бива, потребно је прво савладати језик којим ћемо се служити. Зато на почетку наводимо основне појмове.

Популација је скуп јединки чије карактеристике изучавамо. Карактеристике које су предмет изучавања називамо *обележјима*. О њима најчешће довољно сазнајемо на основу неког подскупа популације који називамо *узорак*. Уколико се

подскуп бира насумично, односно ако сваки подскуп има неку вероватноћу да буде извучен, говоримо о *случајном узорку.* Поред случајности његова важна особина је *репрезентативност.* Као што и сама реч каже, потребно је да се на основу њега може закључити о расподели обележја на читавој популацији, као и да одабир чланова узорка не зависи од вредности обележја тих чланова.

Поставља се питање зашто уопште узимамо узорак уколико је цела популација доступна. Много је разлога за то: трошкови прикупљања, анализе података, време за које је потребно извршити анализу од постављеног задатка, и многи други.

Пример 1.0.1. *Претпоставимо да желимо да видимо какво је мишљење нације непосредно пре реализације референдума о неком, за државу кључном питању, и претпоставимо да је на постављено питање могуће одговорити само са "да" или "не". Тада је популација гласачко тело државе - пунолетни грађани, док је обележје одговор на питање. Имајући у виду предзнање из вероватноће, најприродније је да тај одговор моделирамо случајном величином X чије су вредности 0 (за "не") или 1 (за "да"). Даље, како немамо услова да испитамо целу популацију непосредно пре референдума, и то чак нема ни смисла, јер би то значило понављање референдума два пута, испитаћемо само неке грађане. То ће бити наш узорак. Ту треба бити опрезан. Имајући у виду да су претходне статистичке анализе показале да многе социо-економске карактеристике утичу на јавно мњење, јасно је да нпр. узорак од 1000 становника руралних средина, или 1000 становника који живе у градским језгрима, ће нас навести на скроз другачије закључке за којим трагамо. Због тога ова два узорка нису репрезентативна. Пример репрезентативног узорка би био неки случајан избор од 1000 чланова популације.*

Пример 1.0.2. *Претпоставимо да је циљ истраживања да се види какво је просечно знање математике ученика средњих школа. Природно је онда дефинисати случајну величину која је број поена на матурском испиту из математике јер би тај број поена требало да осликава знање ученика, а да је оно што нас занима заправо математичко очекивање те променљиве.*

Нерепрезентативни узорак би свакако био узорак који садржи претежно ђаке Математичке гимназије.

У преходним примерима се издвајала случајна величина дефинисана на популацији. Њу називамо *обележјем.* Најчешће нам је циљ да на основу неког узорка закључимо о неком конкретном параметру те популације (односно параметру расподеле обележја). Тај параметар оцењујемо неком функцијом од чланова узорка. Та функција се назива *статистика.* Дакле, статистиком називамо сваку функцију од чланова узорка која не зависи од непознатих параметара. У претходним примерима би параметар који се оцењује била очекивана вредност посматраног обележја.

Каква ће бити даља анализа узорка највише зависи од типа обележја. Разликујемо:

- категоричко (квалитативно) обележје:
  - номинално: крвна група, пол, сексуално опредељење, верска припадност итд.;
  - ординално: платни разред, степен стручне спреме, интензитет бола, статус студената (буџетски, самофинансирајући) итд.;
- нумеричко (квантитативно) обележје:
  - дискретно: број деце, оцена на испиту, број искоришћених дана одмора, број позива хитној помоћи у току ноћи итд.;
  - непрекидно: тежина, висина, време чекања у реду у банци, годишњи приходи, време преживљавања после операције итд.

У примеру 1.0.1 имали смо категоричко обележје (променљиву). Иако је скуп вредности био $\{0, 1\}$ ради се само о кодирању. Полазно обележје (карактеристика популације) коју смо посматрали је имало две категорије. Такође, ради се о номиналном обележју јер су категорије равноправне. Између њих не постоји поредак. Пример 1.0.2 илуструје нумеричко обележје.

Познавање типова података је кључно за добру организацију базе података која се користи у анализи! Поред тога, многи статистички пакети захтевају да се дефинише класа обележја са којим се ради.

Основни кораци у статистичкој анализи су:

1. осмишљавање експеримента;
2. узорковање и прикупљање података;
3. прелиминарна анализа:
   - одређивање типова података;
   - графички приказ података;
   - дескриптивне статистике;
4. идентификација аутлајера;
5. конструкција статистичког модела:
   - закључивање о вредностима непознатих параметара;
   - тестирање статистичких хипотеза;
   - испитивање квалитета модела;
6. прогноза.

Циљ овог уџбника је управо да се науче методолошке основе сваког корака понаособ.

1.1 Осмишљавање експеримента и узорковање

Пре прикупљања података добро је бити упознат са циљем истраживања и статистичким методама које ће се даље користити. Самим тим је препоручљиво да статистичар буде укључен у све фазе статистичке анализе, а не тек након што се прикупе подаци. Тиме се штити од потенцијалног проблема да се на податке не могу применити неке методе. Поред тога, прикупљање непотребних податка ће бити избегнуто.

Случајни узорак је узорак у коме сваки од чланова популације има могућност да се нађе у узорку. Ако су сви узорци истог обима једнако вероватни, узорак називамо *прост случајан узорак*. Две основне врсте узорковања су *узорак без враћања* и *узорак са враћањем*. У случају да имамо коначну популацију од N елемената вероватноћа да се извуче узорак обима n без враћања је $(\binom{N}{n}n!)^{-1}$, док је са враћањем та вероватноћа N^{-n}. Сваки од ових приступа има и мана и предности. На пример, ако је популација мала, у случају узорковања са враћањем велика је вероватноћа да ће се неки чланови популације поновити. С друге стране, уколико се узорковање врши са враћањем можемо сматрати да су чланови узорка независне и једнако расподељене случајне величине.

У даљем тексту ћемо претпостављати да је популација велика и да се ради о узорку са враћањем. Тада, ако је X посматрано обележје, са $X_1, X_2, ..., X_n$ ћемо означити прост случајан узорак (низ независних и једнако расподељених случајних величина). Јасно је да су то случајне величине јер ми унапред не знамо које ће бити вредности посматраног обележја на случајно изабраним члановима популације. Малим словима $x_1, ..., x_n$ ћемо означавати реализован узорак (регистроване вредности). Ову терминологију користићемо у даљем тексту. Другим методама узорковања се бави грана статистике *теорија узорка*. Више детаља се може наћи у [22]. Важно је напоменути да се у раду са узорцима добијеним другим методама подразумева добро познавање методологије за прост случајан узорак.

1.2 Прелиминарна анализа

Овај корак је веома важан за проналажење одговарајућег статистичког модела. Саставни делови прелиминарне анализе су графички приказ података и приказ сумарних статистика које нам могу дати више информација о нумеричким карактеристикама обележја. За приказ узорка у случају категоричког обележја можемо користити следеће:

- *табеларни приказ:* приказује се учесталост по категоријама;

- *тракасти дијаграм* [1]*:* фреквенције приказана у табели се приказују у виду трака (стубова) на графику;
- *кружни дијаграм*[2]: круг се дели на "исечке" који су пропорционални фреквенцијама приказаних у табели.

Следећим примером ћемо илустровати наведене начине приказа.

Пример 1.2.1. *Посматраћемо саобраћајне несреће које су се догодиле у једној држави у којој су аутономна возила пуштена у саобраћај у периоду од 2012. до 2016. године. Оно што нас посебно занима је да установимо шта то све утиче на исход несреће. Узет је узорак од 200 несрећа које су забележене у полицијским станицама. Обележја која ћемо посматрати су тип несреће који се десио, тип пута на коме се десила несрећа и да ли се несрећа десила у раскрсници . Ради се о категоричким променљивама.*

Ради прегледности прикупљени подаци су кодирани. Различити типови несрећа (typeC) *означени су бројевима од 1 до 8. Број 1 означава чеони судар, број 2 упоредну вожњу, број 3 означава сустизање, број 4 бочни судар, број 5 ударац у објекат, број 6 превртање, број 7 означава учешће пешака, а број 8 остале типове несрећа. Тип пута* (typeR) *је кодиран бројевима од 1 до 3 (једносмерни, двосмерни и физички раздвојен). Треће посматрано обележје* (crossR) *је кодирано бројевима 0 и 1 (не и да). Део прикупљене базе података изгледа овако:*

	typeC	typeR	crossR
1	*4*	*1*	*1*
2	*3*	*1*	*1*
3	*8*	*1*	*1*
4	*3*	*1*	*1*
5	*3*	*2*	*1*
6	*3*	*1*	*1*

Табеларни приказ фреквенција за свако од обележја приказан је у табелама 1.1.

[1] енгл. barplot
[2] енгл. pie chart

typeC	1	2	3	4	5	6	7	8
	7	34	79	34	5	1	29	11

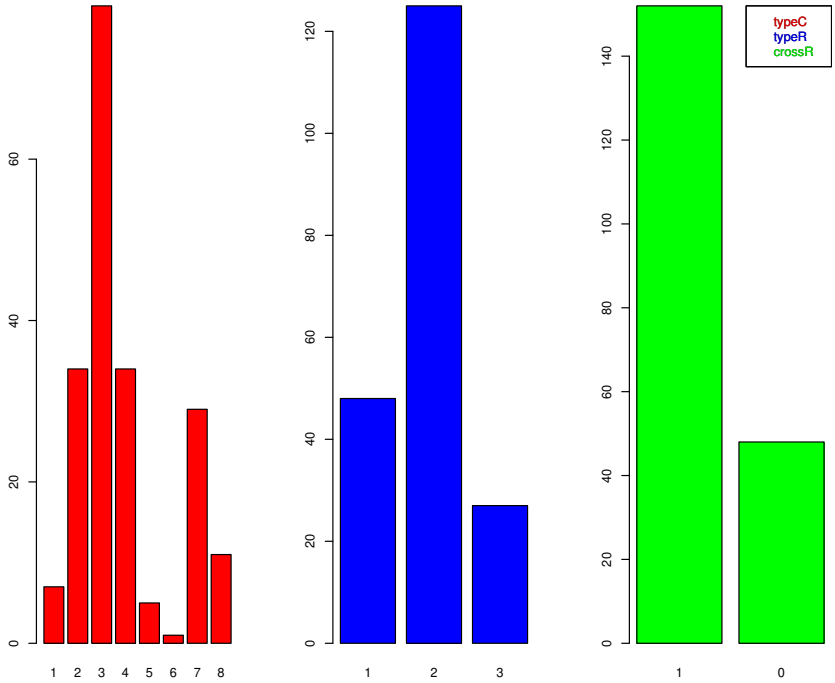
typeR	1	2	3
	48	125	27

crossR	0	1
	48	152

Табела 1.1: Табеларни приказ фреквенција обележја

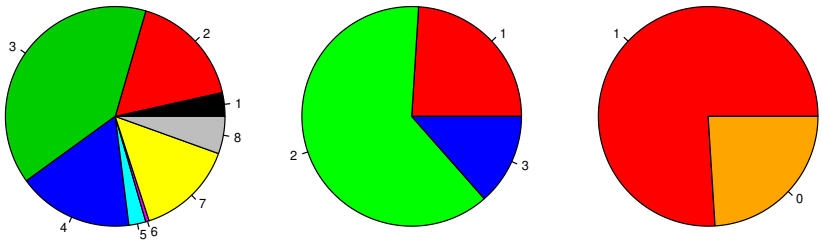
Иако се, када се обележја приказују одвојено, губи заједничка расподела, а самим тим и међусобни однос обележја, неки закључци се ипак могу донети. Видимо да је најчешћа несрећа "сустизање", што је вероватно резултат неправилног претицања. Најређа је превртање, што се може објаснити тиме да су возила довољно унапређена да до тога не дође. Можемо да приметимо да се већина несрећа догодила у раскрсницама што значи да има смисла улагати новац у побољшање сигнализације. Већи број несрећа се догодио на двосмерном путу, што је донекле и очекивано јер је повећана интеракција између возила, односно возача.
На сликама 1.1 и 1.2 су приказани тракасти дијаграми и кружни дијаграми за свако од обележја. Са тих дијаграма се може више закључити о томе како су обележја расподељена. На пример, не можемо моделирати тип несреће случајном величином која узима све вредности са једнаком вероватноћом. Још један од прелимираних закључака је да је вероватноћа да се деси несрећа у раскрсници 0.75. Оно што не можемо да закључимо са ових графичких приказа је да ли постоји нека веза између посматраних обележја. За то нам је потребно да посматрамо расподелу обележја "одједном", односно да видимо каква је њихова заједничка расподела што ћемо и учинити.

```
baza=read.csv("D:/bazaNesrece.txt", sep="")
baza$typeC=factor(baza$typeC)
baza$typeR=factor(baza$typeR)
baza$crossR=factor(baza$crossR,levels=c(1,0))
head(baza)
```



Слика 1.1: Тракасти дијаграми за типове несрећа, типове пута, и присуство раскрснице

```
table(baza$typeC)
table(baza$typeR)
table(baza$crossR)
par(mfrow=c(1,3))
barplot(table(baza$typeC),col='red')
barplot(table(baza$typeR),col='blue')
barplot(sort(table(baza$crossR),decreasing = T),col='green')
legend('topright',legend=c('typeC','typeR','crossR'),
text.col=c('red','blue','green'))
```



Слика 1.2: Кружни дијаграми

```
par(mfrow=c(1,3),mar=c(0,0,0,0))
pie(table(baza$typeC),col=1:8,main='typeC')
pie(table(baza$typeR),col=c('red','green','blue'),main='typeR')
pie(table(baza$crossR),col=c('red','orange'),main='crossR')
```

Заједничке расподеле (у паровима) посматраних обележја, приказане су у табелама 1.2. Ради се о табеларном приказу дводимензионалног обележја.

Питање: Шта можете да кажете о вероватноћи да се реализује прва врста несреће на физички одвојеном путу?

```
table(baza$typeC,baza$typeR)
table(baza$typeR,baza$crossR)
table(baza$crossR,baza$typeC)
```

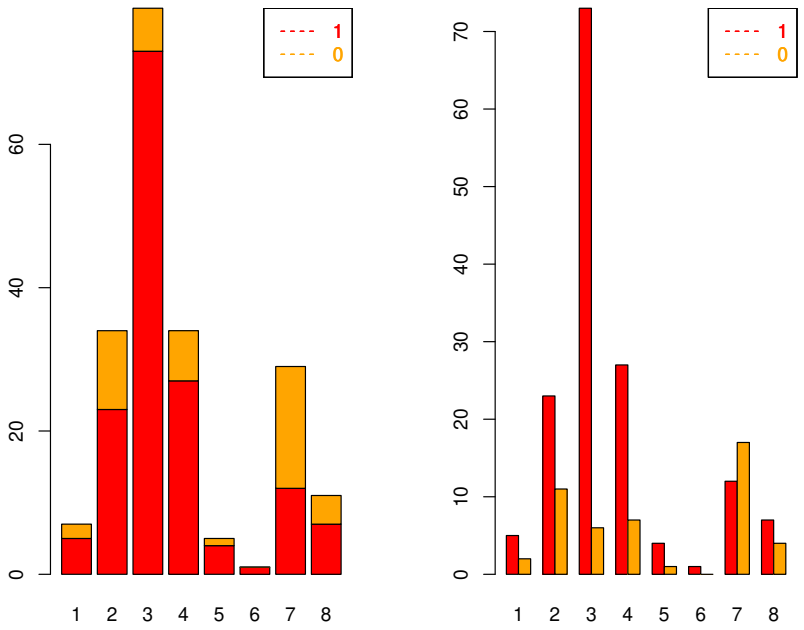
	typeR		
typeC	1	2	3
1	1	6	0
2	5	24	5
3	27	38	14
4	8	21	5
5	0	5	0
6	1	0	0
7	5	23	1
8	1	8	2

	crossR	
typeR	1	0
1	44	4
2	85	40
3	23	4

	typeC							
crossR	1	2	3	4	5	6	7	8
1	5	23	73	27	4	1	12	7
0	2	11	6	7	1	0	17	4

Табела 1.2: Табеларни приказ фреквенција парова обележја

Тракасти дијаграми су веома погодни за приказ вишедимензионалног обележја. Приказаћемо како изгледа расподела типова несрећа на раскрсницама и изван њих.



Слика 1.3: Тракасти дијаграми типова несрећа на раскрсницама и ван њих (лево-наслагани дијаграм, десно-груписани дијаграм)

```
par(mfrow=c(1,2))
barplot(table(baza$crossR,baza$typeC),col=c('red','orange'))
legend('topright',text.col=c('red','orange'),legend=c('1','0'),
lty=2,col=c('red','orange'))
barplot(table(baza$crossR,baza$typeC),col=c('red','orange'),
beside=TRUE)
```

За приказ нумеричог обележја најчешће се користе *хистограми*, док је стандард да се у оквиру прелиминарне анализе одреде и *мере централне тенденције* и *мере расејања*.

За конструкцију хистограма потребно је да узорак групишемо у категорије (интервале), тако да сваки елемент узорка припада тачно једној категорији и одредимо број елемената из узорка који се налази у свакој од категорија. О броју и положају категорија одлучујемо ми као статистичари. Препорука је да има барем 5 категорија и да је број категорија $[\log_2 n]+1$, где је n обим узорка. Како би се избегло да се подаци налазе на граници између категорија и да тако долазимо у ситуацију да нисмо сигурни којој категорији елемент узорка треба да припадне, за леву границу првог интервала не треба узети минималну вредност узорка већ мало мању вредност. Поред тога, пожељно је да величине категорија буду на једну децималу више него што су дати подаци. Други начин да се реши поменути проблем је да величина категорија буде на исти број децимала као подаци, али да границе буду задате на једну децималу више.

Величину категорије одређујемо на основу *распона* узорка $R = X_{(n)} - X_{(1)}$. Када знамо које су реализоване вредности статистика поретка говоримо о реализованом распону узорка. Са $X_{(k)}$ смо означили k-ти по реду елемент сортираног узорка. Случајна величина $X_{(k)}$ се назива *статистика поретка*, а низ $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ се назива *варијациони низ*. Уколико имамо k категорија, њихова приближна величина је $\frac{R}{k}$. Категорије не морају да буду једнаке величине али је то најчешће случај.

Нека је n_i број елемената из узорка који се налазе у i-тој категорији. *Хистограм* представља графички приказ учесталости елемената узорка по категоријама. Уколико су на y-оси фреквенције, односно ако је за сваки од интервала приказан правоугаоник висине n_i, говоримо о *хистограму апсолутних фреквенција*, уколико је приказан тај број подељен са величином узорка, односно ако је висина i-тог правоугаоника $\frac{n_i}{n}$, говоримо о *хистограму релативних фреквенција*, док ако је висина i-тог правоугаоника $\frac{n_i}{nd_i}$ где је d_i дужина i-тог интервала, говоримо о *хистограму густине*. Потенцијални недостаци хистограма су то да његов изглед зависи од

положаја и броја категорија, као и да функција коју цртамо је део-по-део константна а не неприкидна. Ти недостаци се превазилазе коришћењем неких софистициранијих метода оцењивања густине, које због своје комплексности, превазилазе оквире овог уџбеника. Више о томе се може прочитати у [27].

Приметимо да хистограм густине заправо представља оцену густине обележја X за које претпостављамо да има апсолутно непрекидну расподелу. Са хистограма густине можемо да видимо које од познатих расподела долазе у обзир за моделирање посматраног обележја, као и о неким особинама које расподеле "кандидати" треба да поседују.

Пример 1.2.2. *Посматрајмо плате у јануару 195 случајно изабраних просветних радника у Србији. Подаци су приказани у табели 7.4. Представићемо их хистограмом густине.*

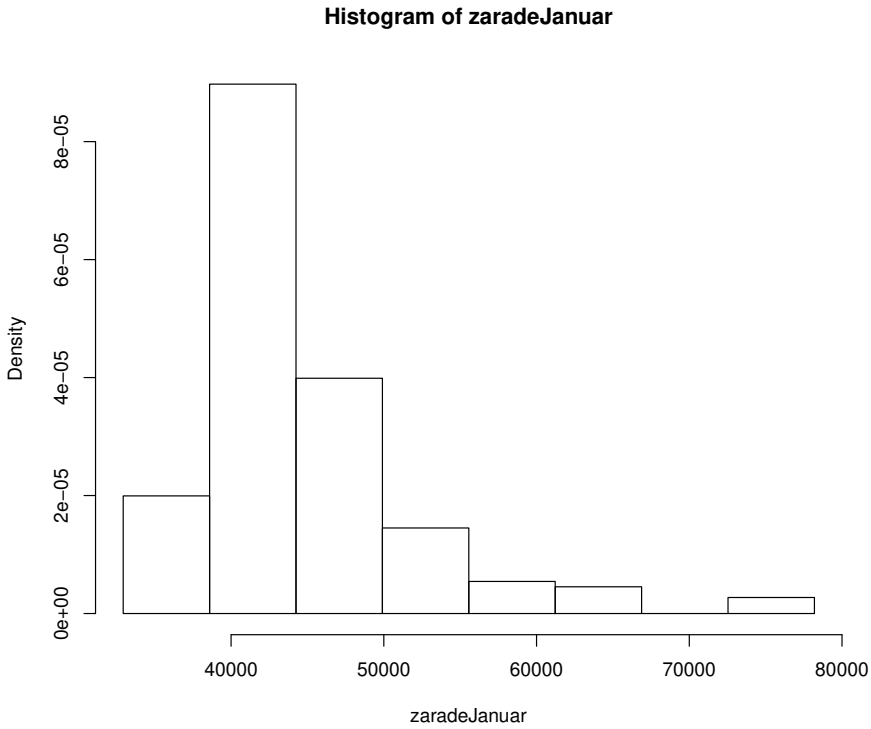
Минимална плата у узорку је 32 936 динара а максимална 78 193 динара. Дакле, узорачки распон је $R = 45257$. *Број категорија које ћемо користи за прављење хистограма је* $k = [\log_2 195] + 1 = 8$. *Величина категорије треба да буде приближно* 5657.125. *Тако добијамо следећи табеларни приказ:*

[32935. 9,38593.1] *22*	*(, 44250.3]* *99*	*(, 49907.5]* *44*	*(, 55564.7]* *16*
(, 61221.9] *6*	*(,66879.1]* *5*	*(,72536.3]* *0*	*(,78193.5]* *3*

На основу табеларног приказа конструишемо хистограм.

На слици 1.4 је приказан хистограм густине. Са њега можемо много тога да закључимо о расподели посматраног обележја. Остала два типа хистограма ће изгледати исто до на вредности на y-оси (јер су величине категорија једнаке).

Питање: Да ли је расподела симетрична? Које од расподела које знате долазе у обзир за моделирање?



Слика 1.4: Хистограм густине месечних зарада у просвети

```
summary(zaradeJanuar)
range(zaradeJanuar)
```

```
hist(zaradeJanuar,breaks=32935.9+(0:8)*5657.2,plot=FALSE)
hist(zaradeJanuar,breaks=32935.9+(0:8)*5657.2,prob=TRUE)
```

За расподелу која има дугачак реп на десној страни кажемо да је *померена удесно*. Ако је дугачак реп на левој страни кажемо да је *померена улево*. У примеру 1.2.2 са хистограма 1.4 видимо да је расподела зарада померена удесно.

О померености расподеле и другим особинама обележја X можемо сазнати и из оцена такозваних параметара централне тенденције:

- очекивана вредност;
- медијана;
- мода.

Природна оцена за очекивану вредност (средњу вредност, математичко очекивање) је средња вредност елемената узорка која се назива *узорачка средина* $\bar{X}_n = \frac{X_1+\cdots+X_n}{n}$, односно њена реализована вредност $\bar{x}_n = \frac{x_1+\cdots+x_n}{n}$. Од сада нећемо наглашавати да се ради о реализованој вредности неке статистике већ ће то бити јасно из контекста. Најчешће ћемо реализоване вредности означавати малим словима.

Медијана расподеле је онај параметар m_e за који је истовремено испуњено

$$P\{X \leq m_e\} \geq 0.5 \text{ и } P\{X \geq m_e\} \geq 0.5,$$

дакле нека "тачка која је у средини". Можемо је доживети и као дубину расподеле (најдубља тачка кад се гледа са обе стране). Оцена за медијану расподеле је *узорачка медијана* дефинисана са

$$\widetilde{X}_{med} = \begin{cases} X_{(k+1)}, & n = 2k+1 \\ \frac{X_{(k)}+X_{(k+1)}}{2}, & n = 2k. \end{cases}$$

Мода расподеле је она вредност у којој функција густине (или закон расподеле) достиже максимум. Уколико функција има више локалних максимума онда кажемо да расподела има више мода. *Узорачка мода* је она вредност која се најчешће појављује у узорку. Уколико је посматрано обележје апсолутно непрекидна случајна величина онда узорачка мода није добра оцена моде расподеле јер су сви елементи узорка различити. Неке оцене моде расподеле се могу наћи у, на пример, [4].

Питање: Које од расподела које знате имају јединствену моду?

Уколико је расподела симетрична, медијана и очекивана вредност се поклапају. Ако је расподела *унимодална*, односно има једну

моду, онда се, у случају симетричних расподела, мода поклапа са претходно наведеним.

Питање: Које су предности и мане наведених оцена параметара централне тенденције?

Поред наведених параметара расподеле важне су и такозване мере расејања:

- распон расподеле;
- стандардно одступање расподеле;
- интерквартилно (међуквартилно) растојање.

На основу узорка *распон расподеле* (носач расподеле, скуп тачака у којима је функција густине или закон расподеле различит од 0) оцењујемо *узорачким распоном* $R = X_{(n)} - X_{(1)}$, и већ из саме дефиниције видимо да то није нарочита мера расејања расподеле јер познавањем истог не знамо много више о самом типу расподеле, али свакако да нам податак о истом може помоћи приликом избора статистичког модела.

Стандардно одступање расподеле $\sigma = \sqrt{E(X - EX)^2}$ нам даје информацију колико случајна величина одступа од свог очекивања. Треба имати у виду да за неке расподеле оно не постоји. Природна оцена за σ^2 је *узорачка дисперзија*

$$\bar{S}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2, \tag{1.2.1}$$

па се σ оцењује са $\bar{S}_n$. Из разлога који ћемо убрзо видети, уместо (1.2.1) се користи *поправљена узорачка дисперзија*[3] дата са

$$\tilde{S}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2. \tag{1.2.2}$$

Питање: Коју знате расподелу са дисперзијом која није коначна?

[3]Већина статистичких софтверских пакета има имплементирану ову оцену дисперзије.

Подсетимо се да је α квантил расподеле она вредност x за коју је $F(x) = P\{X \leq x\} = \alpha$. У случају расподела које нису апсолутно непрекидне ова дефиниција се може уопштити. На основу узорка, α квантил се може оценити тако што се направи варијациони низ и онда одреди вредност c_α таква да је $\alpha \times 100\%$ чланова узорка мање (или једнако) од c_α. На тај начин се одређују, први узорачки квартил (0.25 квантил) Q_1 и трећи узорачки квартил (0.75 квантил) Q_3 који имају важну улогу приликом дефинисања још једне узорачке мере расејања (са q_1 и q_3 ћемо означити њихове реализоване вредности). Међутим, претходно описан метод не даје нам прецизан алгоритам оцењивања квантила (јер тачка c_α за коју важи да је приближно $\alpha \times 100\%$ чланова узорка мања од c_α није јединствена).

Један начин за одређивање Q_1 и Q_3 је следећи: одредимо M_e узорачку медијану полазног низа, а затим медијане поднизова које прави узорачка медијана у низу, односно, уколико је $n = 2k + 1$ онда је Q_1 медијана низа $X_{(1)}, ..., X_{(k)}$ а q_3 узорачка медијана низа $X_{(k+2)}, ..., X_{(2k+1)}$, уколико је $n = 2k$ онда је Q_1 медијана низа $X_{(1)}, .., X_{(k)}$ а Q_3 медијана низа $X_{(k+1)}, ..., X_{(2k)}$.

Сада је природна оцена за *интерквартилно растојање* (међуквартилно растојање) $IQR = Q_3 - Q_1$. Са iqr ћемо означити реализовану вредност статистике IQR.

Пример 1.2.3. *У случају посматраних зарада из претходног примера добија се*

$$\bar{x}_{195} = 44687.38 \quad \widetilde{x}_{med} = 43056$$
$$\tilde{s}_{195} = 7099.882 \quad iqr = 6112.$$

```
mean(zaradeJanuar)
median(zaradeJanuar)
sd(zaradeJanuar)
IQR(zaradeJanuar)
```

Питање: На основу претходне анализе података шта закључујете о расподели зарада, да ли је симетрична?

1.3 Идентификација аутлајера

Аутлајер, односно изузетак, је појам који се не може строго дефинисати. Најприближнији опис био би да је то члан узорка који се не уклапа у постојећи статистички модел, односно одступа од осталих чланова узорка. Те тачке су јако важне и не смемо их априори избацивати. Треба испитати да ли оне представљају неку грешку и какав је њихов утицај на модел. Уколико се ради о грешци онда би тај елемент узорка требало избацити. У даљем тексту ћемо приказати једноставан метод за идентификацију аутлајера. Још неке методе биће споменуте у наредним поглављима.

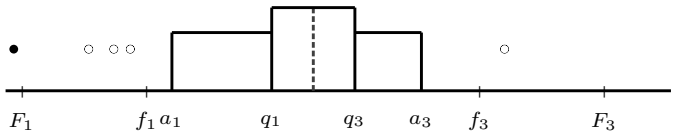
Један начин да се представе подаци је такозвани кутијасти дијаграм.[4] На слици 1.5 је приказан један овакав дијаграм. Ознаке на графику су следеће:

- q_1, q_3 су први и трећи (реализовани) узорачки квартили;
- $f_1 = q_1 - 1.5iqr, \quad f_3 = q_3 + 1.5iqr$;
- $F_1 = q_1 - 3iqr, \quad F_3 = q_3 + 3iqr$;
- a_1 најмањи елемент узорка који је већи од f_1, a_3 је највећи елемент узорка који је мањи од f_3;
- испрекиданом линијом је означена узорачка медијана.

Елементи узорка који су између f_1 и F_1, односно f_3 и F_3 су *благи аутлајери* док они изван ових граница, који нису у "кутији" су *прави аутлајери.*

Пример 1.3.1. *Кутијасти дијаграм за зараде је приказан на слици 1.6.* $F_3 = 64853.5$, $f_3 = 55685.5$. *Дакле, аутлајери јасно постоје а на нама је да одлучимо да ли ћемо их задржати или не у даљој анализи. Поред тога, не уочава се јасно одступање од симетричне расподеле. До истог закључка можемо доћи и поређењем узорачке средине и медијане. Уколико се значајно разликују, претпоставка да је расподела обележја симетрична је неоснована (што у овом примеру није случај). Међутим, на наш*

[4] енгл. box plot



Слика 1.5: Кутијасти дијаграм

закључак, уколико га доносимо поређењем узорачке медијане и узорачке средине, може утицати и присуство аутлајера (посебно на вредност узорачке средине).

```
boxplot(zaradeJanuar,horizontal = TRUE,col='red')
```

Питање: Како се промени узорачка средина и узорачка медијана када се из узорка избаце аутлајери?

Кутијасти дијаграми, као што смо видели у примеру 1.3.1, осим за идентификацију аутлајера, могу послужити да се установи да ли је расподела обележја симетрична или не. Наиме, уколико је расподела симетрична медијана ће бити приближно на средини кутије и све ознаке ће бити симетричне у односу на њу. Поред тога кутијасти дијаграми су веома корисни за визуелно поређење обележја две или више популација. Илустроваћемо то наредним примером.

Пример 1.3.2. *Резултати теста из математике по 20 случајно одабраних ученика из две основне школе су приказани у наредној табели.*

I	79	85	70	71	97	76	98	91	85	75
	77	82	70	82	74	85	83	94	79	78
II	100	47	47	71	74	69	66	72	49	73
	69	72	77	70	66	47	69	60	73	100