СТАТИСТИКА

Садржај

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 самном и да овладате са што већим бројем познатих модела, и 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 током курса служити. Зато на почетку наводимо основне појме.

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" и "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 одгвор моделирамо случајном величином X чије су вредности 0 (за "не") и 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 мишљење јавног мнјења, 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 учиника,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 обележја). Тај параметар оцењујемо неком функцијом од чланова узорка. Та функција се назива *статистика*. У претходним примерима би тај параметар била на пример средња вредност посматраног обележја.

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

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

У примеру 1.0.1 имали смо категоричко обележје (променљиву). Иако је скуп вредности био $\{0, 1\}$ ради се само о кодирању, али полазна карактеристика коју смо посматрали је имала две категорије. Такође, ради се о номиналној променљивој 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;
   - дескриптивна статистика;
4. идентификација аутлајера
5. закључивање о вредностима непознатих параметара;
6. тестирање статистичких хипотеза;
7. прогноза.

Циљ овог удб2ника је управо да се науче методолошке основе сваког корака.

1.1 Осмишљавање експеримента и узорковање

Пре прикупљања података добро је бити упознат са циљем истраживања и статистичким методама које ће се даље користити, него укључити статистичара тек након што се прикупе подаци. Тако се штитимо од потенцијалне проблема да се на податке не могу применити неке методе. Поред тога, нећемо прикупљати податке који нам нису потребни.

Случајни узорак је узорак у коме сваки од чланова популација има могућност да се нађе у узорку. Ако су сви узорци истог обима једнако вероватни, узорк 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 $\frac{1}{\binom{N}{n}}$, док је са враћањем $\frac{1}{N^n}$. 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 овако 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. Другим методама узорковања се бави грана статистике *теорија узорка*. 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. Више детаља се може наћи у [21]. Важно је напоменути да дад 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 математичког модела. Графички приказ у многоме помаже. За приказ узорка у случају категоричког обележја можемо користити следеће:

- *табеларни приказ:* приказује се учесталост по категоријама;
- *тракасти дијаграм* [1]*:* фреквенција приказана у табели се приказује у виду трака на графику;
- *кружни дијаграм*[2]: круг се дели на "исечке" који су пропорционални фреквенцијама приказаних у табели.

Следећим примером ћемо илустровати наведене начине приказа.

[1] енг. barplot
[2] енг. пие чарт

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 које 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. На пример, број 1 означава чеони судар, број 2 упоредну вожњу и тако даље. Тип пута (typeR) је кодиран бројевима од 1 до 3 (једносмерни, двосмерни, и физички раздвојен), да ли се несрећа догодила у раскрсници (crossR) је кодиран 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*

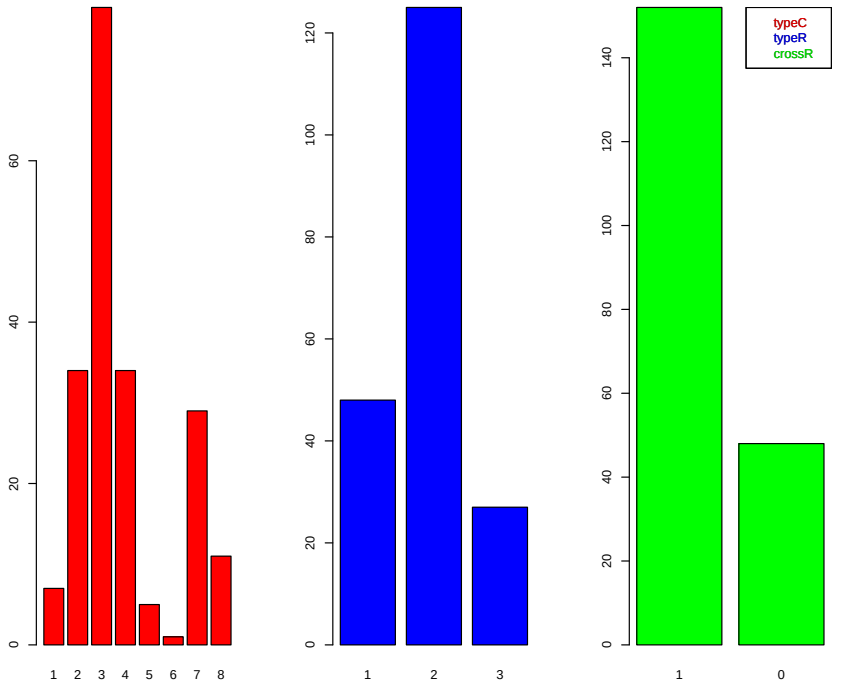
Табеларни приказ фреквенција *за свако од обележја изгелда овако:*

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

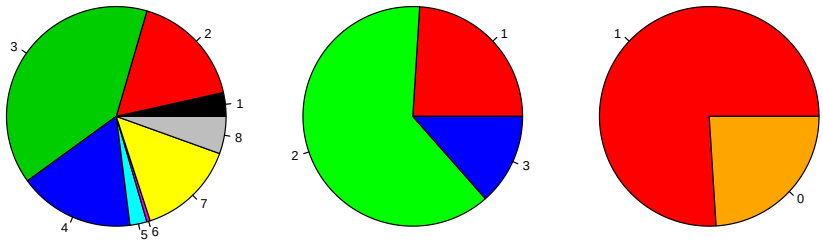
typeR	*1*	*2*	*3*
	48	*125*	*27*

crossR	*0*	*1*
	48	*152*

Иако, када се обележја приказују одвојено, се 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.



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



Слика 1.2: Кружни дијаграми

На сликама 1.1 и 1.2 су приказани тракасти дијаграми *и* кружни дијаграми *за свако од обележја. Са њих 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 њихива заједничка расподела.*

Заједничке расподеле (у паровима) посматраних обележја,

приказане су у следећим табелама (ради се о табеларном приказу *дводимензионог обележја):*

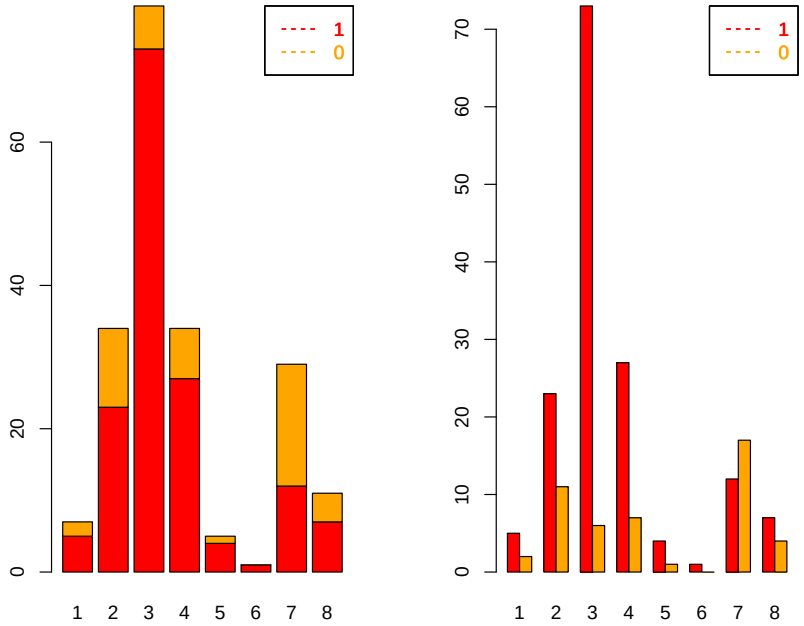
typeC \ *typeR*	*1*	*2*	*3*
1	*1*	*6*	*0*
2	*5*	*24*	*5*
3	*27*	*38*	*14*
4	*8*	*21*	*5*
5	*0*	*5*	*0*
6	*1*	*0*	*0*
7	*5*	*23*	*1*
8	*1*	*8*	*2*

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

crossR \ *typeC*	*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*

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

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



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

```
bazaCas$typeC=factor(bazaCas$typeC)
bazaCas$typeR=factor(bazaCas$typeR)
bazaCas$crossR=factor(bazaCas$crossR)

head(bazaCas)
summary(bazaCas)
table(bazaCas$typeC)
table(bazaCas$typeR)
table(bazaCas$crossR)
```

```
par=c(mfrow=c(1,3))
barplot(table(bazaCas$typeC),col='red')
barplot(table(bazaCas$typeR),col='blue')
barplot(table(bazaCas$crossR),col='green')
legend('topright',legend=c("typeC","typeR","crossR"),
text.col=c('red','blue','green'))
```

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

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

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

Што се тиче нумеричог обележја најчешће се за графички приказ 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.

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 $X_{(1)} \leq X_{(2)} \leq \cdots X_{(n)}$ 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 иза узорка који се налазе у i-тој категорији. *Хистограм* управо 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 $\frac{d_i}{n}$ где је d_i дужина i-тог интервала, говоримо о *хистограму густине*.

Питање: Зашто се последње поменути хистограм назива баш хистограм густине?

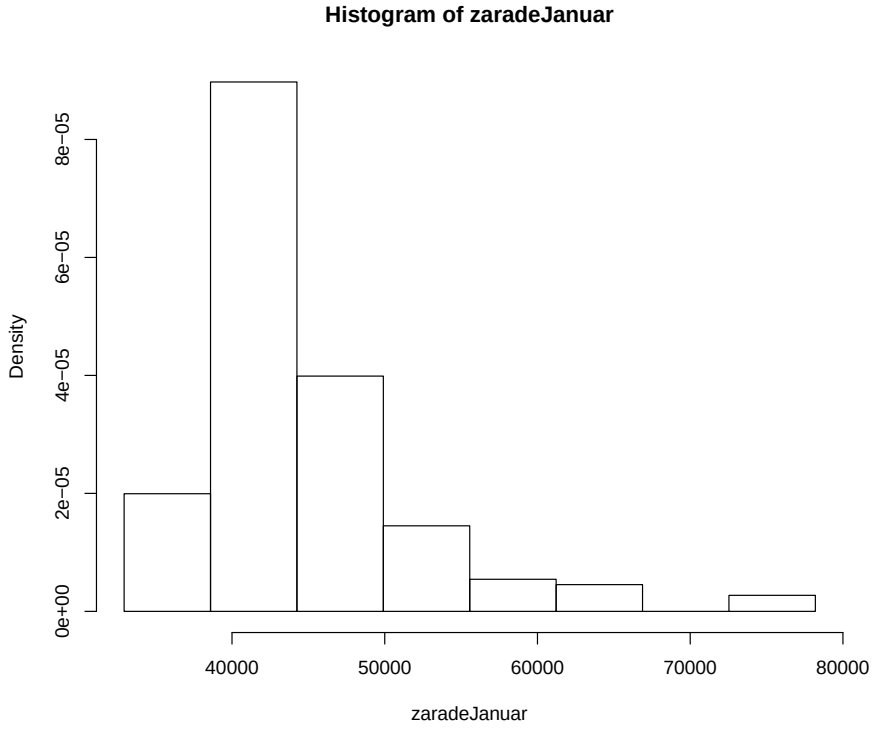
Приметимо да хистограм заправо представља оцену густине обележја X за које претпостављамо да има апослутну непрекидну расподелу. Са хистограма густине можемо да видимо које од познатих расподела долазе у обзир за моделирање посматраног обележја.

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

Минимална плата у узорку 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.

[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.



Слика 1.4: Хистограм густине месечних зарада у просвети

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.

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

```
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 из 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* и означава са $\bar{X}_n = \frac{X_1+\cdots+X_n}{n}$, односно њена реализована вредност $\bar{x} = \frac{x_1+\cdots+x_n}{n}$. Од сада нећемо нагалшавати да се ради о реализованој вредности неке статистике [3].
Медијана расподеле је онај параметар μ за који је истовремено испуњено

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

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

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

[3] свака функција од узорка која не зависи од непознатих параметара се назива се статистика

Мода расподеле је она вредност у којој функција густине (или закон расподеле) достиже максимум. *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 једниственву моду?
Уколико је расподела симетрична медијана и очекивана вредност се поклапају. Ако је расподела *унимодална*[4] онда се и она поклапа са претходно наведеним, у случају симетричних расподела.
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* оцењујемо *узорачким распоном*, и већ из саме дефиниције видимо да то није нарочита мера расејања расподеле јер познавањем истог не знамо много више о самом типу расподеле.
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*

$$\bar{S}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2, \tag{1.1}$$

па се σ оцењује са $\bar{S}_n$. Из разлога који ћемо убрзо видети, умести

[4] има једну моду

(1.1) се користи *поправљена узорачка дисперзија*[5] дата са

$$\tilde{S}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2. \tag{1.2}$$

Питање: Коју знате расподелу са дисперзијом која није коначна?

На основу узорка одређујемо 0.25 квантил расподеле са q_1 и 0.75 квантил расподеле са q_3 расподеле тако да приближно 25% чланова узорка је мање од q_1, односно веће од q_3. Један начин за то 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 $X_{(1)}, ..., X_{(k+1)}$ а q_3 узорачка медијана низа $X_{(k+1)}, ..., X_{(2k+1)}$, 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* $IQR = q_3 - q_1$.

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

$$\bar{x} = 44687.38 \quad m_e = 43056$$
$$\tilde{s} = 7099.882 \quad IQR = 6112.$$

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

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

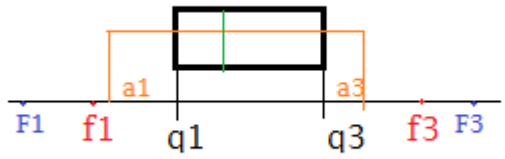
[5] Већина статистичких софтверских пакета има имплементирану ову оцену дисперзије.

1.3 Идентификација аутлајера

Аутлајер 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. 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.

Један начин да се представе подаци је такозвани кутијасти дијаграм.[6] На слици 1.5 је приказан један овакав дијаграм. Ознаке на графику су следеће:

- q_1, q_3 су први и трећи узорачки квартил;
- $f_1 = q_1 - 1.5IQR, \quad f_3 = q_3 + 1.5IQR$;
- $F_1 = q_1 - 3IQR, \quad F_3 = q_3 + 3IQR$;
- a_1 најмањи елемент узорка који је већи од f_1, a_3 је највећи елемент узорка који је мањи од f_3;
- зеленом бојом је означена узорачка медијана;

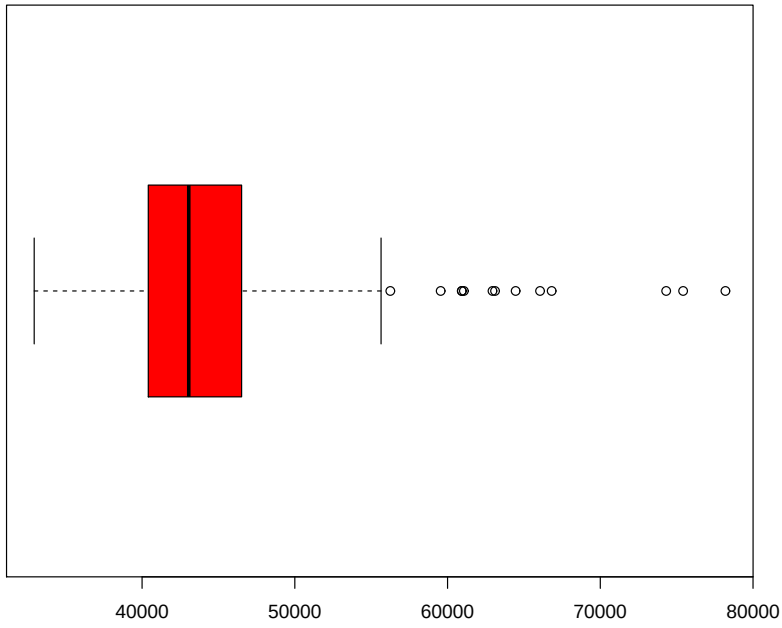


Слика 1.5: Бокс плот дијаграм

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*.

[6] енг. box plot

Пример 1.3.1. *Кутијасти дијаграм за зараде је приказан на слици 1.6.* $F_3 = 64853.5,\ f_3 = 55685.5.$



Слика 1.6: Бокс плот дијаграм зарада

Дакле, аутлајери јасно постоје а на нама је да оцлучимо да ли ћемо их задржати или не у даљој анализи.

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

Питање: *Како се промени узорачка средина и узорачка медијана када се из узорка избаце аутлајери?*
Питање: *Да ли у овом примеру има смисла избацивати аутлајере?*

Кутијасти дијаграми, осим за идентификацију аутлајера, могу послужити да се установи да ли 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.

У случају зарада из претходног примера, јасно се уочава 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.

1.4 Особине узорачке средине и узорачке дисперзије

Већ смо споменули да је природна оцена за EX, на основу п.с.у. узорка $X_1, ..., X_n$, баш узорачка средина $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Очекивана вредност те оцене је

$$E(\bar{X}_n) = \frac{1}{n}\sum_{i=1}^{n} EX_i = EX.$$

Последња једнакост важи јер претпостављамо да се ради о простом случајном узорку односно да X_i има исту расподелу као X, па и математичко очекивање. Дакле очекивана вредност оцене коју користимо је баш средња вредност популације. Колико је средње квадратно одступање те оцене од средње вредности видимо из

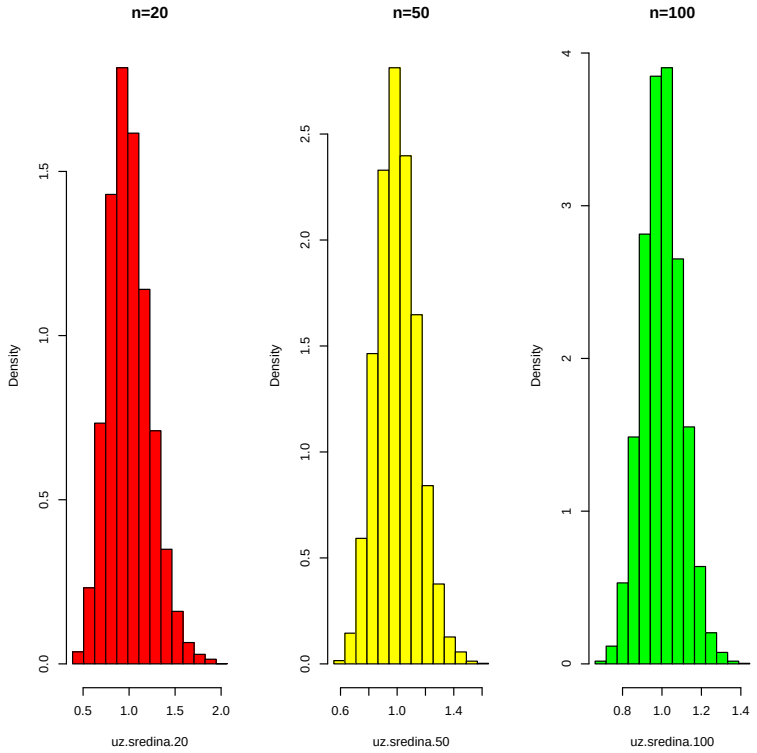
$$D(\bar{X}_n) = \frac{1}{n^2}\sum_{i=1}^{n} D(X_i) = \frac{D(X)}{n}.$$

Ако X има коначну дисперзију онда ће дисперзија оцене опадати како n расте, што је свакако једна од особина које желимо да оцена поседује.

Пример 1.4.1. *Претпоставићемо да X има експоненцијалну $\mathcal{E}$ расподелу. Тада је $EX = 1$ и $DX = 1$. Генерисаћемо "пуно" узорака ($N = 10000$) обима $n \in \{20, 50, 100\}$ из $\mathcal{E}(1)$ расподеле. На основу сваког од узорака оценићемо параметар средње вредности*

са узорачком средином. Хистограми тих оцена, за различите обиме узорка, приказани су на слици 1.7.

Јасно се уочава да се са порастом обима узорка смањује одступање оцене од стварне вредности 1. Примећујемо да са порастом обима узорка расподела оцене све више подсећа на нормалну расподелу. Објашњење тога лежи у Централној граничној теореми (услови теореме су задовољени јер имамо низ независних и једнако расподељених случајних величина са коначном дисперзијом).



Слика 1.7: Хистограм оцена средње вредности

```
N=10000
uz.sredina.20=rep(0,N)
uz.sredina.50=rep(0,N)
uz.sredina.100=rep(0,N)
for(i in 1:N)
{
x20=rexp(20)
x50=rexp(50)
x100=rexp(100)
uz.sredina.20[i]=mean(x20)
uz.sredina.50[i]=mean(x50)
uz.sredina.100[i]=mean(x100)
}
hist(uz.sredina.20,breaks=0.386+0:14*0.12,col='red',prob=TRUE,
main='n=20')
hist(uz.sredina.50,breaks=0.553+0:14*0.078,col='yellow',
prob=TRUE,main='n=50')
hist(uz.sredina.100,breaks=0.661+0.056*0:14,col='green',
prob=TRUE,main='n=100')
```

Напомена: *У коду изнад нисмо иницијализовали "семе" генератора који се користи за генерисање случајних бројева. Ово је важно напоменути јер се при следећем извршавању програма хистограми могу незнатно променити.*

Овај пример је изузетно важан јер нам илуструје како можемо да особинама неких оцена закључујемо *емпиријски*, на основу података. Закључке које добијемо емпиријском студијом би требало, с математичке тачке гледишта, и формално теоријски показати. Емпиријски закључци нам заправо дају смерницу шта да покажемо, и обично упућују на то да ли су наше хипотезе тачне или не. С практичне тачке гледишта, често је довољно донети закључке на основу емпиријске студије.

Већ смо напоменули да се уместо узорачке дисперзије често користи поправљена узорачка дисперзија. Разлог томе је следећа једнакост:

$$E(n\bar{S}_n^2) = (n-1)DX. \tag{1.3}$$

Увођењем фактора корекције добијамо да је очекивана вредност поправљене оцене баш једнака дисперзији обележја, односно параметру који се оцењује. Ради једноставности извођења једнакости (1.3) увешћемо следеће ознаке. Нека је $EX_1 = m$ и $DX_1 = \sigma^2$. Тада је

$$EX_iX_j = \begin{cases} EX_i^2 = m^2 + \sigma^2, & \text{за } i = j \\ EX_i \cdot EX_j = m^2. & \text{за } u \neq j \end{cases}$$

Одавде следи да је

$$\begin{aligned} E\bar{X}_n^2 &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} EX_iX_j = \frac{1}{n^2}\sum_{i=1}^{n} EX_i^2 + \frac{2}{n^2}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} EX_iX_j \\ &= \frac{1}{n^2}\cdot nm^2 + \frac{2}{n^2}\cdot\frac{n(n-1)}{2}(m^2+\sigma^2) = m^2 + \frac{(n-1)}{n}\sigma^2 \end{aligned}$$

Даље важи следећи низ једнакости:

$$\begin{aligned} E(n\bar{S}_n^2) &= E(\sum_{i=1}^{n}(X_i - \bar{X}_n)^2) = E(\sum_{i=1}^{n} X_i^2 - 2\bar{X}_n\sum_{i=1}^{n} X_i + n\bar{X}_n^2) \\ &= E(\sum_{i=1}^{n} X_i^2 - n\bar{X}_n^2) = n(m^2 + n\sigma^2) - nm^2 - (n-1)\sigma^2 \\ &= (n-1)\sigma^2. \end{aligned}$$

Задатак 1.4.1. Приказати $D(n\bar{S}_n^2)$, преко EX_1, EX_1^2, EX^3 и EX_i^4 (необавезно свих). Направити емпиријску студију о особинама узорачке и поправљене узорачке дисперзије по угледу на пример 1.4.1 у којој ћете нацртати хистограме оцена $\bar{S}_n^2$ и $\widetilde{S}_n^2$ за различите обиме узорка под претпоставком да X има експоненцијалну $\mathcal{E}(1)$ расподелу. Шта можете да кажете о расподели оцена? Како се мења прецизност оцена са порастом обима узорка?

Задатак 1.4.2. Направити емпиријску студију о особинама узорачке медијане по угледу на пример 1.4.1 у којој ћете нацртати хистограме узорачке медијане и узорачке средине за различите обиме узорка под претпоставком да X има нормалну$\mathcal{N}(0,1)$ расподелу. Шта можете да кажете о расподели оцена? Која од ове две оцене боље представља централну тенденцију узорка?

1.5 Емпиријска функција расподеле

У претходном поглављу видели смо како можемо да оценимо функцију густине, односно закон расподеле случајне величине X. У овом поглављу ћемо приказати како можемо да оценимо њену функцију расподеле F.

Нека је $X_1, X_2, ..., X_n$ прост случајан узорак из популације на којој посматрамо обележје X. Како је, за $x \in \mathbb{R}$, $F(x) = P\{X \leq x\} = E(I\{X \leq x\})$, где је случајна величина $I(\cdot)$ индикатор догађаја (видети 6.2), природна оцена функције расподеле је

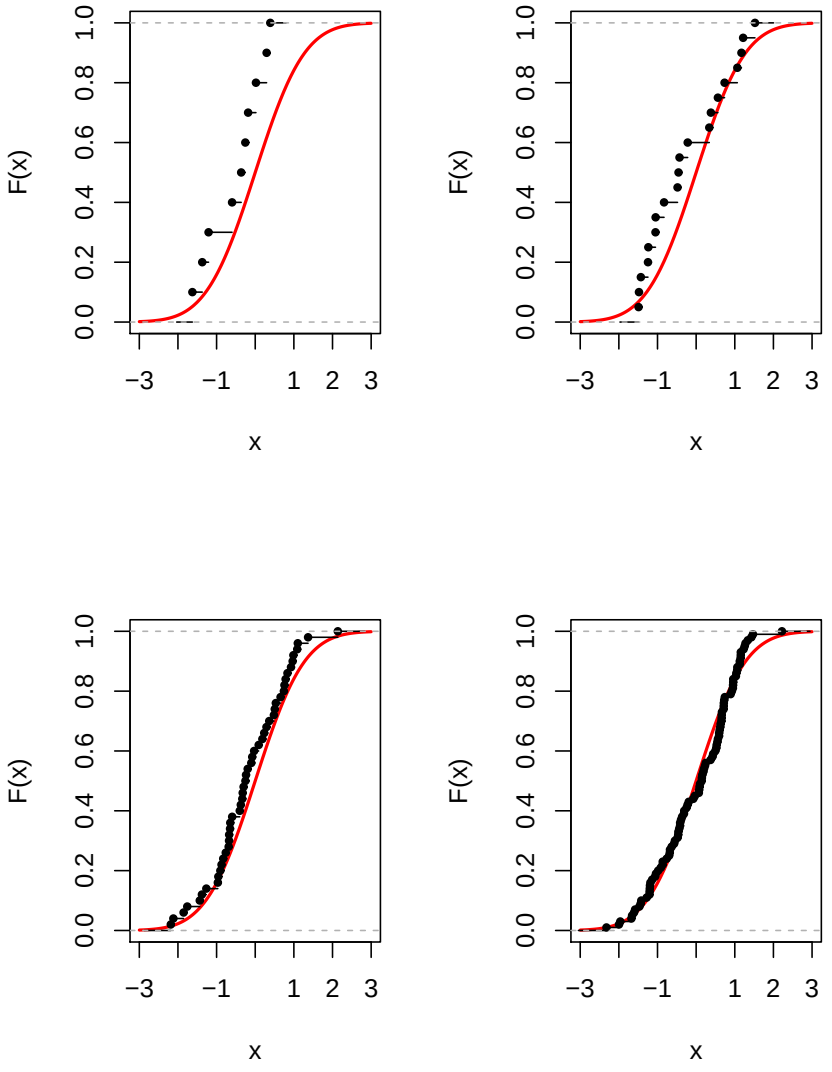
$$F_n(x) = \frac{\sum_{i=1}^{n} I\{X_i \leq x\}}{n}.$$

Приметимо да је $nF_n(x)$ сума n независних и једнако расподељених индикатора и као таква има Биномну $\mathcal{B}(n, F(x))$ расподелу па је $E(F_n(x)) = F(x)$ и $D(F_n(x)) = \frac{F(x)(1-F(x))}{n}$. Одавде видимо да како расте обим узорка n тако се и емпиријска функција расподеле F_n "приближава" правој функцији расподеле F. Ово запажање садржано је у следећој теореми која је позната још и као централна теорема статистике.

Теорема 1.5.1 (Гливенко-Кантелијева теорема). *Нека је $X_1, X_2, ..., X_n$ н.с.у. из популације са обележјем X са функцијом расподеле $F(x)$. Даље, нека је $F_n(x)$ одговарајућа емпиријска функција расподеле. Тада*

$$P\{\sup_x |F_n(x) - F(x)| \to 0, \text{ кад } n \to \infty\} = 1.$$

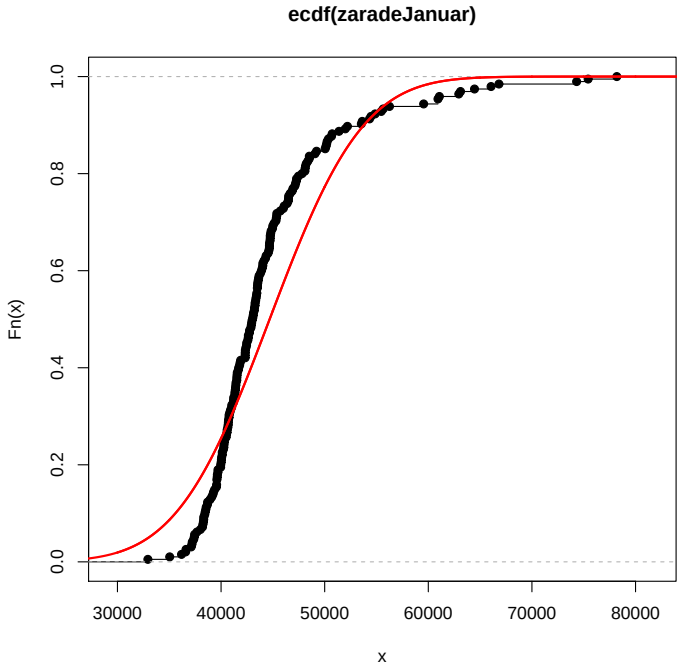
На графику 1.8 приказане су емпиријске функције расподеле узорака различитог обима из обележја са нормалном $\mathcal{N}(0, 1)$ расподелом. Црвеном линијом приказана је одговарајућа функција расподеле. Јасно се види да је већ за обим узорка $n = 50$ оцена функције расподеле довољно добра оцена стварне функције расподеле.



Слика 1.8: Емпиријска функција расподеле на основу узорка обима $n = 10$ (горе лево), $n = 20$ (горе десно), $n = 50$ (доле лево) и $n = 100$ (доле десно)

Емпиријска функција расподеле је доста погодна за "имитирање" расподеле посматраног обележја (уколико је потребно да узоркујемо из стварне расподеле F коју не знамо, можемо узорковати из F_n), али из њеног графичког приказа се не закључује о расподели лако као са хистограма, па се у ту сврху најчешће не користи. С друге стране уколико имамо већ неку претпоставку о расподели, лакше ћемо одбацити исту на основу емпиријске функције расподеле него на основу хистограма.

Пример 1.5.1. *Посматрајмо податке из примера 1.2.2. Желимо да видимо да ли има смисла претпоставити да је расподела нормална. Параметре расподеле оценићемо са узорачком средином и узорачком дисперзијом. На истом графику (види слику 1.9) приказаћемо емпиријску функцију расподеле и претпостављену функцију расподеле.*



Слика 1.9: Емпиријска функција расподеле месечних зарада у просвети

Са графика се види јасно види да нормална расподела није добар избор.

```
plot(ecdf(zaradeJanuar))
xniz=seq(from=25000,to=90000,by=5)
lines(xniz,pnorm(xniz,mean = mean(zaradeJanuar),
sd=sd(zaradeJanuar)),col='red',lwd=2)
```

Задатак 1.5.1. На основу реализованог простог случајног узорка $1, 0, 1, 1, 1, 0, 2, 3, 2$ одредити емпиријску функцију расподеле.

Задатак 1.5.2. Нека је $F_9(\cdot)$ емпиријска функција расподеле из претходног задатка и нека је Y дискретна случајна величина са функцијом расподеле $F_9(\cdot)$. Одредити закон расподеле случајне величине Y.

Задатак 1.5.3. Нека је $X_1, X_2, ..., X_n$ п.с.у. из популације на којој посматрамо апсолутно непрекидну случајну величину X са расподелом F. Показати да се за велико n, за свако $x \in \mathbf{R}$, случајна величина $\sqrt{n}(F_n(x) - F(x))$ се може апроксимирати нормалном расподелом. Одредити параметре те расподеле.

2

Оцењивање непознатих параметара расподела

До сада смо се упознали са неким статистичким оценама као што су хистограм, емпиријска функција расподеле, узорачка средина, узорачка дисперзија и друге. Тада нисмо водили рачуна из које је расподеле обележје X. Такав приступ припада домену *непараметарске статистике*. Сада претпостављамо да је статистички модел одређен до на непознате параметре расподеле. То значи да имамо претпоставку о расподели обележја и остаје нам само да оценимо параметре те расподеле. Овакав приступ припада домену *параметарске статистике*. И један и други приступ имају своје предности и мане и један од циљева овог уџбеника је да се са њима упознамо.

2.1 Тачкасте оцене параметара

Као што смо у претходном поглављу видели, један од основних статистичких задатака је да на основу доступног узорка апроксимирамо вредност неких параметара популације. У овом поглављу бавићемо се такозваним *тачкастим оценама*. Приказаћемо два основна метоода за њихово добијање и навести основна својства која квалитетана оцена треба да поседује. Поред тога, показаћемо и како, када на располагању имамо више могућности за оцену, да их међусобно упоредимо.

Нека је статистички модел одређен до на непознат параметар θ за који знамо да припада скупу Θ, при чему параметар θ може да буде и вишедимензионалан. Тај скуп називаћемо *скупом допустивих вредности за непознат параметар* θ. Најчешће ћемо претпостављати да обележје X има функцију расподеле $F(\cdot;\theta)$, или густину $f(\cdot;\theta)$, или закон расподеле $p(\cdot,\theta)$ где је θ непознат параметар. У даљем тексту, приказаћемо два основна метода за добијање тачкастих оцена параметра θ. Кад год није наглашено другачије, претпостављамо да на располагању имамо п.с.у. $X_1, X_2, ..., X_n$.

2.1.1 Метод момената

Пре него што се упознамо са овим методм увешћемо још неке појмове.

Дефиниција 2.1.1. *Математичко очекивање* EX^k, $k \in \mathbf{N}$, *се назива* k*-ти моменат расаподеле, а математичко очекивање* $E(X - EX)^k$ k*-ти центрирани моменат расаподеле.*

Дакле, EX је заправо први моменат, а k-ти моменат расподеле случајне величине X можемо посматрати и као први моменат случајне величине X^k. Зато је природна оцена за EX^k k-*т*и узорачки моменат деифинисан са $\frac{\sum_{i=1}^n X_i^k}{n}$. На исти начин долазимо и до k-*т*ог центрираног узорачког момената. Приметимо, да је други центрирани узорачи моменат заправо узорачка дисперзија. Оцене непознатих параметара се добијају као решење система једначина који се добије кад се изједначе теоријски моменти са одговарајућим узорачким моментима (в. табелу 2.1). То је илустровано наредним примерима.

Пример 2.1.1. *Нека* X *има нормалну* $\mathcal{N}(m, \sigma^2)$ *расподелу. Имамо два непозната параметра па су нам потребне две једначине. Најједноставније је да поставимо једначине у коме фигуришу прва два момента, и то, имајући у виду шта представљају непознати параметри за нормалну расподелу, очекивање и дисперзију. Тако добијамо систем*

$$m = EX = \bar{X}$$
$$\sigma^2 = DX = \bar{S}^2.$$

теор. м.	узор. м.	теор. цент. м.	узор. цент. м.
EX	$\bar{X}_n$	$--$	$--$
EX^2	$\frac{\sum_{i=1}^n X_i^2}{n}$	DX	$\bar{S}_n^2$
EX^3	$\frac{\sum_{i=1}^n X_i^3}{n}$	$E(X-EX)^3$	$\frac{\sum_{i=1}^n (X_i-\bar{X}_n)^3}{n}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
EX^k	$\frac{\sum_{i=1}^n X_i^k}{n}$	$E(X-EX)^k$	$\frac{\sum_{i=1}^n (X_i-\bar{X}_n)^k}{n}$

Табела 2.1: Теоријски и одговарајући узорачки моменти

Одавде тривијално следи да је $(\hat{m}_n, \hat{\sigma}_n^2) = (\bar{X}_n, \bar{S}_n^2)$.

Пример 2.1.2. *Нека* X *има експоненцијалну* $\mathcal{E}(\lambda)$ *расподелу, односно густин у расподеле* $f(x;\lambda) = \lambda e^{-\lambda x}, x > 0$, *где је* $\lambda > 0$ *непознат параметар. Сада нам је потребна само једна једначина па опет узимамо најједноставнију.*

$$\frac{1}{\lambda} = EX = \bar{X}_n.$$

Одавде добијамо да је $\hat{\lambda}_n = \frac{1}{\bar{X}}$.

Оцену смо могли да добијемо и из једначине

$$DX = \frac{1}{\lambda^2} = \bar{S}_n^2.$$

Одавде је $\hat{\lambda} = \frac{1}{S}$. *Наравно, ове две оцене неће бити идентичне, али не би требало много да се разликују уколико је наша претпоставка о расподели обележја исправна.*

Пример 2.1.3. *Нека* X *има* $\mathcal{U}[0,\theta]$ *расподелу где је* $\theta > 0$, *непознат параметар. Оцену добијамо из једначине*

$$EX = \frac{\theta}{2} = \bar{X}_n.$$

Одавде је $\hat{\theta}_n = 2\bar{X}_n$.

Пример 2.1.4. *Нека је* X *индикатор са вероватноћом успеха* p. *Параметар* $p \in (0,1)$ *је непознат. Сада је*

$$p = EX = \bar{X}_n,$$

одакле одмах видимо да је $\hat{p}_n = \bar{X}_n$.

Пример 2.1.5. *Нека X има Пуасонову $\mathcal{P}(\lambda)$ расподелу, где је $\lambda > 0$, непознат параметар. Сада је*

$$\lambda = EX = \bar{X}_n,$$

па је $\hat{\lambda}_n = \bar{X}_n$.

Пример 2.1.6. *У [16] описан је експеримент у коме је испитиван квалитет воде. Узето је 103 узорка и 58 није било контаминирано. У осталим узорцима се нашла бар по једна пијавица. Резултати су приказани у следећој табели.*

бр. пијавица	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*≥9*
бр. узорака	*58*	*25*	*13*	*2*	*2*	*1*	*1*	*0*	*1*	*0*

Уколико са X-означимо број пијавица у узорку, један од могућих статистичких модела је да X има Пуасонову $\mathcal{P}(\lambda)$ расподелу, али свакако не и једини. На основу претходног примера 2.1.5 добијамо да је $\hat{\lambda} = \bar{x}_{103} = 0.816$. Приметимо да пошто имамо реализован узорак $\hat{\lambda}$ није више случајна величина већ њена реализација.

Метод момената је заправо специјални случај такозваног *метода замене* код кога се систем једначина прави изједначавајући неке функције од узорка са њиховим узорачким "парњацима."

Пример 2.1.7. *Нека X има експоненцијалну $\mathcal{E}(\lambda)$. Непознат параметар λ можемо добити и изједначавајући медијану расподеле са узорачком медијаном.*
Медијану расподеле добијамо из

$$0.5 = 1 - F(\mu) = 1 - e^{-\lambda\mu}.$$

Одавде је $\mu = -\frac{\log(0.5)}{\lambda}$. Сада добијамо једначину

$$-\frac{\log(0.5)}{\lambda} = m_e,$$

одакле је $\hat{\lambda}_n = \frac{\log 2}{m_e}$.

Задатак 2.1.1. На основу п.с.у. $X_1, ..., X_n$ одредити оцену непознатих параметра у случају да:

- X има Биномну $\mathcal{B}(3,p)$ расподелу, $p \in (0,1)$ је непознат параметар;
- X има закон расподеле

$$X: \begin{pmatrix} -1 & 0 & 2 \\ \theta & \theta & 1-2\theta \end{pmatrix},$$

  $\theta < 0.5$ је непознат параметар;
- X има Гама $\gamma(2,\ \beta)$, расподелу, $\beta > 0$, је непознат параметар;
- X има померену експоненцијалну расподелу, односно

$$f(x) = \begin{cases} \lambda e^{-\lambda(x-\mu)}, & x \geq \mu \\ 0, & x < \mu \end{cases} \tag{2.1}$$

  $\mu \in \mathbf{R}$ и $\lambda > 0$ су непознати параметри.

2.1.2 Метод максималне веродостојности

Основни принцип овог метода је да је оцена непознатог параметра (који може бити вишедимензионални) вредност која максимизира функцију веродостојности. Интуитивно, то би била вредност параметра за коју је највероватније да "се деси" баш наш реализован узорак.

Дефиниција 2.1.2. *У случају дискретног обележја функција веродостојности је*

$$L(\theta) = P_\theta\{X_1 = x_1, ..., X_n = x_n\}.$$

У случају простог случајног узорка

$$L(\theta) = \prod_{i=1}^{n} P_\theta\{X_i = x_i\}.$$

У случају апсолутно непрекидног обележја функција веродостојности је

$$L(\theta) = f_\theta(X_1, ..., X_n).$$

У случају простог случајног узорка

$$L(\theta) = \prod_{i=1}^{n} f_\theta(x_i).$$

Дакле, оцена $\hat{\theta}_n$ је

$$\hat{\theta}_n = \text{argmax}_{\theta\in\Theta} L(\theta).$$

Из саме дефиниције, видимо да оцена добијена овим методом не мора бити јединствена, а не мора да ни постоји. Али, уколико важе неки додатни услови може се показати да постоји и да је јединствена.

Веома често је лакше маскимизирати неку монотону трансформацију функције веродостојности. Најчешће се максимизира логаритам функције веродостојности $l(\theta) = \log L(\theta)$. Неколико наредних примера илуструје примену овог метода.

Пример 2.1.8. *Обележје X је индикатор са вероватноћом успеха p. Тада функцију веродостојности можемо написати у облику*

$$L(p) = \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i}.$$

$$l(p) = \sum_{i=1}^{n} X_i \log p + \sum_{i=1}^{n}(1-X_i)\log(1-p).$$

Функција је диференцијабилна за $p \in (0,1)$ па ћемо тако тражити њен максимум. Решавамо једначину

$$\frac{\partial l(p)}{\partial p} = 0,$$

и добијамо да је $\hat{p}_n = \bar{X}_n$. Треба још проверити да ли је добијено решење баш максимум посматране функције. То можемо лако закључити на основу понашања функције $\frac{\partial l(p)}{\partial p}$ која је позитивна за $p < \hat{p}_n$, а негативна за $p > \hat{p}_n$. Због тога је $l(p)$ растућа на $(0, \hat{p}_n)$ а опадућа на $(\hat{p}_n, 1)$ па је $\hat{p}_n$ заиста тачка у којој $l(p)$ постиже максимум, односно $\hat{p}_n$ је оцена непознатог параметра методом максималне веродостојности. Приметимо да смо исту оцену добили и применом метода момената.

Пример 2.1.9. *Нека X има експоненцијалну $\mathcal{E}(\lambda)$ расподелу. Претпостављамо да је реализованпрост случајан узорак $x_1, ..., x_n$.*

Функција веродостојности је

$$L(\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i}.$$

Сада је

$$l(\lambda) = \log L(\lambda) = n \log \lambda - \lambda \sum_{i=1}^{n} x_i.$$

Ова функција је диференцијална за $\lambda > 0$ па максимум можемо добити из једначине

$$\frac{\partial l(\lambda)}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0.$$

Одавде је $\hat{\lambda}_n = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{\bar{x}_n}$.

Како је $\frac{\partial^2 l(\lambda)}{\partial \lambda^2}|_{\lambda=\hat{\lambda}_n} < 0$ закључујемо да $\hat{\lambda}_n$ јесте оцена максималне веросотојности. То је њена вредност на основу реализованог узорка $x_1, ..., x_n$. За сваки узорак добићемо другу вредност, па се оцена, као случајна величина може написати у облику

$$\hat{\lambda}_n = \frac{1}{\bar{X}_n}.$$

Пример 2.1.10. *Нека X има $\mathcal{U}[0, \theta]$ расподелу. Функција веродостојности је*

$$L(\theta) = \prod_{i=1}^{n} \frac{1}{\theta} I\{X_i \leq \theta\} = \frac{1}{\theta^n} I\{X_{(n)} \leq \theta\}.$$

Ова функција није диференцијабилна по θ, па се не може тражити максимум на уобичајан начин. Приметимо да је $L(\theta) = 0$ кад год је индикатор који се појављује у изразу за функцију веродостојности једнак 0, односно кад је $\theta < X_{(n)}$, зато максимум тражимо на скупу $\theta \geq X_{(n)}$. Даље, приметимо да је $\frac{1}{\theta^n}$ опадајућа функција по θ па ће максимум достићи за најмање могуће θ, што је у нашем случају $X_{(n)}$. Зато је $\hat{\theta}_n = X_{(n)}$.

За разлику од оцене добијене методом момената, сада нам се не може десити да је највећа вредност у узорку већа од оцене горње границе интервала. То је једна од очигледних предности метода максималне веродостојности.

Као што смо већ навели, оцена добијена овим методом не мора бити јединствена. То можемо видети у наредном примеру.

Пример 2.1.11. *Нека X има униформну $U[\theta-1,\theta+1]$ расподелу. Функција веродостојности је*

$$L(\theta)=\prod_{i=1}^{n}2^{-1}I\{X_i\in[\theta-1,\theta+1]\}=2^{-n}I\{X_{(n)}\leq\theta+1,X_{(1)}\geq\theta-1\}.$$

Видимо да је вредност функције веродостојности 2^{-n} кад год је индикатор једнак 1 па се тако максимум достиже за свако θ за које је то испуњено. То је еквивалентно са $\theta\in[X_{(n)}-1,X_{(1)}+1]$, односно све вредности из интервала представљају оцену максималне веродостојности. Једна могућа оцена била би

$$\hat{\theta}_n=\frac{|X_{(1)}|}{|X_{(1)}|+|X_{(2)}|}\cdot(X_{(n)}-1)+(1-\frac{|X_{(1)}|}{|X_{(1)}|+|X_{(2)}|})\cdot(X_{(1)}+1).$$

Оцене добијене овом методом имају следеће лепо својство: *Нека је g нека функција. Уколико је $\hat{\theta}_n$ оцена методом максималне веродостојности за θ онда је $g(\hat{\theta}_n)$ оцена методом максималне веродостојности за $g(\theta)$.*

Пример 2.1.12. *Оцена максималне веродостојности за дисперзију индикатора је $\hat{p}_n(1-\hat{p}_n)$, где је $\hat{p}_n=\bar{X}_n$ оцена максималне веродостојности за p.*

Задатак 2.1.2. Нека X има Биномну $\mathcal{B}(N,p)$ расподелу, при чему је N познато. Одредити оцену за p методом максималне веродостојности. Како се мења оцена уколико је познато да је $p>0.5$.?

Задатак 2.1.3. Нека X има $\Gamma(8,\beta)$ расподелу, где је $\beta>0$ непознат параметар. Оценити га методом максималне веродостојности.

Задатак 2.1.4. Сматра се да се број голова на фудбалсикм утакмицама може моделирати Пуасоновом $\mathcal{P}(\mu)$ расподелом. Узет је узорак од 10 утакмица једног тима у току претходне две сезоне и добијено је да је просечан број голова био 1.5. Методом максималне веродостојности оценити вероватноћу да на утакмици у којој игра посматрани тим, нема голова.

2.2 Особине оцена

У претходним поглављима видели смо да различити методи оцењивања могу резултовати и различитим оценама. Зато се природно јавља потреба да се испита њихов квалитет. Интуитивно, квалитетна оцена непознатог параметра θ би била она статистика за коју можемо да кажемо да је у просеку блиска стварној вредности параметра θ и да се налази у његовој близини са великом вероватноћом. Сада ћемо те особине формализовати.

Нека је $\hat{\theta}_n$ оцена непознатог параметра θ на основу узорка $X_1, X_2, ..., X_n$.

Дефиниција 2.2.1. *Уколико је $E(\hat{\theta}_n) = \theta$ за оцену $\hat{\theta}_n$ кажемо да је непристрасна оцена параметра θ.*

Дефиниција 2.2.2. *Уколико је $\lim_{n\to\infty} E(\hat{\theta}_n) = \theta$, оцена $\hat{\theta}_n$ је асимптотски непристрасна оцена параметра θ.*

Дефиниција 2.2.3. *Уколико $\forall \varepsilon > 0$, $\lim_{n\to\infty} P\{|\hat{\theta}_n - \theta| > \varepsilon\} = 0$ оцена $\hat{\theta}_n$ је постојана оцена параметра θ.*

Својство 2.2.1 је важно али уколико оцена "много осцилира" око стварне вредности, није од суштинског значаја. Зато је битно испитати и колика је дисперзија оцене. Својство 2.2.3 нам заправо каже да се за довољно велики узорак оцена непознатог параметра може наћи у произвољно малој околини стварне вредности параметра θ, са великом вероватноћом. Коришћењем Чебишовљеве неједнакости добијамо да је довољан услов да ово важи да је

$$\lim_{n\to\infty} E(\hat{\theta}_n - \theta)^2 = 0. \tag{2.1}$$

Због једноставности провере, најчешће су управо овај услов и проверава. У случају да је оцена непристрасна услов (2.1) је еквивалентан са условом

$$\lim_{n\to\infty} D(\hat{\theta}_n) = 0.$$

Уколико је више оцена непристрасно (или бар асимптотски) и постојано, намеће се питање коју од њих одабрати. Да бисмо на то одговорили потребно је да уведемо неку меру квалитета оцена. Јасно је да та мера квалитета треба да буде заснована на неком растојању праве вредности и њене оцене, при чему морамо да водимо рачуна о томе да је оцена случајна величина. Једна могућност је да посматрамо средње квадратно одступање оцене од праве вредности параметра.

Дефиниција 2.2.4. *Нека су $\hat{\theta}_n$ и $\hat{\hat{\theta}}_n$ две оцене параметра θ. Казаћемо да је $\hat{\theta}_n$ боља од $\hat{\hat{\theta}}_n$ у средње квадратном, уколико је*

$$E(\hat{\theta}_n - \theta)^2 < E(\hat{\hat{\theta}}_n - \theta)^2. \tag{2.2}$$

Поред средњеквадратног, често се користи и средње апсолутно одступање $E(|\hat{\theta} - \theta|)$ али је његово израчунавање обично доста компликованије па ово растојање није први избор.
Врло често, када није могуће наћи расподелу оцена, квалитет се испитује Монте Карло методом (очекивања која је потребно одредити се оцењују на основу великог броја понављања експеримента у којима се оцењује θ, под истим условима).
Алгоритам којим бисмо добили оцене средњеквадратног одступања је следећи:

1. Генеришемо узорак $\boldsymbol{x} = (x_1, ..., x_n)$ обима n из расподеле $F(\theta)$;

2. На основу узорка $\boldsymbol{x}$ одредимо $\hat{\theta}_n(\boldsymbol{x})$;

3. Поновимо кораке 1 и 2 N пута и на тај начин добијемо низ оцена $\hat{\theta}_n^{(1)}, \hat{\theta}_n^{(2)}, ..., \hat{\theta}_n^{(N)}$;

4. Одредимо квадратно одступање за сваку од добијених оцена, односно $(\hat{\theta}_n^{(1)} - \theta)^2, (\hat{\theta}_n^{(2)} - \theta)^2, ..., (\hat{\theta}_n^{(N)} - \theta)^2$;

5. Средње квадратно одступање $E(\hat{\theta}-\theta)^2$ оцењујемо са

$$\frac{1}{N}\sum_{i=1}^{N}(\hat{\theta}_n^{(i)}-\theta)^2. \tag{2.3}$$

Оцена средње квадратног одступања 2.3 има све "лепе" особине узорачке средине које смо показали у претходном поглављу. Како би се постигла што мања дисперзија оцене средњеквадратног растојања, потребно је да N буде велико. На пример, уколико желимо тачност оцене на две децимале N мора бити бар 10000.
Приметимо још да применом овог алгоритхма можемо добити и расподелу оцене. Наиме, у кораку 3 имамо низ оцена на основу кога се може оценити функција расподеле (или функција густине) случајне величине $\hat{\theta}_n$.

Пример 2.2.1. *У примеру 2.1.1 смо добили да су оцене методом момената за m и σ^2 редом $\hat{m}_n=\bar{X}_n$ и $\sigma_n^2=\bar{S}_n^2$. Из $E(\hat{m}_n)=m$ закључујемо да је $\hat{m}_n$ непристрасна оцена за m, док из $E\bar{S}_n^2=\frac{n-1}{n}\sigma^2$ закључујемо да је $\hat{\sigma}_n^2$ асимптотски непристрасна за σ^2. Што се постојаности тиче, имајући у виду коначност момената нормалне расподеле, постојаност $\hat{m}_n$ следи из $D(\hat{m}_n)=\frac{\sigma^2}{n}$, док се за $\hat{\sigma}_n^2$ може показати да $E(\hat{\sigma}_n^2-\sigma^2)^2\to 0$, па је и ова оцена постојана.*

Пример 2.2.2. *Испитајмо непристрасност оцене $\hat{\lambda_n}$ из примера 2.1.9, за $n>3$.*

$$E(\hat{\lambda}_n)=E\Big(\frac{n}{\sum_{i=1}^{n}X_i}\Big).$$

Присетимо се да збир n независних случајних величина са $\mathcal{E}(\lambda)$ има $\gamma(n,\lambda)$ расподелу. Зато је потребно да одредимо $E(\frac{1}{Y})$, где је $Y\sim\gamma(n,\lambda)$. Треба водити рачуна да $E(\frac{1}{Y})$ није исто што и $\frac{1}{EY}$.

$$\begin{aligned}E\Big(\frac{1}{Y}\Big)&=\int_0^\infty\frac{1}{x}\frac{x^{n-1}\lambda^n e^{-\lambda x}}{\Gamma(n)}dx=\int_0^\infty\frac{x^{n-2}\lambda^n e^{-\lambda x}}{\Gamma(n)}dx\\&=\int_0^\infty\frac{x^{n-2}\lambda^{n-1}e^{-\lambda x}}{\Gamma(n-1)}dx\cdot\frac{\lambda\Gamma(n-1)}{\Gamma(n)}=1\cdot\frac{\lambda}{n-1}.\end{aligned}$$

Одавде је

$$E(\hat{\lambda}_n)=\frac{n}{n-1}\lambda.$$

Оцена није непристрасна али јесте асимптотски непристрасна. Одавде видимо и да је $\widetilde{\lambda}_n = \frac{n-1}{n}\hat{\lambda}_n$ *непристрасна оцена параметра* λ.
За испитивање постојаности оцена одредићемо

$$E(\hat{\lambda}_n - \lambda)^2 = E(\hat{\lambda}_n^2) - 2\lambda E(\hat{\lambda}) + \lambda^2.$$

Даље је

$$\begin{aligned} E(\hat{\lambda}_n^2) &= n^2 E\big(\frac{1}{Y^2}\big) = n^2 \int_0^\infty \frac{x^{n-3}\lambda^n e^{-\lambda x}}{\Gamma(n)} dx \\ &= n^2 \int_0^\infty \frac{x^{n-3}\lambda^{n-2} e^{-\lambda x}}{\Gamma(n-2)} dx \cdot \frac{\lambda^2 \Gamma(n-2)}{\Gamma(n)} = 1 \cdot \frac{\lambda^2 n^2}{(n-1)(n-2)}, \end{aligned}$$

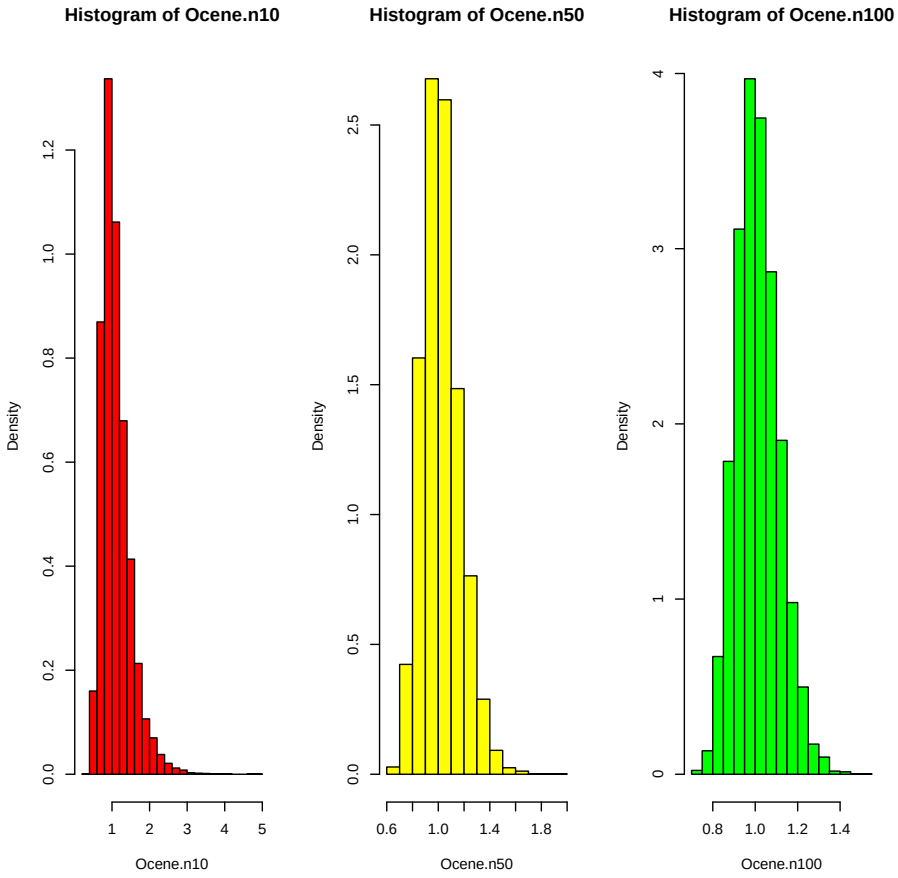
па је

$$\begin{aligned} E(\hat{\lambda}_n - \lambda)^2 &= \lambda^2 \Big(\frac{n^2}{(n-1)(n-2)} - \frac{2n}{n-1} + 1 \Big) \\ &= \frac{\lambda^2}{(n-1)(n-2)} (n^2 - 2n(n-2) + n^2 - 3n + 2) \\ &= \frac{\lambda^2 (n+2)}{(n-1)(n-2)} \to 0, \quad n \to \infty. \end{aligned}$$

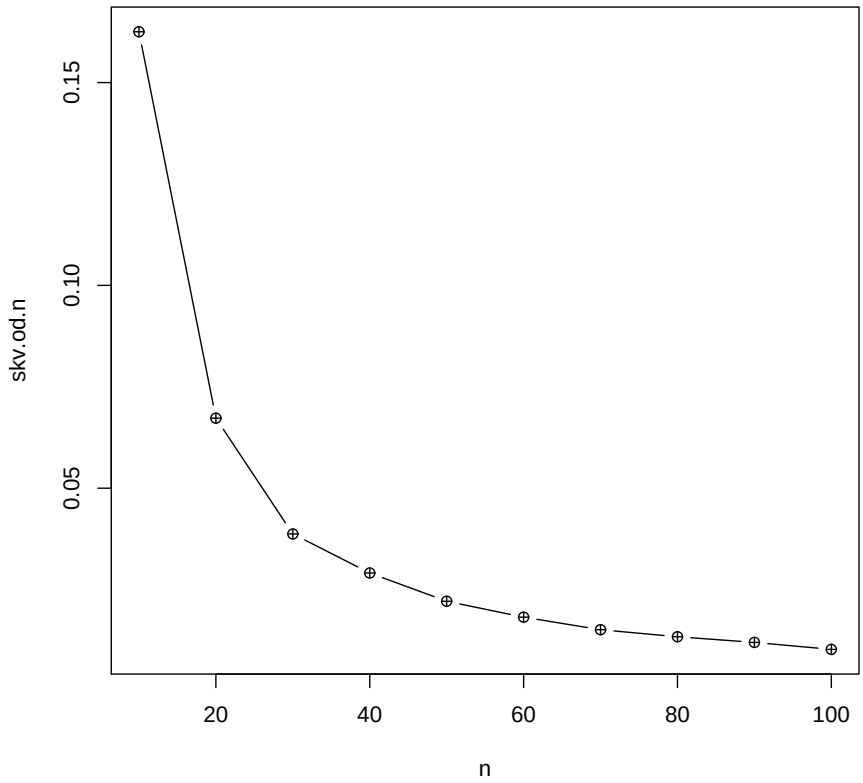
Одавде закључујемо да је $\hat{\lambda}_n$ *постојана оцена за* λ. *Слично важи и за оцену* $\widetilde{\lambda}_n$. *Наиме важи,*

$$\begin{aligned} D(\widetilde{\lambda}_n) &= E(\widetilde{\lambda}_n^2) - \lambda^2 = \frac{(n-1)^2}{n^2} \cdot \frac{\lambda^2 n^2}{(n-1)(n-2)} - \lambda^2 \\ &= \frac{\lambda^2}{n-2} \to 0, \quad n \to \infty. \end{aligned}$$

У случају да се нисмо сетили да Y *има* $\gamma(n, \lambda)$ *расподелу, о особинама оцене* $\hat{\lambda}_n$ *(и* $\widetilde{\lambda}_n$*) бисмо могли ипак нешто закључити примењујући кораке 1-3 алгоритма 2.2. На слици 2.1 су приказани хистограми густина оцена* $\hat{\lambda}_n$ *које се добијају када је права вредност параметра* $\lambda = 1$. *Можемо да уочимо својство асимптотске непристрасности и постојаности. Како* n *расте вредности које се добијају су "све ближе" правој вредности* λ. *То се још боље види на слици 2.2*



Слика 2.1: Хистограм оцена за λ



Слика 2.2: Средње квадратно одступање $\hat{\lambda}$ од λ за узорке различитих обима

```
# функција која понавља поступак оцењивања N пута на
узорцима генерисаним из E(λ) расподеле
OceneExpLambda<–function(n,lambda,N)
{
ocene=rep(0,N)
for(i in 1: N)
{
uzorak=rexp(n,lambda)
ocene[i]=1/mean(uzorak)
}
return(ocene)
}

set.seed(10)
n=10*(1:10)

# конструишемо низове потребне за оцене средњеквадратног
одступања за различите обиме узорка
Ocene.n10=OceneExpLambda(n=10,lambda=1,N=10000)
Ocene.n20=OceneExpLambda(n=20,lambda=1,N=10000)
Ocene.n30=OceneExpLambda(n=30,lambda=1,N=10000)
Ocene.n40=OceneExpLambda(n=40,lambda=1,N=10000)
Ocene.n50=OceneExpLambda(n=50,lambda=1,N=10000)
Ocene.n60=OceneExpLambda(n=60,lambda=1,N=10000)
Ocene.n70=OceneExpLambda(n=70,lambda=1,N=10000)
Ocene.n80=OceneExpLambda(n=80,lambda=1,N=10000)
Ocene.n90=OceneExpLambda(n=90,lambda=1,N=10000)
Ocene.n100=OceneExpLambda(n=100,lambda=1,N=10000)

# код за цртање хистограма приказаних на слици 2.1
hist(Ocene.n10,breaks=16,prob="TRUE",col='red')
hist(Ocene.n50,breaks=14,prob="TRUE",col='yellow')
hist(Ocene.n100,breaks=14,prob="TRUE",col='green')
```

```
skv.od.n10=mean((Ocene.n10-1)^2)
skv.od.n20=mean((Ocene.n20-1)^2)
skv.od.n30=mean((Ocene.n30-1)^2)
skv.od.n40=mean((Ocene.n40-1)^2)
skv.od.n50=mean((Ocene.n50-1)^2)
skv.od.n60=mean((Ocene.n60-1)^2)
skv.od.n70=mean((Ocene.n70-1)^2)
skv.od.n80=mean((Ocene.n80-1)^2)
skv.od.n90=mean((Ocene.n90-1)^2)
skv.od.n100=mean((Ocene.n100-1)^2)
skv.od.n=c(skv.od.n10,skv.od.n20,skv.od.n30,skv.od.n40,skv.od.n50,
skv.od.n60,skv.od.n70,skv.od.n80,skv.od.n90,skv.od.n100)

plot(n,skv.od.n,type='b',pch=10,main="")
```

У наредном примеру видећемо како се пореде оцене.

Пример 2.2.3. *Нека X има $\mathcal{U}[0,\theta]$ расподелу. У примерима 2.1.3 и 2.1.10 смо видели да су оцене добијене методом момената и методом маскималне веродостоности редом*

$$\hat{\theta}_n = 2\bar{X}_n, \quad \hat{\hat{\theta}}_n = X_{(n)}.$$

Сад ћемо их упоредити. Прво, није тешко показати да су обе оцене постојане, при чему је прва и непристрасна, а друга асимптотски непристрасна. Одредићемо средње квадратна одступања за сваку од оцена.

$$E(\hat{\theta}_n - \theta)^2 = D(\hat{\theta}_n) = D(2\bar{X}_n) = \frac{\theta^2}{3n} \tag{2.4}$$

Да бисмо одредили друго средњеквадратно одступање потребно је

да нађемо расподелу за $X_{(n)}$. *Имамо да је*

$$F_{X_{(n)}}(x) = P\{X_{(n)} \le x\} = P\{X_i \le x,\ i = 1, ..., n\} = \prod_{i=1}^{n} P\{X_i \le x\}$$
$$= (F_X(x))^n = \frac{x^n}{\theta^n},\ \ x \in [0, \theta],$$
$$f_{X_{(n)}}(x) = \frac{nx^{n-1}}{\theta^n},\ \ x \in [0, \theta].$$

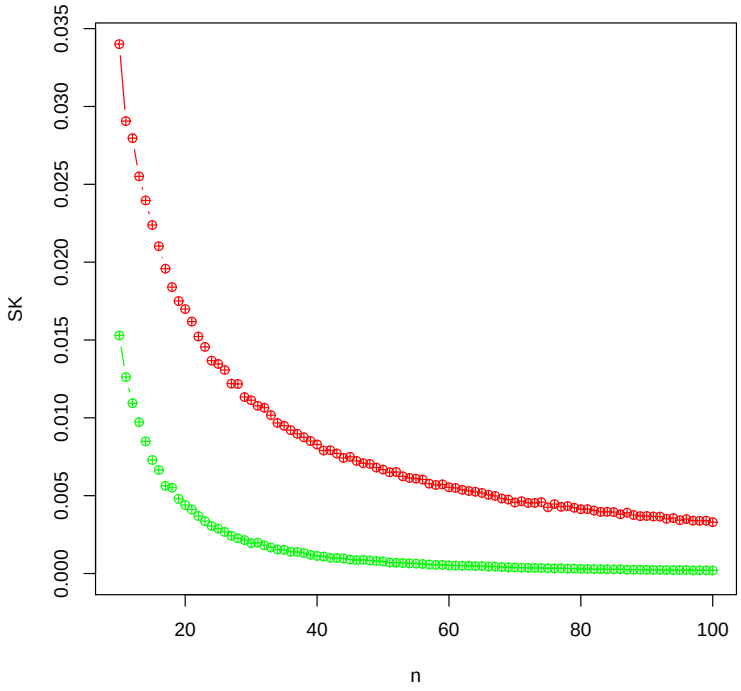
Одавде је

$$EX_{(n)} = \int_0^\theta \frac{nx^n}{\theta^n} dx = \frac{n\theta}{n+1},$$
$$EX_{(n)}^2 = \int_0^\theta \frac{nx^{n+1}}{\theta^n} dx = \frac{n\theta^2}{n+2}.$$

Одавде је

$$E(\hat{\hat{\theta}}_n - \theta)^2 = EX_{(n)}^2 - 2\theta EX_{(n)} + \theta^2$$
$$= \frac{n\theta^2}{n+2} - 2\frac{n\theta^2}{n+1} + \theta^2 = \frac{\theta^2}{(n+2)(n+1)}. \qquad (2.5)$$

Из (2.4) *и* (2.5) *видимо да је оцена добијена методом максималне веродостојности боља у средњеквадратном смислу. То се види и на слици 2.3 на којој је приказана оцена средњеквадратног одступања Монте Карло приступом за ове оцене и различите обиме узорка. Напомињемо да приликом поређења средњеквадратног одступања поредимо квалитет оцена за свако* n *а не само за велике обиме узорка (испитивање асимптотских својстава попут постојаности и асимптотске непристрасности) што је посебно важно за примену у пракси.*



Слика 2.3: Средње квадратна одступања $\hat{\theta}$ (црвено) и $\hat{\hat{\theta}}$ (зелено) од θ за узорке различитих обима

```
#код који користимо за цртање слике 2.3
n=10:100
N=10000
SK1=c()
SK2=c()
# за сваки обим узорка примењујемо Монте Карло методу,
односно понављамо експеримент N пута
for(i in 1:length(n))
{
ocene1=rep(0,N)
ocene2=rep(0,N)
```

```
for(j in 1:N){
uzorak=runif(n[i],0,1)
ocene1[j]=2*mean(uzorak)
ocene2[j]=max(uzorak)
}
SK1=c(SK1,mean((ocene1-1)^2))
SK2=c(SK2,mean((ocene2-1)^2))
}
plot(n,SK1,type='b',col='red',ylim=c(0,max(SK1)),pch=10,
ylab='SK')
lines(n,SK2,type='b',col='green',pch=10)
```

Задатак 2.2.1. Обележје X има следећи закон расподеле:

$$X : \begin{pmatrix} -1 & 0 & 1 \\ \theta & 1-2\theta & \theta \end{pmatrix},$$

где је $\theta \in (0, 0.5)$ непознат параметар. Наћи оцене које се добијају методом момената и максималне веродостојности и испитати њихову непристрасност и постојаност. Која од оцена је боља у средњеквадратном смислу? Приказати график зависности оцењеног средњеквараног одступања оцена од стварне вредности параметра $\theta = 0.2$, од обима узорка. Урадити то за обе посматране оцене. Шта закључујете са тог графика?

Задатак 2.2.2. Упоредити квалитет оцена добијених методом момената и методом максималне веродостојности за сваки од непознатих параметара померене експоненцијалне расподеле са густином (2.1). Емпиријски утврдити њихове расподеле.

Задатак 2.2.3. Претпоставимо да се радни век лед сијалица може моделирати експоненцијалном λ расподелом. Тестирано је 8 сијалица и добијене су следеће вредности (у годинама): 2,1,3,5,5,4,7,4. Одредити оцену непознатог параметра λ. Одабир метода за оцењивање образложите.

Задатак 2.2.4. За оцену параметра средње вредности нормалне расподеле, на основу п.с.у. $X_1, X_2, ...X_n$ предлажу се две оцене:

$\hat{m}_n = \bar{X}_n$ и $\widetilde{m}_n = \frac{X_1}{2} + \frac{1}{2}\sum_{i=2}^{n-1}\frac{X_i}{n-1}$. Која од оцена је боља у средњеквадратном смислу?

Задатак 2.2.5. Циљ истраживање је да се пронађе одговарајући статистички модел који би описао време преживљавања након лечења неког карцинома. Време трајања студије је ограничено на 5 година. Једна од могућих опција за моделирање је експоненцијална расподела $\mathcal{E}(\lambda)$. Оценити параметар λ ако на располагању имате само информацију да ли је особа која се налази у студији доживела крај истраживања или не, односно на располагању имате узорак 0,0,1,0,1,1,1,0,1,1,1,1,0,1,1,1,0,1,1,1 (1 означава да је особа преживела, а 0 да је умрла).
Како бисте имали информацију о квалитету оцене, односно о грешци која се прави приликом оцењивања, урадити мини емпиријску студију у којој ће те је оценити. Приликом узорковања из експоненцијалне расподеле можете користити оцењену вредност λ на основу датог узорка.

Задатак 2.2.6. На основу п.с.у. $X_1, X_2, ..., X_n$ из дискретне униформне расподеле на $\{1, 2, ..., N\}$ оценити N методом максималне веродостојности и испитати непристрасност оцене.

2.3 Интервалне оцене параметара

У претходном поглављу смо видели како се добијају тачкасте оцене параметара. У овом поглављу упознаћемо се са интервалним оценама истих. Као и до сада, претпостављамо да на располагању имамо п.с.у. из расподеле $F(\theta)$.

Дефиниција 2.3.1. *Нека је θ непознат параметар. Нека су L_n и U_n статистике за које је $P\{L_n \leq U_n\} = 1$ и $P\{L_n \leq \theta \leq U_n\} = \beta$. Интервал $[L_n, U_n]$ се назива $\beta\%$* двострани интервал поверења за параметар θ, *а β* ниво поверења.

Аналогно се дефинишу једнострани доњи и једнострани горњи интервали поверења.
Нека су $\hat{L}_n$ и $\hat{U}_n$ реализоване вредности статистика. Тада је $(\hat{L}_n, \hat{U}_n)$ реализовани интервал поверења. Важно је да у интерпретацији интервалних оцена не дође до забуне. **Није тачно да је $P\{\theta \in (\hat{L}_n, \hat{U}_n)\} = \beta$ јер параметар θ није случајна величина!** На основу неког другог узорка добићемо други реализовани интервал поверења, па је исправна интерпретација нивоа поверења да ће се у $\beta\%$ случајева стварна вредност параметра налазити у реализованом интервалу поверења.
За налажење интервала поверења потребно је наћи неку функцију од узорка и непознатог параметра чија расподела не зависи од непознатог параметра коју називамо *стожерна величина*. За параметре неких расподела већ постоји устаљена процедура које функције од узорка треба користити и коју расподелу имају. Илустроваћемо неке од њих за неке од најкоришћенијих расподела, али свакако да треба имати у виду да то нису једини могући избори стожерних величина и да се избором неких других могу добити и други интервали поверења.

2.3.1 Закључивање у моделу са нормалном расподелом

Претпоставимо да обележје X има нормалну $\mathcal{N}(m, \sigma^2)$ расподелу. Интервали поверења за m и σ^2 се могу добити коришћењем следећих тврђења.

Теорема 2.3.1. *Нека је* $X_1, X_2, ..., X_n$ *n.c.y. из* $\mathcal{N}(m, \sigma^2)$ *расподеле. Тада*

1. $\frac{\bar{X}_n - m}{\frac{\sigma}{\sqrt{n}}}$ *има нормалну* $\mathcal{N}(0,1)$ *расподелу;*

2. $\frac{(n-1)\widetilde{S}_n^2}{\sigma^2} = \frac{n\bar{S}_n^2}{\sigma^2}$ *има* χ^2_{n-1} *расподелу;*

3. *Случајне величине* $\bar{X}_n$ *и* $\widetilde{S}_n^2$ *су независне;*

4. $\frac{\sqrt{n}(\bar{X}_n - m)}{\widetilde{S}_n}$ *има Студентову* t_{n-1} *расподелу.*

Формалан доказ теореме изостављамо али ћемо навести неколико главних корака у доказу. Наиме 1. је последица својстава нормалне расподеле описаних у Додатку 6. Када бисмо у изразу за узорачку дисперзију $\widetilde{S}_n^2$ уместо $\bar{X}_n$ имали m онда би $\frac{(n-1)\widetilde{S}_n^2}{\sigma^2}$ била сума квадрата независних случајних величина са стандардном нормалном расподелом, и као таква имала χ^2_n расподелу. Међутим, како је m оцењено, број степени се смањује за 1. Комбинацијом 1., 2. и 3. се добија 4. на основу дефиниције Студентове расподеле.

Сада можемо одредити $\beta\%$ интервал поверења за поменуте параметре. Разликоваћемо неколико случајева.

Интервал поверења за m када је σ^2 познато

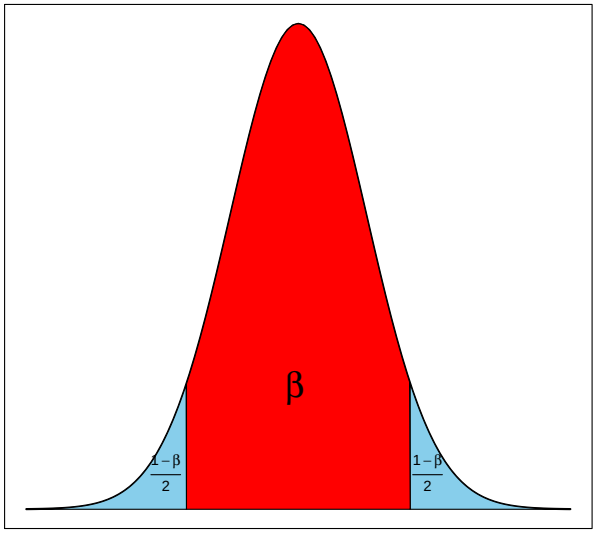
Потребно је прво да нађемо помоћну функцију од узорка чију расподелу знамо, а у којој се јавља m. Једна могућност, на основу теореме 2.3.1, је

$$T_n = \frac{\bar{X}_n - m}{\frac{\sigma}{\sqrt{n}}},$$

за коју знамо да има $\mathcal{N}(0,1)$ расподелу. Зато можемо одредити константу C тако да је $P\{|T_n| \leq C\} = \beta$ (види слику 2.4). Због симетричности нормалне расподеле $C = \Phi^{-1}(\frac{1+\beta}{2})$ (Φ је функција расподеле случајне величине са стандардном нормалном расподелом). Даље, неједнакост $|T_n| \leq C$ је еквивалентна са $\bar{X}_n - C\frac{\sigma}{\sqrt{n}} \leq m \leq \bar{X}_n + C\frac{\sigma}{\sqrt{n}}$. Одавде видимо да су статистике L_n

и U_n које смо тражили редом $L_n = \bar{X}_n - C\frac{\sigma}{\sqrt{n}}$ и $U_n = \bar{X}_n + C\frac{\sigma}{\sqrt{n}}$, односно да је интервална оцена параметра m једнака

$$\left(\bar{X}_n - C\frac{\sigma}{\sqrt{n}}, \bar{X}_n + C\frac{\sigma}{\sqrt{n}}\right). \tag{2.1}$$



Слика 2.4: Конструкција $\beta\%$ интервала поверења

*Н*апомена: Интервал поверења не мора бити симетричан, али се у случају симетрично расподељене стожерне величине узима најчешће баш такав.

Пример 2.3.1. *У једној фабрици кондиторских производа, приликом контроле квалитета, узет је узорак од 10 чоколада и измерена тежина паковања и добијена просечна вредност 2.4 грама. На основу претходних истраживања познато је да се тежина паковања може моделирати $\mathcal{N}(m, 0.4)$ расподелом. Одредићемо 90% интервал поверења за m.*

Константа C која нам је потребна је $C = \Phi^{-1}(0.95) = 1.64$. *Поред тога имамо да је* $\bar{x}_{10} = 2.4$ *и* $n = 10$. *Сада, користећи* (2.1), *добијамо интервал* $(2.08, 2.73)$.

Следећим примером ћемо илустровати суштину интервала поверења како не би долазило до грешака у његовој интерпретацији.

Пример 2.3.2. *Генерисаћемо узорке из* $\mathcal{N}(0,1)$ *(стварна вредност параметра* $m = 0$*) и правићемо* 95% *интервале поверења за сваки од 10000 узорака и проверити да ли садрже стварну вредност параметра* m.

Добили смо да је у 9515 случајева интервал поверења садржао стварну вредност параметра m *што је у складу са тим да је задатки ниво поверења* 95%.

```
# функција која нам враћа интервал поверења за m када
је σ^2 познато за генерисани узорак из нормалне N(m,σ^2)
расподеле
interval.poverenja.m1 <- function(n,m,sigma,beta) {
xsr=mean(rnorm(n,m, sigma))
L=xsr - qnorm((1+beta)/2)*sigma/ sqrt(n)
U =xsr + qnorm((1+beta)/2)*sigma/ sqrt(n)
return(c(L, U))
}

set.seed(1)
brojac=0
N=10000

for(i in 1:N)
{
intpov=interval.poverenja.m1(n=10,0,1,0.95)
if ((intpov[[1]]<=m)&&(intpov[[2]]>=m)) brojac=brojac+1
}

>brojac
[1] 9515
```

Дужина интервала (2.1) је $2C\frac{\sigma}{\sqrt{n}}$ и не зависи од узорка, већ само од његовог обима. Последица овога биће да можемо да, за задати ниво поверења, одредимо обим узорка који нам је потребан да би дужина интервала била мања од неке одређене вредности.

Пример 2.3.3. *Уколико у примеру 2.3.1 желимо да направимо 90\$ интервал поверења који је ужи од 0.05 минимални обим узорка који нам је потребан добијамо из неједнакости*

$$2 \cdot 1.64\sqrt{\frac{0.4}{n_0}} \leq 0.5.$$

Добија се да је $n_0 = 17$.

Задатак 2.3.1. За интервалну оцену параметра m предлаже се интервал $[\bar{X}_n - 1.64 \cdot \frac{\sigma}{\sqrt{n}}, \bar{X}_n]$. Одредити ниво предложеног интервала поверења, а затим конструисати симтерични интервал истог нивоа поверења на начин описан у овом поглављу. Упоредити дужину предложеног са дужином конструисаног интервала поверења.

Интервал поверења за m када је σ^2 непознато

Сада се за стожерну величину може узети

$$T_n = \frac{\bar{X}_n - m}{\frac{\tilde{S}_n}{\sqrt{n}}},$$

за коју знамо да има t_{n-1} расподелу. Као и малопре, можемо одредити константу C тако да је $P\{|T_n| \leq C\} = \beta$. Због симетричности Студентове расподеле $C = F^{-1}_{t_{n-1}}(\frac{1+\beta}{2})$. Неједнакост $|T_n| \leq C$ је еквивалентна са $\bar{X}_n - C\frac{\tilde{S}_n}{\sqrt{n}} \leq m \leq \bar{X}_n + C\frac{\tilde{S}_n}{\sqrt{n}}$. Одавде видимо шта су статистике L_n и U_n које смо тражили.

Пример 2.3.4. *Издавачка кућа жели да избаци на тржиште нову књигу и треба да утврди њену цену. Како би се цена што боље прилагодила тржишту, врши се истраживање о просечној цени сличних књига. Циљ је одредити* 90% *интервал поверења за* EX. *Узет је узорак од 25 насумично одабраних књига и добијено да је* $\bar{x}_{24} = 14.5$ *и* $\tilde{s}_{24} = 3.5$. *На основу неких претходних*

истраживања је установљено да се може сматрати да цена књиге има нормалну расподелу.

Због претпоставке о нормалности обележја можемо применити резултат из овог поглавља да нађемо одгварајући интервал поверења. Прво одређујемо C. Добија се да је $C = F^{-1}_{t_{24}}(\frac{1+0.9}{2}) = 1.71$, па је доња граница интервала $14.5 - \frac{3.5 \cdot 1.71}{\sqrt{25}} = 13.30$. Слично се добија да је горња граница 15.70. Сада издавачка кућа може да искористи ту информацију за формирање цене.

У случају непознате дисперзије дужина интервала је $2C\frac{\widetilde{S}_n}{\sqrt{n}}$ и зависи од узорка, тако да, уколико желимо да контролишемо дужину интервала, не можемо да поновимо процедуру која се ради у случају познате дисперзије. Свакако једна могућност је да узмемо неки "пробни" узорак и да на основу њега оценимо σ^2 а онда након тога, на основу добијене оцене за n_0 извучемо узорак одговарајуће величине. Имајући у виду случајну природу оцене за σ може се свакако десити да овакав приступ не да увек одговрајуће резултате, али свакако може да помогне при стицању осећаја ког реда величине треба да буде минимални обим узорка.

Интервал поверења за σ^2

Једна од могућности за стожерну величину у овом случају је

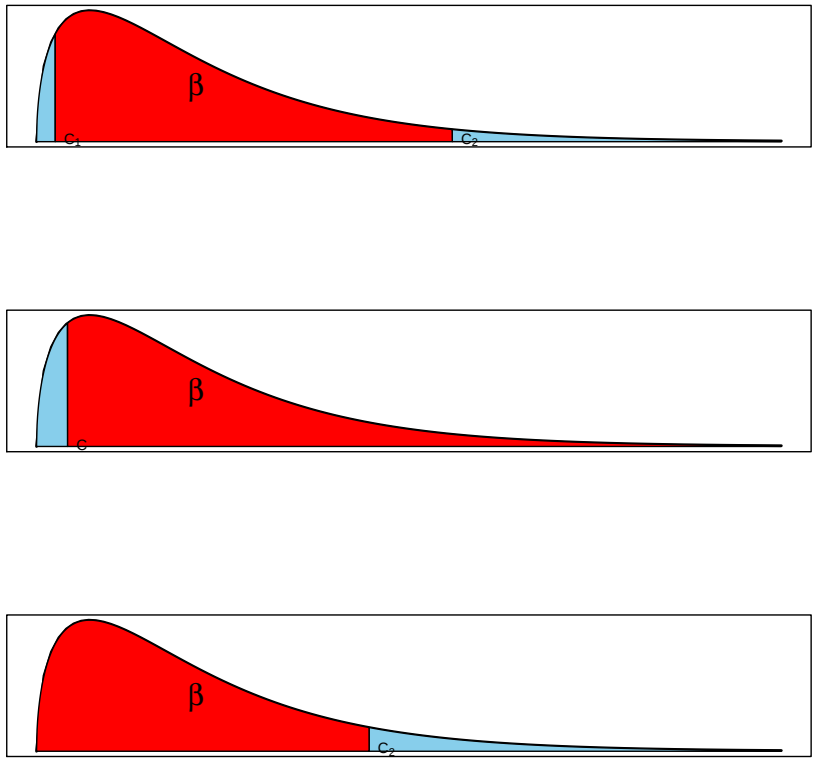
$$T_n = \frac{(n-1)\widetilde{S}_n^2}{\sigma^2}, \tag{2.1}$$

која има χ^2_{n-1} расподелу. Као што знамо, ова расподела није симетрична, иако се за велико n може апроксимирати нормалном расподелом, па двострани интервал поверења не можемо правити на претходно описани начин када користимо симетричност T. Уобичајно се интервал поверења прави тако да је "$\frac{1-\beta}{2}\%$ лево од доње границе, и исто толико са десне границе", односно одреди се C_1 и C_2 тако да је $P\{T < C_1\} = \frac{1-\beta}{2}$ и $P\{T > C_2\} = \frac{1-\beta}{2}$. Тада је $P\{C_1 \leq T \leq C_2\} = \beta$, па су константе $C_1 = F^{-1}_{\chi^2_{n-1}}(\frac{1-\beta}{2})$ и $C_2 = F^{-1}_{\chi^2_{n-1}}(\frac{1+\beta}{2})$. Неједнакост $C_1 \leq T \leq C_2$ је еквивалентна са

$$\frac{(n-1)\widetilde{S}_n^2}{C_2} \leq \sigma^2 \leq \frac{(n-1)\widetilde{S}_n^2}{C_1}, \tag{2.2}$$

односно добили смо двострани интервал поверења за σ^2.

Често нам је битно да нађемо горњу, односно доњу границу за σ^2, тј. да нађемо једностране интервале поверења $[L_n, \infty)$ или $(0, U_n]$. За конструкцију $[L_n, \infty)$ потребно је да одредимо C тако да је $P\{T_n \leq C\} = \beta$, док за $(0, U_n]$ је потребно да одредимо C тако да је $P\{T_n \geq C\} = \beta$. Илустрација како се одређују константе C неопходне за конструкцију интервала, приказана је на слици 2.5.



Слика 2.5: Конструкција $\beta\%$ интервала поверења

Пример 2.3.5. *У једном истраживању о попуњености места у градском превозу на некој одређеној линији, узет је узорак од 18 полазака у дневном режиму вожње у току такозваног "шпица", у периоду од месец дана. Добијено је да је просечан број људи* $\bar{x}_{18} = 52.2$*, са узорачким стандардним одступањем* $\widetilde{s}_{18} = 3.9$*. На основу претходних истраживања постоји податак да се ово обележје може моделирати нормалном расподелом, зато за одређивање 95% интервала поверења за параметар* σ^2 *можемо користити* (2.1)*.*

Добијамо да је $C_1 = F^{-1}_{\chi^2_{17}}(0.025) = 12.79$ *и* $C_2 = F^{-1}_{\chi^2_{17}}(0.975) = 30.19$ *па је, на основу* (2.2) *одговарајући двострани интервал поверења за* σ^2 *једнак* $(8.6, 20.2)$*. Уколико нам је битна само горња граница можемо одредити једнострани интервал. Из једнакости* $P\{T_{18} \geq C\} = 0.95$ *добијамо да је* $C = 8.7$ *па је тражени интервал* $(0, 29.7)$*.*

Задатак 2.3.2. Приликом трговања на берзи инвеститору је важно да у сваком тренутку може да процени колико су ризичне акције које поседује. Једна од могућих мера ризика је дисперизја расподеле дневних приноса дефинисаних, у тренутку t са

$$R_t = \frac{S_t - S_{t-1}}{S_{t-1}},$$

где је S_t цена акције у тренутку t. Уз претпоставку да су дневни приноси међусобно независне и једнако расподељене случајне величине, дисперзија расподеле приноса се може оценити поправљеном узорачком дисперзијом. Честа је и претпотавка да се дневни приноси могу моделирати нормалном расподелом. Управо уз такву претпоставку узет је узорак од 20 радних дана и одређена просечна вредност приноса $\bar{r}_{20} = 1.5$ и њихова узорачка дисперзија $\widetilde{s}^2_{20} = 0.7$. На основу добијених података одредити 99% двострани интервал поверења за дисперзију дневних приноса, као и одговарајуће једностране интервале поверења.

Задатак 2.3.3. На адреси www.finance.yahoo.com/ могу се пронаћи цене акција разних компанија. Одаберите неке две компанијщ и период од 20 радних дана и на основу тога, уз претпоставке из претходног задатка, одредите интервале поверења за

дисперзију дневних приноса, и на основу тога закључите која од компанија је ризичнија. Поред тога одредитии интервале поверења за просечне дневне приносе. У коју од компанија бисте инвестирали?

2.3.2 Закључивање у моделу са нормалном расподелом - случај два узорка

Често се појављује потреба за проналажењем интервалне оцене за разлику очекивања две независне случајне величине са нормалним $\mathcal{N}(m_1, \sigma_1^2)$ и $\mathcal{N}(m_2, \sigma_2^2)$ расподелама. Најчешће се ради о посматрању једног обележја на две популације (разлика у висинама особа мушког и женског пола, разлика коефицијента интелигенције код различитих популација и сл.). У тој ситуацији нам помаже следећа теорема коју наводимо без доказа.

Теорема 2.3.2. *Нека су* $X_1, ..., X_{n_1}$ *и* $Y_1, ..., Y_{n_2}$ *два независна п.с.у. из* $\mathcal{N}(m_1, \sigma_1^2)$ *и* $\mathcal{N}(m_2, \sigma_2^2)$*, редом. Тада важи:*

- $\frac{\bar{X}_{n_1} - \bar{Y}_{n_2} - (m_1 - m_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ *има* $\mathcal{N}(0,1)$ *расподелу;*

- *уколико је* $\sigma_1^2 = \sigma_2^2 = \sigma^2$ *онда* $\frac{\bar{X}_{n_1} - \bar{Y}_{n_2} - (m_1 - m_2)}{S\sqrt{\frac{1}{n_1} + \frac{1}{n2}}}$*, где је* $S^2 = \frac{(n_1-1)\widetilde{S}_{n_1}^2 + (n_2-1)\widetilde{S}_{n_2}^2}{n_1+n_2-2}$*, има* $t_{n_1+n_2-2}$ *расподелу;*

- *уколико је* $\sigma_1^2 \neq \sigma_2^2$ *онда* $\frac{\bar{X}_{n_1} - \bar{Y}_{n_2} - (m_1 - m_2)}{\sqrt{\frac{\widetilde{S}_{n_1}^2}{n_1} + \frac{\widetilde{S}_{n_2}^2}{n_2}}}$ *има* t_ν *расподелу, где је*

$$\nu = \frac{\left(\frac{\widetilde{S}_{n_1}^2}{n_1} + \frac{\widetilde{S}_{n_2}^2}{n_2}\right)^2}{\frac{\left(\frac{\widetilde{S}_{n_1}^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{\widetilde{S}_{n_2}^2}{n_2}\right)^2}{n_2 - 1}}; \tag{2.1}$$

- $\frac{\frac{\widetilde{S}_{n_1}^2}{\sigma_1^2}}{\frac{\widetilde{S}_{n_2}^2}{\sigma_2^2}}$ *има Фишерову* $\mathcal{F}_{n_1-1, n_2-1}$ *расподелу.*

Сада, уколико су нам потребни двострани интервали поверења за разлику очекивања можемо користити стожерну величину

$$T_{n_1,n_2} = \frac{\bar{X}_{n_1} - \bar{Y}_{n_2} - (m_1 - m_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}, \tag{2.2}$$

или

$$T_{n_1,n_2} = \frac{\bar{X}_{n_1} - \bar{Y}_{n_2} - (m_1 - m_2)}{S\sqrt{\frac{1}{n_1} + \frac{1}{n2}}} \tag{2.3}$$

или

$$T_{n_1,n_2} = \frac{\bar{X}_{n_1} - \bar{Y}_{n_2} - (m_1 - m_2)}{\sqrt{\frac{\widetilde{S}_{n_1}^2}{n_1} + \frac{\widetilde{S}_{n_2}^2}{n_2}}}, \tag{2.4}$$

у зависности од тога да ли су нам дисперзије обележја познате или не, и ако нису, да ли знамо да су једнаке или не. Пдређујемо C тако да је $P\{|T_{n_1,n_2}| < C\} = \beta$, при чему користимо резултате из теореме 2.3.2.

Најчешће, у пракси дисперзије нису познате, а како да формално испитамо ли су једнаке или не, видећемо ускоро. У томе нам може помоћи интервална оцена за количник дисперзија оцена. Уколико се 1 налази у њему можемо сматрати да су обележја једнака.

Да бисмо нашли интервалну оцену за количник дисперзија користићемо стожерну величину

$$Q_{n_1,n_2} = \frac{\frac{\widetilde{S}_{n_1}^2}{\sigma_1^2}}{\frac{\widetilde{S}_{n_2}^2}{\sigma_2^2}} \tag{2.5}$$

за коју знамо да има F_{n_1-1,n_2-1} расподелу. Ова расподела је асиметрична и поступак одређивања интервала поверења је исти као код интервалне оцене за дисперзију једног обележја. На пример, ако желимо да одредимо двострани $\beta\%$ интервал поверења, одредићемо константе C_1 и C_2 тако да је $P\{Q_{n_1,n_2} < C_1\} = \frac{1-\beta}{2}$ и $P\{Q_{n_1,n_2} > C_2\} = \frac{1-\beta}{2}$, па се интервал поверења добија из интервала $C_1 \leq Q_{n_1,n_2} \leq C_2$.

Пример 2.3.6. *Желимо да нађемо* 95% *интервал поверења за разлику просечне количине кофеина у кафи код два произвођача. Због тога су узета два узорка, од првог и другог произвођача, и то редом 15, односно 12 паковања. Добијени су следећи резултати*

Узорак	*I*	*II*
$\bar{X}$ *mg*	*80*	*77*
$\widetilde{S}$ *mg*	*5*	*6*

На основу неких претходних истраживања може се претпоставити да посматрана обележја имају нормалне расподеле са неједнаким дисперзијама.

Користећи (2.1) *добијамо да је* $\nu = 21.42 \approx 21$. *Константа* C *је онда* $C = F_{t_{21}}^{-1}(0.975) = 2.08$ *па се за леву границу интервала добија* $(80-77) - 2.08 \cdot \sqrt{\frac{5^2}{15} + \frac{6^2}{12}} = -1.49$, *а за десну* $(80-77) + 2.08 \cdot \sqrt{\frac{5^2}{15} + \frac{6^2}{12}} = 7.49$.

Пример 2.3.7. *У једном истраживању о факторима који утичу на повишен крвни притисак измерене су вредности систолног и дијастолнг крвног притиска 6 мушких и 4 женска испитаника. Резултати тестирања су следећи:*

	мушки		*женски*	
	сред. вред.	*сд. одступ.*	*сред. вред.*	*сд. одступ.*
систолни	*126*	*9.5*	*115*	*11.4*
дијастолни	*72*	*7*	*71*	*7.9*

Одредићемо 95% интервал поверења за разлику очекиваних вредности систолног притиска између полова.

На основу добијених резултата добијамо да је интервал поверења за $\frac{\sigma_2^2}{\sigma_1^2}$ *у случају систолног притиска* $(0.18, 21.43)$ *па закључујемо да ћемо интервал поверења за за разлику очекивања правити у случају кад је* $\sigma_1^2 = \sigma_2^2$. *Добијамо да је* $C = F_{t_8}^{-1}(0.975) = 2.33$ *и* $S = 10.25$ *па је тражени интервал поверења* $(-4.26, 26.26)$.

Задатак 2.3.4. Одредити 95% интервал поверења за разлику очекиваних вредности дијасистолног притиска између полова.

Задатак 2.3.5. И једном истраживању проучавано је како веганска особа утиче на телесну тежину. Узет је узорак од 20 особа које су више од 5 година на веганском режиму и 20 особа које нису и измерена је њихова тежина. Све особе су истог пола. Резултати су следећи:

вегани	84	63	81	81	91	92	59	71	94	98	84	66	87	77	80	68	87	54	57	83
остали	93	104	79	78	100	95	68	106	88	91	90	88	78	76	83	78	76	80	80	85

Одредити 90% интервал поверења за разлику очекиваних вредности телесне тежине у посматране две групе. Шта мислите, да ли тип ишране утиче на телесну тежину?

Задатак 2.3.6. Направити мини емпиријску студију у којој ћете испитати квалитет интервала поверења за количник дисперзија два обележја. Можете претпоставити да су стварне расподеле обележја $\mathcal{N}(1,1)$ и $\mathcal{N}(0,1)$, и 10000 пута поновите експеримент у коме генеришете по један узорак из свааке од расподела, одредити 95% интервал поверења за количник дисперзија и проверити да ли стварни количник (који је у овом случају једнак 1) припада конструисаном интервалу. На крају одредите проценат интервала који су садржали стварну вредност. Који сте проценат очекивали да добијете пре истраживања, а који сте добили?

2.3.3 Закључивање у моделу са Биномном $\mathcal{B}(1,p)$ расподелом

Претпостављамо да је X индикатор и да је вероватноћа успеха p. Поред ово интерпретације p се може и интерпетирати као пропорција популације која задовољава неку услов. Као и до сада, да бисмо одердили интервал поверења за p, треба да нађемо неку функцију чију расподелу знамо. Најчешће се користи

$$T_n = \frac{\bar{X}_n - p}{\sqrt{\frac{p(1-p)}{n}}}. \tag{2.1}$$

За велико n гранична расподела од T_n је нормална $\mathcal{N}(0,1)$ и оно што је и исто јако важно је да је конвергенција брза па то можемо користити већ за $n > 20$. Имајући ово у виду, за тражење $\beta\%$

интервала поверења потребно је одредити C тако да је $P\{|T_n| \leq C\} = \beta$. Приметимо да је неједнакост $|T_n| < C$ еквивалента са $T_n^2 < C^2$, односно

$$\left(\frac{\bar{X}_n - p}{\sqrt{\frac{p(1-p)}{n}}} \right)^2 \leq C^2,$$

што је даље еквивалентно са

$$\bar{X}_n^2 - 2p\bar{X}_n + p^2 \leq C^2 \frac{p}{n} - C^2 \frac{p^2}{n},$$

односно са

$$p^2 \left(\frac{C^2}{n} + 1 \right) - p \left(\frac{C^2}{n} + 2\bar{X}_n \right) + \bar{X}_n^2 \leq 0. \tag{2.2}$$

Неједнакост (2.2) је квадратна неједначина по p и, имајући у виду да се уз p^2 налази позитиван коефицијент, њено решење је скуп $[p_1, p_2]$, где су p_1, p_2 решења одговарајуће квадратне једначине. Тако добијамо да је тражени интервал поверења $[p_1, p_2]$. Приликом формирања интервала поверења треба водити рачуна да је p вероватноћа, односно да $p \in [0, 1]$. Уколико се добије p_1 мање од 0, и/или p_2 веће од 1, интервал треба редуковати.

За велико n и када p није блиско 0, односно 1 (односно кад $n\hat{p}$, или $n(1-\hat{p})$, није много мало, стандардни праг је 5), може се и дисперзија $\bar{X}_n$, која је једнака $\frac{p(1-p)}{n}$, оценити својом оценом максималне веродостојности, односно са $\frac{\bar{X}_n(1-\bar{X}_n)}{n}$. Тада имамо да

$$T_n = \frac{\bar{X}_n - p}{\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}} \tag{2.3}$$

има приближно нормалну $\mathcal{N}(0,1)$ расподелу, па примењујући исти поступак као у случају интервалне оцене за m у Нормалном моделу, добија се интервал поверења

$$\left(\bar{X}_n - C\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}, \bar{X}_n + C\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}} \right). \tag{2.4}$$

Ову формулу можемо користити и када желимо да одредимо приближан обим узорка који ће нам обезбедити да нам дужина интервала буде ужа од унапред задате вредности. Наиме, дужина интервала је

$$d = 2C\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Пошто не знамо колико је n не можемо да одредимо $\hat{p}$. Имамо два начина да превазиђемо ову препреку. Први, конзервативан, је да искористимо да је највећа могућа вредност производа $\hat{p}(1-\hat{p}) = \frac{1}{4}$, и онда одредимо n тако да је $C\sqrt{n} < d$. Други начин је да прво узмемо узорак неке, унапред одабране, величине, нпр. 20, (под претпоставком да можемо) и да на основу тог узорка оценимо $\hat{p}$, а онда у скалду са тим, одредимо n тако да је $d < C$. Након тога извучемо нови узорак величине n на основу кога ћемо одредити тражени интервал поверења. Овакав приступ има ограничење у томе да некада није могуће узети такозвани "пилот" узорак.

Пример 2.3.8. *Компанија за производњу играчака жели да пласира нови производ на тржиште. Направљен је пилот пројекат у коме је направљено 1000 играчака и поклоњено случајно одабраним породицама. Једина обавеза срећних добитника је била да одговори на питање да ли им се производ допао или не. Сакупљени су резултати и добијено је да је позитиван одговор дало 800 породица. На први поглед, то заиста одаје утисак да је одзив позитиван, али ради потпуније слике (и комплетнијег извештаја руководству компаније), направљен је* 99% *интервал поверења за* p*. То је урађено на следећи начин.*

Прво је оцењена вероватноћа, односно $\hat{p} = \frac{800}{1000} = 0.8$*. Како је* $n(1-\hat{p}) = 200$ *што је заиста велики број, па може да се користи* (2.3)*. Добијамо да је* $C = \Phi^{-1}\left(\frac{1+0.99}{2}\right) = 2.58$*, па је на основу* (2.4) *интервал поверења је* $\left(0.8 - 2.58\sqrt{\frac{0.2\cdot 0.8}{1000}}, 0.8 + 2.58\sqrt{\frac{0.2\cdot 0.8}{1000}}\right) = (0.76, 0.83)$.

Задатак 2.3.7. Како би се проценио проценат оболелих од вируса $\mathrm{SARS-CoV-2}$ који немају симптоме болести који би их навели на то да се пријаве надлежној здравственој установи, узет је узорак од 500 случајно одабраних грађана. Тестови су изузетно скупи па

зато није било могућности за узимањем већег узорка. Добијено је да је 47 грађана било позитивно на поменути вирус. Наћи 95% интервал поверења за проценат оболелих.

Задатак 2.3.8. Циљ једног истраживања био је да се утврди проценат људи који поштује обавезу ношења појаса у возњи. Случајно је заустављено 40 возила од којих су у 32 путници били адекватно везани. Наћи 99% интервал поверења за поменути проценат становништва и одредити његову дужину. Колико је било потребно зауставити кола уколико бисмо желели да имамо интервал краћи од 0.3?

Случај два узорка

Уколико имамо два обележја X и Y са расподелама

$$X : \begin{pmatrix} 0 & 1 \\ 1-p_1 & p_1 \end{pmatrix} \quad Y : \begin{pmatrix} 0 & 1 \\ 1-p_2 & p_2 \end{pmatrix},$$

и два независна узорка обима n_1 и n_2 који су велики онда можемо опет искористити Централну граничну теорему на основу које имамо да

$$T_{n_1,n_2} = \frac{\bar{X}_{n_1} - \bar{Y}_{n_2} - (p_1 - p_2)}{\sqrt{\frac{\bar{X}_{n_1}(1-\bar{X}_{n_1})}{n_1} + \frac{\bar{Y}_{n_2}(1-\bar{Y}_{n_2})}{n_2}}} \tag{2.1}$$

има граничну нормалну $\mathcal{N}(0,1)$ расподелу, па интервал поверења за разлику $p_1 - p_2$ добијамо из услова да је $P\{|T_{n_1,n_2}| < C\} = \beta$.

Пример 2.3.9. *Директор једног супермаркета треба да донесе одлуку да ли ће задржати оба снабдевача лимуна или ће повећати потражњу од једног а другом неће продужити уговор. Узео је извештаје неколико случајно одабраних набавки у претходном двогодишњем периоду и израчунао укупан проценат оштећених воћки.*

Снабдевач	*А*	*Б*
оштећено	*8%*	*8.5%*
укупан бр. воћки	*9000*	*10000*

Разлика између произвођача очигледно постоји, али да ли је она довољна да се не продужи уговор једном од снабдевача?

Директор је одлучио да направи 95% интервал поверења за разлику вероватноћа да се достави оштећена воћка. На основу (2.1) добија се да је тражени интервал поверења

$$(\bar{X}_{n_1} - \bar{X}_{n_2} - C \cdot A, \bar{X}_{n_1} - \bar{X}_{n_2} + C \cdot A),$$

где је $A = \sqrt{\frac{\bar{X}_{n_1}(1-\bar{X}_{n_1})}{n_1} + \frac{\bar{Y}_{n_2}(1-\bar{Y}_{n_2})}{n_2}}$ и $C = \Phi^{-1}(0.975) = 1.96$, односно добија се интервал (-0.012, 0.003). На основу тога, пошто је разлика била мала (интервал поверења садржи нулу) овога пута директор је решио да продужи уговор обема компанија али само на годину дана и након тога понови истраживање.

Задатак 2.3.9. Једна феминистичка група започела је истраживање чији је циљ да се покаже да су студенти женског пола "вреднији". Један од параматера је и проценат студената који положи све испите које је предвиђено за 2. годину студија. Случајно је узето 200 студената женског и 150 студената мушког пола. Резултати анкете су следећи: 82 студената мушког и 70 студената женског пола су претходну годину "дали у року". Наћи 90% интервал поверења за разлику пропорција студената мушког и женског пола.

Задатак 2.3.10. Циљ једног истраживања је да се види да ли возачи жене или мушкарци више поштују пешаке на пешацком прелазу. Случајно јеодабрано неколико прометних раскрсница и бележени полови вожача који су прописно сачекали пешаке и оних који то нису и добијено је да је од 41 од 50 жена сачекало пешаке да пређу улицу, док је од 82 мушараца свега 40 то учинило. Одредити 90% интервал поверења за разлику вероватноћа између полова да се прописно испоштују пешаци на прелазу.

Задатак 2.3.11. Колике минималне узорке је потребно узети из две популације, под претпоставком да се узимају узорци истог обима, да би дужина интервала поверења за разлику пропорција била мања од 0.4?

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

Као и у случају Биномног модела, основа за заключивање у Пуасоновом моделу је Централна гранична теорема. Уколико су

$X_1, ..., X_n$ независне и једнако расподељене случајне величине са $\mathcal{P}(\lambda)$ онда

$$T_n = \frac{\bar{X}_n - \lambda}{\sqrt{\frac{\lambda}{n}}} \tag{2.1}$$

има, за велико n, нормалну $\mathcal{N}(0,1)$ расподелу. Тада важи:

$$\begin{aligned}\beta = P\{|T_n| \leq C\} = P\{T_n^2 \leq C^2\} = P\left\{(\bar{X}_n - \lambda)^2 \leq C^2 \frac{\lambda}{n}\right\} \\ = P\left\{\lambda^2 - \lambda\Big(\frac{C^2}{n} + 2\bar{X}_n\Big) + \bar{X}_n^2 \leq 0\right\}.\end{aligned}$$

Решење ове квадратне неједначине је тражени интервал поверења. Треба водити рачуна да је $\lambda > 0$ па у случају да то није испуњено интервал треба редуковати. У случају заиста великог обима узорка може се користити и стожерна величина

$$T_n = \frac{\bar{X}_n - \lambda}{\sqrt{\frac{\bar{X}_n}{n}}}, \tag{2.2}$$

и тада је интервал поверења

$$\left(\bar{X}_n - C\sqrt{\frac{\bar{X}_n}{n}}, \bar{X}_n + C\sqrt{\frac{\bar{X}_n}{n}}\right). \tag{2.3}$$

Пример 2.3.10. *Познато је да се број голова на фудбалској утакмици често може моделирати Пуасоновом расподелом $\mathcal{P}(\lambda)$, где је λ очекиван број голова на једној утакмици. Како би се нашао 95% интервал поверења за λ за утакмице које се одигравају у "нокаут" фази Лиге шампиона, узет је узорак од 20 утакмица у претходних 5 година и добијено да је $\hat{\lambda} = \bar{X}_{20} = 2.5$..*
Квадратна једначина коју треба решити је

$$\lambda^2 - \lambda \cdot 5.192 + 6.25 = 0.$$

Добија се интервал $[1.897, 3.295]$.
Да смо користили апроксимацију (2.3) *добили бисмо интервал* $[1.807, 3.193]$ *који је веома сличан претходном па видимо да ову апроксимацију има смисла користити и за узорке средњег обима.*

Задатак 2.3.12. Број позива Хитној помоћи у интервалу унапред одређене дужине, се често може моделирати Пуасоновом $\mathcal{P}(\lambda)$ расподелом. Одредити 90% интервал поверења за λ уколико је посматрани временски интервал ноћ у току радног дана, а узорак се састоји од случајно одабраних 30 дана и чији је просек 11.]

2.3.5 Случај два узорка

Нека су X и Y два обележја са Пуасоновим $\mathcal{P}(\lambda_1)$ и $\mathcal{P}(\lambda_2)$ расподелама, и нека су узорци који су нам на располагању обима n_1 и n_2. За велико n_1 и n_2 из Централне граничне теореме добијамо да

$$T_{n_1,n_2} = \frac{\bar{X}_{n_1} - \bar{Y}_{n_2} - (\lambda_1 - \lambda_2)}{\sqrt{\frac{\bar{X}_{n_1}}{n_1} + \frac{\bar{Y}_{n_2}}{n_2}}}, \tag{2.1}$$

има нормалну $\mathcal{N}(0,1)$ расподелу, па интервал поверења за $\lambda_1 - \lambda_2$ добијамо из услова $P\{|T_{n_1,n_2}| \leq C\} = \beta$. Као и до сада $C = \Phi^{-1}(\frac{1+\beta}{2})$ а одговарајући интервал

$$\left(\bar{X}_{n_1} - \bar{Y}_{n_2} - C\sqrt{\frac{\bar{X}_{n_1}}{n_1} + \frac{\bar{Y}_{n_2}}{n_2}}, \bar{X}_{n_1} - \bar{Y}_{n_2} + C\sqrt{\frac{\bar{X}_{n_1}}{n_1} + \frac{\bar{Y}_{n_2}}{n_2}} \right). \tag{2.2}$$

2.3.6 Интервал поверења за средњу вредност

Тражили смо интервале поверења за p у Биномном моделу, и за λ у Пуасоновом моделу. Оно што повезује та два примера је да се ради о параметрима који представљају средње вредности за посматрана обележја. Зато се природно намеће питање да ли исту идеју можемо да искористимо да нађемо интервал поверења за средњу вредност обележја из произвољне расподеле?

Тачкаста, непараметарска оцена за $m = EX$ је $\bar{X}_n$. Зато је природно да у стожерној величини за којом трагамо фигуришe разлика $\bar{X}_n - m$. На основу Централне граничне теореме знамо да, ако је $DX = \sigma^2 < \infty$ онда, за велико n

$$T_n = \frac{\bar{X}_n - m}{\frac{\sigma}{\sqrt{n}}}$$

има нормалну $\mathcal{N}(0,1)$ расподелу. Имајући у виду да је n велико, можемо непознато σ заменити постојаном оценом $\hat{\sigma} = \widetilde{S}_n$, и онда као и у случају Биномног и Пуасоновог модела, одредити интервал поверења за m.

На сличан начин се може конструисати и интервал поверења за разлику очекивања два обележја. Наравно, под претпоставком да на располоагању имамо узорке довољно великих обима.

Задатак 2.3.13. На основу п.с.у. великог обима из популације на којој обележје X има Геометријску $\mathcal{G}(p)$ расподелу, одредити $\beta\%$ интервал поверења за p.

Задатак 2.3.14. Потребно је наћи $\beta\%$ интервал поверења за λ у случају да X има $\mathcal{P}(\lambda)$ расподелу. Међутим, уместо уобичајног п.с.у. $X_1, ..., X_n$ на располагању имате само узорак одговарајућих индикатора $\mathrm{I}\{X_1 = 0\}, \mathrm{I}\{X_2 = 0\}, ..., \mathrm{I}\{X_n = 0\}$, односно само информацију да ли је X_i једнако нули или не. Можете претпоставити да је обим узорка довољно велик.

3

Тестирање статистичких хипотеза

До сада смо се бавили оцењивањем параметара и то је свакако један од најважнијих статистичких задатака. Њима сродан је проблем тестирања статистичких хипотеза о вредностима параметара (необавезно параметара расподела, може се радити и о непознатим функцијама). За разлику од класичног тестирања, код овог вида тестирања резултат је стохастичке природе па тако сваки закључак има и своју вероватноћу.

Основни састојци сваког статистичког теста су:

- Нулта хипотеза (H_0) и алтернативна хипотеза (H_1) (хипотеза која се прихвата уколико одбацујемо H_0); Одабир нулте и алтернативне хипотезе спада у дизајн експеримента и томе треба посветити посебну пажњу. Јако је важно да "знамо шта хоћемо" и то правилно формулишемо;
- Тест статистика - статистика[1] на основу чије реализоване вредности доносимо закључак;
- Критична област W (нека врста правила). Уколико реализована вредност тест статистике упадне у критичну област одбацујемо хипотезу;
- Вероватноћа грешке коју допуштамо.

[1] функција од узорка која не зависи од непознатих параметара

Најважније је да се добро поставе хипотезе јер од тога зависе сви даљи закључци. Оно што заправо желимо да покажемо је најбоље да буде у алтернативној хипотези, и то највише због тога што да бисмо нешто одбацили довољно је да је једном (применом једног статистичког теста) добијемо негативан резултат тестирања, док да бисмо потврдили хипотезу потребно је то урадити са доста тестова. Тако да се тестирање заправо врши да би се одбацила нулта хипотеза у корист прихватања алтернативне хипотезе. Због тога се алтернативна хипотеза често назива *радном хипотезом.*

Пример 3.0.1. *Сматра се да студенти Математичког факултета спадају у натпросечне грађане. С циљем да се ове тврдње оправдају насумично је одабрано 30 студената Математичког факултета и измерен им је IQ.*

У овој ситуацији је природно да нулта хипотеза буде да је просечан IQ студената Математичког факултета 100 против алтернативе да је већи од 100.

Приликом статистичког закључивања могуће је направити грешке. То је илустровано у табели 3.1. Знаком " $-$ " означена је грешка у закључивању, а знак " $+$ " представља одсуство грешке.

H_0	тачна	нетачна
прихватимо	$+$	$-$
одбацимо	$-$	$+$

Табела 3.1: Резултат статистичког тестирања

Уколико одбацимо нулту хипотезу која је тачна направили смо грешку прве врсте. Уколико не одбацимо нетачну нулту хипотезу направили смо грешку друге врсте. Вероватноћа грешке прве врсте се најчешће означава са α. Вероватноћа грешке друге врсте се означава са β. $1-\beta$ представља моћ теста. Тест је моћнији уколико боље одбацује нетачне хипотезе.

Добар пример за илустрацију врсте грешака и њиховог значаја је суђење оптуженику при чему ако се докаже да је крив, следује му смртна казна. Свако је невин док се не докаже супротно. Дакле, H_0 је да је оптужени невин, а H_1 да је крив. Грешка прве врсте би била да невин човек страда, док би грешка друге врсте била да је кривац на слободи.

Задатак 3.0.1. Уколико је нулта хипотеза да особа није заражена вирусом SARS − CoV − 2 а алтернативна да јесте, објаснити последице грешке прве и последице грешке друге врсте.

Још треба напоменути да хипотезе могу бити просте и сложене. Просте су оне за које је скуп одређен хипотезом једночлан, док ако је вишечлан говоримо о сложеним хипотезама. У случају сложених нултих хипотеза важно је још увести два нова појма, *ниво значајности теста* и *мера теста* чија се дефиниција упрошћава у случају простих хипотеза.
Мера теста је α за које је $\sup_{H_0} P\{\text{грешка } I \text{ врсте}\} = \alpha$, док је ниво значајности теста задато ограничење за вероватноћу грешке прве врсте. Из дефиниције видимо да је мера теста увек мања или једнака задатком нивоу значајности који се у пракси најчешће исто означава са α. У случају просте хипотезе, уколико је расподела тест статистике под нултом хипотезом непрекидна, онда се мера теста, ниво значајности теста и вероватноћа грешке прве врсте поклапају. Ако расподела није непрекидна онда се свакако може десити да се мера теста и задати ниво значајности не поклапају.

Ниво значајности теста се увек задаје пре тестирања. Најчешће вредности су $0.1, 0.05$ и 0.01. Дакле, вероватноћа грешке прве врсте је контролисана! Следећи корак је да се за задати ниво значајности теста одреди облик критичне области. Обично или велике или мале вредности тест статистике упућују на одбацивање хипотезе па су најчешћи облици критичних области $W = \{T_n \leq C\}$, или $W = \{T_n \geq C\}$, или $W = \{T_n \leq C_1\} \cup \{T_n \geq C_2\}$. За одређивање константе у критичним областима је потребно одредити расподелу тест статистике под нултом хипотезом (уколико је хипотеза сложена онда узорак одређен нултом хипотезом може имати различите расподеле па и тест статистика) јер су константе одређене условом да је грешка прве врсте ограничена нивоом значајности теста α. Уколико је расподела тест статистике под нултом хипотезом једнозначна, непрекидна и знамо је, онда је константе могуће одредити. Међутим, чак и у том најједноставнијем случају, расподелу тест статистике под H_0 није увек једноставно одредити, а некада чак није ни могуће. Зато се тада често расподела оцењује Монте Карло методама, слично као кад смо испитивали квалитет оцена.

Алгоритам је следећи:

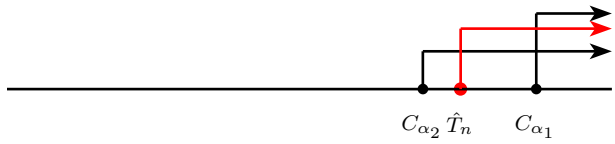
1. Претпоставимо облик критичне области;
2. Генеришемо узорак из расподеле која је одређена нултом хипотезом $\boldsymbol{X} = (X_1, ..., X_n)$;
3. Одредимо вредност тест статистике $T_n = T_n(\boldsymbol{X})$.
4. Поновимо кораке 2 и 3. N пута. На тај начин добијамо низ вредности тест статистике $T_n^{(1)}, ..., T_n^{(N)}$ на основу ког одредимо емпиријску расподелу тест статистике;
5. На основу низа из корака 4 одредимо критичну област. На пример, ако је критична областоблика $W = \{T_n \geq C\}$ и ако знамо да је $P_{H_0}\{T_n \geq C\} = \alpha$, онда је емпиријски квантил реда $1-\alpha$, односно $F_N^{-1}(1-\alpha)$, где је F_N емпиријска функција расподеле за добијени "узорак" вредности тест статистике.

За одређивање моћи теста потребно нам је да знамо расподелу статистике кад је узорак из расподеле одређене алтернативном хипотезом. Уколико расподеле не знамо, можемо је оценити применом Монте Карло метода. Међутим, како је моћ теста $P_{H_1}\{T_n \in W\} = E_{H_1}(\mathrm{I}\{T \in W\})$, онда не морамо оценити читаву расподелу статистике већ само очекивања индикатора да је тест статистика упала у критичну област. Очекивање оцењујемо са просечним уделом успешно остварних експеримената, односно са s/N, где је s број експеримената у којима је статистика одбачена (упала у критичну област).

Још један битан појам у статистичким тестирањима је p-вредност теста. То је најмањи ниво значајности теста за који ћемо, на основу датог узорка, одбацити H_0. Тако да, ако је $p \leq \alpha$ онда одбацујемо хипотезу, у супротном је прихватамо. Приметимо да p-вредност зависи од узорка, па је то заправо једна случајна величина. Један, доста интуитиван начин да се интерпретира p-вредност је да је то заправо вероватноћа да се реализује узорак "гори од нашег", па уколико смо добили баш мали број то значи да је наш узорак заправо "лош" и да треба одбацити хипотезу. Овај број је стандардни излаз за већину тестова у разним софтверским алатима тако да се статистичко закључивање у великом броју

случаја врши управо на основу добијене p-вредности. У наредном примеру ћемо приказати како у неким конкретним ситуацијама можемо израчунати p-вредност.

Пример 3.0.2. *Претпоставимо да имамо тест заснован на статистици T_n као и да је критична област за тестирање облика $W = \{T_n \geq C\}$. Реализована вредност тест статистике је $\hat{T}_n$. На слици 3.0.2 приказане су критичне области за $\alpha_1 < \alpha_2$. Види се да за ниво значајности α_1 реализована вредност тест статистике не упада у критичну област, док кад се ниво значајности повећа и кад је нпр. једнак α_2 упада. Нама је циљ да нађемо најмањи виво значајности теста за који ће $\hat{T}_n$ (означена црвеном бојом) упасти у критичну област . Јасно је да је α баш једнако $P\{T \geq \hat{T}_n\}$, што је и тражена p-вредност. Напомињемо да се ова вероватноћа рачуна у случају важења нулте хипотезе.*

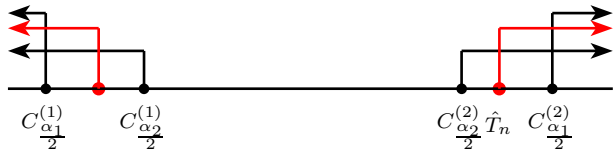


Слика 3.1: Одређивање p-вредности у случају да је $W = \{T_n \geq C\}$

Уколико је критична област облика $W = \{T_n \leq C\}$ на сличан начин као и малопре, добијамо да је p-вредност једнака $P\{T_n \leq \hat{T}_n\}$.

Уколико је критична област облика $W = \{T_n \leq C_1\} \cup \{T_n \geq C_2\}$ прво је важно напоменути да, како не желимо да фаворизујемо неки део критичне области, оне се праве тако да под нултом хипотезом буду једнако вероватне. Због тога се константе C_1 и C_2 одређене условом $P_{H_0}\{T_n \leq C_1\} = \frac{\alpha}{2} = P_{H_0}\{T_n \geq C_2\}$. На слици 3.0.2 се види како изгледају критичне области за $\alpha_1 < \alpha_2$ и као и у претходном случају, за α_1 се прихвата, а за α_2 одбацује хипотеза. Треба да одредимо гранични ниво значајности. Дакле,

како повећавамо α *границе ће се приближавати реализованој вредности тест статистике. Уколико је* $\hat{T}_n$ *ближа десном делу критичне области (као што је то случај на слици 3.0.2) онда ће за гранично* α *важити* $\frac{\alpha}{2} = P\{T_n \geq \hat{T}_n\}$ *у супротном је* $\frac{\alpha}{2} = P\{T_n \leq \hat{T}_n\}$*, па закључујемо да је p-вредност* $2\min(P\{T_n \geq \hat{T}_n\}, P\{T_n \leq \hat{T}_n\})$*. Уколико је расподела тест статистике под нултом хипотезом симетрична око нуле, онда се ова p-вредност своди на* $2P\{T_n \geq |\hat{T}_n|\} = P\{|T_n| \geq |\hat{T}_n|\}$.



Слика 3.2: Одређивање p-вредности у случају да је $W = \{T_n \leq C_1\} \cup \{T_n \geq C_2\}$

Постоји много начина да класификујемо статистичке тестове. Најприроднији је по томе да ли је расподела узорка позната до на непознат параметар или не, односно на *параметарске* и *непараметарске* тестове. Непараметарски тестови се могу поделити по типу хипотеза које се тестирају. Најпознатији су следећи:

- тестови сагласности са расподелом;
- тестови симетрије (да ли је расподела симетрична);
- тестови независности два или више обележја;
- тестови о једнакој расподељености два узорка.

Сваки од поменутих типова тестова ћемо у наредним поглављима детаљније описати.

3.1 Параметарски тестови

Нулта хипотеза код ових тестова се може приказати у облику

$$H_0 : \theta \in \Theta_0,$$

а алтернативна

$$H_1 : \theta \in \Theta_1.$$

Уколико је Θ_0 једночлан, одосно $\Theta_0 = \{\theta_0\}$ кажемо да је нулта хипотеза *проста*, у супротном да је *сложена*.
Овакви тестови се најчешће примењују у следећим ситуацијама:

- познато нам је да обележје X има неку расподелу $F(\theta)$, желимо да проверимо да је $\theta = \theta_0$ (односно да покажемо да то не важи);

- имамо неко вишедимензионо обележје (X, Y) за које знамо облик зависности Y од X до на непознате параметре;

- посматрамо два или више обележја за која знамо да имају расподелу из исте класе расподела, које зависе од непознатих параметара и желимо да тестирамо хипотезу да између тих параметара постоји нека функционална веза (нпр. да су сви једнаки).

У наредним одељцима представићемо неке од најважнијих параметарских тестова.

3.1.1 Тестови у нормалном моделу

Претпоставимо да обележје X има нормалну $\mathcal{N}(m, \sigma^2)$ расподелу, при чему на располагању имамо прост случајан узорак $X_1, ..., X_n$. Често се јавља потреба за тестирањем да m има баш неку одређену вредност m_0 (односно да нема ту вредност). Дакле, потребно је тестирати $H_0 : m = m_0$ против неке алтернативе. Сам облик алтернативе ће утицати на облик критичне области. Формално, за тест статистику можемо узети било коју статистику од узорка за коју знамо како да одредимо расподелу уколико важи нулта хипотеза. Поред тога,, пожељно је да уколико важи нулта хипотеза тест

статистика са вероватноћом α упада критичну област, а да уколико узорак није из нулте хипотезе да је вероватноћа да упадне у критичну област већа од α, и да са што већом вероватноћом одбаци нетачну нулту хипотеза (има велику моћ), као и да моћ достиже вредност 1 за довољно велики обим узорка (постојан је). У даљем тексту, приказаћемо тест статистике које се најчешће користе у овом случају.

Уколико је σ^2 познато онда за тестирање можемо користити тест статистику

$$T_n = \frac{\bar{X}_n - m_0}{\frac{\sigma}{\sqrt{n}}}, \tag{3.1}$$

за коју знамо да, уколико је тачна нулта хипотеза, да је узорак из нормалне $\mathcal{N}(m_0, \sigma^2)$ расподеле, има нормалну $\mathcal{N}(0,1)$ расподелу. Уколико σ^2 није познато онда за тестирање можемо користити тест статистику

$$T_n = \frac{\bar{X}_n - m_0}{\frac{\tilde{S}_n}{\sqrt{n}}}, \tag{3.2}$$

која у случају нулте хипотезе, има Студентову t_{n-1} расподелу.

Најчешће три алтернативне хипотезе су:

- $$H_1 : \ m \neq m_0; \tag{3.3}$$
- $$H_1 : \ m < m_0; \tag{3.4}$$
- $$H_1 : \ m > m_0. \tag{3.5}$$

Алтерхнативна хипотеза је јако битна за одређивање облика критичне области. Како је $\bar{X}_n$ је оцена за m, ако је H_0 није тачна T_n неће бити "довољно блиско нули". У случају (3.3) природно је да је критична област облика $W = \{|T_n| \geq C\}$ јер је расподела тест статистике под нултом хипотезом симетрична, а много мале и много велике вредности тест статистике упућују на ту алтернативну хипотезу. Константу C одређујемо из услова $P_{H_0}\{|T_n| \geq C\} = \alpha$. Овај услов се може записати у облику $\Phi(C) = 1 - \frac{\alpha}{2}$, па је $C = \Phi^{-1}(1 - \frac{\alpha}{2})$.

Сличним разматрањем закључујемо да је облик критичне области за алтернативну хипотезу (3.4) $W = \{T_n \leq C\}$ па је $C = \Phi^{-1}(\alpha)$, док је за алтернативну хипотезу (3.5) $W = \{T_n > C\}$ и $C = \Phi^{-1}(1-\alpha)$.

Приметимо да се критична област може добити инвертовањем интервала поверења чији је ниво $\beta = 1 - \alpha$. Свакако важи и обрнут закључак, да се интервал поверења може добити инвертовањем критичне области. На пример, двострани $(1-\alpha)\%$ интервал поверења за m_0 је $(\bar{X}_n - C\frac{\sigma}{\sqrt{n}}, \bar{X}_n + C\frac{\sigma}{\sqrt{n}})$, где је $C = \Phi^{-1}(\frac{1+1-\alpha}{2}) = \Phi^{-1}(1-\frac{\alpha}{2})$. Критична област која одговара овом интервалу поверења је

$$W = \mathbb{R} \setminus (\bar{X}_n - C\frac{\sigma}{\sqrt{n}}, \bar{X}_n + C\frac{\sigma}{\sqrt{n}}) = \{|T_n| \geq C\}.$$

Последња једнакост следи из следећег низа еквиваленција:

$$m_0 \leq \bar{X}_n - C\frac{\sigma}{\sqrt{n}} \Longleftrightarrow T_n \geq C$$
$$m_0 \geq \bar{X}_n + C\frac{\sigma}{\sqrt{n}} \Longleftrightarrow T_n \leq -C.$$

Одавде закључујемо да све функције од узорка (стожерне величине) које смо користили у претходном поглављу можемо користити као тест статистике за тестирање да посматрани параметри имају неку одређену вредност.

Пример 3.1.1. *У току ванредног стања већина* IT *компанија је прешла на такозвани "рад од куће". Власници компанија су стекли утисак да радници више раде кад су код куће и да би можда требало да наставе са таквим видом рада и након завршетка ванредног стања. Међутим како су свесни да је на сам рад утицало много фактора, у неколико компанија је спроведено мини истраживање, пре кога је установљено да је просечно ефективно радно време "у канцеларији" пре преласка на "рад од куће" било 6.2 сата. Случајно је одабрано неколико тимова којима је речено да настављају рад од куће и након укидања ванредног стања, и бележено ефективно радно време у току наредних 5 дана. Тако је добијен узорак обима 100 и узорачка средина* $\bar{x}_{100} = 6.4$ *и узорачко одступање* $\widetilde{s}_{100} = 0.5$. *На основу разултата треба закључити да ли да се настави рад од куће, при чему је дозвољена греска* $\alpha = 0.05$.

Први корак је да поставимо хипотезе. С обзром на то да власници компаниј желе да провере свој закључак, имамо да је $H_0 : \ m = 6.2$ *против* $H_1 : \ m > 6.2$. *Како је претпоставка да се радно време може моделирати нормалном расподелом смислена, и при том још имамо и велики узорак на располагању, можемо користити статистику* (3.2), *за* $m_0 = 6.2$, *која ако је нулта хипотеза тачна има* t_{99} *расподелу која се због великог броја степена слободи може апроксимирати стандардном нормалном расподелом. Критична област је облика* $W = \{T_{100} \geq C\}$ *где је* $C = \Phi^{-1}(0.95) = 1.64$, *док је реализована вредност тест статистике*

$$\hat{T}_{100} = \frac{6.4 - 6.2}{\frac{0.5}{\sqrt{100}}} = 4,$$

па како припада критичној области закључујемо да хипотезу треба одбацити и да, посматрано из угла власника компанија, има смисла наставити са радом од куће.

Задатак 3.1.1. Одредити p-вредност теста из примера 3.1.1.

Задатак 3.1.2. У једном истраживању проучаван је утицај уводјења ванредног стања на број продатих кесица сувог квасца у продавницама средње величине. Пре увођења ванредног стања у једној таквој продавници просечан број продатих кесица у току једног дана је био 11. Случајно је одабрано 13 дана и просечан број је био 19, док је узорачка дисперзија 2.1. Да ли можете закључити да је повећање статистички значајно? Одговор образложити. Можете сматрати да број продатих кесица се може добро моделирати нормалном расподелом.

Задатак 3.1.3. Одредити интервал поверења чијим инвертовањем се добија критична област за тестирање $H_0 : \ m = m_0$ против алтернативе 3.4 на основу статистике (3.1).

Сада ћемо одредити моћ теста када је $H_0 : m = m_0$ и $H_1 : m \neq m_0$. Означимо са $M(\theta)$ моћ теста када је $m = \theta$. Тада је

$$
\begin{aligned}
M(\theta) &= P_\theta\left\{\Big|\frac{\bar{X}_n - m_0}{\frac{\sigma}{\sqrt{n}}}\Big| > C\right\} = 1 - P_\theta\left\{\Big|\frac{\bar{X}_n - m_0}{\frac{\sigma}{n}}\Big| \le C\right\} \\
&= 1 - P_\theta\left\{-C \le \frac{\bar{X}_n - m_0}{\frac{\sigma}{\sqrt{n}}} \le C\right\} \\
&= 1 - P_\theta\left\{m_0 - \frac{C\sigma}{\sqrt{n}} \le \bar{X}_n \le m_0 + \frac{C\sigma}{\sqrt{n}}\right\} \\
&= 1 - P_\theta\left\{\frac{m_0 - \frac{C\sigma}{\sqrt{n}} - \theta}{\frac{\sigma}{\sqrt{n}}} \le \frac{\bar{X}_n - \theta}{\frac{\sigma}{\sqrt{n}}} \le \frac{m_0 + \frac{C\sigma}{\sqrt{n}} - \theta}{\frac{\sigma}{\sqrt{n}}}\right\} \\
&= 1 - P_\theta\left\{-C + \frac{m_0 - \theta}{\frac{\sigma}{\sqrt{n}}} \le \frac{\bar{X}_n - \theta}{\frac{\sigma}{\sqrt{n}}} \le C + \frac{m_0 - \theta}{\frac{\sigma}{\sqrt{n}}}\right\} \\
&= 1 - \Phi\Big(C + \frac{m_0 - \theta}{\frac{\sigma}{\sqrt{n}}}\Big) + \Phi\Big(-C + \frac{m_0 - \theta}{\frac{\sigma}{\sqrt{n}}}\Big). \qquad (3.1)
\end{aligned}
$$

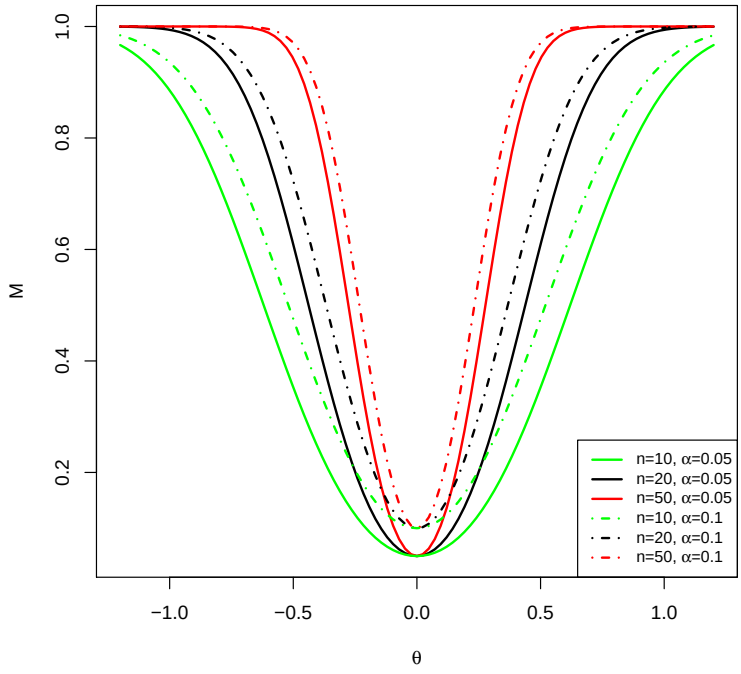
Уколико је $\theta = m_0$ онда је $M(\theta) = \alpha$ што и треба да важи јер "смо тада у области важења нулте хипотезе". Уколико је $\theta > m_0$ онда једнакост (3.1) се може написати на следећи начин:

$$
\begin{aligned}
M(\theta) &= 1 + \underbrace{\Big[\Phi(C) - \Phi\Big(C + \frac{m_0 - \theta}{\frac{\sigma}{\sqrt{n}}}\Big)\Big]}_{A} - \Phi(C) \\
&\quad - \underbrace{\Big[\Phi(-C) - \Phi\Big(-C + \frac{m_0 - \theta}{\frac{\sigma}{\sqrt{n}}}\Big)\Big]}_{B} + \Phi(-C) \\
&= \alpha + A - B.
\end{aligned}
$$

Из облика густине стандардне нормалне расподеле лако се закључује да је $A \ge B$ па добијамо да је $M(\theta) \ge \alpha$. Уколико је $\theta < m_0$ онда једнакост (3.1) можемо написати у облику

$$
\begin{aligned}
M(\theta) &= 1 - \underbrace{\left[\Phi\Big(C + \frac{m_0 - \theta}{\frac{\sigma}{\sqrt{n}}}\Big) - \Phi(C)\right]}_{D} - \Phi(C) \\
&\quad + \underbrace{\left[\Phi\Big(-C + \frac{m_0 - \theta}{\frac{\sigma}{\sqrt{n}}}\Big) - \Phi(-C)\right]}_{E} + \Phi(-C) \\
&= \alpha - D + E.
\end{aligned}
$$

Из облика густине стандардне расподеле се закључује да је $E \geq D$ па и у овом случају добијамо да је $M(\theta) \geq \alpha$.



Слика 3.3: Моћ двостраног теста за $H_0 : m = 0$, кад је $\sigma = 1$ познато

Из једнакости (3.1) се могу закључити још неке особине овог теста. Наиме, моћ је растућа функција по n, и кад $n \to \infty$, $M(\theta) \to 1$. Такође, смањујући грешку прве врсте смањујемо и моћ (види слику 3.3). Исти закључци се могу добити и у случају једностраних алтернативних хипотеза.

Једна могућност да се тестира нулта хипотеза да је $\sigma^2 = \sigma_0^2$ је да се искористи тест статистика

$$T_n = \frac{(n-1)\tilde{S}_n^2}{\sigma_0^2} \tag{3.2}$$

за коју знамо да, уколико је нулта хипотеза тачна, има χ^2_{n-1} расподелу.

Најчешће алтернативе су

- $$H_1 : \sigma^2 \neq \sigma_0^2; \tag{3.3}$$
- $$H_1 : \sigma^2 < \sigma_0^2; \tag{3.4}$$
- $$H_1 : \sigma^2 > \sigma_0^2. \tag{3.5}$$

Посмтрајмо алтернативу (3.3). Уколико је она тачна $\tilde{S}_n^2$ ће бити или значајно мање или значајно веће од σ_0^2 па је природно да критична област буде облика $W = \{T_n \leq C_1\} \cup \{T_n \geq C_2\}$. Константе C_1 и C_2 ћемо бирати тако да је $\frac{\alpha}{2} = P_{H_0}\{T_n \leq C_1\} = P_{H_0}\{T_n \geq C_2\}$. Одавде је $C_1 = F^{-1}_{\chi^2_{n-1}}(\frac{\alpha}{2})$ и $C_2 = F^{-1}_{\chi^2_{n-1}}(1-\frac{\alpha}{2})$.
Уколико се ради о алтернативама (3.4), или (3.5), критичне области су редом облика $\{T_n \leq C\}$, односно $\{T_n \geq C\}$.

Пример 3.1.2. *Компанија која се бави производњом батерија тврди да рок употребе батерије има нормалну расподелу са дисперзијом* 0.9^2. *Како би се проверила тврдња произвођача, узето је 10 батерија и одређено њихово време трајања. Добијено је да је* $\tilde{s}_{10} = 1.2$. *Да ли се на основу тога може закључити да време трајања батерије има дисперзију већу него што произвођач тврди? Дозвољена вероватноћа грешке прве врсте је* $\alpha = 0.05$.

Из постављеног проблема може се закључити да је $H_0 : \sigma^2 = 0.9^2$ *против алтернативе да је* $\sigma^2 > 0.9^2$. *Критична област за тестирање је* $\{T_{10} \geq C\}$ *при чему је* $C = F^{-1}_{\chi^2_9}(0.95) = 16.92$. *Реализована вредност статистике је* $\hat{T}_{10} = \frac{9 \cdot 1.2^2}{0.9^2} = 16$, *па не упада у критичну област. Закључак је да не одбацујемо тврдњу произвођача.*

До истог закључка смо могли да дођемо одређивањем $p-$*вредности теста која у овом случају износи* $P_{H_0}\{T_{10} \geq \hat{T}_{10}\} = P_{H_0}\{T_{10} \geq 16\} = 0.07$.

Из овог примера можемо да видимо да ниво значајности теста, одређен пре почетка тестирања, може одиграти кључну улогу у закључивању. Да смо допустили веће α од 0.07 одбацили бисмо хипотезу.

Случај два узорка

Претпоставимо да имамо два независна узорка $X_1, ..., X_{n_1}$ и $Y_1, ..., Y_{n_2}$ која су из нормалних $\mathcal{N}(m_1, \sigma_1^2)$ и $\mathcal{N}(m_2, \sigma_2^2)$. Најчешће желимо да тестрамо хипотезе $H_0 : \ m_1 - m_2 = m_0$ и $H_0 : \ \frac{\sigma_1^2}{\sigma_2^2} = A$. С обзиром на то да смо објаснили везу између интервала поверења и тест статистика, природно се намећу статистике које смо користили за прављење интервала поверења за $m_1 - m_2$ и $\frac{\sigma_1^2}{\sigma_2^2}$. У случају $H_0 : \ m_1 - m_2 = 0$ три случаја:

1. σ_1^2 и σ_2^2 су познати параметри. Тада користимо статистику

$$T_{n_1,n_2} = \frac{\bar{X}_{n_1} - \bar{X}_{n_2} - m_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}.$$

   Уколико је нулта хипотеза тачна, на основу теореме 2.3.2, T_{n_1,n_2} има нормалну $\mathcal{N}(0,1)$ расподелу.

2. $\sigma_1^2 = \sigma_2^2 = \sigma^2$, при чему σ^2 није познато. Тада користимо статистику

$$T_{n_1,n_2} = \frac{\bar{X}_{n_1} - \bar{X}_{n_2} - m_0}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \tag{3.6}$$

где је $S^2 = \frac{(n_1-1)\widetilde{S}^2_{n_1}+(n_2-1)\widetilde{S}^2_{n_2}}{n_1+n_2-2}$, за коју знамо на основу теореме 2.3.2 да, уколико је H_0 тачна, има Студентову $t_{n_1+n_2-2}$ расподелу.

3. $\sigma_1^2 \neq \sigma_2^2$ и оба параметра су непозната. Тада користимо

$$T_{n_1,n_2} = \frac{\bar{X}_{n_1} - \bar{Y}_{n_2} - m_0}{\sqrt{\frac{\widetilde{S}^2_{n_1}}{n_1} + \frac{\widetilde{S}^2_{n_2}}{n_2}}}, \tag{3.7}$$

која, ако је H_0 тачна, на основу теореме 2.3.2 има t_ν расподелу, где је

$$\nu = \frac{\left(\frac{\widetilde{S}^2_{n_1}}{n_1} + \frac{\widetilde{S}^2_{n_2}}{n_2}\right)^2}{\frac{\left(\frac{\widetilde{S}^2_{n_1}}{n_1}\right)^2}{n_1-1} + \frac{\left(\frac{\widetilde{S}^2_{n_2}}{n_2}\right)^2}{n_2-1}}. \tag{3.8}$$

У сва три случаја, ако је $H_1 : m_1 - m_2 \neq m_0$ критична област је $W = \{|T_{n_1,n_2}| \geq C\}$, ако је $H_1 : m_1 - m_2 < m_0$ критична област је $W = \{T_{n_1,n_2} \leq C\}$, а ако је $H_1 : m_1 - m_2 > m_0$ онда је критична област $W = \{T_{n_1,n_2} \geq C\}$. Константе C се одређују из услова да је $P_{H_0}\{T_n \in W\} = \alpha$.

Да бисмо одлучили да ли да користимо статистику за случај кад су дисперзије једнаке или за случај кад су различите треба да тестирамо $H_0 : \ \sigma_1^2 = \sigma_2^2$. Приметимо да је, грешка друге врсте у овом тестирању да су дисперзије различите а да ми ту хипотезу не одбацимо и она нам је у овом конкретном случају битна јер је последица одлуке који ћемо тест користити за даље тестирање. Да бисмо смањили ту грешку допустићемо већу грешку прве врсте, односно, уместо уобичајног $\alpha = 0.05$ тестирање можемо извршити за $\alpha = 0.1$ или чак за $\alpha = 0.2$.

За тестирање $H_0 : \ \frac{\sigma_1^2}{\sigma_2^2} = 1$ можемо користити статистику

$$T_{n_1,n_2} = \frac{\tilde{S}^2_{n_1}}{\tilde{S}^2_{n_2}}, \tag{3.9}$$

за коју знамо да, уколико је H_0 тачно има Фишерову F_{n_1-1,n_2-1} расподелу. У случају да је $H_0: \frac{\sigma_1^2}{\sigma_2^2} = A$ статистика коју користимо се лако модеификује.

Критична област у случају алтернативе $H_1: \sigma_1^2 \neq \sigma_2^2$ је облика $W = \{T_{n_1,n_2} \leq C_1\} \cup \{T_{n_1,n_2} \geq C_2\}$ а константе C_1 и C_2 се одређују из услова $P_{H_0}\{T_{n_1,n_2} \leq C_1\} = P\{T_{n_1,n_2} \geq C_2\} = \frac{\alpha}{2}$. Уколико је $H_1: \sigma_1^2 > \sigma_2^2$ онда је $W = \{T_{n_1,n_2} \geq C\}$, а ако је $H_1: \sigma_1^2 < \sigma_2^2$ онда је $W = \{T_{n_1,n_2} \leq C\}$.

Пример 3.1.3. *Желимо да тестирамо да ли постоји значајна разлика (уз допуштену вероватноћу грешке 0.05) између просечног систолног притиска код мушкараца и жена, и да је извршено истраживање описано у примеру 2.3.7. Дакле, нулта хипотеза је $m_1 - m_2 = 0$ против алтернативне $m_1 - m_2 \neq 0$. До закључка можемо одмах доћи имајући у виду поменуту везу између интервала поверења и тестирања хипотеза. Како 95% интервал поверења одређен у примеру 2.3.7 садржи $m_0 = 0$ немамо разлога да одбацимо хипотезу. Из методолошких разлога приказаћемо и цео процес тестирања уколико не желимо да се ослањамо на претходно добијене резултате .*
Како немамо податке о дисперзијама посматраног обележја, прво ћемо тестирати $H_0: \sigma_1^2 = \sigma_2^2$, користећи статистику (3.9). Њена реализована вредност је $\hat{T}_{6,4} = \frac{9.5^2}{11.4^2} = 0.69$ па је p-вредност $2\min(P\{T_{6,4} \geq 0.69\}, P\{T_{6,4} \leq 0.69\}) = 2F_{\mathcal{F}_{5,3}}(0.69) = 0.67$ па хипотезу не одбацујемо и у даљем тестирању користимо тест статистику (3.6). Даље је $S = 10.25$ па је реализована вредност тест статистике $\hat{T}_{n_1,n_2} = \frac{126-115}{10.25\cdot\sqrt{\frac{1}{6}+\frac{1}{4}}} = 1.66$. Одавде се добија да је p-вредност $2(1 - F_{t_8}(1.66)) = 0.13$ па немамо довољно основа да хипотезу одбацимо.

Спарени тест

Могуће је да се деси да посматрана обележја нису независна,и да имамо п.с. узорак парова $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$ и нека је $EX_i = m_1$ и $EY_i = m_2$. Претпостављамо и да $D_i = X_i - Y_i$ има $\mathcal{N}(m_D, \sigma_D^2)$ расподелу, при чему σ_D^2 је непознато. Желимо да тестирамо $H_0: m_1 - m - 2 = m_0$ против неке од стандардних

алтернатива. Ово је еквивалентно са $H_0 : m_D = m_0$ па за ово тестирање можемо искористити

$$T_n = \frac{\bar{D}_n - m_0}{\frac{\tilde{S}_n}{\sqrt{n}}}, \tag{3.10}$$

за коју знамо да ако важи H_0, има t_{n-1} расподелу (за више детаља видети (3.2)).
Овакав тест се често примењује како би се установили ефекти терапије. Илустроваћемо то следећим примером.

Пример 3.1.4. *Једна фармацеутска компанија је направила препарат за који гарантује да смањује ниво холестерола у крви константном употребом у периоду од две недеље. Како би се провериле тврдње узет је узорак од 20 особа, измерен ниво холестерола у крви, а затим и након третмана. Добијено је да је просечна разлика* $\bar{d}_{20} = 1.2$ *и узорачко одсупање разлика* $s_{20} = 0.1$. *Урадићемо тестирање са нивоом значајности* $\alpha = 0.05$.

Нека је m_1 *је ниво холестерола пре третмана, а* m_2 *је ниво холестерола после третмана. Имамо да је* $H_0 : \ m_D = m_2 - m_1 = 0$ *против* $H_1 : \ m_1 - m_2 > 0$, *па је критична област облика* $W = \{T_{20} \geq C\}$. *Добијамо да је реализована вредност тест статистике* $\hat{T}_{20} = \frac{1.2}{\frac{0.1}{\sqrt{20}}} = 1.34$, *па можемо израчунати p-вредност. Добијамо да је она једнака p-вредност је* $P_{H_0}\{T \geq 1.34\} = 1 - F_{t_{19}}(1.34) = 0.098$ *па нећемо одбацити хипотезу, односно компанија нема "основане" гаранције.*

Задатак 3.1.4. На тржишту се појавио нови препрат за мршављење, за кога произвођач тврди да се редовном употребом од 14 дана телесна тежина смањује бар 4 килограма. Тврдња је поткрепљена студијом у коме је учествовало 10 жена са прекомерном телесном тежином до 20 килограма и добило се да је просечна разлика -4.2 килограма, уз узорачку стандардно одступање 0.5 килограма. Уз претпоставку о нормалности разлика тежина тестирати тврдњу произвођача са нивоом значајности $\alpha = 0.05$.

3.1.2 Тестови у Биномном моделу

Нека је X индикатор са вероватноћом успеха p. Желимо да тестирамо $H_0 : p = p_0$. Као и до сада разматраћемо алтернативне

хипотезе

- $$H_1 : p \neq p_0; \tag{3.1}$$

- $$H_1 : p < p_0; \tag{3.2}$$

- $$H_1 : p > p_0. \tag{3.3}$$

Уколико имамо велики узорак можемо искористити статистику

$$T_n = \frac{\bar{X}_n - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}, \tag{3.4}$$

и чињеницу да, уколико је H_0 тачна T_n има нормалну $\mathcal{N}(0,1)$. Имајући у виду да је $\bar{X}_n$ тачкаста оцена за p, у случају алтернативне хипотезе (3.1) критична област је облика $W = \{|T_n| \geq C\}$, у случају алтернативне хипотезе (3.2) је $W = \{T_n \leq C\}$, док је у случају алтернативне хипотезе (3.3) критична област облика $W = \{T_n \geq C\}$.

Пример 3.1.5. *Како би се показало да је вероватноћа да се човек излечи од сезонског грипа за мање од недељу дана, већа од 0.5, узет је узорак од 500 грађана који су имали сезонски грип ове године. Од тога је 320 пријавило излечење у поменутом временском року. Сада је природно* $H_0 :; p = 0.5$ *против* $H_1 : p > 0.5$. *Добијамо да је реализована вредност тест статистике* $\hat{T}_{500} = \frac{\frac{320}{500} - 0.5}{\frac{0.5}{\sqrt{500}}} = 6.26$ *па је p-вредност теста* $P_{H_0}\{T_{500} \geq 6.26\} = 1 - \Phi(6.26) \approx 0$. *Одавде закључујемо да треба одбацити* H_0, *односно да је вероватноћа да се човек излечи од сезонског грипа за мање од недељу дана, већа од 0.5.*

Једно од могућих питања је и, ако је дошло до грешке у закључку ког типа је грешка. С обзиром на то да смо овде нулту хипотезу одбацили, грешка до које је могло доћи је одбацивање тачне нулте хипотезе, односно ради се о грешци прве врсте.

Задатак 3.1.5. Упоредити квалитет теста за различите обиме узорка (3.4) са тестом заснованим на статистици

$$T_n' = \frac{\bar{X}_n - p_0}{\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}},$$

који се исто користи за тестирање $H_0 : \ p = p_0$ против $H_1 \ p \neq p_0$. Приликом испитивања проверите да ли је вероватноћа грешке добро исконтролисана. Поред тога одредите моћ за $p \neq p_0$. Уколико је потребно, можете до закључака доћи и емпиријски, коришћењем Монте Карло метода.

Када немамо мали узорак можемо да искористимо да $S_n = X_1 + \cdot + X_n$, у случају нулте хипотезе, има Биномну $\mathcal{B}(n, p_0)$. Уколико је алтернативна хипотеза (3.1) критчна област ће бити облика $W = \{S_n \leq C_1\} \cup \{S_n \geq C_2\}$ а C_1 и C_2 одређујемо тако да је

$$\sum_{i=0}^{C_1} \binom{n}{i} p_0^i (1-p_0)^{n-i} \leq \frac{\alpha}{2} \text{ и } \sum_{i=0}^{C_1+1} \binom{n}{i} p_0^i (1-p_0)^{n-i} > \frac{\alpha}{2},$$

као и

$$\sum_{i=C_2}^{n} \binom{n}{i} p_0^i (1-p_0)^{n-i} \leq \frac{\alpha}{2} \text{ и } \sum_{i=C_2-1}^{n} \binom{n}{i} p_0^i (1-p_0)^{n-i} > \frac{\alpha}{2}.$$

Слично поступамо у случају једностраних хипотеза. Наиме, ако је алтернативна хипотеза (3.2) онда је критична област облика $W = \{S_n \leq C\}$ а константу c одређујемо из услова

$$\sum_{i=0}^{C} \binom{n}{i} p_0^i (1-p_0)^{n-i} \leq \alpha \text{ и } \sum_{i=0}^{C+1} \binom{n}{i} p_0^i (1-p_0)^{n-i} > \alpha.$$

Ако је алтернативна хипотеза (3.3) онда је критична област облика $W = \{S_n \geq C\}$ а константу C одређујемо из услова

$$\sum_{i=C}^{n} \binom{n}{i} p_0^i (1-p_0)^{n-i} \leq \alpha \text{ и } \sum_{i=C-1}^{n} \binom{n}{i} p_0^i (1-p_0)^{n-i} > \alpha.$$

Пример 3.1.6. *Желимо да покажемо да вероватноћа да се деси дефект у изради прстена из Тути-фрути колекције, кад се ради по мери, мања од 20%. Направљено је 20 комада, без грешке. Да ли имамо довољно разлога да верујемо у тврдњу произвођача?*

Дакле, имамо да је $H_0 : p = 0.2$ *против* $H_1 : p < 0.2$ *и јасно је да немамо узорак великог обима на располагању јер је се ради о специјалним дизајнерским колекцијама за којима, због цене, није велика потражња. Критична област је облика* $W = \{S_{20} \leq C\}$ *па је p-вредност* $P\{S_{20} \leq 0\} = 0.8^{20} = 0.01$ *па закључујемо да ћемо одбацити хипотезу у корист произвођача. До сличног закључка можемо доћи одређивањем критичне области. Наиме, ако је* $C = 0$ *добија се ероватноћа грешке прве врсте 0.01, док уколико је* $C = 1$*, она износи 0.07. Одавде закључујемо да ако нам је грешка ограничена са 0.05 морамо узети да је* $C = 0.01$*. Приметимо још да тада 0.05 није мера теста јер не можемо формирати критичну област тако да се тај ниво значајности достиже.*

Случај два узорка

Претпоставимо да је X индикатор са вероватноћом успеха p_1 и Y индикатор са вероватноћом успеха p_2, и да имамо на располагању два независна узорка обима n_1 и n_2, редом. Тада, за тестирање $H_0 : p_1 - p_2 = p_0$ можемо искористити статистику

$$T_{n_1,n_2} = \frac{\bar{X}_{n_1} - \bar{Y}_{n_2} - p_0}{\sqrt{\frac{\bar{X}_{n_1}(1-\bar{X}_{n_1})}{n_1} + \frac{\bar{Y}_{n_2}(1-\bar{Y}_{n_2})}{n_2}}} \tag{3.1}$$

која, у случају да су оба узорка велика, ако важи H_0, има $\mathcal{N}(0,1)$ расподелу. Критичне области формирамо као и до сада, у зависности од алтернативне хипотезе.

У специјалном случају, када је $p_0 = 0$, онда је оцена за p_1, односно p_2 (пошто су једнаке), се може добити на основу обједињеног узорка и износи

$$\hat{p_1} = \frac{\bar{X}_{n_1} n_1 + \bar{Y}_{n_2} n_2}{n_1 + n_2},$$

па је алтернативна веризја тест статистике

$$T'_{n_1,n_2} = \frac{\bar{X}_{n_1} - \bar{Y}_{n_2} - p_0}{\sqrt{\hat{p}_1(1-\hat{p}_1)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \tag{3.2}$$

Алтернативну верзију можемо конструисати и у општем случају кад је $H_0 : p_1 - p_2 = p_0$. Тада се можемо p_1 и p_2 оценити методом максималне веродостојности уз услов да је њихова разлика p_0. Међутим, тада се одређивање оцене значајно компликује, па како се овај случај ретко јавља у пракси, нећмо га приказати.

Пример 3.1.7. *Циљ истраживања је да се покаже да постоје разлике у проценту пушача међу средњошколцима у центру града и на периферији. Резултати истраживања су следећи: од 125 испитаника из центра града њих 47 је пушило, док је од 153 испитаника са периферије, њих 52 пушило. Какав је закључак на основу одговарајућег тестирања?*

Нека је p_1 вероватноћа да средњошколац из центра града пуши, док p_2 вероватноћа да средњошколац који није из центра града, пуши. H_0 је да је $p_1 = p_2$, а алтернативна хипотеза $p_1 \neq p_2$. Оцена за p_1 на основу обједињеног узорка је

$$\hat{p} = \frac{\frac{47}{125}+\frac{52}{153}}{125+153} = 0.356.$$

Реализована вредност тест статистике је

$$\hat{T} = \frac{\frac{47}{125}-\frac{52}{153}}{\sqrt{0.356\cdot 0.644(\frac{1}{125}+\frac{1}{153})}} = 0.6258741.$$

P-вредност теста је сад $2P\{T \geq 0.626\}) = 0.533$ па не одбацујемо H_0, односно закључујемо да нема разлика у центру града и на периферији. Да смо користили статистику (3.1) добили бисмо да је $\hat{T} = 0.624861$, што се незнатно разликује од претходног и резултат би свакако био исти.

Задатак 3.1.6. Направите самостално истраживање у ком ћете проверити да ли постоје разлике у пропорцијама оних који су положили све испите у једној академској години, у односу на пол испитаника. Водите рачуна о обиму узорака који вам је потребан да бисте применили резултате из овог одељка.

3.1.3 Тестови у Пуасоновом моделу

Претпоставимо да X има Пуасонову $\mathcal{P}(\lambda)$ расподелу. Желимо да тестирамо $H_0 : \lambda = \lambda_0$. Уколико на располагању имамо велики узорак можемо искористити тест статистику

$$T_n = \frac{\bar{X}_n - \lambda_0}{\sqrt{\frac{\lambda_0}{n}}}, \tag{3.1}$$

која, уколико је нулта хипотеза тачна, има $\mathcal{N}(0,1)$ расподелу, а кртичне области формирамо у зависности алтернативе. Уколико је $H_1 : \lambda \neq \lambda_0$ онда је критична област је облика $W = \{|T_n| \geq C\}$,у случају $H_1 : \lambda < \lambda_0$ имамо да је $W = \{T_n \leq C\}$, док је у случају $H_1 : \lambda > \lambda_0$ критична област облика $W = \{T_n \geq C\}$. Поред статистике (3.1) можемо користити и статистику

$$T_n' = \frac{\bar{X}_n - \lambda_0}{\sqrt{\frac{\bar{X}_n}{n}}}, \tag{3.2}$$

Када је узорак мали можемо искористити статистику

$$S_n = X_1 + \cdots + X_n \tag{3.3}$$

за коју знамо да, уколико је нулта хипотеза тачна, има Пуасонову $\mathcal{P}(n\lambda_0)$ расподелу. Друга могућност је да користимо статистику (3.1) или (3.2), али да расподелу тест статистике, уколико важни нулта хипотеза, одредимо емпиријски, коришћењем Монте Карло метода. Треба имати у виду да нема разлога очекивати да је расподела симетрична, па је тада у случају $H_1 : \lambda \neq \lambda_0$, облика $W = \{T_n \leq C_1\} \cup \{T_n \geq C_2\}$.

Пример 3.1.8. *Желимо да проверимо да ли је дошло до побољсања квалитета једног фудбалсог тима и да ли у овој сезони постиже просечно више голова на утакмици него у прошлој сезони када је просек био 2.3. Узет је узорак од 8 утакмица и добијено да је просечан број голова 2.5. Проверићемо да ли је повећање значајно. Претпостављамо да се број голова може моделирати Пуасновом расподелом.*

Имамо да је нулта хипотеза $H_0 : \lambda = 2.3$ *против* $H_1 : \lambda > 2.5$. *Како имамо мали узорак користићемо статистку* (3.3).

Критична област је облика $W = \{S_n \geq C\}$*, реализована вредност тест статистике* $\hat{S}_8 = 8 \cdot 2.5 = 20$ *па је p-вредност* $P\{S_8 \geq 20\} - 1 - F_{\mathcal{P}(8.4)}(19) = 0.38$ *што не упућује на значајно повећан квалитет посматраног тима.*

Задатак 3.1.7. Извршите тестирање за проблем у примеру 3.1.8 користећи асимптотску апроксимацију. Поред тога, урадите то и користећи емпиријску расподелу тест статистике под нултом хипотезом и упоредите резултате.

Случај два узорка

Претпоставимо да имамо обележја X и Y са Пуасоновим $\mathcal{P}(\lambda_1)$ и $\mathcal{P}(\lambda_2)$ расподелама. Желимо да тестирамо $H_0 : \lambda_1 - \lambda_2 = \lambda_0$. против алтернатива $H_1 : \lambda_1 - \lambda_2 \neq \lambda_0$, $H_1 : \lambda_1 - \lambda_2 > \lambda_0$, и $H_1 : \lambda_1 - \lambda_2 < \lambda_0$. Уколико имамо велике узорке можемо користити статистику

$$T_{n_1,n_2} = \frac{\bar{X}_{n_1} - \bar{Y}_{n_2} - \lambda_0}{\sqrt{\frac{\bar{X}_{n_1}}{n_1} + \frac{\bar{Y}_{n_2}}{n_2}}}, \tag{3.1}$$

док критичне области за посматране алтернативне хипотезе су редом $W = \{|T_{n_1,n_2}| \geq C\}$, $W = \{T_{n_1,n_2} \geq C\}$ и $W = \{T_{n_1,n_2} \leq C\}$. И овде, можемо модификовати статистику тако да оценимо $\frac{\lambda_1}{n_1} + \frac{\lambda_1}{n_2}$ узимајући у обзир услов $\lambda_1 - \lambda_2 = \lambda_0$, али је то изван домашаја овог уџбеника. У специјалном случају кад је $\lambda_0 = 0$ дисперзија разлике узорачких средина се може се оценити на основу обједињеног узорка, па је тада алтернативна верзија статистике (3.1)

$$T'_{n_1,n_2} = \frac{\bar{X}_{n_1} - \bar{Y}_{n_2}}{\sqrt{\frac{\bar{X}_{n_1} n_1 + \bar{Y}_{n_2} n_2}{n_1+n_2} \cdot (\frac{1}{n_1} + \frac{1}{n_2})}}. \tag{3.2}$$

Пример 3.1.9. *Желимо да упоредимо просечне вредности броја примљених голова две екипе исте класе које се такмиче у различитим категоријама. Претпоставићемо да се број примљених голова може моделирати Пуасоновом расподелом. Узети су узорци од 20 утакмица једне и друге екипе и добијено је да је* $\bar{x}_{20} = 1.2$ *а* $\bar{y}_{20} = 1.8$ *и извршено тестирање са нивоом значајности* $\alpha = 0.05$.

Како је овде природно алтернативна хипотеза $H_1 : \lambda_1 - \lambda_2 \neq 0$ *критична област је облика* $W = \{|T_{20,20}| \geq C\}$. *Добијамо да је реализована вредност тест статистике* $\hat{T}_{20,20} = -1.55$ *па је p-вредност* $2P\{T_{20,20} \geq 1.55\} = 0.12$ *што је мало, али ипак не одбацујемо* H_0. *Да смо користили статистику* (3.2) *добили бисмо да је реализована вредност* -1.22 *што нас наводи на исти закључак.*

Задатак 3.1.8. Сматра се да се у неким ситуацијама број секундарних потреса након земљотреса може моделирати Пуасоновом расподелом. Циљ једног истраживања је да се установи да ли постоје разлике у учесталости секундарних потреса у случају да је примарни земљотрес великог и средњег интензитета. Узет је узорак који се састоји од 30 земљотреса средњег и 20 земљотреса јаког интензитета, у последњих 15 година и добијено је да су узорачке учесталости редом $\bar{x}_30 = 3.2$ и $\bar{y}_20 = 3.5$. Тестирати да ли добијена разлика значајна, уколико је праг значајности $\alpha = 0.05$. Уколико је у тестирању направљена грешка, да ли је она прве или друге врсте?

3.2 Непараметарски тестови

У претходном поглављу смо приказали неке од најчшће коришћених параметарских тестова. Оно што је било заједничко за све поменуте тестове је да смо се приликом конструкције тест статистика водили делимично тиме да знамо да одредимо њену расподелу под нултом хипотезом (која је била условљена расподелом посматраног обележја). На пример, за тестирање да $m = m_0$ у случају обележја X са нормалном $\mathcal{N}(m, \sigma^2)$ расподелом, користили смо статистику

$$T_n = \frac{\bar{X}_n - m_0}{\frac{\widetilde{S}_n}{n}}, \tag{3.1}$$

за коју знамо, да ако важи H_0, има нормалну $\mathcal{N}(0,1)$ расподелу. Свакако да исту статистику можемо корстити и у случају да желимо да тестирамо да је $EX = m_0$ у случају неке друге расподеле, али критична област неће више бити иста па можемо направити грешку много већу од дозвољене.

Пример 3.2.1. *Претпоставимо да имамо узорак обима 10 из експоненцијалне $\mathcal{E}(\lambda)$ расподеле и желимо да тестирамо H_0 : $EX = \frac{1}{\lambda} = 1$, против H_1 : $\lambda \neq 1$ са нивоом значајности $\alpha = 0.05$. Уколико бисмо користили статистику* (3.1) *и расподелу која је последица нормалности узорка, односно Студентову t_9, добили бисмо да је критична област $W = \{|T_{10}| \geq C\}$, где је $C = F_{t_9}^{-1}(0.975) = 2.26$. Међутим, у нашем случају неће бити задовољена једнакост $P_{H_0}\{T_{10} \in W\} = \alpha$. Да бисмо то показали, потребно је да знамо расподелу тест статистику уколико важи H_0. Испоставља се да ју је егзактно веома комплоквано одредити. Зато ћемо оценити $P_{H_0}\{|T_{10}| \geq 2.26\}$ коришћењем Монте Карло симулација. Подсетимо се, овај метод се састоји у понављању експеримента под истим условима који се састоји од узорковања и одређивања вредности тест статистике, велики број пута и одређивања процента понављања у којима је тест статистика упала у критичну област.*

```
set.seed(15) s=0 for(i in 1:10000){
uzorak=rexp(10,1)
T=(mean(uzorak)-1)/sd(uzorak)*sqrt(10)
if (abs(T)>=2.26) s=s+1
}
s/10000
[1] 0.1033
```

Добили смо да је оцена вероватноће грешке прве врсте 0.1, а не 0.05, па тест незадовољава полазне претоиставке. Један начин да се реши ова ситуација је да се одредити емпиријски расподела тест статистике под H_0, на и да се онда закључивање врши на основу реализоване вредности тест статистике. Но у таквом приступу је важно да имамо информацију да је расподела полазног узорка експоненцијална и тада тест припада домену параметарске статистике.

Овај пример лепо илуструје колико је значајно да је полазна претпоставка о моделу тачна. Зато уколико нисмо сигурни у њу много је боље осмислити тестирање тако да се приликом њега нигде не користи претпоставка о некој класи расподела коју има посматрано обележје. Такво тестирање припада домену непараметарске статистике. У остатку поглавља представићемо неколико најпознатијих непараметарских тестова који представљају алтернативу параметарским тестовима из претходног поглавља. Поред тога описаћемо и неколико најважнијих тестова сагласности са расподелом и тестова независности два обележја.

3.2.1 Тест знакова

Тестирамо H_0 $m_e = m_{e0}$, где је m_e медијана расподеле коју има обележје X. Уколико се присетимо дефиниције медијане расподеле природно долазимо до следеће тест статистике

$$T_n = \sum_{i=1}^{n} I\{X_i > m_{e0}\}. \tag{3.2}$$

Овај тест познат је још под називом тест знакова[2]. Често се уместо тест статистике (3.2) користи и

$$T_n' = \sum_{i=1}^{n} I\{X_i < m_{e0}\}.$$

Уколико је нулта хипотеза тачна број чланова узорка који су мањи од m_e треба да буде приближно једнак броју који су већи, односно T_n је приближно $\frac{n}{2}$. Приметимо још да, уколико је хипотеза H_0 тачна T_n има $\mathcal{B}(n, \frac{1}{2})$ јер је $P\{X_i > m_{e0}\}$ тада баш $\frac{1}{2}$, па се могу одредити одговарајуће критичне области.

Поред (3.2) можемо користити и "центрирану" верзију тест статистике

$$T_n^c = \sum_{i=1}^{n} I\{X_i > m_{e0}\} - \frac{n}{2}.$$

За велико n[3] се може користити нормална апроксимација, односно да статистика

$$T_n^* = \frac{T_n - \frac{n}{2}}{\sqrt{\frac{n}{4}}}$$

има нормалну $\mathcal{N}(0,1)$ расподелу.

Критичну област формирамо као и до сада, у зависности од алтернативне хипотезе. На пример. уколико је $H_1 : m_e \neq m_{e0}$ критична област за тестирање $W = \{|T_n^*| \geq C\}$ или $W = \{|T_n^c| \geq C\}$ (јер је расподела тест статистике, ако важи H_0, симетрична). Приметимо још да, ако је расподела обележја X симетрична онда се H_0 своди на H_0 $m = m_0$, где је $m = EX$.

Пример 3.2.2. *Фабрика крема за негу лица је одлучила да избаци нови производ на тржиште. Како се раде о изузетно скупом производу испитивање о томе како ће је корисници прихватити, рађено је на малом узорку. Случајно је одабрано 7 корисника великог ланца парфимерија, за које је утврђено да купују сличне производе, и дато им је по једно паковање нове креме. Њихов једини задатак је да након две недеље коришћења оцене производ*

[2] енг. Sign test

[3] ова апроксимација се може користити већ за $n > 10$

оценом од 1 до 5. Добијене су следеће оцене: 2,5,5,4,1,4,5. Да ли на основу ових резултата, са нивоом значајности 0.05, *можемо закључити да је рејтинг 3?*

Сада на располагању имамо мали узорак, и немамо никаквог разлога да верујемо да се ради о обележју са нормалном расподелом. С обзиром на то да тражимо доказе да потврдимо нашу хипотезу, из посматраног проблема можемо закључити да је $H_0: m_e = 3$ *против* $H_1: m_e \neq 3$. *На основу узорка закључујемо да је реализована вредност тест статистике* $\hat{T}_7 = \sum_{i=1}^{7} I\{X_i > 3\} = 5$. *Критична област је облика* $W = \{|T_7^c| \geq C\}$ *па је p-вредност* $P\{|T_7^c| \geq 1.5\} = P\{T_7 \geq 5\} + P\{T_7 \leq 2\} = 0.45$ *па прихватамо* H_0.

Напомена: Овај тест се често користи и за тестирање симетрије око нуле јер ако је расподела обележја симетрична онда је одговарајућа медијана једнака нули.

Задатак 3.2.1. Генерисати узорак обима 50 из експоненцијалне $\mathcal{E}(1)$ расподеле и на основу њега тестирати H_0, са нивоом значајности $\alpha = 0.01$, да је медијана узорка једнака медијани $\mathcal{E}(1)$ расподеле.

Задатак 3.2.2. Тестирати да ли је медијана висине плате просветних радника из примера 1.2.2 45000 динара против алтернативе да је већа од 45000 динара.

Задатак 3.2.3. Инжењери животне средине открили су да проценат активних бактерија у узорцима из отпадних вода, када постројење за пречишћавање ради исправно, има расподелу са медијаном 40%. Ако је проценат већи од 40% потребно је извршити неке промене у систему. Узето је 10 узорака и добијен је проценат бактерија 40,42,33,50,46,37,40,40. Да ли добијени подаци указују да је потребно извршити промене?

3.2.2 Случај два узорка

Посматрамо дводимензионо обележје (X, Y). Претпоставимо да на располагању имамо узорак $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$. Прво ћемо направити низ разлика $D_i = X_i - Y_i$. Уколико на основу

тог новог узорка не можемо закључити да је претпоставка о нормалности испуњена, можемо користити Тест знакова примењен на нови узорак. Нулта хипотеза за овај тест је да медијана обележја $D = X - Y$ има неку фиксну вредност (најчешће да је нула). Важно је напоменути да то није исто што и да се медијане оба обележја поклапају. Разлог томе је што медијана разлике две случајне величине није разлика одговарајућих медијана. Међутим, уколико је расподела D симетрична, онда се нулта хипотеза теста своди на то да је разлика очекивања посматраних обележја једнака некој фиксној вредности. Ово важи јер се у случају симетричне расподеле медијана поклапа са математичким очекивањем, а очекивање разлике је разлика одговарајућих очекивања. Овај тест је непараметарска алтернатива спареном t-тесту (3.10).

Пример 3.2.3. *У старачком дому група испитаника је била изложена новом леку за који се сматра да успорава деменцију. Мерена је вредност једног мождданог параметра пре третмана (X) и након третмана (Y). Сматра се да лек нема ефекта уколико је $P\{X > Y\} = \frac{1}{2}$ као и да је повећана вредност тог параметра управо разлог деменције. Резултати су приказани у следећој табели*

Особа	X	Y
1	*15*	*13*
2	*12.5*	*10*
3	*12*	*11*
4	*12.5*	*12*
5	*12*	*14*
6	*13*	*12.5*
7	*13*	*12.5*
8	*13*	*12*
9	*14*	*12*
10	*12*	*12.5*
11	*12*	*8*

Јасно је да је алтернативна хипотеза да је $P\{X > Y\} > 1/2$ па ће критична област бити облика $\{T_{11} \geq C\}$ где је T_{11} број случајева када је $X_i > Y_i$, односно $T_{11} = \sum_{i=1}^{11} I\{X_i > Y_i\}$. Из студије видимо да је било укупно 9 таквих случајева, односно

реализована вредност тест статистике је $\hat{T}_{11} = 9$. *Тада је p-вредност теста* $P\{T_{11} \geq 9\} = 0.03$, *при чему смо користили да, уколико је* H_0 *тачна,* T_{11} *има Биномну* $\mathcal{B}(11, 0.5)$ *расподелу.*
У случају да смо користили нормалну апроксимацију добили бисмо да је $\hat{T}^*_{11} = 2.11$, *и да је p-вредност 0.02.*

Дакле, закључак је да нови лек стварно има ефекта на успоравање деменције.

3.2.3 Вилкоксонов тест заснован на ранговима и знаковима

Претпостављамо да је расподела обележја X апсолутно непрекидна и симетрична око своје медијане и желимо да тестирамо да математичко очекивање (медијана) има неку вредност, односно $H_0 : m = m_0$. Уколико је нулта хипотеза тачна $X - m_0$ има исту расподелу као $m_0 - X$.

Пре него што представимо Вилкоксонову тест статистку за тестирање ове хипотезе увешћемо појам *ранга*. Ранг неког елемента у узорку представља његов редни број кад се елементи узорка поређају у растућем поретку. На пример уколико је реализован узорак $(x_1, ..., x_4) = (8, 9, 1, 2)$, одговарајући варијациони низ је $(x_{(1)}, .., x_{(4)}) = (1, 2, 8, 9)$ што значи да је 8 трећи по реду, 9 четврти итд., односно да је низ разнгова $(r_1, r_2, r_3, r_4) = (3, 4, 1, 2)$. Уколико у полазном узорку има елемената који имају исту вредност онда је њихова вредност једнака просечној вредности рангова које би имали ти елементи кад би били различити. На пример, уколико имамо узорак $(x_1, ..., x_7) = (1, 1, 7, 5, 5, 6, 1)$ чији је варијациони низ $(x_{(1)}, .., x_{(7)}) = (1, 1, 1, 5, 5, 6, 7)$ добијамо да је одговарајући низ рангова $(2, 2, 2, 4.5, 4.5, 6, 7)$.

Означимо са r_i ранг елемента $|X_i - m_0|$ у узорку $|X_1 - m_0|, ..., |X_n - m_0|$. Тест статистика коју је предложио Вилкоксон је

$$T_n = \sum_{i=1}^{n} r_i I\{X_i - m_0 \geq 0\}. \tag{3.1}$$

Показаћемо да је, у случају да је нулта хипотеза тачна,

$$ET_n = \frac{n(n+1)}{4} \tag{3.2}$$

$$DT_n = \frac{n(n+1)(2n+1)}{24}. \tag{3.3}$$

Прво, приметимо да тада T_n има исту расподелу као случајна величина $T_n' = \sum_{i=1}^{n} I_i$, где су I_i међусобно независне случајне величине за које важи да је $P\{I_i = i\} = 0.5 = P\{I_i = 0\}$ (у суми (3.1) се сваки од рангова појављује са вероватноћом 0.5, а индикатори су међусобно независни). Тада је

$$ET_n = \sum_{i=1}^{n} EI_i = \sum_{i=1}^{n} \frac{i}{2} = \frac{n(n+1)}{4}$$

$$DT_n = \sum_{i=1}^{n} DI_i = \sum_{i=1}^{n} \frac{i^2}{4} = \frac{n(n+1)(2n+1)}{24}.$$

Овде смо користили да међу ранговима нема понављања јер је X апсолутно непрекидно обележје. У случају да то није испуњено, уз изнад описану процедуру одређивања рангова, очекивање остаје исто, а дисперзију треба кориговати. међутим, испоставља се да у случају већих узорака је у реду користити и (3.2).

Егзактна расподела тест статисыике под нултом хипотезом се може наћи и за њу постоје таблице. Уколико је $n > 12$, за одредјивање критичне области може се корстити нормална апроксимација, тј. да $\frac{T_n - ET_n}{\sqrt{DT_n}}$ има стандардну нормалну расподелу.

Приметимо да T_n заправо представља збир рангова елемената низа који су већи од m_0, дакле не мере се вредности елемената узорка већ њихов ранг, чиме се постиже неосетљивост на присуство аутлајера. Ово својство је карактеристично за све тестове који су засновани на ранговима, тако да када их очекујемо, чак и кад имамо информацију о расподели обележја, њихова употреба је препоручљива.

Случај два узорка

Вилкоксонов тест се може лако адаптирати на случај два спарена, као и два независна узорка када желимо да тестирамо да два

обележја имају исти параметар локације. У случају спареног узорка, он се примењује на узорак $D_i = X_i - Y_i,\ i = 1, 2..., n$ и нормална апроксимација се може примењивати за већ за обим узорка $n > 12$.

У случају независних узорака, мора се увести додатна претпоставка да обележја X и Y имају исту расподелу до на константу, тј. да је расподела обелжја X иста као расподела обележја $Y + c$ за неко c. Тада је заправо $H_0 :\ c = 0$, и овај тест се може користити као непараметарска алтернатива t-тесту за једнакост очекивања обележја са нормалним расподелама са истим дисперзијама.

Претпоставимо да на располагању имамо два независна узорка $X_1, ..., X_n$ и $Y_{n+1}, ..., Y_{n+m}$. Један од главних корака приликом конструкције тест статистике је да се два узорка се обједине у један узорак (зато смо их тако и означили).

Статистика, у овом случају је

$$T_{n,m} = \sum_{i=1}^{n} r_i,$$

где је r_i ранг i-тог елемената из узорка $X_1, X_2, ..., X_n$ у обједињеном узорку. Ова статистика се често записује и у облику

$$T_{n,m} = \sum_{i=1}^{n+m} r_i z_i,$$

где је $z_i = 1$ ако је i-ти елемент из првог узорка, у супротном $z_i = 0$.
Показаћемо да је , ако је H_0 тачна, $ET_{n,m} = \frac{n(n+m+1)}{2}$ и да је $D(T_{n,m}) = \frac{nm(n+m+1)}{12}$, а може се показати и да се за $n, m > 10$ може користи нормална апроксимација, међутим то је изван домашаја овог уџбеника.

$$E(T_{n,m}) = \sum_{i=1}^{n} E(r_i) = n \cdot E(r_1) = n \cdot \sum_{i=1}^{n+m} \frac{i}{n+m} = \frac{n(n+m+1)}{2}. \tag{3.4}$$

Дисперзију рачунамо на основу формуле

$$D(T_{n,m}) = ET_{n,m}^2 - (ET_{n,m})^2.$$

$$E(T_{n,m}^2) = E(\sum_{i=1}^{n} r_i^2) + E(\sum_{i\neq j} r_i r_j) = nE(r_1^2) + \sum_{u\neq j} E(r_i r_j).$$

У суми изнад разликујемо посебно сабирке кад је $i = j$ и кад је $i \neq j$ јер рангови нису међусобно независне случајне величине!

$$E(r_1^2) = \sum_{i=1}^{n+m} \frac{i^2}{n+m} = \frac{(m+m+1)(2n+2m+1)}{6}$$

$$E(r_1 r_2) = \sum_{i=2}^{n+m}\sum_{j=1}^{i-1} \frac{ij}{\binom{n+m}{2}} = \sum_{i=2}^{n+m} \frac{i}{\binom{n+m}{2}} \cdot \frac{(i-1)i}{2}$$
$$= \frac{(m+n+1)(3m+3n+2)}{12}.$$

Даље је

$$E(T_{n,m}^2) = n \cdot \frac{(n+m+1)(2n+2m+1)}{6} \tag{3.5}$$
$$+ n(n-1) \cdot \frac{(m+n+1)(3m+3n+2)}{12}. \tag{3.6}$$

Замењујући (3.4) и (3.5) у израз за дисперзију, добијамо да је $D(T_{n,m}) = \frac{nm(n+m+1)}{12}$.

Напомена: Обележја се обично нумеришу тако да је узорак чију суму рангова посматрамо онај мањег обима.

До истог закључка се може доћи и из следећег облика ове статистике. Наиме, природна оцена за $P\{X > Y\}$ је

$$U_{n,m}^{\star} = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} I\{X_i > Y_{n+j}\}.$$

Приметимо да израз изнад заправо означава удео парова (X_i, Y_{n+j}) за које је први елемент већи од другог. Даље је

$$U_{n,m}^{\star} = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} I\{X_{(i)} > Y_{(n+j)}\}$$
$$= \frac{1}{nm}\sum_{i=1}^{n}(r_i - i) = \frac{1}{nm}T_{n,m} - \frac{(n+1)}{2m}. \tag{3.7}$$

Статистика $U_{n,m} = mnU^{\star}_{n,m}$ је позната и као Ман-Витнијева[4] статистика. Из израза (3.7) видимо да постоји линеарна веза између ње и Вилкоксонове статистике, тако да се суштински ради о истом тесту.

Уколико је $X > Y$ (у расподели) онда ће вредност статистике $T_{n,m}$ ($U_{n,m}$) бити већа него што очекујемо јер ће елементи узорка $X_1, .., X_n$ имати углавном већи ранг од оних из узорка $Y_{n+1}, ..., Y_{n+m}$. Слично, ако је $X < Y$ онда ће вредности тест статистике бити мање него што очекујемо. На основу овог разматрања можемо лако доћи до облика критичне област у зависности од алтернативне хипотезе.

Пример 3.2.4. *За израду каблова коришћене су две различите технике и како би се упоредио квалитет производа, одређене су њихове максималне силе издржљивости (у њутнима).*

први кабл: 10854, 9106, 10325, 11627, 10051, 10001, 10000, 13720, 11632, 11222

други кабл: 11000, 11072, 8851, 10245, 11000, 10030, 11197, 10959, 9157, 11513, 9540, 10856.

Имајући у виду природу материјала може се претпоставити да су расподеле издржљивости оба кабла исте до на константу, али да немају нормалну расподелу.

Обједињен узорак је 10854, 9106, 10325, 11627, 10051, 10001, 10000, 13720, 11632, 11222, 11000, 11072, 8851, 10245, 11000, 10030, 11197, 10959, 9157, 11513, 9540, 10856. Одговарајући рангови су 11, 2, 10, 20, 8, 6, 5, 22, 21, 18, 14.5, 16, 1, 9, 14.5, 7, 17, 13, 3, 19, 4, 12. Одавде добијамо да је реализована вредност тест статистике $\hat{T}_{10,12} = 123$*, односно нормализована вредност је* $\frac{123-115}{\sqrt{230}} = 0.528$*. На основу тога се добија да је p-вредност 0.60 па не одбацујемо* H_0*, односно нема разлике у изради каблова.*

Ради веће прецизности апроксимације, приликом нормализације тест статистике врши се корекција непрекидности јер су вредности које узима T *целобројне, односно рачуна се*

$$\frac{T_{10,12} - ET_{10,12} - 0.5}{\sqrt{DT_{10,12}}}.$$

[4] енг. Mann-Whitney

И управо та вредност је имплементирана у R-у. Поред тога, одузета је и минимална сума рангова која износи $\frac{10 \cdot 11}{2}$.

Задатак 3.2.4. Циљ једног клиничког испитивања био је да се увиди ефикасност нове антиретровирусне терапије за пацијенте који имају ХИВ. Пацијенти који учествују у истраживању, подељени су на случајан начин у две групе. Пацијенти из прве групе су наставили да примају стандардну антиретровирусну терапију, док су пацијенти из друге групе били подвргнути новој терапији. Након шестомесечне терапије измерен је копија ХИВ-а у милиметру крви и добијени су следећи резултати:

стандардна терапија	нова терапија
7200	400
7000	250
2000	800
550	1400
1250	8000
1000	7400
2250	1020
6300	5000
3400	920
6300	1400
9100	2700
970	4200
1040	5200
670	4100
400	500

Са нивоом значајности теста $\alpha = 0.05$ проверити да ли нова терапија даје боље резултате.

3.2.4 Тестови сагласности са расподелом

Важна класа непараметарских тестова су тестови сагласности са расподелом, односно проверавање хипотезе да је модел на коме заснивамо даља статистичка закључивања исправан.

Претпоставимо да обележје X им функцију расподеле F. У овом поглављу бавићемо се тестовима који се односе на тестирање нулте хипотезе $H_0 : \ F = F_0$, при чему F_0 може зависити од непознатих параметара. Приказаћемо неколико најједноставнијих класа ових тестова.

Тестови засновани на емпиријској функцији расподеле

Прва група тестова које смо поменули заснована је на униформној конвергенцији емпиријске функције расподеле ка правој функцији расподеле обележја, односно на тврђењу Гливенко-Кантелијеве теореме (видети 1.5.1). За сада претпостављамо да је F_0 апсолутно непрекидна функција расподеле. Такође, претпостављамо да F_0 не зависи од непозатих параметара.

- Тест Колмогоров-Смирнова:

$$D_n = \sup_x |F_n(x) - F_0(x)|$$

  У случају алтернативе хипотезе $F \neq F_0$ критична област за тестирање је облика $W = \{D_n \geq C\}$. У R-у користимо функцију *ks.test*.

  За мале вредности обима узорка може се егзактно исвести расподела (и постоје одговарајуће таблице критичних вредности), док је за велико n нађена асимптотска расподела статистике $\sqrt{n}|F_n(x) - F_0(x)|$. Ова статистика има једно јако лепо својство, а то је да, ако је H_0 тачна, њена расподела не

зависи од F_0. Ово ћемо и показати.

$$D_n = \sup_x |F_n(x) - F_0(x)| = \sup_{y=F_0(x)} |F_n(F_0^{-1}(y)) - y|$$
$$= \Big|\frac{1}{n}\sum_{i=1}^{n} I\{X_i \le F_0^{-1}(y)\} - y\Big| = \Big|\frac{1}{n}\sum_{i=1}^{n} I\{F_0(X_i) \le y\} - y\Big|$$
$$= \Big|\frac{1}{n}\sum_{i=1}^{n} I\{U_i \le y\} - y\Big|,$$

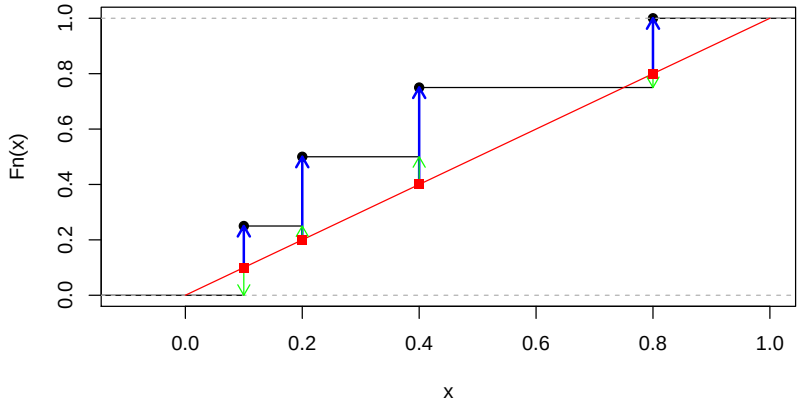
где је $U_1, ..., U_n$ низ независних случајних величина са униформном $\mathcal{U}[0,1]$ расподелом. Ово следи из познатог тврђења да ако X има функцију расподеле F_0 онда $F_0(X)$ има униформну $\mathcal{U}[0,1]$ расподелу. Дакле, показали смо да, ако је H_0 тачна D_n има исту расподелу као да тестирамо $H_0 : F(y) = y$, за $y \in [0,1]$. Управо због овог својства довољно је да једном одредимо расподелу и после је можемо користити за тестирање сагласности са било којом апослутно непрекидном расподелом (која не зависи од непознатих параметара).

Поред овога, како бисмо лакше одредили вредност тест статистике, приметимо да се D_n може написати у облику

$$D_n = \max_{1\le k\le n} \Big(\Big|F_0(X_i) - \frac{k-1}{n}\Big|, \Big|\frac{k}{n} - F_0(X_i)\Big|\Big).$$

До овог закључка смо дошли јер је $F_n(x)$ део по део константна функција, а F_0 непрекидна па треба проверити промене у тачкама скока, и са леве и са десне стране.

Пример 3.2.5. *Дат је узорак* $0.1, 0.2, 0.4, 0.8$. *На основу њега тестираћемо хипотезу да је обележје из униформне* $\mathcal{U}[0,1]$ *расподеле. Одредићемо вредност статистике Колмогоров Смирнова. На цртежу 3.2.5 су зеленом бојом приказана одступања емпиријске од теоријске функције расподеле с леве стране, а плавом бојом одступања с десне стране. Тест статистика чију вредност тражимо представља дужину максималног одступања.*



Слика 3.4: Рачунање тест статистике Колмогоров-Смирнова

То је приказано и у наредној табели.

X_i	$\lvert F_0(X_i) - \frac{k-1}{n} \rvert$	$\lvert \frac{k}{n} - F_0(X_i) \rvert$
0.1	*0.10*	*0.15*
0.2	*0.05*	*0.30*
0.4	*0.10*	*0.35*
0.8	*0.05*	*0.20*
$\hat{D}_4$	*0.35*	

Видимо да је максимална вредност посматраног одступања 0.35.

- Тест Крамер-фон Мизеса

$$\omega_n^2 = \int_{-\infty}^{\infty} (F_n(x) - F_0(x))^2 dF_0(x). \tag{3.1}$$

Вредности подинтегралне функције ће бити блиске нули уколико је H_0 тачна, док ако није, вредности ће бити веће од нуле. Због тога је критична област облика $W = \{\omega_n^2 \geq C\}$

На сличан начин може се показати да расподела тест статистике под нултом хипотезом не зависи од F_0.

Функција у R-у коју користимо је *cvm.test* из пакета *goftest*. Уколико је потребно одредити вредности статистике без коришћења програмског језика R потребно је да се (3.1) напише на погоднији начин. Увођењем смене $y = F_0(x)$ израз (3.1) постаје

$$\omega_n^2 = \int_0^1 (F_n(F_0^{-1}(y)) - y)^2 dy.$$

Како је $F_n(F_0^{-1}(y) = \frac{1}{n}\sum_{i=1}^n I\{X_i \leq F_0^{-1}(y)\}$ део по део константна функција израз изнад постаје

$$\begin{aligned}
\omega_n^2 &= \sum_{i=1}^n \int_{F_0(X_{(i-1)})}^{F_0(X_{(i)})} (F_n(F_0^{-1}(F_0(X_{i-1}))) - y)^2 dy \\
&= \sum_{i=1}^n \int_{F_0(X_{(i-1)})}^{F_0(X_{(i)})} (F_n(X_{(i-1)}) - y)^2 dy = \sum_{i=1}^n F_n^2(X_{(i-1)}) \\
&- \sum_{i=1}^n F_n(X_{(i-1)})(F_0^2(X_{(i)}) - F_0^2(X_{(i-1)})) + \frac{1}{3} \\
&= \sum_{i=1}^n \frac{(i-1)^2}{n^2} - \sum_{i=1}^n \frac{i}{n}(F_0^2(X_{(i)}) - F_0^2(X_{(i-1)})) + \frac{1}{3}.
\end{aligned}$$

С обзиром на то да је, уколико важни нулта хипотеза, нађена гранична расподела за статистику $n\omega_n^2$ у већини софтверских пакета је имплементирана управо та вредност тест статистике (и за мале и за велике вредности обима узорка). За мале обиме узорка се таблица критичних вредности може једноставно добити Монте Карло методама, а свакако да уколико се користе вец имплементиране функције, то је већ аутомацки урађено.

- Тест Андерсон-Дарлинга

$$A_n = \int_{-\infty}^{\infty} \frac{(F_n(x) - F_0(x))^2}{F_0(x)(1 - F_0(x))} dF_0(x). \tag{3.2}$$

И овај тест припада групи тестова чија расподела, уколико је H_0 тачна, не зависи од F_0. Вредности подинтегралне функције ће бити блиске нули уколико је H_0 тачна, док ако није, вредности ће бити веће од нуле. Због тога је критична област облика $W = \{A_n \geq C\}$. Функција у R-у коју користимо је *ad.test* из пакета *goftest*. Уколико је потребно, тест статистика (3.2) се може написати у погодној форми за изражунавање на исти начин као што смо то урадили у случају статитике (3.1). Поред тога, како, уколико важни H_0, статистика nA_n има граничну расподелу, управо је та верзија најчешће имплементирана у софтверским пакетима.

У случају да желимо да тестирамо хипотезу $H_0 : \ F = F_0(\theta)$, где је θ непознат параметар (или више њих), можемо адаптирати претходно описане тестове и то на следећи начин: оценимо θ методом максималне веродостојности, а затим применимо претходне тестове за тестирање сагласности са $F_0 = F_0(\hat{\theta})$. У неким (честим ситуацијама) кад је θ параметар скалирања или локације, може се опет показати да расподела статистика под нултом хипотезом не зависи од θ, али је *различита од оне која се добија без оцењивања параметара.* Најједноставнији начин да дођемо до те расподеле је да је оценимо симулацијама (Монте-Карло методом). У случају тестирања сагласности са нормалном $\mathcal{N}(m, \sigma^2)$ расподелом ова модификација теста Колмогоров-Смирнова се назива још и Лилифорсова модификација. У R-у се за примену исте може користити функција LcKS из пакета KScorrect.

Слећим примером илустроваћемо како можемо извршити тестирање користећи сва три поменута теста сагласности када расподела тест статистике у случају важења нулте хипотезе, не зависи од тих параметара.

Пример 3.2.6. *Тестираћемо да ли заради просветних радника из примера 1.2.2 имају нормалну расподелу. За то ћемо применити класичне тестове, али, пошто немамо претпоставку о параметрима, мораћемо да искористимо њихове модификације са оцењеним параметрима. Прво морамо да оценимо расподеле*

тест статистика под нултом хипотезом. Нека је $\alpha = 0.05$ *за који ћемо да извршимо тестирање. Прво ћмо одредити критичне области за тестирање.*

```
library(goftest)
N=10000
alfa=0.05
Tks0=rep(0,N)
Tcvm0=rep(0,N)
Tad0=rep(0,N)
n=195
set.seed(1)
for(i in 1:N)
{
uzorak=rnorm(n,0,1)
Tks0[i]=ks.test(uzorak,'pnorm',mean(uzorak),sd(uzorak))$stat
Tcvm0[i]=cvm.test(uzorak,'pnorm',mean(uzorak),sd(uzorak))$stat
Tad0[i]=ad.test(uzorak,'pnorm',mean(uzorak),sd(uzorak))$stat
}
# одређујемо критичне вредности
(Cks=quantile(Tks0,1-alfa))
(Ccvm=quantile(Tcvm0,1-alfa))
(Cad=quantile(Tad0,1-alfa))
#реалзиована вредности тест статистика са оцењеним параметрима
m=mean(zaradeJanuar)
so=sd(zaradeJanuar)
(Tks=ks.test(zaradeJanuar,'pnorm',m,so)$stat)
(Tcvm=cvm.test(zaradeJanuar,'pnorm',m,so)
(Tad=ad.test(zaradeJanuar,'pnorm',m,so)$stat)
```

Добили смо да су критичне области за тестирање редом $\{D_{195} \geq 0.06\}$, $\{195\omega^2_{195} \geq 0.13\}$ *и* $\{195A_{195} \geq 0.76\}$[5] *а реализоване вредности тест статистика* $\hat{D}_{195} = 0.18$, $195 \cdot \hat{\omega}^2_{195} = 1.84$ *и* $195\hat{A}_{195} = 10.52$ *па хипотезу нормалности одбацујемо. До истог закључка смо могли доћи оцењивањем p-вредности теста за чије рачунање користимо оцењену расподелу тест статистике по нултом хипотезом.*

[5] у *R*-у су имплементиране статисгтике $n\omega^2_n$ и nA_n

```
(1-ecdf(Tks0)(Tks))
(1-ecdf(Tcvm0)(Tcvm))
(1-ecdf(Tad0)(Tad))
```

У сва три случаја добили смо да је p-вредност 0 што јасно упућује на одбацивање нулте хипотезе нормалности.

Задатак 3.2.5. Извести погодан израз за израчунавање Андерсон-Дарлингове статистике, по угледу на исти у случају статистике Крамер-фон Мизеса.

Задатак 3.2.6. Напаравити таблице критичних вредности за тестирање просте хипотезе сагласности са расподелом за статистике наведене у овом поглављу.

Задатак 3.2.7. Тест статистика Колмогоров-Смирнова (и остале наведене у овом поглављу) се често користе за тестирање сагласности са експоненцијалном $\mathcal{E}(\lambda)$ расподелом. Показати да, уколико се λ оцењује методом максималне веродостојности, расподела тест статистике под нултом хипотезом не зависи од λ.

Задатак 3.2.8. Направите емпиријску студију у којој ћете упоредити моћи тестова наведених у овом поглављу за тестирање H_0 $X \sim \mathcal{E}(1)$ уколико су алтернативне расподеле $\mathcal{E}(1.5)$ и $\gamma(1.5, 1)$ и оим узорка $n = 20$.

Задатак 3.2.9. Тестирајте H_0 да обележја из задатка 3.2.4 имају нормалне расподеле са произвољним параметрима.

χ^2-тест сагласности са расподелом

За разлику од тестова из претходног одељка, овај тест се може применити и када X није апсолутно непрекидна случајна величина. Пре конструктције тест статистике потребно је да се скуп вредности обележја X подели у k дисјунктних категорија а затим, да се преброји број елемената из узорка у свакој од категорија. Тест статистика ће управо бити заснована на разлици овог броја и очекиваног броја елемената из узорка у свакој од категорија.

Нека је M_j број елемената у j-тој категорији. Приметимо да тада M_j има биномну $\mathcal{B}(n, p_j)$, при чему ако је нулта хипотеза тачна је $p_j = P_{H_0}\{X$ је у j-тој категорији$\}$. Одавде је $EM_j = np_j$. Сада можемо конструисати тест статистику

$$T_n = \sum_{j=1}^{k} \frac{(M_j - np_j)^2}{np_j}.$$

Познато је да уколко је H_0 тачно, за велике обиме узорка, T_n има χ^2_{k-1} расподелу. Како, уколико је нулта хипотеза тачна, очекиван број елемената у свакој од категорија не би требало много да се разликује од реализованог, критична област је природно облика $W = \{T_n \geq C\}$, осим уколико не желимо и да се штитимо од "намештања података" (када су нам превише мале вредности такође сумњиве).

*Н*апомена: Уколико је $np_j < 5$ треба спојити категорије. Поред тога треба водити рачуна да је овај тест предвиђен за веће обиме узорка.

Пример 3.2.7. *Желимо да проверимо да ли је коцкица за игру заиста хомогена, односно да ли је вероватноћа да падне било који од бројева* $\{1, 2, 3, 4, 5, 6\}$ *заиста* $\frac{1}{6}$. *Дакле, обележје* X *које посматрамо је број који се добије у бацању коцкице, а нулта хипотеза је* $H_0 : P\{X = k\} = \frac{1}{6},\ k = 1, 2..., 6$. *Алтернативна хипотеза је да* H_0 *не важи. Бацили смо коцкицу 60 пута и добили следећи резултат:*

	1	*2*	*3*	*4*	*5*	*6*
M_j	*9*	*12*	*10*	*10*	*9*	*11*
np_j	*10*	*10*	*10*	*10*	*10*	*10*

Реализована вредност тест статистике је

$$\hat{T}_{60} = \frac{(9-10)^2}{10} + \frac{(12-10)^2}{10} + \frac{(10-10)^2}{10} + \frac{(10-10)^2}{10} + \frac{(9-10)^2}{10} + \frac{(11-10)^2}{10} = \frac{7}{10}.$$

Како је критична област облика $W = \{T_{60} \geq C\}$, *и* T_{60}, *ако је* H_0 *тачна има* χ^2_5 *расподелу, добијамо да је p-вредност теста*

$p = P\{T \geq C\} = 1 - F_{\chi^2_5}(0.7) = 0.98$, *па прихватамо* H_0. *За тестирање можемо користити и уграђену функцију* chisq.test *у R-у, али је за коришћење исте потребно да имамо низ фрквенција по категоријама, као низ и вероватноћа да се припадне свакој од категорија.*

```
chisq.test(x=c(9,12,10,10,9,11),p=rep(1/6,6),correct=FALSE)
```

Ако F_0 зависи од непознатих параметара онда прво те параметре оцењујемо методом максималне веродостојности а затим вероватноће p_j одређујемо користећи управо те оцењене параметре. Тест статистика остаје иста али је сада њена расподела за велие обиме узорка, уколико важи H_0, $\chi^2_{k-1-\text{број оцењених параметара}}$.

Пример 3.2.8. *Желимо да тестирамо да ли узорак зарада из примера 1.2.2 упућује на то да се зараде просветних радника могу моделирати нормалном* $\mathcal{N}(m, \sigma^2)$ *расподелом. Како немамо претпоставку о параметрима расподеле, оценићемо их методом максималне веродостојности. Добијамо да је*

$$\hat{m} = \bar{x}_{195} = 44687.38, \quad \hat{\sigma} = \bar{s}^2_{195} = 7063.472.$$

Поделићемо узорак у категорије. Ту имамо слободу како да то урадимо. Један начин је да извршимо поделу домена $\mathbb{R}$ *на исти начин као кад формирамо хистограм. Вероватноће* $p_1, p_2, p_3, p_4, p_5, p_6$ *и* p_7 *смо одредили на следећи начин:*

$$\begin{aligned}
p_1 &= P\{X \leq 38593.1\} = \Phi\Big(\frac{38593.1 - \hat{m}}{\hat{\sigma}}\Big) = 0.1947366 \\
p_2 &= P\{38593.1 < X \leq 44250.3\} \\
&= \Phi\Big(\frac{44250.3 - \hat{m}}{\hat{\sigma}}\Big) - \Phi\Big(\frac{38593.1 - \hat{m}}{\hat{\sigma}}\Big) = 0.2806563 \\
p_3 &= P\{44250.3 < X \leq 49907.5\} \\
&= \Phi\Big(\frac{49907.5 - \hat{m}}{\hat{\sigma}}\Big) - \Phi\Big(\frac{44250.3 - \hat{m}}{\hat{\sigma}}\Big) = 0.2940863
\end{aligned}$$

$$
\begin{aligned}
p_4 &= P\{49907.5 < X \leq 55564.7\} \\
&= \Phi\Big(\frac{55564.7 - \hat{m}}{\hat{\sigma}}\Big) - \Phi\Big(\frac{49907.5 - \hat{m}}{\hat{\sigma}}\Big) = 0.1682498 \\
p_5 &= P\{55564.7 < X \leq 61221.9\} \\
&= \Phi\Big(\frac{61221.9 - \hat{m}}{\hat{\sigma}}\Big) - \Phi\Big(\frac{55564.7 - \hat{m}}{\hat{\sigma}}\Big) = 0.05249501 \\
p_6 &= P\{61221.9 < X \leq 66879.1\} \\
&= \Phi\Big(\frac{66879.1 - \hat{m}}{\hat{\sigma}}\Big) - \Phi\Big(\frac{61221.9 - \hat{m}}{\hat{\sigma}}\Big) = 0.008912812 \\
p_6 &= P\{X \geq 66879.1\} = 1 - \Phi\Big(\frac{66879.1 - \hat{m}}{\hat{\sigma}}\Big) = 0.0008631106.
\end{aligned}
$$

Тако добијамо

	$[-\infty, 38593.1]$	$(, 44250.3]$	$(, 49907.5]$	$(, 55564.7]$	$(, 61221.9]$	$(, 66879.1]$	$(, \infty]$
M_j	22	99	44	16	6	5	3
np_j	37.97	54.73	57.35	32.81	10.24	1.74	0.17

На основу табеле 3.2.8 закључујемо да последње три категорије треба спојити.

	$[-\infty, 38593.1]$	*(, 44250.3]*	*(, 49907.5]*	*(,55564.7]*	$(, \infty)$
M_j	*22*	*99*	*44*	*16*	*14*
np_j	*37.97*	*54.73*	*57.35*	*32.81*	*12.14*

Тест статистика, под нултом хипотезом има χ^2_2 *расподелу. Критична област је облика* $W = \{T_{195} \geq C\}$. *За ниво значајности теста* $\alpha = 0.05$ *добијамо да је* $C = 5.99$. *Реализована вредност тест статистике је* $\hat{T}_{195} = 54.535$ *па одбацујемо хипотезу нормалности. До истог закључка можемо доћи рачунањем p-вредности теста за коју се добија да је мања од* 10^{-11}.

Задатак 3.2.10. Желимо да тестирамо да ли се број позива у току дана може моделирати Пуасоновом расподелом. резултати истраживања приказани су у следећој табели:

бр. позива	$[0, 4]$	$(4, 6]$	$(6, 8]$	$[8, 10]$
	15	13	10	7

Применом χ^2 теста, са нивоом значајности теста 0.05 проверити хипотезу.

Задатак 3.2.11. Циљ једног истраживања био је да се установи да ли чланови породице гласају независно један од другог на политичким изборима. Случајно је одабрано 280 породица са по 3 члана која имају могућност изласка на изборе и постављено им је питање на које могу да одговоре са ДА или НЕ, и за сваку породицу је забележен број позитивних одговора. Уколико постоји независност при гласању у оквиру исте породице онда број позитивних одговора можемо моделирати одговарајућом биномном расподелом. Добијени су следећи резултати:

бр. позитивних одговора	0	1	2	3
бр. породица	50	80	85	65

Применом χ^2 теста, са нивоом значајности теста 0.05 проверити хипотезу.

Задатак 3.2.12. Како би се проверио квалитет радном генератора који је имплементиран у програму R, генерисано је 200 псеудослучајних бројева. Са нивоом значајности теста $\alpha = 0.001$ тестирати хипотезу да генерисан узорак потиче из униформне $\mathcal{U}[0,1]$ расподеле. Генерисан узорак је дат у табели испод.

0.28	0.18	0.20	0.50	0.25
0.00	0.44	0.68	0.33	0.22
0.51	0.91	0.36	0.41	0.49
0.01	0.85	0.35	0.20	0.65
0.06	0.73	0.06	0.81	0.33
0.95	0.57	0.48	0.64	0.86
0.09	0.48	0.40	0.28	0.64
0.29	0.33	0.02	0.10	0.01
0.88	0.16	0.13	0.26	0.53
0.12	0.48	0.40	0.06	0.83

Задатак 3.2.13. Проверити, са нивоом значајности $\alpha = 0.1$, да ли се следећи подаци могу моделирати Геометријском $\mathcal{G}(p)$ расподелом.

4	8	4	2	2	2	1	2	8
5	2	8	4	7	6	4	7	2
2	6	2	3	5	1	3	2	3
4	1	2	6	3	3	4	9	2
1	5	7	1	5	4	3	4	7

Задатак 3.2.14. На претходним изборима су учествовале четири политичке странке и финална расподела гласова (у процентима) је

странка	I	II	III	IV
удео	54	13	20	13

Ускоро се ближе нови избори, странке које ће се кандидовати су исте као на претходним изборима. Случајно је одабрано 200 грађана који су анкетирани о томе за кога би гласали, и након што су избачени одговори попут " не желим да се изјасним" добијени су следећи резултати

странка	I	II	III	IV
број	85	20	22	13

Да ли на основу тога закључујемо да је дошло до промене мишљења јавног мњења?

3.2.5 Тестови о једнакој расподељености два узорка

Један од тестова из ове категорије је и већ описани Вилкоксонов тест који смо користили за тестирање нулте хипотезе да два независна обележја, за који важи да имају исту класу расподеле до на непознат параметар локације, имају и исти параметар локације. Уколико се хипотеза одбаци она може упућивати и на то да је сама полазна претпоставка о припадности истој класи расподела није тачна.

Поред овог теста, за тестирање нулте хипотезе о једнакости расподела два обележја, можемо направити аналогне тестове класичним тестовима сагласности, само за два узорка. Нека су F и G функције расподела апсолутно непрекидних обележја X и Y. Тада се нулта хипотеза H_0 : "обележја X и Y су једнако расподељена",

може формулисати и преко њихових функција расподела, односно $F(x) = G(x)$ скоро за свако $x \in \mathbb{R}$. Зато је природан начин да се конструише тест управо на основу разлика емпиријских функција расподела $F_{n_1}(x)$ и $G_{n_2}(x)$. Нека је $N = n_1 + n_2$ величина обједињеног узорка. Неки од тестова који користе ову идеју су:

- Тест Колмогоров-Смирнова

$$T_{n_1,n_2} = \sup_{x \in \mathbb{R}} |F_{n_1}(x) - G_{n_2}(x)|.$$

  За $\sqrt{\frac{n_1 n_2}{N}} T_{n_1,n_2}$ је нађена гранична расподела под нултом хипотезом.

- Крамер-фон Мизесов тест

$$T_{n_1,n_2} = \int_{-\infty}^{\infty} (F_{n_1}(x) - G_{n_2}(x))^2 dH_N(x),$$

  где је $H_N(x)$ емпиријска функција расподеле обједињеног узорка. За $\frac{n_1 n_2}{N} T_{n_1,n_2}$ је нађена гранична расподела под нултом хипотезом.

За велике обиме узорака, за одређивање критичне области, се користе критичне вредности одређене на основу граничних расподела, док су за мале вредности обима узорака одређене егзактне расподеле. У R-у се за први тест користи иста функције као у једнодимензионом случају, док се за се Крамер-фон Мизесов и користи функција $\mathrm{cvm_t est}$ из пакета twosamples.

Пример 3.2.9. *Циљ истраживања био је да се утврди да ли су расподеле времена чекања између позива у две такси станице исте. Случајно је одабран једночасовни интервал у току једног дана и посматрана су времена између позива, а затим из сакупљених резултата узети случајни узорци обима 20. Добијени су следећи резултати (у секундама):*

А:	*16.0*	*139.9*	*0.4*	*84.1*	*49.3*
	17.1	*1.7*	*25.7*	*9.4*	*2.9*
	16.0	*16.7*	*23.0*	*21.9*	*55.7*
	27.7	*25.7*	*0.2*	*20.1*	*26.8*
Б:	*13.1*	*17.2*	*86.6*	*19.2*	*24.8*
	36.3	*49.0*	*35.1*	*11.9*	*12.5*
	25.6	*79.8*	*16.9*	*14.1*	*30.5*
	55.8	*12.4*	*14.4*	*22.3*	*24.6*

Применићемо оба описана теста.

```
A=c(16.0,139.9,0.4,84.1,49.3,
17.1,1.7,25.7,9.4,2.9,
16.0,16.7,23.0,21.9,55.7,
27.7,25.7,0.2, 20.1,26.8)
B=c(13.1,17.2,86.6,19.2,24.8,
36.3,49.0,35.1,11.9,12.5,25.6,
79.8,16.9,14.1,30.5,
55.8,12.4,14.4,22.3,24.6)
ks.test(A,B)
library("twosamples")
cvm_test(A, B)
```

У случају Колмогоров-Смирнов теста добија се p-вредност 0.67, док у случају Крамер-фон Мизесовог теста, добија се p-вредност 0.66, па нулту хипотезу нећемо одбацити.

Задатак 3.2.15. Генерисати два независна обима $n = 20$ узорка из $\mathcal{N}(0,1)$ и $\mathcal{N}(1,1)$ и проверити да ли ће тестови описани у овом поглављу одбацити хипотезу о једнакој расподељености ова два обележја.

Задатак 3.2.16. Проверити, на основу следећих података, да ли су расподеле поена мушких и женских студената на једном испиту исте.

мушки:	77	84	89	76	81	85	90
	90	69	84	77	82	90	94
	83	83	64	84	63	85	70
	66	58	70	85			
женски:	79	82	127	73	59	95	83
	95	80	78	88	100	49	77
	62	81	83	69	69	76	75
	91	49	68	81			

За податке приказати и кутијасте дијаграме. Да ли графички приказ указује на резултат тестирање?

3.2.6 Тестови независности

χ^2 тест

Желимо да тестирамо H_0 да су обележја X и Y независна. Подсетимо се да су обележја X и Y независна ако свака два скупа A и B $P\{X \in A, Y \in B\} = P\{X \in A\}P\{Y \in B\}$. Зато ћемо формирати $K \times L$ категорија (K категорија $A_1, ..., A_k$ за скуп вредности X и L категорија $B_1, ..., B_L$ за вредности Y). На располагању имамо п.с.у. $(X_1, Y_1), ..., (X_n, Y_n)$. Означимо са M_{ij} број елемената из узорка чија се X-компонента налази у i-тој категорији, и Y-компонента у j-тој категорији. Случајна величина M_{ij} има Биномну $\mathcal{B}(n, p_{ij})$ расподелу, где је $p_{i,j} = P\{X \in A_i, Y \in B_j\}$, а матрица $\{M_{ij}\}$ се назива *табелом контигенције*. Приметимо да, ако је хипотеза H_0 тачна, онда је $p_{ij} = p_{i,\cdot} \cdot p_{\cdot,j}$, где су $p_{i,\cdot} = P\{X \in A_i\}$ и $p_{\cdot,j} = P\{Y \in A_j\}$ маргиналне вероватноће. Тада сличним резоновањем као у χ^2-тесту сагласности са расподелом, формирамо тест статистику

$$T_n = \sum_{i=1}^{K} \sum_{j=1}^{L} \frac{(M_{ij} - n\hat{p}_{ij})^2}{n\hat{p}_{ij}},$$

при чему је $\hat{p}_{ij}$, оцењено на основу узорка под претпоставком да важи H_0, једнако

$$\hat{p}_{ij} = \hat{p}_{i,\cdot}\hat{p}_{\cdot,j} = \frac{\sum_{j=1}^{L} M_{ij}}{n} \cdot \frac{\sum_{i=1}^{K} M_{ij}}{n}.$$

Уколико важи H_0, тест статистика T_n има асимптотски $\chi^2_{(K-1)(L-1)}$ расподелу јер је број категорија $K \cdot L$ а број оцењених параметара $K-1+L-1$ јер за маргиналне вероватноће постоји ограничење

$$\sum_{i=1}^{K} p_{i,\cdot} = 1 = \sum_{j=1}^{L} p_{\cdot,j}.$$

*Н*апомена: И овде морамо извршити груписање категорија уколико је $n\hat{p}_{ij} < 5$.

У *R*-у се за примену овог теста користи функција chisq.test.

Пример 3.2.10. *Циљ истраживања је да се испита да ли постоји веза између учесталости физичке активности и пушења на студентској популацији. Разликују се следеће категорије:*

- *X (физичка активност):*
  1. *често*
  2. *понекад*
  3. *никад;*
- *Y (пушење):*
  1. *интензивно*
  2. *свакодневно*
  3. *понекад*
  4. *никад;*

Узет је узорак од 236 студената једног универзитета и добијени су следећи резултати:

$Y \setminus X$	*1.*	*2.*	*3.*	$\sum$
1.	*7*	*1*	*3*	*11*
2.	*87*	*18*	*84*	*189*
3.	*12*	*3*	*4*	*19*
4.	*9*	*1*	*7*	*17*
$\sum$	*115*	*23*	*98*	*236*

Одговарајућа табела очекиваних вредности у свакој од категорија, изгледа овако:

$Y \setminus X$	*1.*	*2.*	*3.*	$\sum$
1.	*5.36*	*1.07*	*4.57*	*11*
2.	*92.10*	*18.42*	*78.48*	*189*
3.	*9.26*	*1.85*	*7.89*	*19*
4.	*8.28*	*1.66*	*7.06*	*17*
$\sum$	*115*	*23*	*98*	*236*

Видимо да морамо да спојимо неке од категорија. Један начин за то је да се споје категорије 2. и 3. обележја X. Сада добијамо следећу табелу очекиваних фреквенција:

$Y \setminus X$	*1.*	*2. и 3.*	$\sum$
1.	*5.36*	*5.64*	*11.00*
2.	*92.10*	*96.90*	*189.00*
3.	*9.26*	*9.74*	*19.00*
4.	*8.28*	*8.72*	*17.00*
$\sum$	*115.00*	*121.00*	*236.00*

Добијамо да је реализована вредност тест статистике

$$\hat{T}_{236} = \frac{(7-5.36)^2}{5.36} + \frac{(4-5.64)^2}{5.64} + \frac{(87-92.10)^2}{192.10} + \frac{(102-96.90)^2}{96.90} + \frac{(12-9.26)^2}{9.26} + \frac{(7-9.74)^2}{9.74} + \frac{(9-8.28)^2}{8.28} + \frac{(8-8.72)^2}{8.72} = 3.23.$$

Како T_{236}, уколико важи H_0, има χ^2_3 расподелу, добијамо да је p-вредност $P\{T_{236} \geq 3.23\} = 0.357$ па не одбацујемо нулту хипотезу. Овај резултат се може објаснити тиме да је узет узорак искључиво из студенцке популације па можда још није дошло до негативног утицаја пушења на здравље.

```
M=matrix(c(7,1,3,87,18,84,12,3,4,9,1,7),ncol=3,by=2)
chisq.test(as.table(M))

chisq.test(as.table(M))$expected # приказујемо табелу очекиваних фреквенција како бисмо знали које категорије да спојимо
M2=cbind(M[,1],M[,2]+M[,3])
```

```
chisq.test(as.table(M2))
```

Задатак 3.2.17. Испитати независност типа пута и типа несрећа које се дешавају на путу на основу података презентованих у примеру 1.2.1.

Задатак 3.2.18. На једној научној конференцији учесници су могли да одаберу боју мајце коју ће добити као саставни део конференцијског материјала. Табеларни приказ жеља изгледа овако:

пол\боја	бела	плава	црвена
мушки	30	30	10
женски	25	25	25

Проверити, са нивоом значајности теста $\alpha = 0.1$ да ли одабир боје мајце зависи од пола.

Задатак 3.2.19. У R пакету FactoMineR налази се база података housetasks у коме се налази расподела кућних послова по половима. Шта можете да закључите о независности типа посла и пола који га обавља?

Задатак 3.2.20. Испитати, на основу података из задатка 3.2.16 да ли су број поена на испиту и пол независна обележја.

Пирсонов и Спирманов тест некорелисаности

Уколико су обележја независна онда су и некорелисана, док обрнуто не важи. Зато уколико желимо да одбацимо хипотезу независности довољно је проверити некорелисаност обележја.

Подсетимо се, две случајне величине (обележја) X и Y су некорелисане уколико је $cov(X,Y) = E((X - EX)(Y - EY)) = EXY - EX \cdot EY = 0$. Ово је еквивалентно са тим да је коефицијент корелације, дефинисан са

$$\rho = \frac{\mathrm{cov(X, Y)}}{\sqrt{DX} \cdot \sqrt{DY}},$$

једнак нули.
Оцена методом замене за ρ је

$$\hat{\rho}_n = \frac{\sum_{i=1}^n (X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{\sqrt{\sum_{i=1}^n (X_i - \bar{X}_n)^2} \cdot \sqrt{\sum_{i=1}^n (Y_i - \bar{Y}_n)^2}}.$$

Ова оцена је позната под називом *Пирсонов коефицијент корелације.*

Како је $|\rho| = 1$ ако и само ако између обележја постоји линеарна веза, вредности блиске ± 1 упућују на јаку корелисаност, а вредности блиске 0 на некорелисаност. Зато је природно да управо $\hat{\rho}_n$ буде статистика за тестирање хипотезе о неорелисаности два обележја. Једина препрека је што расподела ове тест стастистике, под нултом хипотезом, евидентно зависи од расподеле обележја X и Y. Познато је да, уколико H_0 важи и ако X и Y имају нормалну расподелу, онда $T_n = \hat{\rho}_n \sqrt{\frac{n-2}{1-\hat{\rho}_n^2}}$ има t_{n-2} расподелу. Уколико обележја долазе из других расподела, расподелу те статистике морамо оценити. Најчешће се то ради неком методм реузорковања, али то излази из оквира овог уџбеника.

Уколико сумњамо на позитивну или негативну корелисаност онда ће критична област бити једнострана, у супротном ће бити облика $W = \{|T_n| \geq C\}$.

Како бисмо се "ослободили" од претпоставке о расподели обележја можемо одредити емпиријски коефицијент корелације за статистике ранга $(R_1, ..., R_n)$ и $(S_1, ..., S_n)$, односно

$$\hat{r}_n = \frac{\sum_{i=1}^n (R_i - \bar{R}_n)(S_i - \bar{S}_n)}{\sqrt{\sum_{i=1}^n (R_i - \bar{R}_n)^2} \cdot \sqrt{\sum_{i=1}^n (S_i - \bar{S}_n)^2}},$$

Овај коефицијент корелације је познат под називом *Спирмнаов коефицијент корелације.* Може се показати да за велико n, и када нема понављања елемената из узорка, уколико су обележја независна, расподела нормализоване статистике се може апроксимирати нормалном. Односно, важи

$$\frac{\hat{r}_n}{\sqrt{\frac{1}{n-1}}} \sim \mathcal{N}(0, 1),$$

док се већ за $n > 10$ може користити апроксимација $\hat{r}_n\sqrt{\frac{n-2}{1-\hat{r}_n^2}}$ Студентовом t_{n-2} расподелом. У пракси се овај тест користи и када има понављања елемената узорка, али се тада рангови додељују на већ описан начин (за исте елементе се рачуна средња вредност рангова).

Важно је приметити главну разлику између Пирсоновог и Спирмановог теста. Наиме, Пирсонов коефицијент корелације је добар да детектује линеарну зависност док је Спирманов добар да детектује када је једна променљива било која монотона трансформација друге променљиве.

У R-у су оба представљена теста имплементирана функцијом cor.test. За мале обиме узорка p-вредности су одређене на основу егзактне расподеле тест статистике, док је у случају великих узорака коришћена асимптотска апроксимација.

Пример 3.2.11. *Генерисаћемо три узорка обима 50: један који представља обележје X са $\mathcal{N}(1,1)$ расподелом други који представља обележје $Y = 0.4X + Z$, где је $Z \sim \mathcal{N}(0, 0.1^2)$ и трећи који представља обележје $Q = \tan(X)$ и одредити Пирсонов и Спирманов коефицијент (и извршити тестирање независности):*

```
set.seed(117)
x=rnorm(50,1,1)
y=0.4*x+rnorm(50,0,0.1)
q=tan(x)
cor(x,y,method=c('pearson'))
cor.test(x,y,method='pearson')
cor(x,y,method=c('spearman'))
cor.test(x,y,method='spearman')
cor(x,q,method=c('pearson'))
cor.test(x,q,method='pearson')
cor(x,q,method=c('spearman'))
cor.test(x,q,method='spearman')
```

Резултати су приказани у следећој табели:

	$\hat{\rho}_{50}$	*p-вред*	$\hat{r}_{50}$	*p-вред*
(X, Y)	*0.97*	$< 10^{-16}$	*0.96*	$< 10^{-16}$
(X, Q)	$2.8 \cdot 10^{-4}$	*1*	*-0.35*	*0.01*

Видимо да је у случају кад су X и Y са нормалним расподелама, Пирсонов коефицијент корелације је нешто ближи 1, док је у случају да то није испуњено, за разлику од Пирсоновог коефицијента корелација, Спирманов коефицијент корелације показао значајна одступања од нуле што упућује на зависност између обележја X и Q.

Задатак 3.2.21. Испитати,коришћењем тестова из овог поглавља, на основу података из задатка 3.2.16 да ли су број поена на испиту и пол независна обележја.

4

Регресиони модели

У претходном поглављу видели смо како можемо да закључимо асоцијацији између два обележја. Сада ћемо да видимо како можемо моделирати зависност једне променљиве у односу на другу.

Са речју "регресија" математичари су се први пут сусрели у раду Ф. Галтона, *Regression toward mediocrity in hereditary stature* из 1855. године. Он је дошао до закључка да синови веома високих очева нису тако високи. Иако је Галтон разлог за то пронашао у генетици, његов пример иницирао је проучавање ове теме од стране статистичара и тако почиње развој ове веома значајне статистичке области.

Случајна величина $f(X) = E(Y|X)$ назива се *регресиона функција*, при чему X може бити вишедимензиона случајна величина. Модели чији је циљ моделовање ове зависности се називају *регресиони модели*. Случајна величина Y се назива зависна променљива, а X независна или предиктор. Приметимо да је улога регресионе функције да опише понашање зависне променљиве када је позната независна. Може се показати да је регресиона функција је права линија ако и само ако случајни вектор $(X, Y)^T$ има вишедимензионална нормалну расподелу. Регресиону праву има смисла конструисати и када знамо да заједничка расподела није нормална. Тада је то права која од свих правих линија најбоље описује зависност између Y и X у смислу средњеквадратног одступања. Ре-

гресиони модел се може представити у облику

$$Y = f(X) + \varepsilon,$$

где је ε случајна промењлива независна од X, најчешће са нормалном $\mathcal{N}(0, \sigma^2)$ расподелом. Како нам је циљ да моделирамо ову зависност можемо X сматрати познатом детерминистичком величином као и да читава случајност Y долази од случајних грешака. Формално овако постављен проблем називамо *контролисана регресија*. Од сада па надаље претпоставићемо да се ради о овом типу регресије.

Дакле, наш главни задатак је да оценимо зависност која постоји између зависне променљиве и предиктора (необавезно једног). Имамо две могућности: да претпоставимо функционалну зависност која зависи од неки параметара и да те параметре оценимо (параметарски приступ), или да непараметарски оценимо ту функцију. У неколико наредних одељака представићемо најједноставније параметарске регресионе моделе.

4.1 Проста линеарна регресија

Претпоставићемо да на располагању имамо узорак $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$ и да је модел који желимо да применимо

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

где је $\{\varepsilon_i\}$ низ случајних величина који задовољава услове[1]:

1. $E(\varepsilon_i) = 0$, за $i = 1, 2, ..., n$;
2. $E(\varepsilon_i \varepsilon_j) = 0$, за $i \neq j$;
3. $D(\varepsilon_i) = \sigma^2 < \infty$.

[1] Ови услови су познати као услови Гаус-Маркова

Поред ових услова, уколико не претпостављамо да се ради о контролисаној регресији, треба додати и услов да су, за свако $i = 1, 2..., n$, случајне величине ε_i и X_i независне. Управо овај услов нам омогућава да без умањења општости можемо претпоставити да се ради о контролисаној регресији.

Пре него што пређемо на оцењивање непознатих параметара, важно је напоменути да се "линеарност" не односи на "линеарност" по предиктору већ по параметрима!

Један од могућих начина да оценимо параметре је да их оценимо оним вредностима који минимизирају суму квадратних одступања оцењене од праве вредности, односно

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - (\beta_1 X_i + \beta_0))^2.$$

Овај приступ је природан јер је $E(Y|X)$ функција $f(X)$ која минимизира растојање $E(Y - f(X))^2$.
Решавамо систем једначина

$$\frac{\partial S(\beta_0, \beta_1)}{\beta_0} = \sum_{i=1}^{n} -2(Y_i - \beta_1 X_i + \beta_0) = 0 \tag{4.1}$$

$$\frac{\partial S(\beta_0, \beta_1)}{\beta_1} = \sum_{i=1}^{n} -2X_i(Y_i - \beta_1 X_i + \beta_0) = 0 \tag{4.2}$$

$$\tag{4.3}$$

Добијамо да су тражене оцене коефицијената $\hat{\beta}_0$ и $\hat{\beta}_1$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} Y_i X_i - n\bar{X}_n\bar{Y}_n}{\sum_{i=1}^{n} X_i^2 - n\bar{X}_n^2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{\sum_{i=1}^{n}(X_i - \bar{X}_n)^2} \tag{4.4}$$

$$\hat{\beta}_0 = \bar{Y}_n - \hat{\beta}_1\bar{X}_n. \tag{4.5}$$

Односно, оцењена регресиона функција је $\hat{\beta}_0 + \hat{\beta}_1 X$. Приметимо да ова права садржи тачку $(\bar{X}_n, \bar{Y}_n)$ што најбоље илуструје да овај приступ моделирању заправо има за циљ да добро опише тачке у близини просека. Приметимо још да полазни модел можемо написати у центрираном облику $Y_i = \beta_1(X_i - \bar{X}_n) + \beta_0 + \beta_1\bar{X}_n + \varepsilon_i$.

Испоставља се да је овај облик погоднији за прогнозирање јер $\hat{Y}_i = \hat{\beta}_1(X_i - \bar{X}_n) + \bar{Y}_n$.

Уколико важе наведени услови за низ грешака оцене $\hat{\beta}_0$ и $\hat{\beta}_1$ су непристрасне и постојане. Како би се ово показало, најбоље је да $\hat{\beta}_1$ прикажемо као линеарну комбинацију Y_i, односно у облику:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n Y_i(X_i - \bar{X}_n) - \bar{Y}_n \sum_{i=1}^n (X_i - \bar{X}_n)}{n\bar{S}_X^2} = \sum_{i=1}^n Y_i \cdot \frac{(X_i - \bar{X}_n)}{n\bar{S}_X^2}. \tag{4.6}$$

Сада је

$$\begin{aligned} E(\hat{\beta}_1) &= \sum_{i=1}^n E(Y_i) \cdot \frac{(X_i - \bar{X}_n)}{n\bar{S}_X^2} = \sum_{i=1}^n (\beta_0 + \beta_1 X_i) \cdot \frac{(X_i - \bar{X}_n)}{n\bar{S}_X^2} \\ &= \beta_0 \sum_{i=1}^n \frac{(X_i - \bar{X}_n)}{n\bar{S}_X^2} + \beta_1 \cdot \frac{\sum_{i=1}^n X_i^2 - n\bar{X}_n^2}{n\bar{S}_X^2} = 0 + \beta_1 \cdot \frac{n\bar{S}_X^2}{n\bar{S}_X^2} = \beta_1. \end{aligned}$$

Облик (4.6) нам је посебно погодан за испитивање свостава оцене јер из претпоставке да су ε_i и ε_j међусобно некорелисани добијамо да су и Y_i међусобно некорелисане случајне величине. Одавде је

$$\begin{aligned} D(\hat{\beta}_1) &= \sum_{i=1}^n D(Y_i) \cdot \frac{(X_i - \bar{X}_n)^2}{n^2\bar{S}_X^4} = \sum_{i=1}^n \sigma^2 \cdot \frac{(X_i - \bar{X}_n)^2}{n^2\bar{S}_X^4} = \sigma^2 \frac{n\bar{S}_X^2}{n^2\bar{S}_X^4} \\ &= \frac{\sigma^2}{n\bar{S}_X^2}. \end{aligned}$$

Јасно је да, уколико је $\sigma^2 < \infty$, $D(\hat{\beta}_1) \to 0$, кад $n \to \infty$, па је оцена постојана. Слично показујемо непристрасност и постојаност $\hat{\beta}_0$, користећи да се $\hat{\beta}_0$ може представити у облику

$$\hat{\beta}_0 = \sum_{i=1}^n \frac{Y_i}{n} - \sum_{i=1}^n Y_i \cdot \frac{\bar{X}_n(X_i - \bar{X}_n)}{n\bar{S}_X^2} = \sum_{i=1}^n \frac{Y_i}{n}\left(1 - \frac{\bar{X}_n(X_i - \bar{X}_n)}{\bar{S}_X^2}\right).$$

Уколико се уведе додатна претпоставка да грешке модела $\{\varepsilon_i\}$ представљају низ независних случајних величина са $\mathcal{N}(0, \sigma^2)$ онда

1) добијене оцене се поклапају са оценама добијеним методом маскималне веродостојности;

2) можемо одредити расподелу добијених оцена.

Да бисмо показали 1) приметимо да уколико ε_i има $\mathcal{N}(0,\sigma^2)$ онда Y_i има $\mathcal{N}(\beta_0+\beta_1 X_i)$ расподелу па је функција веродостојности

$$L(\beta_0,\beta_1)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(Y_i-(\beta_0+\beta_1X_i))^2}{2\sigma^2}},$$

односно њен логаритам је

$$l(\beta_0,\beta_1)=-n\log(\sqrt{2\pi})-\frac{n}{2}\log(\sigma^2)-\frac{S(\beta_0,\beta_1)}{2\sigma^2}.$$

Одавде се види да су вредности које максимизирају ову функцију управо оне које минимизирају $S(\beta_0,\beta_1)$. Поред тога добијамо да је оцена за σ^2

$$\tilde{\sigma}_n^2=\frac{S(\hat{\beta}_0,\hat{\beta}_1)}{n}=\frac{\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2}{n}. \tag{4.7}$$

Како су $\hat{\beta}_0$ и $\hat{\beta}_1$ линеарне комбинације нормално расподељених случајних величина, онда оне имају редом $\mathcal{N}(E(\hat{\beta}_0),D(\hat{\beta}_0))$ и $\mathcal{N}(E(\hat{\beta}_1),D(\hat{\beta}_1))$ расподеле. Имајући у виду шта су очекивања и дисперзије оцена добијамо да

$$\frac{\hat{\beta}_1-\beta_1}{\frac{\sigma}{\sqrt{n}S_X}}\sim\mathcal{N}(0,1)\quad\text{и}\quad\frac{\hat{\beta}_0-\beta_0}{\frac{\sigma}{\sqrt{n}}\sqrt{\left(1+\frac{\bar{X}_n^2}{S_X^2}\right)}}\sim\mathcal{N}(0,1).$$

Сада је јасно да можемо искористити управо ове функције од узорка уколико желимо да направимо интервале поверења за β_0 и β_1, и да тестирамо хипотезе да параметри имају неку одређену вредност. На пример, уколико желимо да видимо да ли постоји утицај предиктора на зависну променљиву тестираћемо $H_0:\ \beta_1=0$. За то, на основу претходног, можемо искористити статистику

$$T_n=\frac{\hat{\beta}_1}{\frac{\sigma}{\sqrt{n}S_X}}, \tag{4.8}$$

која, уколико је H_0 тачна, има $\mathcal{N}(0,1)$ расподелу. Међутим, ту наилазимо на препреку. Наиме, σ^2 је необзервабилан параметар

(не знамо га), па морамо да га оценимо. Једна могућност је да искористимо оцену (4.7). Међутим, чак и да немамо оцену методом максималне веродостојности, како је то параметар који представља дисперзију грешака, природно је да његова оцена буде у вези са дисперзијом оцењених грешака (резидуала модела). Означимо са $e_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_i X_i)$ резидуал i-те обзервације. Узорачка дисперзија резидуала је

$$\frac{1}{n}\sum_{i=1}^{n}(e_i - \bar{e})^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2 = \tilde{\sigma}_n^2.$$

Може се показати да је $E(\sum_{i=1}^{n} e_i^2) = (n-2)\sigma^2$ па оцена (4.7) није непристрасна. Зато ћемо σ^2 оценити са

$$\hat{\sigma}_n^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2,$$

која на основу претходног, представља непристрасну оцену дисперзије грешака модела. Тада

$$\frac{\hat{\beta}_1 - \beta_1}{\frac{\hat{\sigma}_n}{\sqrt{n}S_X}}$$

има Студентову t_{n-2} расподелу (доказ изостављамо). Зато ћемо, за тестирање $H_0: \beta_1 = 0$, уместо (4.8) користити статистику

$$T_n = \frac{\hat{\beta}_1}{\frac{\hat{\sigma}}{\sqrt{n}S_X}}. \tag{4.9}$$

Слично,

$$\frac{\hat{\beta}_0 - \beta_0}{\frac{\hat{\sigma}_n}{\sqrt{n}}\sqrt{1 + \frac{\bar{X}_n^2}{S_X^2}}}$$

има t_{n-2} расподелу па се на основу тога може извести интервал поверења за β_0 или тестирати хипотезе у вези са вредношћу параметра β_0.

Оцењена вредност за зависну променљиву Y_0 (и за регресиону функцију) када предиктор узима вредност X_0 је $\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 X_0$.

Како су и $\hat{\beta}_0$ и $\hat{\beta}_1$ линеарне комбинације нормално расподељених случајних величина, и $\hat{Y}_0$ ће то бити. Заиста,

$$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 X_0 = \frac{1}{n}\sum_{i=1}^{n} Y_i - \bar{X}_n \cdot \sum_{i=1}^{n} Y_i \cdot \frac{(X_i - \bar{X}_n)}{n\bar{S}_X^2}$$
$$+ X_0 \cdot \sum_{i=1}^{n} Y_i \cdot \frac{(X_i - \bar{X}_n)}{n\bar{S}_X^2} = \sum_{i=1}^{n} Y_i \cdot \Big(\frac{1}{n} + (X_0 - \bar{X}_n) \cdot \frac{(X_i - \bar{X}_n)}{n\bar{S}_X^2}\Big).$$

Сада је

$$E(\hat{Y}_0) = E(\hat{\beta}_0) + E(\hat{\beta}_1)X_0 = \beta_0 + \beta_1 X_0,$$
$$D(\hat{Y}_0) = \sum_{j=1}^{n} \Big(\frac{1}{n} + (X_0 - \bar{X}) \cdot \frac{(X_j - \bar{X}_n)}{n\bar{S}_X^2}\Big)^2 \cdot \sigma^2 = \sigma^2\Big(\frac{1}{n} + \frac{(X_0 - \bar{X}_n)^2}{n\bar{S}_X^2}\Big).$$

Добијамо да, за оцењену вредност $\hat{Y}_0$, на основу предиктора X_0, важи:

$$\frac{\hat{Y}_0 - EY_0}{\hat{\sigma}_n\sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X}_n)^2}{nS_x^2}}} \sim t_{n-2}$$

На основу овог својства можемо лако направити интервал поверења за **средњу вредност** зависне променљиве уколико је предиктор једнак X_0. Као и у претходним поглављима, можемо одредити константу C тако да је $\{|T_n| < C\} = \beta$ (јер је Студентова расподела симетрична). Добијамо да је $C = F^{-1}_{t_{n-2}}\big(\frac{1+\beta}{2}\big)$. Одавде је $\beta\%$ интервал поверења

$$\left(\hat{Y}_0 - C \cdot \hat{\sigma}_n\sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X}_n)^2}{nS_x^2}}, \hat{Y}_0 + C \cdot \hat{\sigma}_n\sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X}_n)^2}{nS_x^2}}\right).$$

Као што смо и очекивали, интервал је најужи за $X_0 = \bar{X}_n$ док како се одаљавамо од "средишта" података он постаје шири. Приметимо још да дужина интервала тежи нули како n тежи бесконачности, што је последица тога да $\hat{Y}_0$ постојана оцена за $\beta_0 + \beta_1 X_0$.

Како је $Y_0 = \beta_0 + \beta_1 X_0 + \varepsilon_0$, добијамо

$$E(\hat{Y}_0 - Y_0) = 0$$
$$D(\hat{Y}_0 - Y_0) = D(\hat{Y}_0) + D(\varepsilon_0) = \sigma^2\Big(\frac{1}{n} + \frac{(X_0 - \bar{X}_n)^2}{n\bar{S}_X^2}\Big) + \sigma^2$$

па важи

$$\frac{\hat{Y}_0 - Y_0}{\hat{\sigma}_n\sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X}_n)^2}{nS_X^2}}} \sim t_{n-2}.$$

Одавде можемо направити интервал поверења за **вредност** зависне променљиве уколико је предиктор једнак X_0. Понављајући исти поступак као приликом прављења интервала поверења за очекивану вредност зависне променљиве, добијамо интервал

$$\left(\hat{Y}_0 - C\cdot\hat{\sigma}_n\sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X}_n)^2}{nS_x^2}}, \hat{Y}_0 + C\cdot\hat{\sigma}_n\sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X}_n)^2}{nS_x^2}}\right).$$

Приметимо да је овај интервал шири од интервала поверења за средњу вредност, као и да његова дужина не опада ка нули, кад обим узорка тежи бесконачности.

До сада смо видели како можемо да закључимо о томе да је утицај независне променљиве да зависну значајан, као и да оценимо вредност зависне променљиве када нам је позната вредност независне. У даљем тексту ћемо приказати како можемо да установимо колико добро наш модел описује посматрану зависност. Уведимо следеће ознаке:

$$SSE = \sum_{i=1}^{n} e_i^2$$
$$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y}_n)^2$$
$$SSTO = \sum_{i=1}^{n} (Y_i - \bar{Y}_n)^2.$$

Показаћемо да је

$$SSTO = SSR + SSE. \tag{4.10}$$

Како оцене задовољавају једначине (4.1) добијамо да важи

$$\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y}_n)e_i = \sum_{i=1}^{n}(\hat{\beta}_0 + \hat{\beta}_i X_i)e_i = \hat{\beta}_1 \sum_{i=1}^{n} X_i(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0. \tag{4.11}$$

(4.12)

Из

$$SSTO = \sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y}_n)^2,$$

користећи (4.11) добијамо да важи (4.10).

Приметимо да $SSTO$ представља укупан варијабилитет зависне променљиве, док SSR представља варијабилитет објашњен моделом. Зато је природно за једну од мера квалитета модела, узети

$$R^2 = 1 - \frac{SSE}{SSTO},$$

који се назива *коефицијент детерминације*. Овај коефицијент представља удео варијабилитета који је објашњен моделом. Из дефиниције следи да је $R^2 \in [0, 1]$. Међутим, како би се избегла замка преприлагођавања модела, најбоље је све податке поделити на два скупа: један на коме се прави модел (тренинг скуп) и један на коме се тестира квалитет модела коришћењем на пример R^2. Уколико се R^2 не рачуна на тренинг подацима већ онима који се користе за тестирање квалитета моде, $R^2 \in (-\infty, 1]$ и јасно је да бољем моделу одговара већи коефицијент детерминације.

Поред коефицијента детерминације који заправо даје информацију о предиктивној моћи модела, уколико нам је главни циљ моделирања закључивање (а не предикција), треба проверити претпоставке на основу којих вршимо закључивања, као што су претпостављена нормалност грешака, константна дисперзија и

њихова некорелисаност. Иако се многе претпоставке могу и формално тестирати, ми ћемо се задржати на графичкој провери истих.

Резидуали модела представљају оцене грешака модела, зато је природно да се управо проучавањем њихових својстава сазнаје имплицитно о особинама грешака. Прво треба приметити да и у случају некорелисаности грешака, резидуали нису међусобно некорелисане случајне величине (на пример може се лако показати да је $\sum_{i=1}^{n} e_i = 0$). Међутим, из особина грешака следе и неке особине резидуала које ћемо у а наредним редовима извести.

На сличан начин као кад смо пручавали особине $\hat{Y}_0$ добијамо да се $\hat{Y}_i$ може приказати у облику $\hat{Y}_i = \sum_{j=1}^{n} h_{ij} Y_i$, где је

$$h_{ij} = \frac{1}{n} + \frac{(X_i - \bar{X}_n)(X_j - \bar{X}_n)}{n\bar{S}_X^2}.$$

Коефицијент h_{ii} се назива *тежина* i-те тачке.[2]. Сада су математичко очекивање и дисперзија i-тог резидуала редом једнаки

$$E(e_i) = E(\hat{Y}_i - Y_i) = 0$$
$$D(e_i) = D(\hat{Y}_i - Y_i) = \sigma^2(1 - h_{ii}).$$

Видимо да што је h_{ii} веће то ће тачке бити ближе правој. Тешким тачкама се сматрају оне чија је тежина бар $\frac{4}{n}$ и за њих се може посебно испитивати колики им је утицај на сам модел. Уместо ових резидуала често се посматрају такозвани стандардизовани резидуали дефинисани са

$$e_i^s = \frac{e_i}{\hat{\sigma}_n\sqrt{1 - h_{ii}}}.$$

Уколико је за грешке модела испуњен услов хомоскедастичности (услов 3.), $\frac{e_i}{\sqrt{1-h_{ii}}}$ ће имати исту дисперзију. Зато се о испуњености овог услова може сазнати на основу графичког приказа стандардизованих резидуала. На њихов графички приказ у зависности од предиктора, одступање од хомоскедастичности се рефлектује

[2] енг. leverage

тако што се не може уочити нека зависност (тенденција раста, пада) већ резидуали "равномерно" осцилују око 0. До истог закључка можемо доћи када представимо зависност између стандардизованих резидуала и оцењене вредности Y.

Стандардизовани резидуали нам омогућавају и да графички проверимо да ли је услов нормалности грешака задовољен. Наиме, како ће $\frac{e_i}{\sqrt{1-h_{ii}}}$, имати нормалну $\mathcal{N}(0,\sigma^2)$ расподелу, онда ће стандардизовани резидуали имати приближно нормалну расподелу, па можемо приказати зависност емпиријских квантила стандардизованих резидуала од теоријских квантила нормалне расподеле[3]. Уколико се добије права линија, онда можемо сматрати да је та претпоставка модела задовољена. Међутим, ако то није случај, онда значи да не можемо вршити тестирање о значајности параметара модела на приказан начин. Често трансформисање зависне променљиве може решити овај проблем.

Стандардизовани резидуали нам могу помоћи у у детекцији аутлајера. Наиме, један од могућих критеријума да би тачка била аутлајер је да је одговарајући стандардизовани резидуал изван интервала $[-2,2]$, у случају узорака малог обима, или изван интервала $[-4,4]$ у случају узорака већих обима. За оне аутлајере са великом тежином треба посебно испитати да ли је њихов утицај на модел велики, и уколико јесте треба размотрити њихово избацивање. Као што је већ било напоменуто, треба бити веома опрезан са избацивањем аутлајера, јер често они садрже важну информацију о подацима. Једна од могућих мера утицаја је такозвано Куково растојање. За i-ту обсервацију дефинисано је са

$$D_i = \frac{(e_i^s)^2}{2} \cdot \frac{h_{ii}}{1-h_{ii}}.$$

За доњу границу утицајних тачака се најчешће узима $\frac{4}{n-2}$ или квантил $F^{-1}_{F_{2,n-2}}(0.5)$, или 1, у случају узорака великог обима.

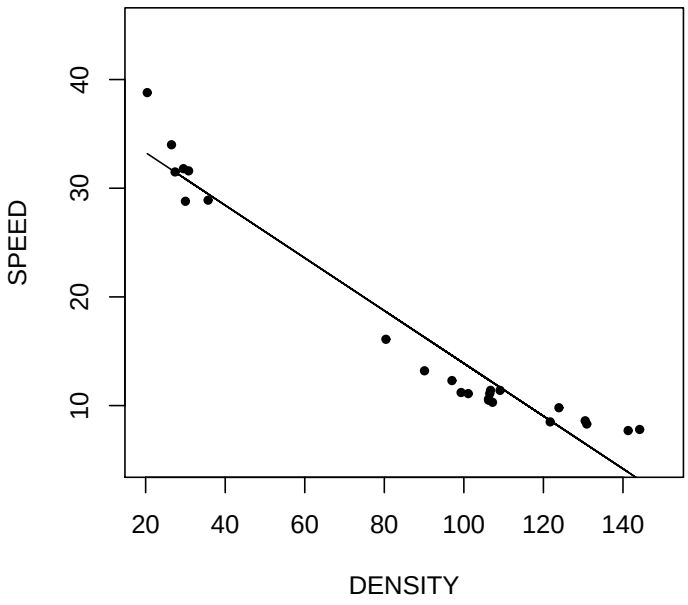
Напомена: У наредним примерима, за графички приказ и анализу модела, две најважније функције које ћемо користити у R-у су:

[3] енг. QQ-plot

- *lm*() - правимо објекат класе линеарни модел;
- *summary*() - уколико је аргумент објекат класе линеарни модел излаз је, између осталог, табела са оцењеним коефицијентима модела, вредностима тест статистика облика (4.9) за тестирање значајности коефицијената, одговарајућу p-вредност, коефицијент детерминације.

Пример 4.1.1. *У циљу истраживања у којој мери број возила на путу утиче на брзину возила сакупљани су подаци о "густини" возила (број аутомобила у једној миљи) и просечној брзини аутомобила (види табелу 7.3). Подаци се налазе у R-пакету* SenSrivastava *у бази* E1.1..

Представићемо податке графички.



Слика 4.1: График зависности брзине возила од густине саобраћаја

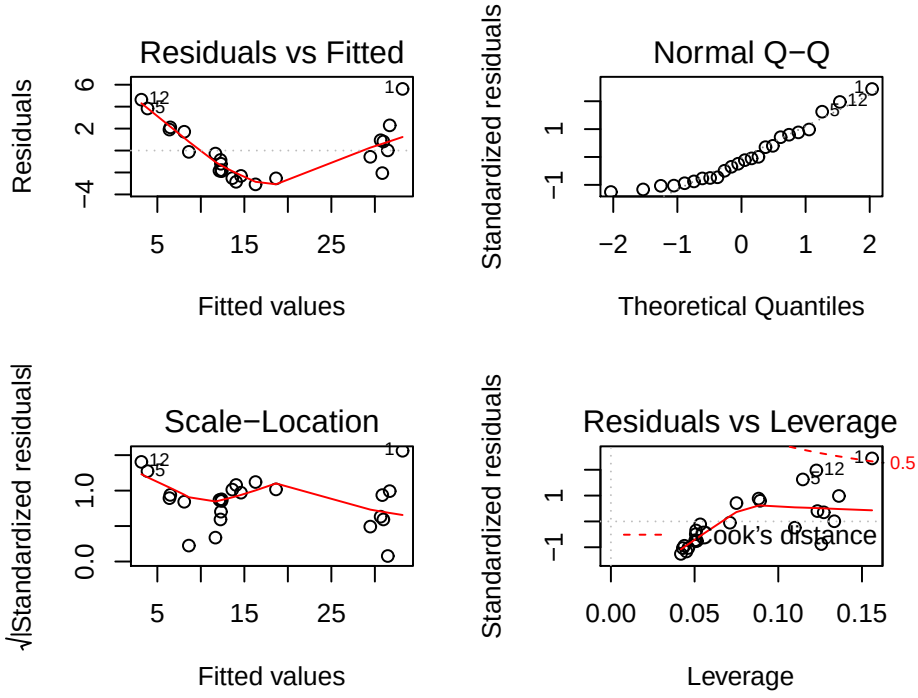
Са графика можемо закључити да са повећањем густине саобраћаја опада брзина истог, што је сасвим очекиван закључак. Најједноставнији модел који би могао да опише податке је линеарна веза, односно $Y = \beta_0 + \beta_1 X + \varepsilon$, *где је* Y *брзина аутомобила а* X *густина собраћаја. Јасно је да у модел морамо да укључимо и неки "шум"* (ε) *који би оправдао то што тачке на графику нису све колинеарне. Оценићемо коефицијенте модела. Из израза* (4.4) *добијамо да је оцењена регресиона права* $\hat{Y} = -0.24X + 38.13$. *За вредности осталих статистика користићемо програмски језик R.*

```
library(SenSrivastava)
attach(E1.1)
plot(E1.1)
model=lm(SPEED~DENSITY,data=E1.1)
# правимо објекат из класе линеарни модел
summary(model)

Call:
lm(formula = SPEED ~ DENSITY, data = E1.1)

   Min.   1st Qu.  Median    Mean   3rd Qu.   Max.
 -3.083  -1.924  -0.425   0.000   1.761   5.617
Coefficients:
                Estimate  Std. Error  t value  Pr(>|t|)
     (Intercept)  38.1295      1.2177    31.31    0.0000
E1.1$DENSITY      -0.2425      0.0126   -19.22    0.0000

Residual standard error: 2.507 on 22 degrees of freedom
Multiple R-squared: 0.9438, Adjusted R-squared: 0.9413
F-statistic: 369.6 on 1 and 22 DF, p-value: 3.041e-15
```

*Из табеле изнад (која је се добија у R-у) видимо вредности тест статистика за тестирање значајности коефицијената (*31.31 *и* -19.22*) као и одговарајуће p-вредности које јасно упућују на то да се коефицијенти значајно разликују од нуле. Пре него што било шта закључимо о утицају густине на брзну саобраћаја, морамо проверити претпоставке модела. Зато ћемо користити*

поменуте "дијагностичке" приказе резидуала и њихових трансформација

```
par(mfrow=c(2,2)) plot(model) # дијагностички графици
```



Слика 4.2: Дијагностички графици

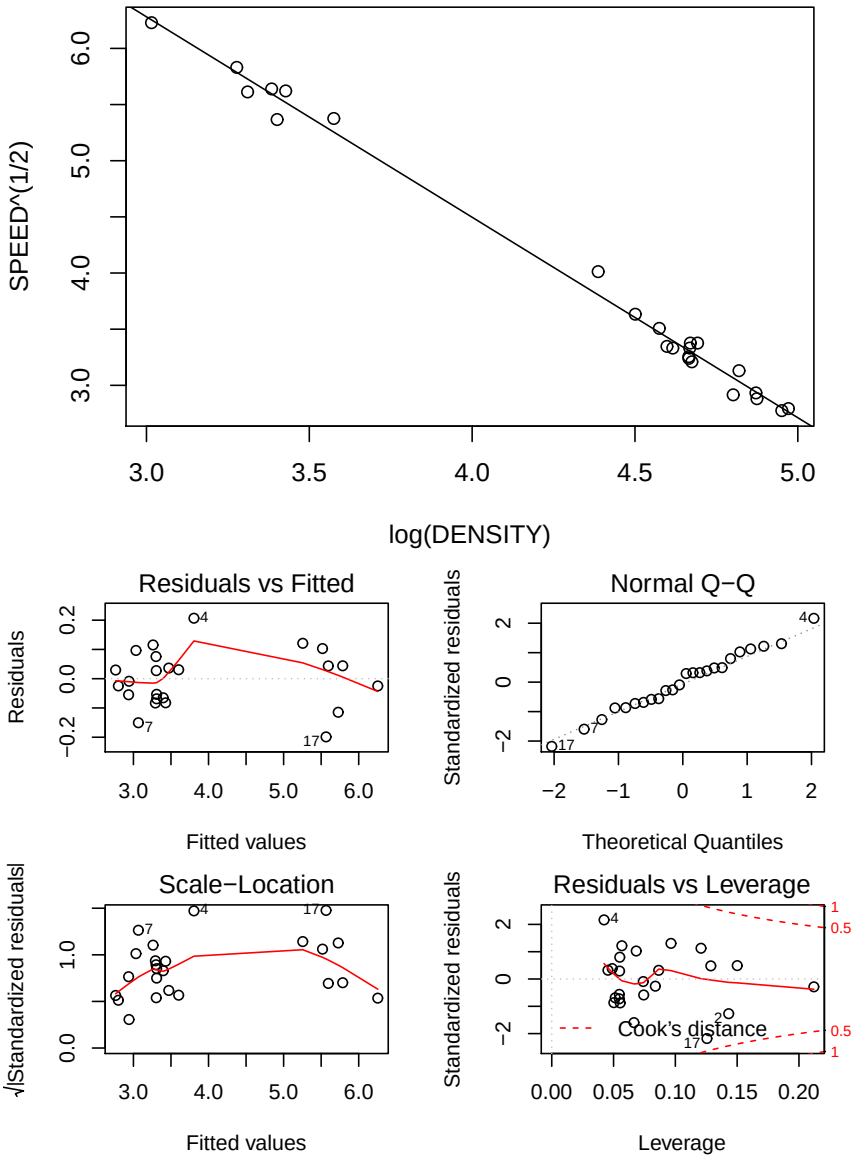
Са графика резидуала од оцењене вредности (график у горњем левом углу) видимо да наша претпоставка о регресионој функцији није најбољи избор јер постоји очита зависност између резидуала и оцењиних вредности модела. На основу тог, и графика испод на ком је представљена зависност стандардизованих резидуала од оцењених вредности, односно $\sqrt{|e^s|}$ *од* $\hat{Y}$*, видимо да претпоставка о хомоскедастичности није испуњена. На основу ових размтрања пробаћемо да неком трансформацијом променљивх*

направимо адекватнији модел. Неке од најчешћих трансформација зависне променљиве су облика $\sqrt{\cdot}$ *и* $\log \cdot$.

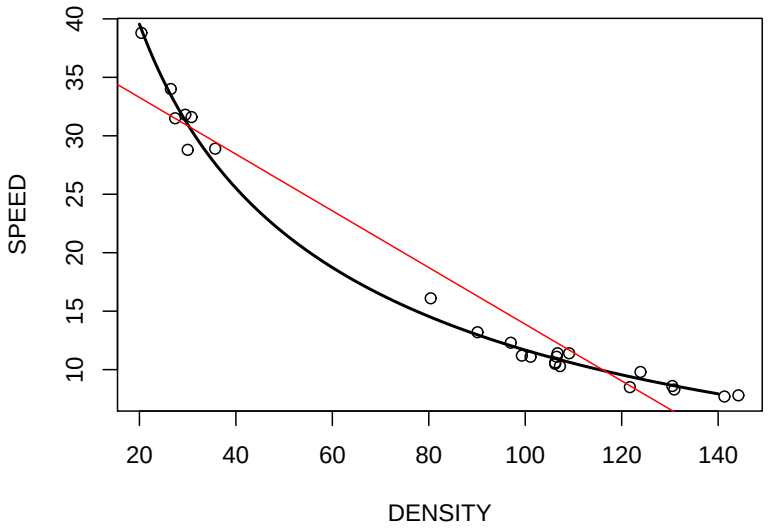
Направићемо модел у коме уместо брзине посматрамо квадратни корен брзине, и уместо густине сабраћаја, посматрамо њен логаритам.

```
modelS=lm(SPEED^(1/2)~log(DENSITY)+DENSITY,data=E1.1)
abline(modelS)
summary(modelS)
Call:
lm(formula = SPEED^{1/2} ~ log(DENSITY), data = E1.1)
     Min.       1st Qu.    Median      Mean     3rd Qu.    Max.
 -0.198849  -0.066165  0.009395  0.052003  0.206610
Coefficients:
                Estimate  Std. Error  t value  Pr(>|t|)
     (Intercept)  11.6361      0.1356    85.79    0.0000
 log(DENSITY)     -1.7849      0.0311   -57.31    0.0000
```

Модел и даље није идеалан, али одступања од претпоставки нису велика као у претходном моделу. То се посебно види на графику 4.1.1.



Слика 4.3: Приказ података након трансформације



Слика 4.4: Поређење модела

Након што смо формирали модел $\sqrt{Y} = 11.64 - 1.78\log(X) + \varepsilon$ *(видимо да су коефицијенти значајни), можемо вршити прогнозу, али треба имати у виду да је* $11.64 - 1.78\log(X)$ *оцењена регресиона функција за* $\sqrt{Y}$*, односно да је* $E(\hat{\sqrt{Y}}) = 11.64 - 1.78\log(X)$*, па применом инверзне трансформације можемо добити добити оцену за очекивану вредност* Y *која неће бити непристрасна јер је* $\hat{E}(\sqrt{Y})^2 \neq E(\hat{\sqrt{Y}}^2)$*. Међутим, може се показати да са порастом узорка та пристрасност нестаје, односно да можемо оценити* $E(Y)$ *са* $\hat{E}(\sqrt{Y})^2$*.*

Уколико желимо да испитамо утицај p предиктора на зависну променљиву, што се најчешће дешава у пракси уместо просте лиенарне регресије посматраћемо модел

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \varepsilon_i, \quad i = 1, 2, ..., n. \tag{4.13}$$

У даљем тексту ћемо укратко представити основе за рад са овим моделима, без много осврта на теоријске резултате. За оне који желе више да знају реферишемо на [24].

Непознати коефицијенти модела (4.13) се оцењују на исти начин као у случају просте регресије, минимизирање суме квадрата одступања, односно:

$$\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p) = \text{argmin}_{\beta_0,\beta_1,....,\beta_p} S(\beta_0, \beta_1, ..., \beta_p),$$

где је

$$S(\boldsymbol{\beta}) = S(\beta_0, \beta_1, ..., \beta_p) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{i1} - \cdots - \beta_p X_{ip})^2. \tag{4.14}$$

Оцене се могу добити решавањем система $\frac{\partial S}{\partial \beta_j} = 0$, за $j = 0, 1, ..., p$. Уз претпоставку да је $\{\varepsilon_i\}$ низ некорелисаних случајних величина са нормалном расподелом са истом дисперзијом, те оцене су и оцене добијене методом максималне вероватноће. Оцена за σ^2 је

$$\widetilde{\sigma}^2 = \frac{S(\hat{\boldsymbol{\beta}})}{n},$$

која није непристрасна. Може се показати да је

$$\widehat{\sigma}^2 = \frac{S(\hat{\boldsymbol{\beta}})}{n - p - 1},$$

непристрасна оцена за σ^2. Поред тога, може се и показати да $(\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p)$ има вишедимензиону нормалну расподелу, односно да $\hat{\beta}_j$ има нормалну расподелу са очекивањем β_j и дисперзијом која се може оценити. На основу тога се могу конструисати интервали за параметере и тестирати хипотезе о значајности коефицијената (коришћењем Валдових статистика као у случају просте регресије). Оцењена вредност за Y_0 и EY_0 за вредност предиктора $(X_{01}, X_{02}, ..., X_{0p})$ је

$$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 X_{01} + \cdots + \hat{\beta}_p X_{0p}.$$

Имајући у виду да се ради о линеарној комбинацији нормално расподељених случајних величина, и $\hat{Y}_0$ има нормалну расподелу

па се на исти начин као у случају просте регресије, могу добити интервали поверења и предвиђања.

За тестирање "угњеждених хипотеза у вези са параметрима" може се користити такозвани F-тест. Наиме, претпоставимо да имамо два угњеждена модела (један се може добити од другог увођењем ограничења на параметре да су једнаки нули) и нека један има p_1+1 непознатих параметара, а други p_2+1 непознатих параметара, при чему је $p_1 < p_2$. Нека су $\hat{\boldsymbol{\beta}}_1$ и $\hat{\boldsymbol{\beta}}_2$ оцене непознатих параметара у та два модела. Желимо да тестирамо H_0 да су ограничења задовољена, то јест да "већи" модел нам не доприноси значајно објашњењу варијабилитета зависне променљиве. Тада, уколико је нулта хипотеза тачна

$$T_n = \frac{\frac{S(\hat{\beta}_1)-S(\hat{\beta}_2)}{p_2-p_1}}{\frac{S(\hat{\beta}_2)}{n-p_2-1}} \tag{4.15}$$

има Фишерову $\mathcal{F}_{p_2-p_1,n-p_2-1}$ расподелу. У R-у је овај тест имплементиран у функцији $anova(model1, model2)$. Пре него што модел вишеструке лиенарне регресије илуструјемо примером, показаћемо на који начин можемо посматрати категоричке предикторе. Категоричке номиналне променљиве се могу представити коришћењем помоћних индикатора[4]. Уколико имамо овележје Z од k категорија можемо га кодирати помоћу $k-1$ помоћна индикатора $(X_1, ..., X_{k-1})$, на пример на следећи начин:

Z	$(X_1, ..., X_{k-1})$
1	$(1, 0, ..., 0)$
2	$(0, 1, ..., 0)$
$\vdots$	$\vdots$
$k-1$	$(0, 0, ..., k-1)$
k	$(0, 0, ..., 0)$

Овим постижемо интерпретабилност коефицијената уз коришћене индикаторе. Коефицијент уз j-ти индикатор представља разлику вредности зависне променљиве између j-те и k-те категорије па се зато k-та категорија често назива и референтном. Наравно,

[4] енг. dummy

редослед категорија се може изменити уколико постоји природна потреба за другим избором референтне категорије.

У наредном примеру ћемо приказати како изгледа пут од података до финалног модела.

Пример 4.1.2. *Посматрамо укупан број поена стечених на предмету Статистика у зависности од броја поена стечених на предмету Вероватноћа и пола студента. Крајњи циљ истраживања је да се донесе одлука о томе да ли предмет Вероватноћа треба да буде услован за предмет Статистика, као и да се установи да ли постоји значајна разлика између полова у погледу разумевања градива из Статистике. Подаци се налазе у бази 7.4, и након учитавања у* R*-у смештени су у бажу* PodaciIspit.

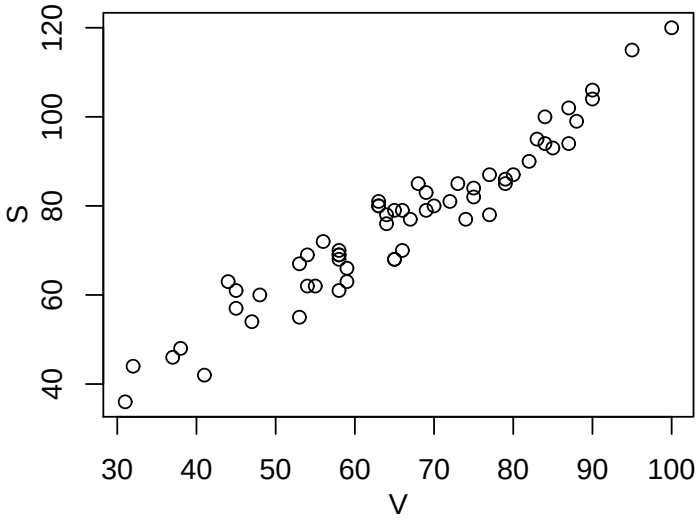
Прво ћемо графички приказати зависност броја поена на предмету Статистика од броја поена на предмету Вероватноћа, као и кутијасте дијаграме поена на предмету Статистика у зависности од пола.

```
PodaciIspit$pol=factor(PodaciIspit$pol)
boxplot(PodaciIspit$ukupanS~PodaciIspit$pol,xlab='pol',ylab='S')
plot(PodaciIspit$ukupanS~PodaciIspit$ukupanV,
xlab='V',ylab='S',mgp=c(2,1,0))
```

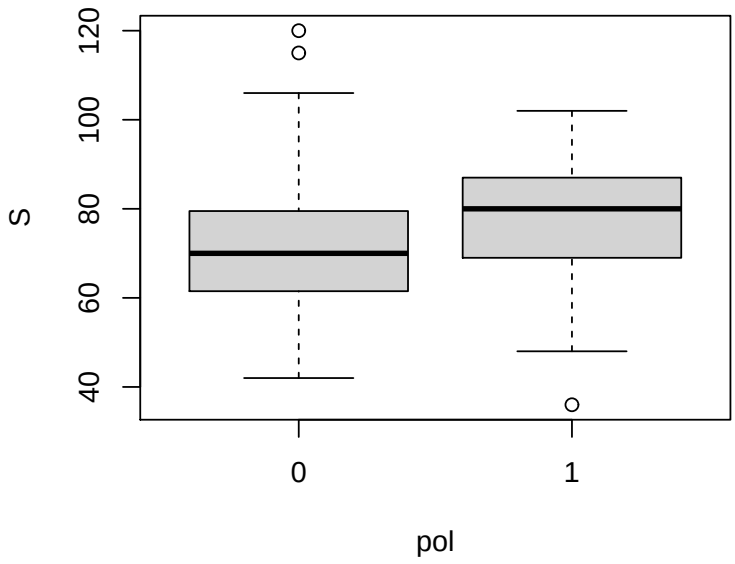
Са графика 4.5 видимо да је веза између остварених поена прилично линеарна, а са графика 4.6 да постоје разлике између расподела поена на предмету Статистика, међу половима. Наиме, две мушке особе се могу идентификовати аутлајером, међутим ако мало боље погледамо резултате, та особе су на оба предмета имала највећи (или приближно највећи) број поена па резултати нису необични и нећемо избацити те особе из даљег разматрања.

Означимо са Y-број поена из Статистике, са X_1 пол студента, и са X_2-број поена из Вероватноће. Посматраћемо модел

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i.$$



Слика 4.5: Зависност броја поена на предмету Статистика од броја поена на предмету Вероватноћа



Слика 4.6: Зависност броја поена на предмету Статистика од пола

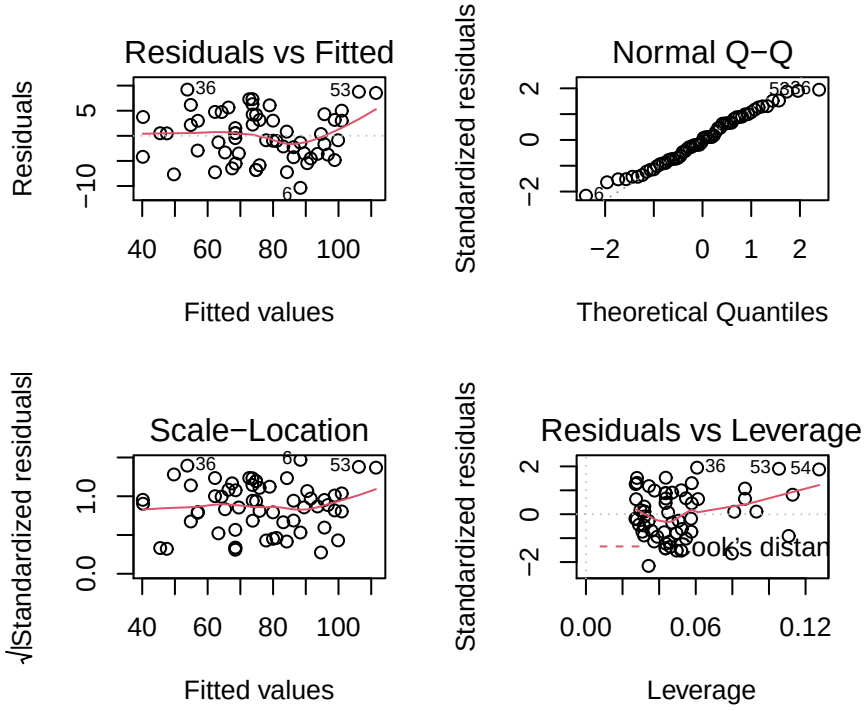
```
model=lm(ukupanS~pol+ukupanV)
summary(model)
Call:
lm(formula = ukupanS ~ pol + ukupanV,data=PodaciIspit)

Residuals:
    Min      1Q  Median      3Q     Max
-10.3577 -3.6294 -0.0507  3.8579  9.2013
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.7405     2.7506    2.45   0.0174
pol1          0.9794     1.2982    0.75   0.4537
ukupanV       1.0472     0.0397   26.39   < 2.2e-16
Residual standard error:4.878 on 57 degrees of freedom
Multiple R-squared: 0.925, Adjusted R-squared: 0.9223
F-statistic: 351.3 on 2 and 57 DF, p-value: < 2.2e-16
```

Пре него што продискутујемо добијене резултате наиратћемо дијагностичке графике на основу којих закључујемо о испуњености модела.

```
par(mfrow=c(2,2))
plot(model)
```

Са графика 4.1.2 видимо да је претпоставка о нормалности задовољена, као и да ништа јасно не упућује на хетероскедастичност грешака модела. Кукова растојања су у границама дозвољеног (и поред изузетно "тешког податка" који смо већ на самом почетку идентификовали потенцијалним аутлајером).

Сада, на основу података о оцењеном моделу, на основу резултата t-теста за значајност коефицијената закључујемо да би требало да су коефицијенти уз предикторе значајни а да слободан члан није значајно различит од 0. Како се ту ради о тестирању две хипотезе истовремено $\beta_1 = \beta_2 = 0$ те резултате ћемо и проверити описаним F тестом. У излазу програма је дат резултат тестирања хипотезе $\beta_1 = \beta_2 = 0$ (реализована вредност статистике (4.15) је 351.3 а p-вредност много мала) закључујемо да се ова хипотеза одбацује. Применом F-теста испитаћемо још



Слика 4.7: Дијагностички графици

да ли коефицијент уз променљиву X_1 значајан. Значајност коефицијента β_2 нема потребе додатно проверавати јер p-вредност t-теста ($< 2e-16$) упућује на изузетну жачајност, а исто се може закључити и са графичког приказа.

Како бисмо видели да ли променљива X_1 додатно објашњава варијабилитет Y направићемо модел без ње и применити $F-test$ (4.15).

```
model1=lm(ukupanS~ukupanV,data=podaciIspit)
anova(model1,model)

Analysis of Variance Table
Model 1: ukupanS ~ ukupanV
Model 2: ukupanS ~ pol + ukupanV
  Res.Df    RSS  Df Sum of Sq    F  Pr(>F)
1     58 1391.41
2     57 1356.51   1    34.89 1.47  0.2309
```

На основу резултата прихватамо нулту хипотезу $\beta_1 = 0$ и закључујемо да пол студената не утиче значајно на остварен број поена.

```
model
Call: lm(formula = ukupanS ~ ukupanV)
Coefficients:
(Intercept) ukupanV
9.732 1.031
```

Финални модел који смо добили је $Y = 9.73 + 1.03X_2$, односно да очекиван број поена на предмету Статистика ће бити корелисан са број поена на курсту Вероватноћа, и тај закључак свакако треба узети у убзир приликом доношења одлуке о условности предмета.

У приказаном примеру се нисмо бавили предиктивном моћи модела, из разлога што је за очекивати да у некој наредној генерацији, и са другим предметним наставницима буде друга расподела поена што би вероватно утицало на сам модел.

Задатак 4.1.1. У бази Kinder (пакет PASWR) налазе се подаци о тежини деце и њиховој висини. Направите одговарајући линеарни модел на основу кога можете оценити тежину на основу висине детета. Обратити пажњу на потенцијалне аутлајере.

Задатак 4.1.2. У бази tree (пакет datasets) се налазе подаци о пречнику, висини и запремини оборених стабала трешње. У бази су ови подаци редом изначени са Girth, Hight и Volume. На основу доступних података направите модел који би био погодан за прогнозирање очекиване вредности запремине стабла на основу висине и пречника.

4.2 Логистичка регресија

У линеарном регресионом моделу смо моделирали зависност $E(Y)$ од предиктора, односно $E(Y|X)$.
Поставља се питање да ли то можемо да урадимо у случају да Y нема нормалну расподелу. Одговор је потврдан, али уз мале модификације. У овом поглављу ћемо разматрати случај када је Y индикатор-неко обележје које узима само две вредности 0 и 1. Разумно је претпоставити да $p_i = P\{Y_i = 1\}$ може зависити од предиктора. Само неки од примера су:

- да ли ће особа добити рак на основу генетеског материјала;
- да ли ће се купцима свидети нови производ на основу података о досадашњој куповини;
- до каквог типа саобраћајне несреће (фаталне или не) ће доћи, у зависности типа возила, типа пута, сигнализације, временских услова;
- да ли ће особа која је узела кредит успети успешно да га врати.

Када бисмо претпоставили да је $p_i = \beta_0 + \beta_1 X_i$ дошли бисмо у опасност да p_i узме вредност изван свог дозвољеног опсега $(0, 1)$. Једна могућност је да трансформишемо p_i тако да трансформисана вредност је у $\mathbf{R}$ а затим извршимо моделирање. Једна од могућих трансформација је $F^{-1}(p_i)$, где је F нека функција расподеле

случајне промењиве дефинисане на $\mathbb{R}$. Следећи пример нам може послужити као мотивациони за коришћење ове трансформације.

Пример 4.2.1. *Претпоставимо да је Y нека зависна промењива чија се средња вредност може моделирати линеарним моделом са нормално расподељеним грешкама, односно да посматрамо модел*

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i.$$

Уместо узорка $Y_1, Y_2, ..., Y_n$ на располагању имамо само информацију да ли је вредност Y_i "прешла неки критични ниво", односно имамо узорак $Y_1^c, ..., Y_n^c$, где је

$$Y_i^c = \begin{cases} 1, & Y_i > c; \\ 0, & Y_i \leq c. \end{cases}$$

Желимо да направимо модел којим ћемо оценити вероватноћу да је Y_i веће од неког нивоа c. Тада је

$$p_i = P\{Y_i^c = 1\} = P\{\varepsilon_i > c - \beta_0 - \beta_1 X_i\} = \Phi\Big(\frac{c - \beta_0 - \beta_1 X_i}{\sigma}\Big)$$
$$= \Phi\Big(\frac{-c + \beta_0 + \beta_1 X_i}{\sigma}\Big).$$

Одавде је

$$\Phi^{-1}(p_1) = \frac{-c + \beta_0 + \beta_1 X_i}{\sigma} = A + BX_i.$$

Дакле трансформација коју смо применили је Φ^{-1}.

Тип регресије приказан у овом примеру (када примењујемо трансформацију $\Phi^{-1}(\cdot)$) се назива *пробит регресија*. У случају да се ради о логистичкој расподели, $F(x) = \frac{1}{1+e^{-x}}$, за $x \in \mathbb{R}$, модел

$$\log \frac{p_i}{1 - p_i} = \beta_0 + \beta_1 X_i$$

се назива *логистички регресиони модел*. Функција $\lambda(X) = \log \frac{p(X)}{1-p(X)}$ логит трансформација. Количник $\frac{p(X)}{1-p(X)}$ се назива *квота*. Управо због интерпретације утицаја предиктора на квоту

логистички регресиони модел се најчешће и користи у случају бинарне зависне променљиве. Наиме, уколико се предиктор повећа за 1, квота се промени за e^{β_1}. Зато оцењујући коефицијент β_1 можемо да видимо какав је утицај промене предиктора на саму промену квоте.

Параметре β_0 и β_1 оцењујемо методом максималне веродостојности. Функција веродостојности дата је са

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} p_i^{Y_i}(1-p_i)^{1-Y_i}.$$

Одавде је

$$\begin{aligned} l(\beta_0, \beta_1) &= \sum_{i=1}^{n} \Big(Y_i \log \frac{p_i}{1-p_i} + \log(1-p_i) \Big) \\ &= \sum_{i=1}^{n} \Big(Y_i(\beta_0 + \beta_1 X_i) + \log \frac{1}{1+e^{\beta_0+\beta_1 X_i}} \Big). \end{aligned}$$

Решавање система $\frac{\partial l(\beta_0,\beta_1)}{d\beta_0} = \frac{\partial l(\beta_0,\beta_1)}{d\beta_1} = 0$, се врши нумерички и нећемо се на томе задржавати. Након што оценимо $\hat{\beta}_0$ и $\hat{\beta}_1$, оцењена логит функција је

$$\hat{\lambda}(X) = \hat{\beta}_0 + \hat{\beta}_1 X,$$

чијом трансформацијом добијамо оцену вероватноће да је $Y = 1$ када је предиктор X је

$$\hat{p}(X) = \frac{1}{1+e^{-(\hat{\beta}_0+\hat{\beta}_1 X)}},$$

односно

$$\hat{p}_i = \frac{1}{1+e^{-(\hat{\beta}_0+\hat{\beta}_1 X_i)}}.$$

Потпуно аналогно поступамо у случају да имамо више од једног предиктора (што се најчешће дешава у пракси).

Након што оценимо модел и видимо да ли је одговарајући, можемо испитати значајност коефицијената коришћењем статистике[5]

$$Z = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)} \tag{4.1}$$

која ако је нулта хипотеза $H_0 : \beta_i = 0$ тачна, има асимптотски нормалну расподелу. Може се показати и да

$$\frac{\hat{\beta}_i - \beta_i}{SE(\hat{\beta}_i)}$$

има асимптотски нормалну расподелу. На основу тога се могу конструисати интервали поверења за коефицијенте модела, а и за логит функцију (јер њена оцена као линеарна комбинација асимптотски нормалних случајних величина има исто нормалну расподелу). Иневерзном трансформацијом интервала поверења за логит функцију можемо добити и интервале поверења за квоту, као и за вероватноћу $p(X)$.

Као што смо већ навели, важан корак у моделирању је да се види колико је модел који смо добили добар. Једна од мера квалитета модела је такозвана *девијација* којом се мери разлика између претпостављеног модела и сатурираног модела (модела код кога је број непознатих параметара једнак броју обзервација, односно када $Y_i \sim \mathcal{B}(1, \theta_i)$). Дефинише се са

$$D = 2(l(y, \hat{\theta}^s) - l(y, \hat{\beta}))$$

где је $\hat{\theta}_s$ оцена у сатурираном моделу. У случају логистичке регресије добија се да је $\hat{\theta}_i^s = Y_i$ и да је $l(y, \hat{\theta}^s) = 0$. Када одредимо девијацију треба да имамо неку вредност са којом ћемо да поредимо. За то је најприродније одредити девијацију модела када нема предиктора (већ само слободан члан). Означимо ту девијацију са D_0. Познато је да $D_0 - D$ има χ^2 расподелу са бројем степени слободе који је једнак разлици броја оцењених параметара у оба модела (практично броју коефицијената уз предикторе). Поред овог, важи и општије тврђење, да ако су D_1 и D_2 девијације два

[5] Валдова статистика

угњездена модела (D_1 се добија од D_2 стављањем услова на коефицијенте модела), онда $D_1 - D_2$ има χ^2 расподелу са бројем степени слободе који је једнак разлици броја оцењених параметара у оба модела. На основу тога можемо закључити да ли увођење додатних предиктора сигнификантно утиче на побољшање модела.

Поред овога може се дефинисати и уопштени коефицијент детерминације

$$R^2 = 1 - \frac{D}{D_0}.$$

Из дефиниције видимо да велике вредности R^2 упућују на добар модел, док уколико је блиска нули, предиктори не доприносе бољем квалитету модела. Међутим, треба имати у виду да су вредности овог коефицијента мање од вредности коефицијента детерминације у линеарном моделу и да не треба очекивати вредности блиске јединици ни у случају доброг модела.

Уколико установимо да се увођењем нових предиктора модел значајно побољшава, или то није случај, адекватност финалног модела се може проверити на неколико начина. Уколико имамо само категоричке предикторе онда можемо конструисати статистику која је заснована на разлици очекиваног броја елемената из узорка за сваку комбинацију предиктора и реализованог броја. Како бисмо једноставније приказали идеју за конструкцију тест статистике предпоставићемо да имамо само један предиктор.

За свако X_j из узорка формирамо подгрупу који чине они елементи узорка чија је независна компонента једнака одабраном X_j. Нека је m_j број елемената у j-тој подгрупи посматраног узорка, $j = 1, 2...J$. У оквиру сваке подгрупе се може оценити условна вероватноћа $P\{Y = 1|X_j\}$. Нека је n_j број елемената у подгрупи за које је вредност зависне променљиве једнака 1. Оцена поменуте вероватноће, на основу логистичког модела је $\hat{p}_j = \hat{P}\{Y = 1|X_j\} = \frac{1}{1+e^{-\hat{\beta}_0 - \hat{\beta}_1 X_j}}$. Тада је очекиван број елемената из узорка чија је вредност зависне променљиве 1, једнака:

$$\hat{n}_j = m_j \hat{p}_j = \frac{m_j}{1 + e^{-\hat{\beta}_0 - \hat{\beta}_1 X_j}}.$$

Нека је

$$r_j = \frac{n_j - m_j\hat{p}_j}{\sqrt{m_j\hat{p}_j(1-\hat{p}_j)}} = \frac{n_j - \hat{n}_j}{\sqrt{\hat{n}_j(1-\frac{\hat{n}_j}{m_j})}}, \tag{4.2}$$

такозвани *Пирсонов j-ти резидуал*. Сада је природна статистика коју можемо користи за тестирање адекватности нашег модела Пирсонова статистика је дефинисана са

$$C = \sum_{j=1}^{J} r_j^2,$$

која има асимптотски χ^2_{J-2} (уколико имамо p предиктора расподела је χ^2_{J-p-1}). Међутим, важно је напоменути да се ова статистика може користити само кад је број елемената у оквиру сваке од категорија велики, односно $\hat{n}_j$ треба да буде бар 5. Приметимо да се r_j може и интерптетирати као разлика између зависне променљиве и оцењене очекиване вредности зависне променљиве и када она има биномну $\mathcal{B}(n_j, p_j)$. Зато се често подаци приказују управо у груписаном облику.

Када је предиктор непрекидна променљива онда се груписање не може извршити на овај начин. Хосмер и Лемеш су предложили следећи поступак: за свако X_j одреди се $\hat{p}_j$, а затим сортирају елементи узорка на основу оцењених вероватноћа, а затим на основу тога изврши подела у групе тако да групе имају приближно једнаке вероватноће. Даље, нека је у тако формираној k-тој групи n'_k број елемената из узорка који јој припадају. Даље, означимо са J_k број полазних група које су уједињене у k-ту подгрупу. Тада се условна вероватноћа у свакој групи оцењује на следећи начин:

$$\bar{P}_k = \sum_{j=1}^{J_k} \frac{m_j}{n'_k}\hat{P}(Y=1|X_j),$$

за свако X_j из k-те групе. Број елеманата у k-тој групи и његова оцена дати су са:

$$O_k = \sum_{j=1}^{J_k} n_j, \quad \hat{O}_k = n'_k\bar{P}_k.$$

Нека је g укупан број новоформираних група. Тада

$$C' = \sum_{k=1}^{g} \frac{(O_k - \hat{O}_k)^2}{n'_k \bar{P}_k (1 - \bar{P}_k)}.$$

C' има χ^2_{g-2} расподелу. Више о овим статистикама, као и неким другим за проверу адекватности модела, се може наћи у [13] и [5]. Приликом њихове примене треба имати у виду да се може упасти у замку преприлагођања модела па ако је циљ да се модел може уопштити и на другим подацима (а не само да се уочи веза између променљивих и идентификују утицају) мере квалитета треба рачунати на подацима на којима модел није конструисан. Међутим, тада поменуте расподеле тест статистика не важе.

На основу оцењеног модела можемо вршити и класификацију, и она, такође, може послужити за одређивање квалитета модела, посебно ако нам је крајњи циљ моделирања управо класификација. Ово је данас и најчешћи начин евалуације модела.

Једна могућност је да ако је $\hat{p}_i > 0.5$ онда је $\hat{Y}_i = 1$, у супротном је 0. Након тога можемо видети проценат добро класификованих података. Ово је природан приступ и као такав најпрепоручљивији. Међутим, понекад, уколико не добијемо задовољавајућу тачност, а битно нам је, на пример због интерпретабилности, да задржимо модел, можемо да померимо праг за класификацију. Дакле,

$$\hat{Y}_i = \begin{cases} 0, & \hat{p}_i < C \\ 1, & \hat{p}_i \geq C \end{cases}$$

а праг за класификацију C бирамо имајући у виду следеће мере тачности које ћемо у даљем тексту описати.

Резултат предвиђања се може приказати такозваном *матрицом конфузије*.

$Y \setminus \hat{Y}$	0	1
0	a	b
1	c	d

Са a смо означили број оних елемената чија је вредност 0 и који су класификовани као 0, b је број 0 које су погрешно оцењени са

1, и тако даље. Тачност класификације[6] је сада $A = \frac{a+d}{a+b+c+d}$. Ова мера није адекватна ако класе нису приближних величина. Уколико је једна класа знатно већа од друге, класификовањем свих елемената узорка тако да припадају тој већој класи постижемо велику тачност, иако је јасно да класификатор није добар. Због тога се дефинишу још неке мере:

- сензитивност (одзив)[7] $TPR = \frac{d}{c+d}$
- специфичност[8] $TNR = \frac{a}{a+b}$
- прецизност $PPV = \frac{d}{b+d}$[9]
- $FPR = 1 - TNR$
- скор $F_1 = 2\frac{PPV \cdot TPR}{PPV+TPR}$.

У зависности од афинитета C се може бирати тако да поменуте мере имају екстремну вредност. Наравно, могу се и комбиновати. На пример, прецизност нам заправо даје податак о броју релевантних података јер обично је догађај чију успешну реализацију означавамо јединицом, тај који посматрамо. С друге стране, одзив нам говори о проценту јединица који су успешно класификовани (који проценат). На пример, ако $Y = 1$ означава да је особа проглашена зараженом неким вирусом, прецизност нам говори колики проценат дијагностикованих можемо сматрати стварно зараженим, док одзив пак, колико се добро на основу улазних параметара, добро дијагностикује особа заражена вирусом. Скор представља хармонијску средину ове две мере и добро га је користити када су оне равноправне. Уколико то нису може се дефинисати и тежинска хармонијска средина између ове две мере при чему ћемо оној важнијој задатки већу тежину.

За одређивање прага се може одредити и такозвана ROC крива која представља зависност између TPR и FPR, односно на y-оси је сензитивност, а на x-оси специфичности. Како нам "тачка" горњи

[6] енг. accuracy
[7] енг. true positive rates (recall)
[8] енг. true negative rates
[9] енг. positive predictive value

леви угао графика (што се у пракси никада не постиже), говори о идеалном класификатору, једна могув1ност је да се одабере тачка са криве која је најближа тачки $(0, 1)$.

Површина испод ROC криве (AUC) се назива *индекс прецизности*. Што је већа површина боља је предиктивна моћ модела и класификатор боље раздваја категорије. Заправо, AUC је оцена вероватноће да при случајном избору два елемента из различитих класа она из класе означене са 0 има мању вредност (на основу које је подељена у класе) од оне из класе означене 1 (подсетимо се дефиниције Вилкоксонове статистике за два независна узорка).

Пример 4.2.2. *Проучавано је како висина утиче на то да ће особе склоне кошарци погодити кош са одређене раздаљине. Узет је узорак од 100 особа и резултати су приказани у табели 7.5.*

Направићемо одговарајућии логистички модел. Како бисмо што непристрасније испитали квалитет модела прво ћемо податке поделити на тренинг скуп (70 особа) а затим тестирати квалитет модела на преосталим подацима (30 особа). Уколико се испостави да је модел задовољавајућег квалитета на комплетном скупу података ћемо направити финални модел.

```
podaciKosarka$pogodak=factor(podaciKosarka$pogodak)
#biramo trening set
indeks.trening=1:70
indeks.test=71:100
podaciKosarkaT=podaciKosarka[indeks.trening,]
modelT=glm(pogodak~visina,family =binomial, data=
podaciKosarkaT)
summary(modelT)
Call:
glm(formula = pogodak ~ visina, family = binomial, data =
podaciKosarkaT)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
 -2.60211  0.06293  0.17823  0.47887  1.89366

Coefficients:
```

	Estimate	*Std. Error*	*z value*	*Pr(>|z|)*
(Intercept)	*-48.6187*	*13.0284*	*-3.73*	*0.0002*
visina	*26.1154*	*6.8822*	*3.79*	*0.0001*

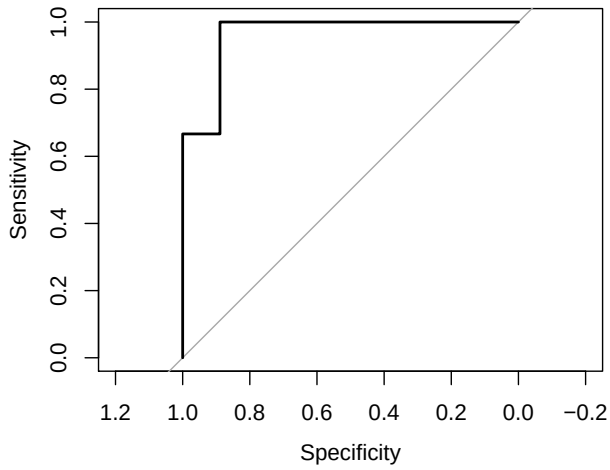
Null deviance: 72.741 on 69 degrees of freedom
Residual deviance: 37.750 on 68 degrees of freedom
AIC: 41.75

На основу модела на тренинг скупу на основу ког закључујемо да је "висина" значајан предиктор (p-вредност на основу (4.1) *је 0.0001), можемо предвидети вероватноће на тест скупу.*

```
podaciKosarkaTest=podaciKosarka[indeks.test,]
podaciKosarkaTest$ver=predict(modelT,
newdata = podaciKosarkaTest,type='response')
```

На основу оцењене вероватноће поготка на тест скупу можемо извршити класификацију на већ описан начин, нацртати ROC криву одредити индекс прецизности. Уколико бисмо желели да размотримо други одабир прага онда би за то најбоље било да га одредимо на тренинг скупу или да у оквиру тренинг скупа одаберемо један подскуп који бисмо користили за то (скуп за валидацију).

```
KosarkasiROC=roc(pogodak~ver,data=podaciKosarkaTest)
auc(KosarkasiROC)
[1] Area under the curve: 0.963
plot(KosarkasiROC)
```



Слика 4.8: *ROC* крива на тест скупу података

Одредићемо и матрицу конфузије.

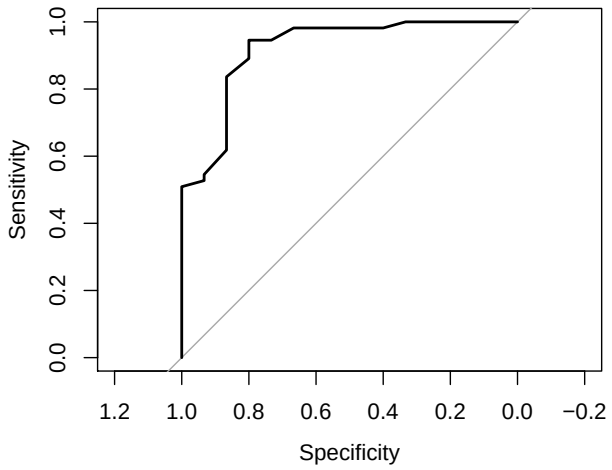
```
podaciKosarkaTest$klasa=ifelse(podaciKosarkaTest$ver>0.5,1,0)
table(podaciKosarkaTest$pogodak,podaciKosarkaTest$klasa)
     0    1
 0   7    2
 1   0   21
```

На тест подацима видимо да за одабран праг класификације (0.5) добијена сензитивност је $21/21 = 1$*, док је специфичност* $7/9 = 0.78$ *што говори у прилог квалитету модела.*

Добијене резултате можемо упоредити са резултатима на тренинг скупу. Модел, уолико је добар, би требало да показује приближан квалитет на оба скупа.

```
podaciKosarkaT$ver=predict(modelT,newdata = podaciKosarkaT,type='response')
KosarkasiROCT=roc(pogodak~ver,data=podaciKosarkaT)
auc(KosarkasiROCT)
Area under the curve: 0.9194
```

```
plot(KosarkasiROCT)
```



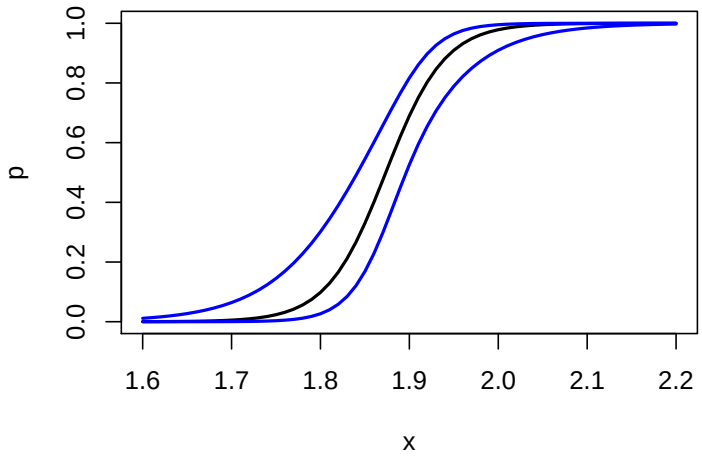
Слика 4.9: *ROC* крива на тренинг скупу података

Након што смо утврдили да не постоје огромне ралике на тренингу и тест скупу направићемо модел на свим подацима и на основу тога приказати оцењену зависност вероватноће (укључујући и интервал поверења) поготка од висине особе. Пре тога, илустрације ради, применићемо Хосмер-Лемешов тест за проверу коректности модела.

```
model=glm(pogodak~visina, family = binomial,data=
podaciKosarka)
hoslem.test(as.numeric(podaciKosarka$pogodak)-1, fitted(model),
g = 10)
    Hosmer and Lemeshow goodness of fit (GOF) test
data: as.numeric(podaciKosarka$pogodak) - 1, fitted(model)
X-squared = 6.5954, df = 8, p-value = 0.5808
```

Резултат теста указује на адекватност модела, што смо и очекивали с обзиром на резултате осталих мера перформанси.

```
xvisina=seq(from=1.6,to=2.2,by=0.01)
probOcena=predict(model,newdata=data.frame(visina=xvisina),
type='response')
# оцењујемо логит функцију
probOcenaL=predict(model,newdata=data.frame(visina=xvisina),
type='link',se.fit=TRUE)
plot(xvisina,probOcena,type='l',lwd=2,xlab='x',ylab='p')

# одређујемо горњу и доњу границу интервала поверења
за логит функцију а затим трансформишемо у границе за
вероватноћу успеха
intG=exp(probOcenaL$fit+1.96*probOcenaL$se.fit)/
(1+exp(probOcenaL$fit+1.96*probOcenaL$se.fit))
intD=exp(probOcenaL$fit-1.96*probOcenaL$se.fit)/
(1+exp(probOcenaL$fit-1.96*probOcenaL$se.fit))
lines(xvisina,intG,col=4,lwd=2)
lines(xvisina,intD,col=4,lwd=2)
```



Слика 4.10: Зависност вероватноће поготка од висине особе

Приликом одређивања интервала користили смо да оцењена логит функција $\hat{\lambda}(X)$ *има асимптотски нормалну расподелу.*

Као и у линеарним логистичким моделима, идентификација аутлајера може бити од великог значаја, и природно је да резидуали модела имају значајну улогу у томе. Класични резидуали $e_i = Y_i - \hat{p}_i$ нису много релевантни у анализи због дискретне структуре зависне променљиве. Много чешће се користе Пирсонови резидуали дефинисани са (4.2). Поред њих користе се још и резидуали девијације које нећемо наводити. Када су подаци груписани, као што је био случај приликом конструкције Пирсонове статистике за проверу адекватности модела, и кад је број елемената из узорка у оквиру сваке групе велики, они имају асимптоски нормалну расподелу, док у супротном не важи. Исто важи и за резидуале девијације. Међутим када се Пирсонови резидуали (и резидуали девијације) стандардизују на одговарајући начин тако да сви имају јединичну дисперзију, њихов графички приказ може помоћи у дијагностици аутлајера. Стандардизован Пирсонов

резидуал дефинисан је са

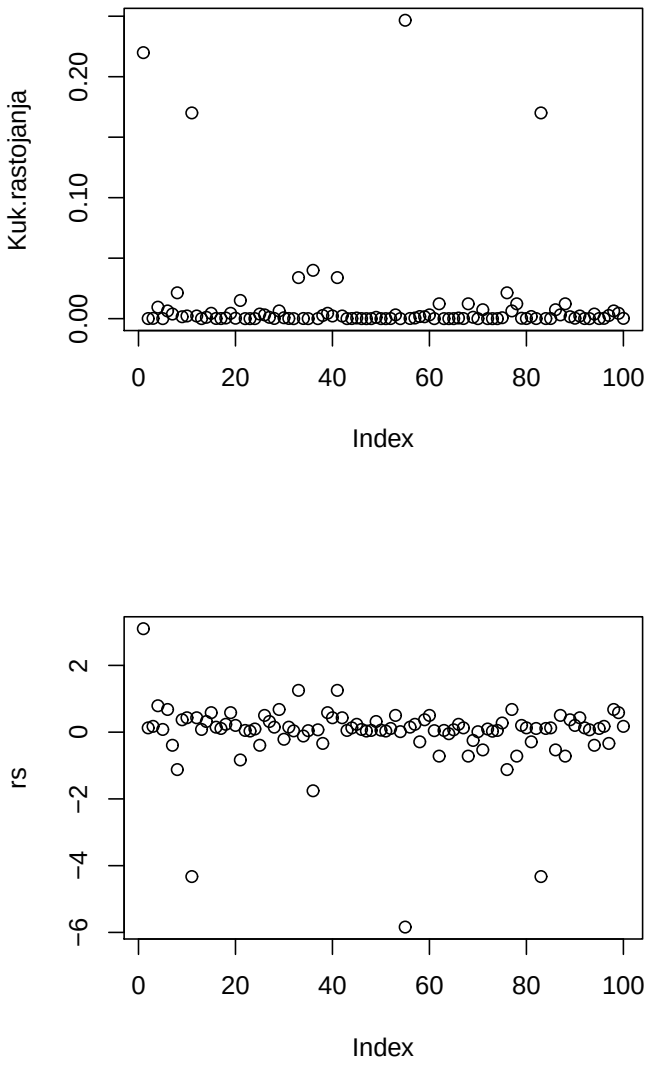
$$r_j^s = \frac{r_j}{\sqrt{1-h_{ii}}},$$

где се тежина добија на сличан начин као у слуцају линеарне регресије. У R-у се за одредивање тежина сваке од обзервација може користити функција *hatvalues*. Имајући ово у виду, комбиновањем графичких приказа стандардизованих резидуала, тежина, Куковог растојања и још неких мера утицаја (видети [5] и [9]) се може доћи до закључка које обзервације треба додатно испитати.

Пример 4.2.3. *Посматрајмо податке из примера 4.2.2 и резидуале из финалног модела. Приказаћемо стандардизоване Пирсонове резидуале, и Кукова растојања.*

```
#одређујемо стандардизоване Пирсонове резидуале
tezine=hatvalues(model)
rs=residuals(model,type='pearson')/sqrt(1-hatvalues(tezine))
#Кукова растојања
Kuk.rastojanja=cooks.distance(model)
plot(Kuk.rastojanja)
plot(rs)
```

Са графика 4.11 видимо да иако неке тацке показују одступања делује да њихов утицај на модел није значајан. Поред тога, ако се осврнемо на то да висина није сигурно једини фактор који утиче на вероватноћу поготка, и да су овакви подаци скроз реалистични, нећемо избацити посматране обзервације из модела.



Слика 4.11: Кукова растојања и стандардизовани Пирсонови резидуали

Следећи пример илуструје како се модел може направити када су подаци груписани.

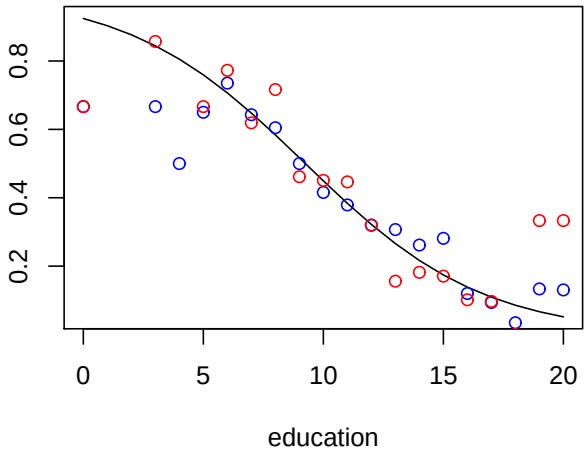
Пример 4.2.4. *У бази* womensrole *пакета* HSAUR *налазе се подаци о мишљењу људи о улози жена у друштву добијени у једном истраживању 1974-1975. године. На питање да ли се слажу да жене треба да буду задужене за породицу а мушарци да се баве државном политиком, могло је да се одговори са ДА и НЕ. Број жена који је одговорио потврдно означен је са* agreee *а оних који се није сложио са* disagree. *Поред тога, у бази се налазе подаци о полу испитаника (променљива* sex*) и година проведених у школовању (променљива* education*). Направићемо логистички регресиони модел и продискутовати како се он може евентуално побољшати.*

```
data("womensrole", package = "HSAUR3")
# правимо модел за груписане податке
modelZene= glm(cbind(agree, disagree)~education+sex,
data = womensrole, family = binomial())
summary(modelZene)
Call:
glm(formula = cbind(agree, disagree)~education + sex, family =
binomial(), data = womensrole)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.72544 -0.86302 -0.06525  0.84340  3.13315
Coefficients:
             Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)    2.5094      0.1839    13.65    0.0000
education     -0.2706      0.0154   -17.56    0.0000
sexFemale     -0.0114      0.0841    -0.14    0.8918
```

На основу добијених резултата видимо да коефицијент уз порменљиву која означава пол испитаника (индикатор да се ради женском полу) није значајан. Проверићемо то и тестом који је заснован на разлици девијација који смо споменули у овом поглављу.

```
modelZene1=glm(cbind(agree, disagree)~education, data = womensrole, family = binomial())
anova(modelZene1,modelZene,test='Chisq')
Analysis of Deviance Table
  Resid. Df  Resid. Dev  Df  Deviance  Pr(>Chi)
1        39       64.03
2        38       64.01   1      0.02    0.8918
```

P-вредност теста упућује на прихватање нулте хипотезе да посматрани коефицијент има вредност нула. Представићемо оцењене вероватноће моделом и реализоване за сваки пол (црвеном бојом за женске особе, а плавом за мушке особе).



Слика 4.12: Мишљење јавног мњења о положају жена-приказ реализованих и предвиђених фреквенција

```
wr=womensrole
wr$ver=predict(modelZene1,newdata=womensrole,type='response')
# издвајамо податке за мушки и женски пол
```

```
wrM=wr[wr$sex=='Male',]
pM=wrM$agree/(wrM$agree+wrM$disagree)
wrF=wr[wr$sex=='Female',]
wrF$ver=predict(modelZene,newdata=wrF,type='response')
pF=wrF$agree/(wrF$agree+wrF$disagree)
plot(wrM$education,wrM$ver,type='l',col='black',xlab='education',
ylab='')
points(wrM$education,pM,col='blue')
points(wrF$education,pF,col='red')
```

Са графика 4.12 видимо се моделом релативно задовољавајуће предвиђају вероватноће, али и да кад бисмо правили моделе одвојено вероватно бисмо добили другачије криве зависности, и поред тога што се коефицијент уз пол показао недовољно значајним. То се може објаснити тиме да разлика између оцењених логит функција није само константна вредност (што се заправо тестира са $H_0 : \beta_0 = 0$). Тај проблем се може решити увођењем додатне променљиве која би описала интеракцију између пола и школованости испитаника. Једна могућност је да формирамо нову променљиву која има вредност 0 кад се ради о референтном полу (мушки), а када се ради о женском полу онда ма вредност променљиве education. *На тај начин правимо модел који узима у обзир различит утицај променљиве* education *у зависности од пола. У даљем тексту направићемо тај модел а читаоцу остављамо да упореди његов квалитет са квалитетом приказаног модела.*

```
wrI=womensrole
wrI$I=(as.numeric(womensrole$sex)-1)*womensrole$education
modelZene2= glm(cbind(agree, disagree) education+sex+I, data
= wrI, family = binomial())
wrI$ver=predict(modelZene2,newdata=wrI,type='response')
```

Задатак 4.2.1. У бази BreastCancer која се налази у пакету mlbench налазе се подаци о резултатима тумора који је класификован као бенингни или малигни. Направити одговарајући логистички модел на основу којег бисте, на основу резултата биопсије, предвиђали малигнитет тумора, а затим и испитати његову

класификациону моћ (одзив, тачност и прецизност). С обзиром да је крајњи циљ класификовање резултата на основу модела, приликом прављења модела и одређивања његовог квалитета водити рачуна о проблему преприлагођавања.

Задатак 4.2.2. У пакету mlbench налази се база података PimaIndiansDiabetes2 о присуству Дијабетеса и неких других клиничких показатеља који потенцијално утичу на присуство. Направити одговарајући модел логистичке регресије који се може користити за предвиђање вероватноће болести, односно за одговарајућу класификацију.

Задатак 4.2.3. Уместо променљиве education из примера 4.2.3 направити нову променљиву која означава да ли је број година школовања мањи од 8, између 8 и 12, и већи од 12. Направити модел који у обзир узима ту нову променљиву и променљиву sex. Утврдити значајност променљиве sex а након тога одрредити Пирсонове резидуале и вредност одговарајуће статистике која се користи за проверу квалитета модела. Да ли је добијен модел задовољавајући?

5

Бајесова статистика

У претходним поглављима приказали смо основне статистичке методе које су засноване на фреквенционистичком схватању вероватноће догађаја, односно да вероватноћа догађаја представља удео остваривања догађаја у великом броју понављања експеримената. У бајесовском приступу, вероватноћа треба да представи степен нашег веровања да ће се неки догађај догодити. Свако ново искуство које стичемо може утицати на промену нашег уверења о реализацији неког догађаја, односно његове вероватноће. Основу за развој метода заснованих на овом приступу представља чувена Бајесова теорема на основу које се може одредити условна вероватноћа неког догађаја. Управо по њеном творцу, енглеском статистичару и теологу, Томасу Бејзу[1] правац статистистике који користи ове методе је назван Бајесова (или бајесовска) статистика. Њене темеље, у 18. веку, поставио је француски математичар Пјер-Симон Лаплас[2], међутим као и свака нова теорија, филозовски веома другачија од класичног приступа, представљала је контраверзу чак и у току 20. века. Томе је допринела и потреба за компликованим израчунавањима. Тек са развојом рачунара уочене су многе предности, и данас Бајесов приступ је свакидашњица у анализи података.

У овом поглављу приказаћемо неке основне статистиче методе попут бајесовског оцењивања и тестирања, с циљем да читаоцу

[1] енг. Thomas Bayes
[2] енг. Pierre-Simon Laplace

приближимо бајесовски приступ статистичог закључивања.

5.1 Бајесово оцењивање

За разлику од класичног (фреквенционистичког) приступа, претпостављамо да су непознати параметри случајне величине које имају неку априорну расподелу. Природно се поставља питање коју?

Могућности су разне и то пре свега зависи од нашег предзнања о неком параметру. Основни мото Бајесовог приступа је да су априроне вероватноће онолике колико ВЕРУЈЕМО да јесу. Природно је да, када се добије нека нова информација о параметру, та расподела промени.

Пример 5.1.1. *Претпоставимо да предавач уђе први пут у учионицу на почетку семестра и постави неко питање које се односи на разумевање првог предавања. Означимо са p вероватноћу тачног одговора. Уколико предавач нема никакво предзнање о студентима јасно је да највише има смисла претпоставити да p има $\mathcal{U}[0,1]$ расподелу. Међутим како одмиче семестар, предавач постаје свестан састава публике и неће више претпостављати да се ради униформној расподелу. Дакле, априорна расподела се модификује. Фреквенционисти би "поновили први час" велики број пута и на основу тога оценили вероватноћу успеха. У пракси је то свакако немогуће јер се састав групе сваке године мења а и први час се дешава једном годишње.*

Ради бољег разумевања навешћемо још један пример.

Пример 5.1.2. *Претпоставимо да смо бацили новчић $N = 3$ пута и да је од тих N пута $k = 3$ пута пало писмо. Желимо да оценимо вероватноћу да падне писмо. Уколико бисмо класично (фреквенционистички) приступили проблему, користећи метод максималне веродостојности добијамо да је $\hat{p} = 1$.*
Наиме функција веродостојности је

$$L(p) \propto p^{\sum_{i=1}^{N} I\{x_i=pismo\}}(1-p)^{N-\sum_{i=1}^{N} I\{x_i=pismo\}} = p^3.$$

Максимум функције на интервалу $[0,1]$ *је баш за* $p=1$. *Сада када бисмо поново бацили новчић очекивали бисмо да поново падне писмо. Што не функционише баш тако у пракси.*

Када примењујемо бајесовски приступ претпостављамо да је p *случајна величина. Како немамо никаквог предзнања о новчићу, претпоставићемо да* p *има* $\mathcal{U}[0,1]$ *расподелу а затим наћи апостериорну расподелу за* p *уколико знамо да је било баш 3 писма. Апостериорна функција густине за* p *је*

$$\begin{aligned} f(p|\sum_{i=1}^{N} I\{x_i = P\} = k) &= \frac{P\{\sum_{i=1}^{N} I\{x_i = P\} = k|p\}\pi(p)}{\int_0^1 P\{\sum_{i=1}^{N} I\{x_i = P\} = k|p\}\pi(p)dp} \\ &= \frac{\binom{N}{k}p^k(1-p)^{N-k}\cdot 1}{\int_0^1 \binom{N}{k}p^k(1-p)^{N-k}\cdot 1 dp} \end{aligned} \tag{5.1}$$

Како је

$$\begin{aligned} g(x) &= \frac{x^k(1-x)^{N-k}}{B(k+1,N-k+1)} = \frac{x^k(1-x)^{N-k}\Gamma(N+2)}{\Gamma(k+1)\Gamma(N-k+1)} \\ &= \frac{x^k(1-x)^{N-k}(N+1)!}{k!(N-k)!} \end{aligned}$$

густина Бета $\beta(k,N-k+1)$ *расподеле добијамо да је интеграл у имениоцу*

$$\int_0^1 \binom{N}{k} p\cdot p^{k-1}(1-p)^{N-k}dp = \int_0^1 \frac{p^k(1-p)^{N-k}(N+1)!}{k!(N-k)!}dp \cdot \frac{\frac{N!}{k!(N-k)!}}{\frac{(N+1)!}{k!(N-k)!}}$$
$$= \frac{1}{N+1}.$$

Одавде је

$$f(p|\sum_{i=1}^{N} I\{x_i = P\} = k) = \frac{\binom{N}{k}p^k(1-p)^{N-k}\cdot 1}{\frac{1}{N+1}} = \frac{p^k(1-p)^{N-k}}{B(k+1,N-k+1)}.$$

Дакле апостериорна расподела је $\beta(k+1,N-k+1)$. *На овом месту напомињемо да интеграл у имениоцу нисмо морали да*

рачунамо јер не зависи од p *па се апостериорна густина одмах може написати у облику*

$$f(p|\sum_{i=1}^{N} I\{x_i = P\} = k) = Cp^k(1-p)^{N-k}, \tag{5.2}$$

и како је C *једнозначно одређено условом да је* $\int_0^1 f(p|\sum_{i=1}^{N} I\{x_i = P\} = k)dp = 1$*, одмах можемо закључити да је апостериорна расподела је* $\beta(k+1, N-k+1)$*. Израз* (5.2) *се краће записује*

$$f(p|\sum_{i=1}^{N} I\{x_i = P\} = k) \propto p^k(1-p)^{N-k},$$

при чему знак $\propto$ *означава пропорционалност. Ова ознака се често среће у литератури.*

Видимо да је класа расподела остала иста али да су се параметри променили. Сада се природно поставља питање шта је оцена за p*. Има смисла да оцена буде баш мода апостериорне расподеле. У том случају добићемо исто што и методом маскималне веродостојности. Међутим, како је природно мислити на просек када описујемо случајне величине, смислена оцена може да буде баш очекивана вредност апостериорне расподеле. Тада је* $\hat{p} = \frac{k+1}{N+2} = \frac{4}{5}$*. Видимо да кад је* N *велико ова оцена и оцена методом максималне веродостојности се не разликују много.*

Ако имамо неко предзнање о новчићу, на пример да је више вероватно да је $p > 0.5$ *и зато претпоставимо да је априорна расподела за* p *бета* $\beta(a, b)$*, за* $a = 3$ *и* $b = 1$*. Тада се аналогним поступком може добити да је апостериорна расподела за* p $\beta(k+a, N-k+b)$ *па је* $\hat{p} = \frac{k+a}{N+a+b} = \frac{6}{7}$*.*

Видимо да оцена прилично зависи од априорне расподеле. Зато се о избору исте мора водити рачуна. Препорука је да када заиста немамо никаквог предзнања о параметру да користимо неку *неинформативну* расподелу. Више о таквим расподелама може се прочитати у нпр. [12].

Осим што зависи од априорне расподеле, оцена зависи и од тога коју нумеричку карактеристику апостериорне расподеле смо

одабрали (имали смо пример моде и пример очекиване вредности). Свако ће природно бирати ону оцену која је по њему најбоља односно она која максимизира "корист", односно минимизира губитак. Управо зато морамо пре оцењивања да дефинишемо функцију губитака. Неке од најчешћих функција губитака су

$$L(\hat{\theta}(X),\theta) = (\theta - \hat{\theta}(X))^2 \tag{5.3}$$

$$L(\hat{\theta}(X),\theta) = |\theta - \hat{\theta}(X)| \tag{5.4}$$

$$L(\hat{\theta}(X),\theta) = 1, \text{ за } |\theta - \hat{\theta}(X)| > \varepsilon \tag{5.5}$$

Одговарајући очекивани губици су

$$U(\hat{\theta},\theta) = E(\theta - \hat{\theta}(X))^2 \tag{5.6}$$

$$U(\hat{\theta},\theta) = E|\theta - \hat{\theta}(X)| \tag{5.7}$$

$$U(\hat{\theta},\theta) = P\{|\theta - \hat{\theta}(X)| > \varepsilon\}. \tag{5.8}$$

Сада можемо дефинисати Бајесов ризик

$$R(\hat{\theta}) = E_\theta U(\hat{\theta},\theta) = \int_{-\infty}^{\infty} U(\hat{\theta},\theta)\pi(\theta)d\theta, \tag{5.9}$$

где је $\pi(\theta)$ априорна густина за θ. У случају да је априорна расподела за θ дискретна, претходни интеграл се своди на суму. **Бајесова оцена биће она вредност параметра која минимизира Бајесов ризик.** У случају (5.3) добија се да је оцена за θ баш $\hat{\theta} = E(\theta|X)$, а у случају (5.4) $\hat{\theta} = \text{med}(\theta|X)$. У случају функције ризика (5.5), кад $\varepsilon \to 0$ добија се да је оцена $\hat{\theta} = \text{argmax} f(\theta|X)$, односно мода апостериорне расподеле. У примеру 5.1.2 смо одредили управо ове оцене.

С обзиром на стохастичку природу параметра θ можемо на основу апостериорне расподеле одредити и *интервале прекривања*, односно интервале у којима се непознати параметар налази са вероватноћом β. Ти интервали су облика (θ_1, θ_2) где су θ_1 и θ_2 одређени из услова $P\{\theta_1 < \theta < \theta_2|x\} = \beta$. Наравно, вероватноће се рачунају на основу апостериорне расподеле параметра θ. Јасно је да *ниво прекривааања* β не одређује једнозначно константе θ_1 и θ_2.

Често се одређују тако да је $P\{\theta_1 > \theta\} = P\{\theta_2 < \theta\} = \frac{1-\beta}{2}$. Поред тога популаран је и приступ такозваних интервала прекривања највеће апостериорне густине (NAG) који су веома често најкраћи могући.

Пример 5.1.3. *Одредићемо* 95% *интервал покривања за* p *из претходног примера. У случају униформне априорне расподеле добили смо да је апостериорна расподела* $\beta(4,1)$. *Одговарајући квантили ове расподеле су* $F^{-1}_{\beta(4,1)}(0.025) = 0.398$ *и* $F^{-1}_{\beta(4,1)}(0.975) = 0.994 = 0.994$. *Дакле, стварна вредност параметра* p *ће припадати интервалу* $(0.398, 0.994)$ *са вероватноћом 0.95.*

У случају да је априорна расподела $\beta(3,1)$ *добијамо да је апостериорна расподела* $\beta(6,1)$. *Сада је 95% интервал прекривања* $(0.541, 0.996)$.

Примећујемо да смо у случају информативне априорне расподеле добили ужи интервал, што је очекивано јер смо, укључивањем предзнања о параметру, на неки начин "сузили" скуп могућности за њега.

Један од основних статистичких задатака је предикција. Када знамо апостериорну расподелу за θ можемо одредити и такозвану *предиктивну расподелу* односно

$$\begin{aligned} f(x|\mathbf{X}) &= \int f(x|\theta, \mathbf{X}) f(\theta|\mathbf{X}) d\theta \\ &= \int f(x|\theta) f(\theta|\mathbf{X}) d\theta, \end{aligned}$$

где је $\mathbf{X}$ реализован прост случајан узорак. Једнакост $f(x|\theta, \mathbf{X}) = f(x|\theta)$ важи јер, за фиксно θ, подаци су међусобно независни.

Пример 5.1.4. *Настављамо пример са новчићем. Уколико је априорна расподела била униформна, добијамо да је закон*

расподеле посматране случајне величине, након експеримента,

$$\begin{aligned} f(x|\mathbf{X}) &= \int_0^1 p^x(1-p)^{1-x}\frac{p^k(1-p)^{N-k}}{\beta(k+1,N-k+1)}dp \\ &= \frac{\beta(x+k+1,N-k-x+2)}{\beta(k+1,N-k+1)}\int_0^1 \frac{p^{x+k}(1-p)^{1-k+N-x}}{\beta(x+k+1,N-k-x+2)}dp \\ &= \frac{\beta(x+k+1,N-k-x+2)}{\beta(k+1,N-k+1)} = \frac{\beta(x+4,2-x)}{\beta(4,1)} \\ &= \frac{(x+3)!(1-x)!4!}{5!3!0!} = \frac{(x+3)!}{30}, \quad \textit{за } x \in \{0,1\}. \end{aligned}$$

Дакле, вероватноћа да падне (једно) писмо у наредном експерименту је $\frac{24}{30} = \frac{4}{5}$. *Слично се може извести расподела за број успеха у наредних* m *бацања. Имамо да је, за* $x \in \{0,1,....,m\}$

$$\begin{aligned} f(x|\mathbf{X}) &= \int_0^1 \binom{m}{x} p^x(1-p)^{m-x}\frac{p^k(1-p)^{N-k}}{\beta(k+1,N-k+1)}dp \\ &= \binom{m}{x}\frac{\beta(x+k+1,N-k+m-x+1)}{\beta(k+1,N-k+1)}. \end{aligned}$$

Често нам крајни циљ истраживања није да нађемо апостериорну расподелу за напознат параметар (већ само средство), него нас занима да оценимо или дођемо до неких других закључака о некој функцији $g(\theta)$ од непознатог параметра θ. На пример, занима нас апостериорна вероватноћа

$$P\{g(\theta) \in A|X\} = \int \mathrm{I}\{g(\theta) \in A\}f(\theta|X)d\theta. \tag{5.10}$$

Приметимо да у случају да нас занима функција расподеле за $g(\theta)$ у тачки y онда је $A = (-\infty, y]$.

Понекад нећемо бити у могућности да експлицитно одредимо (5.10). Тада можемо да генеришемо велики узорак $\theta^{(1)}, ..., \theta^{(N)}$ из апостериорне расподеле и онда (5.10) оценимо са

$$\frac{1}{N}\sum_{i=1}^{N} I\{g(\theta^{(i)}) \in A\}.$$

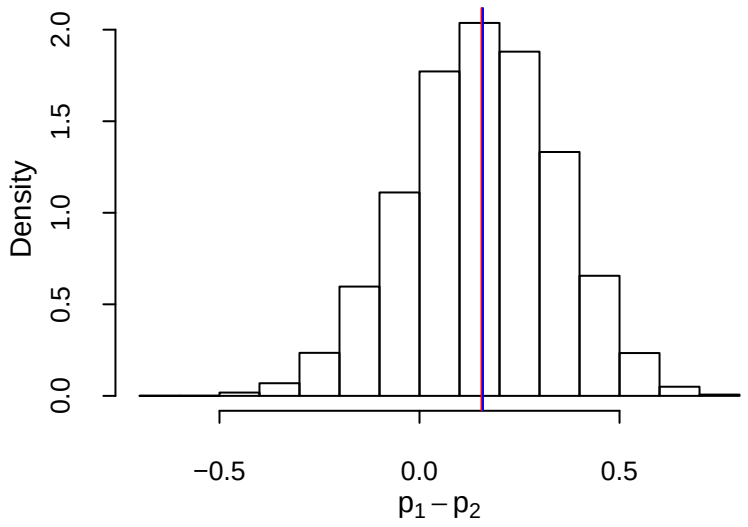
На сличан начин можемо из добијене емпиријске апостериорне расподеле за $g(\theta)$ да добијемо и интервале прекривања.

Пример 5.1.5. *Посматрамо број пацијената који су преживели више од годину дана од тренутка обољевања од једне смртоносне болести у случају да су примали нову експерименталну терапију* (X) *и када је давана стандардна терапија* (Y)*. У првој групи се налазило* n_1 *пацијаната, а у другој* n_2*. Јасно је да тада да се* X *може моделовати са Биномном* $\mathcal{B}(n_1, p_1)$ *а* Y *са Биномном* $\mathcal{B}(n_2, p_2)$ *расподелом. На основу реализованих вредности* x *и* y *циљ је да се оцени* $p_1 - p_2$*.*
Ради једноставности, претпоставићемо да је априорна расподела за $(p_1.p + 2)$ *униформна на* $[0,1] \times [0,1]$*. Тада је апостериорна расподела за* (p_1, p_2)

$$\begin{aligned} f(p_1, p_2|X, Y) &\propto f_{X,Y}(X, Y|p_1, p_2) \cdot \pi(p_1, p_2) \\ &\propto p_1^X (1-p_1)^{n_1 - X} p_2^Y (1-p_2)^{n_2 - Y} \cdot 1 \\ &\propto f(p_1|X) \cdot f(p_2|Y), \end{aligned}$$

па су p_1*, и* p_2 *независне случајне величине у односу на апостериорну расподелу. Из* $f(p_1|X) \propto p_1^X (1-p_1)^{n_1 - X}$ *закључујемо да је апостериорна расподела за* p_1 *у ствари Бета* $\mathcal{B}(X+1, n_1 - X + 1)$*. Слично, добијамо и да је апостериорна расподела за* p_2 *Бета* $\mathcal{B}(Y+1, n_2 - Y + 1)$*. Сада можемо да генеришемо два велика узорка* $p_1^{(1)}, ..., p_1^{(N)}$ *и* $p_2^{(1)}, ..., p_2^{(N)}$ *из добијених апостериорних расподела, на основу којих добијамо узорак* $p_1^{(1)} - p_2^{(1)}, ..., p_1^{(N)} - p_2^{(N)}$ *на основу ког се може апроксимирати апостериорна расподела за* $p_1 - p_2$ *и добити одговарајућа тачкаста и интервална оцена.*

Примера ради нека је $n_1 = 10$ *и* $n_2 = 12$ *и нека су бројеви преживелих редом једнаки* 6 *и* 5*. Тада су апостериорне расподеле за параметре* p_1 *и* p_2 *редом Бета* $\mathcal{B}(7, 5)$ *и* $\mathcal{B}(6, 8)$*. На графику 5.1 је приказан хистограм на основу* $p_1 - p_2$*. Црвена и плава линија (које су веома блиске) представљају редом вредности узорачке средине* 0.155 *и узорачке медијане* 0.159*. Управо те оцене могу бити тачкасте оцене за разлику вероватноћа преживљавања. Што се тиче интервалне оцене, добили смо да је* 95% *интервал прекривања* $(-0.22, 0.51)$*. Интервал је широк али је то било и очекивано с обзиром на неинформативну априорну расподелу параметара и мале обиме узорака. из добијених резултата се може ипак закључити да новодобијена терапија даје неке резултате.*



Слика 5.1: Хистограм генерисаног узорка $p_1 - p_2$ из апостериорне расподеле

```
set.seed(101)
p1=rbeta(10000,7,5)
p2=rbeta(10000,6,8)
hist(p1-p2,prob=TRUE,main="",xlab=expression(p[1]-p[2]))
# потенцијалне тачкасте оцене
abline(v=mean(p1-p2),col='red')
abline(v=median(p1-p2),col='blue')
# 95% интервал прекривања
quantile(p1-p2,c(0.025,0.975))
```

```
2.5%          97.5%
−0.2206679 0.5102768
```

Задатак 5.1.1. Већ је било напоменуто да се број голова на фудбалској утакмици често може моделовати Пуасоновом $\mathcal{P}(\lambda)$ расподелом. С обзиром на то да се састав екипе мења од сезоне до сезоне смислено је претпоставити да је λ случајна величина. Ове године је посматрани тим на 3 узаступне утакмице постигао редом 2,0,1 гол. На основу тога, одредити апостериорну расподелу за λ и $\hat{\lambda}$ уколико је априорна расподела

$$\lambda : \begin{pmatrix} 0.5 & 1 & 1.5 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix}$$

Задатак 5.1.2. Случајна величина X има Биномну $\mathcal{B}(1,p)$ расподелу. На основу узорка 0,0,1,0,1 одредити 90% интервал прекривања за $\log(\frac{p}{1-p})$. Приликом одабира априорне расподеле за p узети у обзир чињеницу да не постоје прелиминарне информације о параметру p.

Задатак 5.1.3. Сматра се да је погодна расподела за моделовање времена (у недељама) које протекне од употребе бицикла до пуцања гума (редовном употребом) Експоненцијална $\mathcal{E}(\lambda)$. Од бициклисте (који зна статистику) је добијена информација да је смислено за априорну расподелу за λ узети Гама $\gamma(6,100)$. Бициклиста је замољен да забележи време које је било потребно да пукну његове тренутне гуме и добијено је 39 и 42. Како се на основу тих нових података променило његово веровање о времену трајања гума? Одредити тачкасту оцену за λ.

Задатак 5.1.4. Сматра се да тежина упаковане чоколаде има нормалну $\mathcal{N}(m, 0.8^2)$ расподелу и да се тежине чоколада могу сматрати низом независних случајних величина. Шеф одсека за контролу квалитета верује да m има нормалну $\mathcal{N}(80, 0.2^2)$ расподелу. Узет је узорак обима 4 и добијене су редом тежине $80, 79, 79.5, 80.1$. Одредити апостериорну расподелу за m и 90% интервал прекривања. Одредити (или апроксимирати симулацијама) предиктивну вероватноћу да тежина паковања буде већа од 81 грама.

5.2 Бајесово тестирање статистичких хипотеза

Бајесов приступ се може користити и за тестирање статистичких хипотеза. На пример, желимо да тестирамо

$$H_0 : \theta = \theta_0 \text{ против } H_1 : \theta \neq \theta_0.$$

Идеја је да се свакој од хипотеза доделе априорне вероватноће $P\{H_0\} = p$ и $P\{H_1\} = 1 - p$ и да се одреде апостериорне вероватноће хипотеза. Прихвата се она хипотеза која има већу апостериорну вероватноћу. Најчешће, осим ако имамо важан разлог да то не урадимо, се хипотезама додељује априорна вероватноћа $p = 0.5$. Тада је

$$P\{H_0|\mathbf{X}\} = \frac{P\{\mathbf{X}|H_0\} \cdot \frac{1}{2}}{P\{\mathbf{X}|H_0\} \cdot \frac{1}{2} + P\{\mathbf{X}|H_1\} \cdot \frac{1}{2}} = \frac{P\{\mathbf{X}|H_0\}}{P\{\mathbf{X}|H_0\} + P\{\mathbf{X}|H_1\}}, \tag{5.1}$$

при чему је $P\{\mathbf{X}|H_i\}$ је функција веродостојности у случају да важи хипотеза H_i. Сада је потребно да задамо неку априорну расподелу за параметар θ тако да је $P\{\theta = \theta_0\} = 0.5$ а густина на скупу H_1 $\pi(\theta)$, при чему је $P\{\theta \neq \theta_0\} = 0.5$. Тада је израз (5.1) једнак

$$\frac{P\{X|\theta_0\}}{P\{X|\theta_0\} + \int P\{X|\theta\}\pi(\theta)d\theta}.$$

Уколико је скуп одређен алтернативном хипотезом једночлан онда је овај израз се своди на

$$\frac{P\{X|\theta_0\}}{P\{X|\theta_0\} + P\{X|\theta_1\}}.$$

Дакле, Бајесово тестирање није ништа друго но поређење два статистичка модела. Уколико желимо да поредимо k модела $M_1, ..., M_k$ који зависе од параметра θ (или више параметара) прво морамо да задамо априорне вероватноће сваког од модела, тј. $\pi_1, ..., \pi_k$ где је $\pi_i = P\{M_i\}$ (вероватноћа да се ради о моделу M_j).

Даље, нека је дата и априорна расодела параметра θ, у ознаци $p_j(\theta|M_j)$ на сваком од посматраних модела. Тада је

$$P\{M_j|X\} = \frac{p(X|M_j)\pi_j}{\sum_{i=1}^{k} p(X|M_i)\pi_i}, \tag{5.2}$$

где је $p(X|M_j) = \int p(X|\theta)p_j(\theta|M_j)d\theta$ (ознака $p(\cdot|\theta)$, у зависности од контекста, означава или условну густину или условни закон расподеле) Сада се два модела M_j и M_l могу поредити посматрањем количника одговарајућих апостериорних вероватноћа, односно

$$\frac{P\{M_j|X\}}{P\{M_l|X\}} = \frac{p(X|M_j)\pi_j}{p(X|M_l)\pi_l}. \tag{5.3}$$

Количник

$$\frac{p(X|M_j)}{p(X|M_l)} = \frac{\int p(X|\theta)p_j(\theta|M_j)d\theta}{\int p(X|\theta)p_l(\theta|M_l)d\theta}. \tag{5.4}$$

се назива *Бајесов фактор*. Како су начешће априорно сви модели једнако вероватни, управо рачунање Бајесовог фактора нам говори који је модел бољи. Поред овога, за поређење модела се може користити и Бајесов информациони критеријум, али о томе овде нећемо говорити.

Пример 5.2.1. *Претпоставимо да желимо да пренесемо поруку (број 0 или 1) преко комуникационог канала. Међутим због, шума не добијамо то већ вредност Y. Дакле, посматрамо модел*

$$Y = X + \varepsilon, \quad X \in \{0, 1\}, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2),$$

где је σ^2 познато, и треба на основу регистроване вредности Y да закључимо да ли је послата вредност $X = 0$ или $X = 1$. Преведено на Бајесов језик, имамо два модела која желимо да упоредимо $M_1 : X = 0$ и $M_2 : X = 1$. Даље, знамо да $Y|M_1 \sim \mathcal{N}(0, \sigma^2)$ расподелу, а да $Y|M_2 \sim \mathcal{N}(1, \sigma^2)$. Сада је Бајесов фактор једнак:

$$\frac{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{y^2}{\sigma^2}}}{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-1)^2}{\sigma^2}}} = e^{-\frac{y^2-(y-1)^2}{\sigma^2}},$$

и у зависности од y и σ^2 можемо донети одлуку. Уколико не знамо параметар σ^2 онда ћемо за њега одабрати неку априорну расподелу и поновити поступак.

5.3 Неколико потенициjалних проблема

У преходном поглављу видели смо да су многе ствари у Бајесовој статистици веома интуитивне. Међутим два потенцијана проблема се очигледно издвајају. Први, како одабрати априорну расподелу, и други, кад је одаберемо, како да вршимо закључивање на основу апостериорне расподеле ако не можемо експлицитно да добијемо облик те расподеле.

Већ смо навели да уколико се ради о функији губитака која представља средњеквадратно растојање оцена за непознат параметар је

$$\hat{\theta} = E(\theta|\mathbf{X}) = \int_{\mathbf{R}} \theta f(\theta|\mathbf{X}) d\theta = \frac{\int_{\mathbf{R}} \theta L(\mathbf{X}|\theta)\pi(\theta)d\theta}{\int_{\mathbf{R}} L(\mathbf{X}|\theta)\pi(\theta)d\theta}.$$

Рачунање интеграла који се јавља у изразу за оцену се може вршити нумерички или Монте Карло интеграцијом, али је то често доста временски захтевно уколико је непознат параметар вишедимензионалан. Тај проблем се најчешће решава Монте Карло методама заснованим на Марковљевим ланцима иза којих стоји идеја да се конструише Марковљев ланац (низовима случајних величина код којих расподела k-тог елемента зависи само од реализоване вредности $k-1$-вог елемента) чија стационарна расподела (расподела коју има n-ти елемент низа за велико n) је баш апостериорна расподела параметра.

Други приступ за апроксимацију апостериорне расподеле је познат под називом "варијационо закључивање" у којем се, грубо речено, коришћењем оптимизационих метода за апостериорну расподелу бира једна из класе допустивих расподела која је, на основу неког критеријума, "најближа" траженој.

Што се тиче самог одабира априорне расподеле, нећемо се много на овоме задржавати. Једна могућност је да се одаберу такозване конјуговане расподеле (априорна и апостериорна припадају истој класи расподела). Списак неких таквих расподела је приказан у табели. Поред тога може се одабрати униформна расподела (цак и неправа расподела-функција чији интеграл није један, је кандидат), или такозвана неинформативна Џефрисова расподела. Више о овоме се може наћи у [12].

X	априорна	апостериорна
$\mathcal{B}(p)$	$\beta(a,b)$	$\beta(a+n\bar{X}_n, b+n(1-\bar{X}_n))$
$\mathcal{P}(\lambda)$	$\gamma(a,b)$	$\gamma(a+n\bar{X}_n, b+n)$
$\mathcal{E}(\lambda)$	$\gamma(a,b)$	$\gamma(a+n, b+n\bar{X}_n)$
$\mathcal{N}(m,\sigma^2)$ σ^2 познато	$\mathcal{N}(m_0,\sigma_0^2)$	$\mathcal{N}((\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2})^{-1})(\frac{m_0}{\sigma_0^2})+\frac{n\bar{X}_n}{\sigma^2}, (\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2})^{-1})$

6

Додатак - важне расподеле и њихове особине

У овом поглављу представићемо важне расподеле које се најчешће користе у статистичком моделирању. За сваку од расподела приказаћемо основне нумеричке карактеристике, важне особине, и неке од најчешћих примена. Ради прегледнијег приказа расподеле су груписане у дискретне и апсолутно непрекидне.

6.1 Дискретне расподеле

- Случајна величина X има Бернулијеву расподелу (познату још и као индикатор неког случајног догађаја), уколико је закон расподеле:

$$X : \begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix}, \quad p \in (0,1). \tag{6.1}$$

  Математичко очекивање и дисперзија случајне величине X су редом $EX = p$ и $DX = p(1-p)$. Из дефиниције се види је да онда погодна за моделовање да ли се неки опит успешно реализовао или не.

- Случајна величина X има Биномну $\mathcal{B}(n,p)$ уколико је њен

закон расподеле

$$P\{X = k\} = \binom{n}{n} p^k (1-p)^{n-k}, \quad k = 0, 1.., n. \tag{6.2}$$

X представља број успешно реализованих опита од n независних покушаја. па је јасно да се X може представити као збир n независних индикатора са расподелом (6.1). Одавде је $EX = np$ и $DX = np(1-p)$.

- Случајна величина X има Геометријску $\mathcal{G}(p)$ расподелу уколико је њен закон расподеле

$$P\{X = k\} = p(1-p)^{k-1}, \quad k = 1, 2... \tag{6.3}$$

X представља број понављања опита до првог успеха (укључујући и последње понављање). Нумеричке карактеристике су $EX = \frac{1}{p}$ и $DX = \frac{1-p}{p^2.}$ У литератури се често може наћи да се случајном величином са Геометријском расподелом назива случајна величина X' за законом расподеле

$$P\{X' = k\} = p(1-p)^k, \quad k = 0, 1..., \tag{6.4}$$

па о томе треба водити рачуна. Из дефиниције закључујемо да је X' има исту расподелу као случајна величина $X - 1$ па се одавде може добити да је $EX' = \frac{1-p}{p}$ и $DX' = \frac{1-p}{p^2.}$ Случајне величине X и X' су често погодне за моделовање броја гађања на кош до првог успеха, броја секундарних потреса након земљотреса, број репордуктивних циклуса до зачећа. Приликом избора ове расподеле један од могућих путоказа може бити и одсуство меморије које од свих дискретних случајних величина поседује једино Геометријска расподела. Наиме, важи

$$P\{X > k + r | X > k\} = P\{X > r\}, \quad k = 1, 2...., r = 1, 2....$$

- Случајна величина X има Пуасонову $\mathcal{P}(\lambda)$, $\lambda > 0$, расподелу уколико је њен закон расподеле

$$P\{X = k\} = \frac{\lambda^k e^{-\lambda}}{k!} \quad k = 0, 1, 2... \tag{6.5}$$

X се може интерпретирати као број неких догађаја у фиксном временском интервалу (нпр. број позива Хитној помоћи у току ноћи, број елементарних непогода, број голова у фудбалу, итд.). Нумеричке карактеристике су $EX = \lambda$ и $DX = \lambda$. Уколико имамо n независних случајних величина са $\mathcal{P}(\lambda_1), ..., \mathcal{P}(\lambda_n)$, онда њихов збир Пуасонову $\mathcal{P}(\lambda_1 + \cdots + \lambda_n)$ расподелу.

- Случајна величина X има Негативну биномну расподелу $\mathcal{NB}(k,p)$ уколико је њен закон расподеле

$$P\{X = n\} = \binom{n-1}{k-1} p^k (1-p)^{n-k}, \quad n = k, k+1, ..., \tag{6.6}$$

где је k природан број. X се може интерпретирати као број понављања експеримената до k-тог успеха (укључујуц2и и последње понављање). На основу тога закључујемо да се може представити као збир независних случајних величина са законом расподеле (6.3). Одавде се једноставно добија да је $EX = \frac{k}{p}$ и $DX = \frac{k(1-p)}{p^2}$. Приметимо да се за $k = 1$ добија баш Геометријска случајна величина (6.3). Као и у случају Геометријске случајне величине и овде постоји неколико алтернативних дефиниција од којих издвајамо ону по којој X' представља број промашаја до k-тог успеха (где је k природан број). Тада је њен закон расподеле

$$P\{X' = n\} = \binom{n+k-1}{k-1} p^k (1-p)^n, \quad n = 0, 1, 2, ... \tag{6.7}$$

Из ове дефиниције следи да се X' може представити као збир k независних случајних величина са Геометријском расподелом (6.4) па је $EX' = \frac{k(1-p)}{p}$ и $DX' = \frac{k(1-p)}{p^2}$.

6.2 Апсолутно непрекидне случајне величине

- Случајна величина X са нормалном $\mathcal{N}(m, \sigma^2)$ расподелом има функцију густине

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-m)^2}{2\sigma^2}}, \quad m \in \mathbb{R},\ \sigma^2 > 0. \tag{6.8}$$

Ова расподела је једна од најчешће коришћних апсолутно непрекидних расподела дефинисаих $\mathbb{R}$ и као таквој ћемо посветити највише пажње.

За њене нумеричке карактеристике важи: $EX = m$ и $DX = \sigma^2$. Поред тога параметар m представља и медијану и моду расподеле. Уколико је $m = 0$ и $\sigma^2 = 1$ ради се о *стандардној Нормалној расподели* чију функцију густине ћемо означавати са $\phi(\cdot)$ а одговарајцу1у функцију расподеле са $\Phi(\cdot)$. Функција $\Phi(x) = \int_{-\infty}^{x} \phi(u)du$ нема аналитички облик али се може израчунати нумерички и у литератури се могу наћи табелиране вредности ове функције.

Може се показати да, ако X има $\mathcal{N}(m, \sigma^2)$ онда случајна величина

$$Z = \frac{X - m}{\sigma}$$

има стандардну Нормалну расподелу. Одавде закључујемо да за функцију расподеле случајне величине X важи

$$F(x) = \Phi\Big(\frac{x - m}{\sigma}\Big).$$

Из израза за густину (6.8) види се да је X симетрична око m, односно да важи да је

$$f(-x + \mu) = f(x - \mu)$$
$$F(x - \mu) = 1 - F(\mu - x).$$

Уколико су $X_1, ..., X_k$ независне случајне променљиве са $\mathcal{N}(m_1, \sigma_1^2), ..., \mathcal{N}(m_k, \sigma_k^2)$ расподелама и $a_1, ..., a_k \in \mathbf{R}$ тако да је $\sum_{i=1}^{k} a_i^2 > 0$ онда

$$a_1X_1 + \cdots + a_kX_k \sim \mathcal{N}(a_1m_1 + \cdots + a_km_k, a_1^2\sigma_1^2 + \cdots + a_k^2\sigma_k^2). \tag{6.9}$$

Поред наведеног својства, које је основа за статистичко закључивање у нормалном моделу, важно је напоменути да се стандардна Нормална расподела јавља као гранична

расподела узорачке средине у случају простог случајног узорка за које посматрано обележје има коначну дисперзију, односно, уколико је $X_1, ..., X_n$ низ независних случајних величина са истом расподелом (прост случајан узорак), онда се за велико n расподела случајне величине

$$\frac{\bar{X}_n - m}{\frac{\sigma}{\sqrt{n}}}$$

може апроксимирати стандардном Нормалном расподелом. Ово тврђење је познато под називом "Централна гранична теорема."

- Случајна величина X са χ^2_n расподелом има функцију густине

$$f(x) = \frac{x^{\frac{n}{2}-1}}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} e^{-\frac{x}{2}}, \quad x > 0. \tag{6.10}$$

Много чешће X се дефинише као **збир n квадрата независних случајних величина са стандардном $\mathcal{N}(0,1)$ расподелом.** Зато је $EX = n$ и $DX = 2n$. Приметимо још да из алтернативне дефиниције следи, на основу Централне граничне теореме, да за велико n

$$\frac{X - n}{\sqrt{2n}} \sim \mathcal{N}(0,1)$$

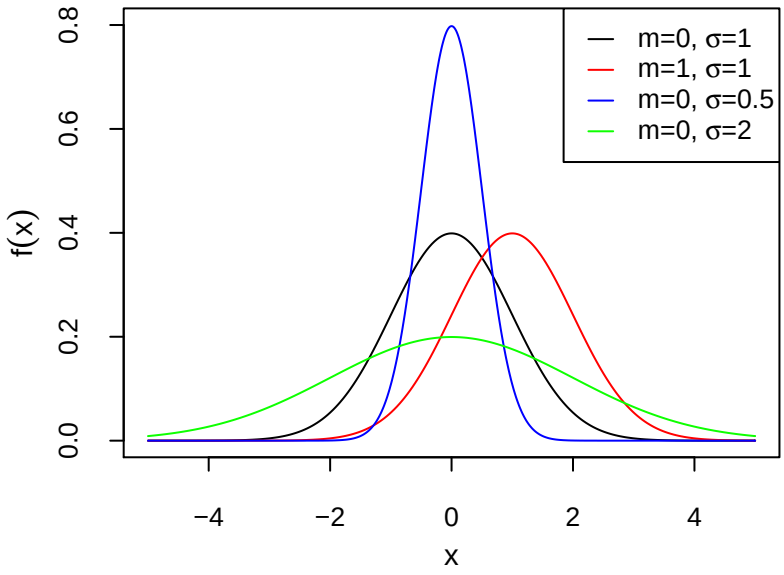
- Случајна величина X са Студентовом расподелом са n степени слободе (у ознаци t_n) има функцију густине

$$f(x) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})} \Big(1 + \frac{x^2}{n}\Big)^{-\frac{n+1}{2}}, \; n \in \mathbf{R}^+, \; x > 0.$$

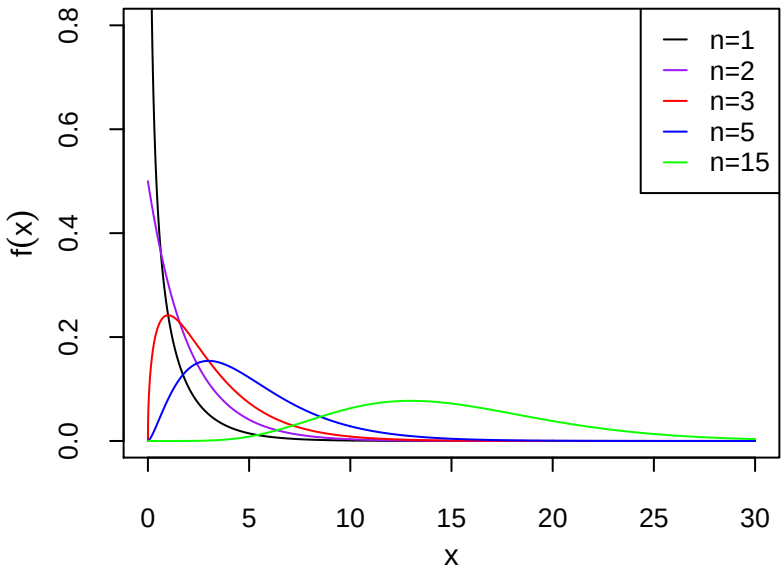
Много чешће се X дефинише као

$$\frac{Z}{\sqrt{\frac{Y}{n}}},$$

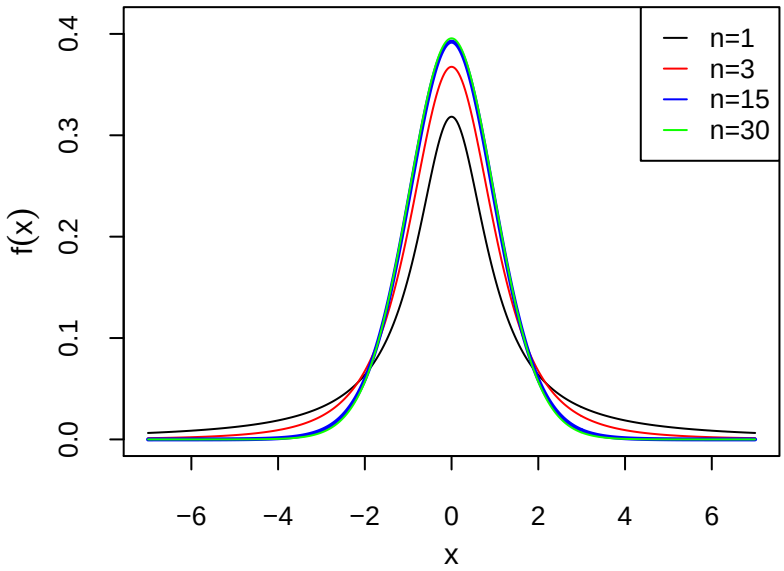
где су Z и Y независне случајне променљиве са $\mathcal{N}(0,1)$ и χ^2_n, редом, расподелама. Може се показати да је $EX = 0$



Слика 6.1: Густина нормалне расподеле за различите вредности параметара m и σ



Слика 6.2: Густина χ^2 расподеле за различите вредности степена слободе n



Слика 6.3: Густина Студентове расподеле за различите вредности степена слободе n

(за $n > 1$, иначе не постоји) и $DX = \frac{n}{n-2}$ (за $n > 2$, иначе не постоји). Ова расподела је симетрична око нуле ($F(x) = 1 - F(-x)$) и као таква веома слична стандардној нормалној расподели, с напоменом да су репови ове расподеле дебљи. Зато је она погодна за моделирање разних обележја која се јављају у финансијама и актуарству (нпр. величина одштета у осигуравајућем друштву). За велико n ($n \geq 30$) се добро апроксимира $\mathcal{N}(0,1)$ расподелом.

- Случајна величина X са експоненцијалном $\mathcal{E}(\lambda)$ расподелом,

$\lambda > 0$, има функцију густине

$$f(x) = \lambda e^{-\lambda x}, \;\; x > 0. \tag{6.11}$$

Функција расподеле је

$$F(x) = 1 - e^{-\lambda x}, \;\; x > 0.$$

Најчешће служи за моделирање трајања нечега. Њене нумеричке карактеристике су $EX = \frac{1}{\lambda}$ и $DX = \frac{1}{\lambda^2}$. Важно својство експоненцијалне расподеле, које је и карактерише у скупу апсолутно непрекидних случајних величина дефинисаних на $\mathbb{R}^+$ је *одсуство памћења*, односно

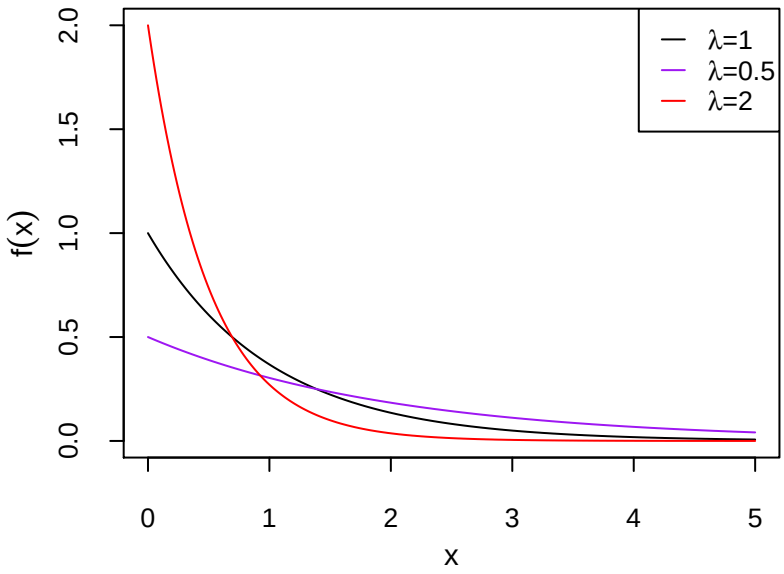
$$P\{X > s + t | X > s\} = P\{X > t\}, \; s, t > 0.$$

Ово својство је некада превише рестриктивно. Зато су предложена разна уопштења којима се, увођењем додатних параметара, овај недостатак превазилази. Једна од таквих расподела је и следећа. Поред наведеног својства може се показати и да $[X]$ има Геометријску расподелу па тако када се време мери у неким одређеним јединицама (на пример данима) уместо Ешпоненцијалне се може користити Геометријска расподела.

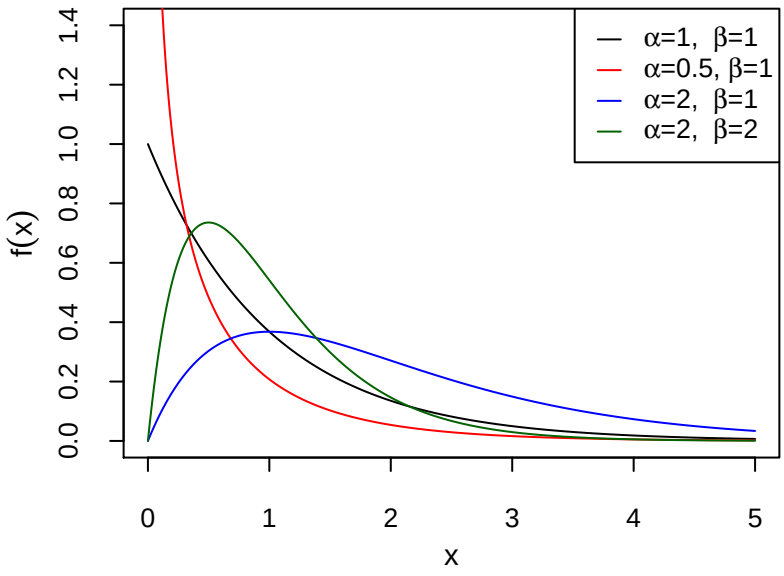
- Случајна величина X са $\gamma(\alpha, \beta)$ расподелом, $\alpha > 0$, $\beta > 0$, има функцију густине

$$f(x) = \frac{x^{\alpha-1}\beta^{\alpha}}{\Gamma(\alpha)} e^{-\beta x}, \;\; x > 0. \tag{6.12}$$

Уколико је $\alpha = 1$ ради се Експоненцијалном $\mathcal{E}(\beta)$. Уколико је $\alpha = \frac{n}{2}$ и $\beta = \frac{1}{2}$ ради се о χ^2_n расподели. Може се показати да је $EX = \frac{\alpha}{\beta}$ и $DX = \frac{\alpha}{\beta^2}$. Уколико је $\alpha = n \in \mathbb{N}$ случајна величина X се може представити као збир n независних случајних величина са Ешпоненцијалном $\mathcal{E}(\beta)$ расподелом. Важи и обрнуто. Збир n независних случајних величина са $\mathcal{E}(\beta)$ расподелом има $\Gamma(n, \beta)$ расподелу. На основу овога закључујемо да је једна од могућих ситуација када можемо искористити ову расподелу моделирање времена чекања док се не реализује унапред задат број догађаја. Флексибилност Гама расподеле можемо уочити и са графика 6.2.



Слика 6.4: Густина ешспоненцијалне $\mathcal{E}(\lambda)$



Слика 6.5: Густина Гама $\gamma(\alpha, \beta)$ раде

- Случајна величина X са Фишеровом F_{n_1,n_2} расподелом, $n_1, n_2 > 0$, има функцију густине

$$f(x) = \frac{\sqrt{\frac{(n_1x)^{n_1}n_2^{n_2}}{(n_1x+n_2)^{n_1+n_2}}}}{xB(\frac{n_1}{2}, \frac{n_2}{2})}, \ x > 0,$$

где је са $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ означена Бета функција. У практичним применама је важно да се X може представити у облику

$$X = \frac{\frac{Y_1}{n_1}}{\frac{Y_2}{n_2}}, \tag{6.13}$$

где су Y_1 и Y_2 независне случајне променљиве са $\chi^2_{n_1}$ и $\chi^2_{n_2}$ расподелама. Може се показати да је $EX = \frac{n_2}{n_2-2}$ (постоји за $n_2 > 2$) и $DX = \frac{2n_2^2(n_1+n_2-2)}{n_1(n_2-2)^2(n_2-4)}$ (постоји за $n_2 > 4$).

- Случајна величина X има Бета $\mathcal{B}(a, b)$ расподелу, $a, b > 0$ уколико је њена густина

$$f(x) = \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}, \ \ x \in [0,1], \ \ a > 0, \ b > 0$$

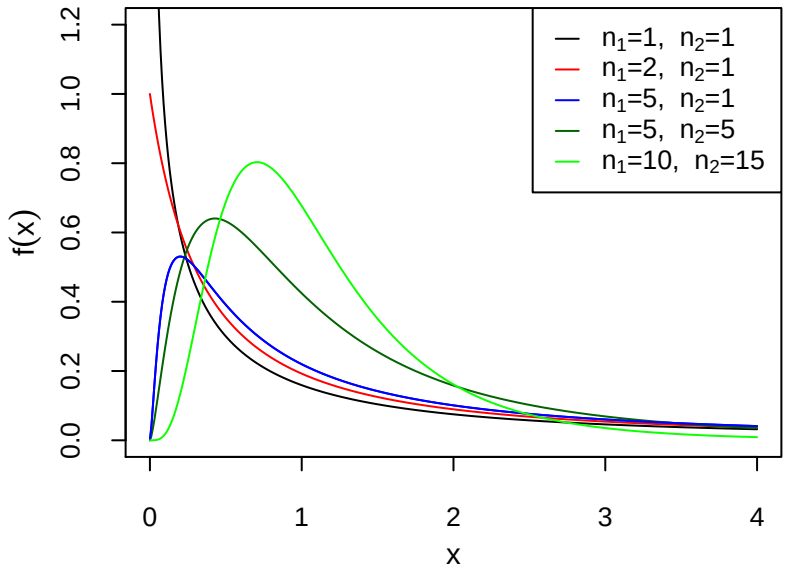
где је $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$. Тада је $E(X) = \frac{a}{a+b}$ и $D(X) = \frac{ab}{(a+b)^2(a+b+1)}$. Када је $a = 1$ и $b = 1$ ради се о униформној $\mathcal{U}[0,1]$ расподели. Ова расподела се због своје флексибилности често користи када је потребно искористи случајну величину дефинисану на $[0,1]$.

- Случајна величина X има Паретову $\mathcal{P}(\alpha, x_0)$ расподелу, $\alpha, x_0 > 0$, уколико је њена густина

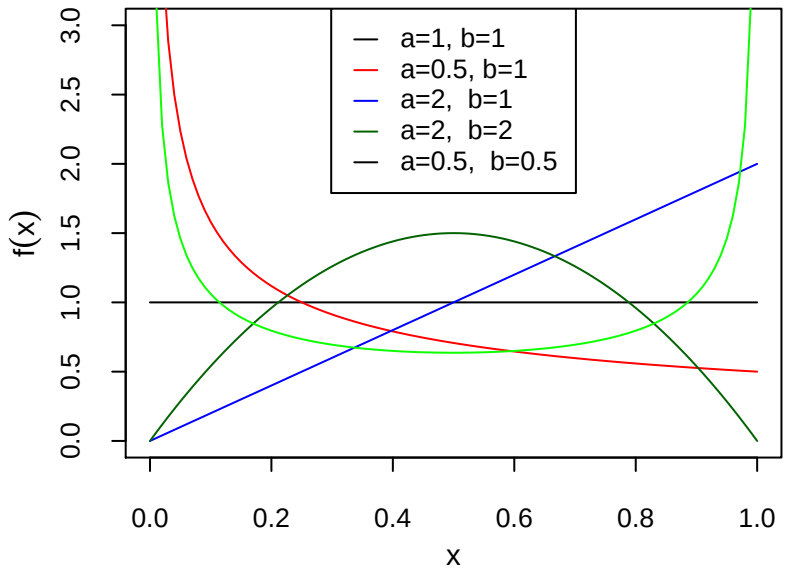
$$f(x) = \frac{\alpha x^\alpha}{x^{\alpha+1}}, \ x > x_0,$$

и функција расподеле

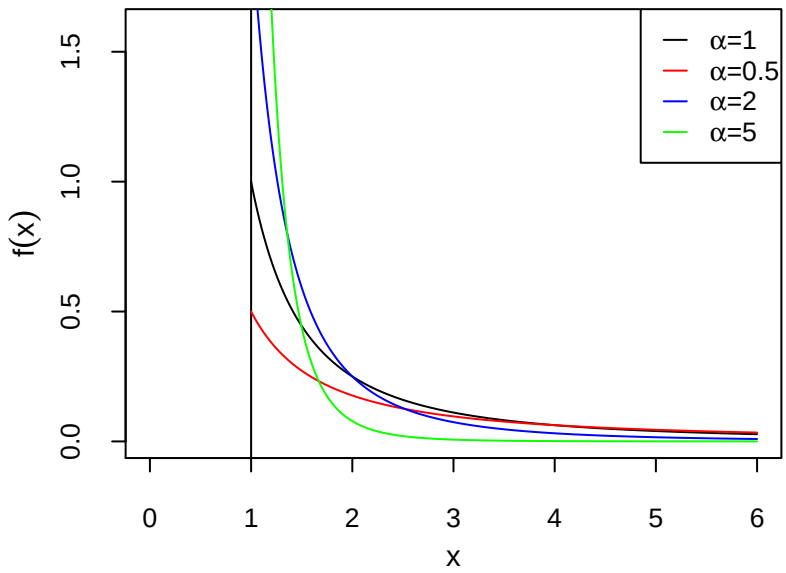
$$F(x) = 1 - \left(\frac{x_0}{x}\right)^\alpha.$$



Слика 6.6: Густина Фишерове F_{n_1,n_2} расподеле



Слика 6.7: Густина Бета $\mathcal{B}(a, b)$ расподеле



Слика 6.8: Густина Парето $\mathcal{P}(\alpha, 1)$ расподеле

Математичко очекивање и дисперзија су $EX = \frac{\alpha x_0}{\alpha - 1}$ (постоји за $\alpha > 1$) и $DX = \frac{x_0^2 \alpha}{(\alpha-1)^2(\alpha-2)}$ (постоји за $\alpha > 2$). Ова расподела се често користи у актуарству за моделовање величине одштета, у хидрологији за моделовање максималне количине падавина у току фиксног временског интервала, и у многим другим сличним ситуацијама. Параметар α је задужен за облик док x_0 има улогу параметра скалирања.

6.3 Важне расподеле у R-у

У програмском језеику R су за већину познатих расподела имплементиране функције које омогућавају генерисање простог

случајног узорка из те расподеле, као израчунавање вредности функције расподеле, закона расподеле/функције густине и инверза функције расподеле. Све функције су облика xdistribution где се уместо x користи слово r за генерисање узорка, p за функцију расподеле, d за функцију густине или закон расподеле, и q за инверз функције расподеле. Основни аргумент функције rdistribution је обим узорка, а осталих наведених вредност у којој се рачуна функција. Све функције за аргументе имају и параметре расподела. Називи одговарајућих функција представљених у овом поглављу приказани су у табели 6.1. Када се ради о дискретним случајним величинама, са $f(x)$ је означена $P\{X = x\}$. У случају Геметријске случајне величине имплементиране су функције за (6.4) па у случају да нам је потребно генерисање случајне величине (6.3) сваком елементу узорка треба додати 1, вредности осталих функција добијамо из једнакости $P\{X = x\} = P\{X' = x - 1\}$. У случају Негативне Биномне расподеле имплементирана је случајна величина са законом расподеле (6.7). У случају Паретове $\mathcal{P}(\alpha, x_0)$ функције се налазе у пакету actuar и треба обратити пажњу на то да се претпоставља да је случајна величина дефинисана на $(0, \infty)$ па генерисане бројеве треба "померити" за x_0, а остале функције адекватно модификовати.

Функција која нам омогућава генерисање дискретне случајне величине са унапред задатим законом расподеле је функција sample. Њени аргументи су величина узорка коју треба генерисати, скуп вредности, низ вероватноћа који представља закон расподеле, и replace који треба да има вредност $TRUE$. У случају да је његова вредност $FALSE$ онда се генерише узорак без понављања из задатког скупа вредности.

расподела	узорак	$F(x)$	$f(x)$	$F^{-1}(x)$
$\mathcal{B}(n,p)$	rbinom	pbinom	dbinom	qbinom
$\mathcal{G}(p)$	rgeom	pgeom	dgeom	qgeom
$\mathcal{P}(\lambda)$	rpois	ppois	dpois	qpois
$\mathcal{NB}(k,p)$	rnbinom	pnbinom	dnbinom	qnbinom
$\mathcal{N}(m,\sigma^2)$	rnorm	pnorm	dnorm	qnorm
χ^2_n	rchisq	pchisq	dchisq	qchisq
t_n	rt	pt	dt	qt
$\mathcal{E}(\lambda)$	rexp	pexp	dexp	qexp
$\gamma(\alpha,\beta)$	rgamma	pgamma	dgamma	qgamma
$\mathcal{F}_{n_1,n_2}$	rf	pf	df	qf
$\mathcal{U}[a,b]$	runif	punif	dunif	qunif
$\mathcal{B}(a,b)$	rbeta	pbeta	dbeta	qbeta
$\mathcal{P}(\alpha,x_0)$	rpareto	ppareto	dpareto	qpareto

Табела 6.1: Позиви функција у R-у

7

Додатак – Подаци

У овом додатку се налазе скупови података који су коришћени у уџбенику. Све ознаке у табелама одговарају ознакама коришћеним извршеним анализама у програмском језику R. База 7.3 из [24], док су остале базе симулиране за потребе курса, али с обзиром на то да су направљене по угледу на неке стварне базе података, би требало да добро осликавају расподеле посматраних података.

ID	typeC	typeR	crossR	ID	typeC	typeR	crossR
1	4	1	1	51	3	1	1
2	3	1	1	52	3	2	0
3	8	1	1	53	4	2	1
4	3	1	1	54	3	1	1
5	3	2	1	55	4	1	1
6	3	1	1	56	3	1	1
7	3	1	1	57	6	1	1
8	3	1	0	58	3	1	1
9	3	2	1	59	3	1	1
10	8	3	1	60	3	1	1
11	2	1	1	61	3	1	1
12	2	3	1	62	3	1	1
13	3	2	1	63	3	1	1
14	5	2	1	64	4	1	1
15	3	2	1	65	7	1	1
16	3	2	1	66	7	1	1
17	3	1	1	67	3	1	1
18	3	1	1	68	7	1	1
19	3	2	1	69	4	1	1
20	2	1	0	70	2	1	1
21	4	1	1	71	4	1	1
22	3	1	1	72	3	1	1
23	2	2	1	73	7	1	1
24	3	1	0	74	4	2	1
25	3	2	1	75	4	2	1
26	2	2	1	76	2	2	1
27	3	1	1	77	3	2	1
28	3	3	1	78	2	2	1
29	2	2	0	79	2	1	0
30	3	2	1	80	3	2	1
31	3	3	1	81	3	2	1
32	3	3	1	82	2	2	1
33	1	2	0	83	7	1	1
34	2	2	1	84	3	1	1
35	3	2	1	85	3	1	1
36	3	3	1	86	3	1	1
37	3	2	1	87	3	1	1
38	3	3	1	88	3	2	1
39	3	3	1	89	3	2	1
40	2	2	0	90	1	1	1
41	3	3	1	91	4	1	1
42	3	2	1	92	4	1	1
43	2	2	0	93	2	1	1
44	3	2	0	94	3	1	1
45	2	2	0	95	3	3	1
46	2	3	0	96	3	3	1
47	3	1	1	97	4	3	1
48	5	2	1	98	4	3	1
49	3	2	1	99	3	3	1
50	3	2	1	100	3	3	1

Табела 7.1: Подаци о саобраћајним несрећама - I део

ID	typeC	typeR	crossR	ID	typeC	typeR	crossR
101	5	2	1	151	4	2	1
102	3	2	1	152	8	2	1
103	8	2	1	153	7	3	0
104	3	2	1	154	2	3	0
105	2	2	1	155	4	3	0
106	3	2	1	156	7	2	0
107	3	2	1	157	7	2	0
108	3	2	1	158	4	2	0
109	3	2	1	159	7	2	0
110	3	2	1	160	3	2	0
111	3	2	1	161	1	2	0
112	3	2	1	162	7	2	0
113	3	2	1	163	2	2	0
114	3	2	1	164	5	2	0
115	3	2	0	165	2	2	0
116	4	2	0	166	4	2	0
117	5	2	1	167	4	2	0
118	2	2	1	168	8	2	0
119	7	2	1	169	7	2	0
120	4	2	1	170	7	2	0
121	4	2	1	171	4	2	0
122	2	2	1	172	7	2	0
123	2	2	1	173	7	2	0
124	7	2	1	174	7	2	0
125	4	2	1	175	7	2	0
126	3	2	1	176	7	2	0
127	3	2	1	177	8	2	0
128	2	2	1	178	7	2	0
129	8	2	1	179	2	2	0
130	1	2	1	180	7	2	0
131	7	2	1	181	7	2	0
132	4	2	1	182	4	2	0
133	7	2	1	183	8	2	0
134	3	2	1	184	7	2	0
135	2	2	1	185	7	2	0
136	3	3	1	186	8	2	0
137	2	3	1	187	4	2	1
138	3	3	1	188	4	2	1
139	4	3	1	189	1	2	1
140	3	3	1	190	1	2	1
141	4	3	1	191	2	2	1
142	2	3	1	192	2	2	1
143	8	3	1	193	4	2	1
144	4	2	1	194	1	2	1
145	7	2	1	195	8	2	1
146	2	2	1	196	4	2	1
147	2	2	1	197	3	2	1
148	7	2	1	198	2	2	1
149	4	2	1	199	4	2	1
150	7	2	1	200	3	2	1

Табела 7.2: Подаци о саобраћајним несрећама - II део

ID	DENSITY	SPEED
1	20.40	38.80
2	27.40	31.50
3	106.20	10.60
4	80.40	16.10
5	141.30	7.70
6	130.90	8.30
7	121.70	8.50
8	106.50	11.10
9	130.50	8.60
10	101.10	11.10
11	123.90	9.80
12	144.20	7.80
13	29.50	31.80
14	30.80	31.60
15	26.50	34.00
16	35.70	28.90
17	30.00	28.80
18	106.20	10.50
19	97.00	12.30
20	90.10	13.20
21	106.70	11.40
22	99.30	11.20
23	107.20	10.30
24	109.10	11.40

Табела 7.3: Број аутомобила у једној миљи и просечна брзина возила

ID	ukupanV	pol	ukupanS	ID	ukupanV	pol	ukupanS
1	70	1	80	31	37	0	46
2	67	1	77	32	69	1	83
3	45	1	61	33	87	1	94
4	59	1	66	34	73	1	85
5	75	1	84	35	45	1	57
6	77	1	78	36	44	1	63
7	48	0	60	37	77	1	87
8	63	1	81	38	38	1	48
9	41	0	42	39	64	0	78
10	65	0	68	40	58	1	69
11	90	0	106	41	90	0	104
12	84	1	94	42	56	1	72
13	66	0	79	43	55	1	62
14	88	1	99	44	85	1	93
15	83	1	95	45	53	0	55
16	72	1	81	46	69	1	79
17	53	0	67	47	74	0	77
18	66	0	70	48	65	0	68
19	87	1	102	49	58	0	61
20	79	1	85	50	82	1	90
21	59	0	63	51	63	1	80
22	31	1	36	52	64	0	76
23	58	1	70	53	95	0	115
24	32	0	44	54	100	0	120
25	47	1	54	55	79	0	86
26	63	0	80	56	84	1	100
27	58	1	69	57	54	1	69
28	54	0	62	58	80	1	87
29	68	1	85	59	58	1	68
30	65	0	79	60	75	1	82

Табела 7.4: Подаци о поенима на испиту

ID	visina	pogodak	ID	visina	pogodak
1	1.80	1	51	2.10	1
2	2.01	1	52	2.02	1
3	1.99	1	53	1.92	1
4	1.89	1	54	2.15	1
5	2.04	1	55	1.99	0
6	1.90	1	56	2.00	1
7	1.81	0	57	1.97	1
8	1.88	0	58	1.79	0
9	1.94	1	59	1.94	1
10	1.93	1	60	1.92	1
11	1.97	0	61	2.08	1
12	1.93	1	62	1.85	0
13	2.04	1	63	2.07	1
14	1.95	1	64	1.67	0
15	1.91	1	65	2.05	1
16	2.00	1	66	1.97	1
17	2.02	1	67	2.01	1
18	1.97	1	68	1.85	0
19	1.91	1	69	1.78	0
20	1.98	1	70	2.16	1
21	1.86	0	71	1.83	0
22	2.07	1	72	2.03	1
23	2.10	1	73	2.11	1
24	2.03	1	74	2.07	1
25	1.81	0	75	1.96	1
26	1.92	1	76	1.88	0
27	1.95	1	77	1.90	1
28	2.00	1	78	1.85	0
29	1.90	1	79	1.98	1
30	1.77	0	80	2.01	1
31	2.00	1	81	1.79	0
32	2.09	1	82	2.02	1
33	1.86	1	83	1.97	0
34	1.73	0	84	2.02	1
35	2.08	1	85	2.01	1
36	1.91	0	86	1.83	0
37	2.05	1	87	1.92	1
38	1.80	0	88	1.85	0
39	1.91	1	89	1.94	1
40	1.93	1	90	1.98	1
41	1.86	1	91	1.93	1
42	1.93	1	92	2.01	1
43	2.07	1	93	2.05	1
44	2.01	1	94	1.81	0
45	1.97	1	95	2.02	1
46	2.04	1	96	1.99	1
47	2.10	1	97	1.80	0
48	2.07	1	98	1.90	1
49	1.95	1	99	1.91	1
50	2.06	1	100	1.99	1

Табела 7.5: Подаци о кошевима људи различитих висина

education	sex	agree	disagree
0	Male	4	2
1	Male	2	0
2	Male	4	0
3	Male	6	3
4	Male	5	5
5	Male	13	7
6	Male	25	9
7	Male	27	15
8	Male	75	49
9	Male	29	29
10	Male	32	45
11	Male	36	59
12	Male	115	245
13	Male	31	70
14	Male	28	79
15	Male	9	23
16	Male	15	110
17	Male	3	29
18	Male	1	28
19	Male	2	13
20	Male	3	20
0	Female	4	2
1	Female	1	0
2	Female	0	0
3	Female	6	1
4	Female	10	0
5	Female	14	7
6	Female	17	5
7	Female	26	16
8	Female	91	36
9	Female	30	35
10	Female	55	67
11	Female	50	62
12	Female	190	403
13	Female	17	92
14	Female	18	81
15	Female	7	34
16	Female	13	115
17	Female	3	28
18	Female	0	21
19	Female	1	2
20	Female	2	4

Литература

[1] Alan Agresti. *Categorical data analysis*, volume 482. John Wiley & Sons, 2003.

[2] Vilijandas Bagdonavicius, Julius Kruopis, and Mikhail S. Nikulin. *Nonparametric tests for complete data*. John Wiley & Sons, 2013.

[3] Guorui Bian, Michael McAleer, and Wing-Keung Wong. A trinomial test for paired data when there are many ties. *Mathematics and Computers in Simulation*, 81(6):1153–1160, 2011.

[4] David R. Bickel and Rudolf Frühwirth. On a fast, robust estimator of the mode: Comparisons to other robust estimators with applications. *Computational Statistics & Data Analysis*, 50(12):3500–3530, 2006.

[5] Yosef Cohen and Jeremiah Y. Cohen. *Statistics and Data with R: An applied approach through examples*. John Wiley & Sons, 2008.

[6] Siddhartha R. Dalal, Edward B. Fowlkes, and Bruce Hoadley. Risk analysis of the space shuttle: pre-challenger prediction of failure. *Journal of the American Statistical Association*, 84(408):945–957, 1989.

[7] Julian J. Faraway. *Linear models with R*. CRC press, 2014.

[8] John Fox and Sanford Weisberg. *An R companion to applied regression, 2nd editon*. Sage Publications, 2019.

[9] Michael Friendly and David Meyer. *Discrete Data Analysis with R: Visualization and Modeling Techniques for Categorical and Count Data*. Boca Raton, FL: CRC Press., 2015.

[10] Francis Galton. Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15:246–263, 1886.

[11] Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian data analysis*. CRC press, 2013.

[12] Jayanta K. Ghosh, Mohan Delampady, and Tapas Samanta. *An introduction to Bayesian analysis: theory and methods*. Springer Science & Business Media, 2007.

[13] David W. Hosmer Jr, Stanley Lemeshow, and Rodney X. Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.

[14] Guido W. Imbens and Donald B. Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

[15] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

[16] John Norman Richard Jeffers. *An introduction to systems analysis: with ecological applications*. University Park Press, 1978.

[17] Vesna Jevremović. *Verovatnoća i statistika za smer Informatika*. Matemativ́cki fakultet, Beograd, 2014.

[18] Richard J. Larsen and Morris L. Marx. *An introduction to mathematical statistics*. Prentice Hall, 2005.

[19] Daniel McFadden. Quantitative methods for analyzing travel behaviour of individuals: Some recent developments (cowles foundation discussion papers no. 474). *Cowles Foundation for Research in Economics, Yale University*, 1977.

[20] Milan Merkle. *Verovatnoća i statistika: za inženjere i studente tehnike*. Akademska misao, 2010.

[21] Ljiljana Petrović. *Teorija uzoraka i planiranje eksperimenata*. Centar za izdavačku delatnost Ekonomskog fakulteta, 2007.

[22] John W. Pratt. Remarks on zeros and ties in the wilcoxon signed rank procedures. *Journal of the American Statistical Association*, 54(287):655–667, 1959.

[23] Sheldon M. Ross. *Introductory statistics*. Academic Press, 2017.

[24] Ashish Sen and Muni Srivastava. *Regression analysis: theory, methods, and applications*. Springer Science & Business Media, 2012.

[25] Olivier Thas. *Comparing distributions*. Springer, 2010.

[26] Larry Wasserman. *All of nonparametric statistics*. Springer Science & Business Media, 2006.