Увод

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 и циљева овог курса.

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 (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 читавој популацији, као и да одабир чланова узорка не зависи од вредности обележја тих чланова. 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. *Претпоставимо да желимо да видимо какво је мисљење 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, најприроднije 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 социоекономске карактеристике утичу на мишљење јавног мњења, 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 неи случајан избор од 1000 чланова популације.

Пример 2. *Претпоставимо да је циљ истраживања да се види какво је 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. Нерепрезентативни узорак би свакако био уозрак 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 (процењујемо) 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, вериска припадност;
  - ординално: разред, интензитет бола, статус студената (будзет, самофинансирајући);
- нумеричко (кватнитативно):

– дискретно: број деце, оцена на испиту, број искоришћених дана одмора...

– непрекидно: тежина, висина, време чекања у реду у банци....

У примеру 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 2 илуструје нумеричко обележје.

Познавање типова података је кључно за добру организацију базе података која се користи у анализи!

Основни кораци у статистичкој анализи

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 курса је управо да се науче методолошке основе сваког корака.

0.0.1 Осмишљавање експеримента

Много је боље у тренутку узимања података бити упозант са статистичким методама које ће се користити, него укључити статистичара тек након што се прикупе подаци. Тако се штитимо о потенцијалне замке да се на податке не могу применити неке методе.

0.0.2 Узорковање

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, узорка називамо *прост случајан узорак*. Различитим типовима узорковања бави се грана статистике *теорија узорка* Две основне врсте узорковања су *без враћања* и *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 це се неки чланови популације поновити. С друге стране, уколико се узорковање врши овако можемо сматрати да су чланови узорка независне и једнако расподљене случајне величине.

Ми ћмо претпостављати да је популација велика и да се ради о узорку са враћањем. Тада, ако је X посматрано обележје, са $X_1, X_2, ..., X_n$ ћемо означити прост случајан узорак (низ независних и једнако расподељених случајних величина). Јасно је да су то случајне величине јер ми унапред не знамо кое 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 током читавог курса.

0.0.3 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]

Пример 3. *Посматраћемо саобраћајне несреће које су се догодиле у Калифорнији у периоду од 2012. до 2016. године. Оно што нас посебно занима је да установимо шта то све утиче на ишод несреће. Узет*

[1] енгл. barplot
[2] енгл. пие чарт

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 означени су бројевима од 1 до 8. На пример, број 1 означава чеони судар, број 2 упоредну вожњу и тако даље. Тип пута је кодиран бројевима од 1 до 3 (једносмерни, двосмерни, и физички раздвојен), да ли се несрећа догодила у раскрсници је кодиран бројевима 0 и 1. 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*

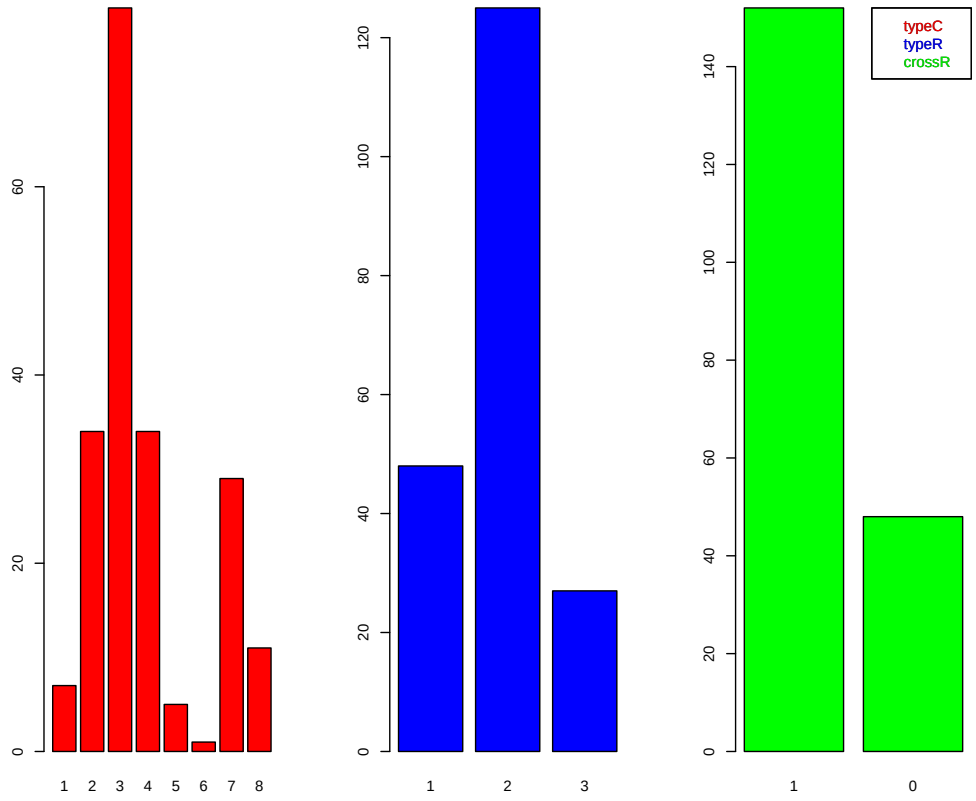
typeR	*1*	*2*	*3*
	48	*125*	*27*

crossR	*0*	*1*
	48	*152*

Иако, када се обележја приказују одвојено, се губи заједничка расподела, неки закључци се ипак могу донети. Видимо да је најче71а 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.

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

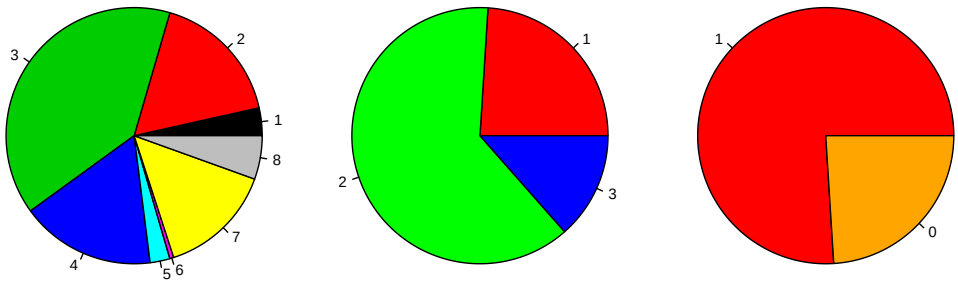
```
head(bazaCas)

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



Слика 1: Тракасти дијаграми за типове несрећа, типове пута, и присуство раскрснице

На цртежу 1 су приказани тракасти 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.



Слика 2: Кружни дијаграми

```
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')
```

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."

```
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'))
```

Заједничке расподеле (у паровима) посматраних обележја, приказане су у следећим табелама:

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*

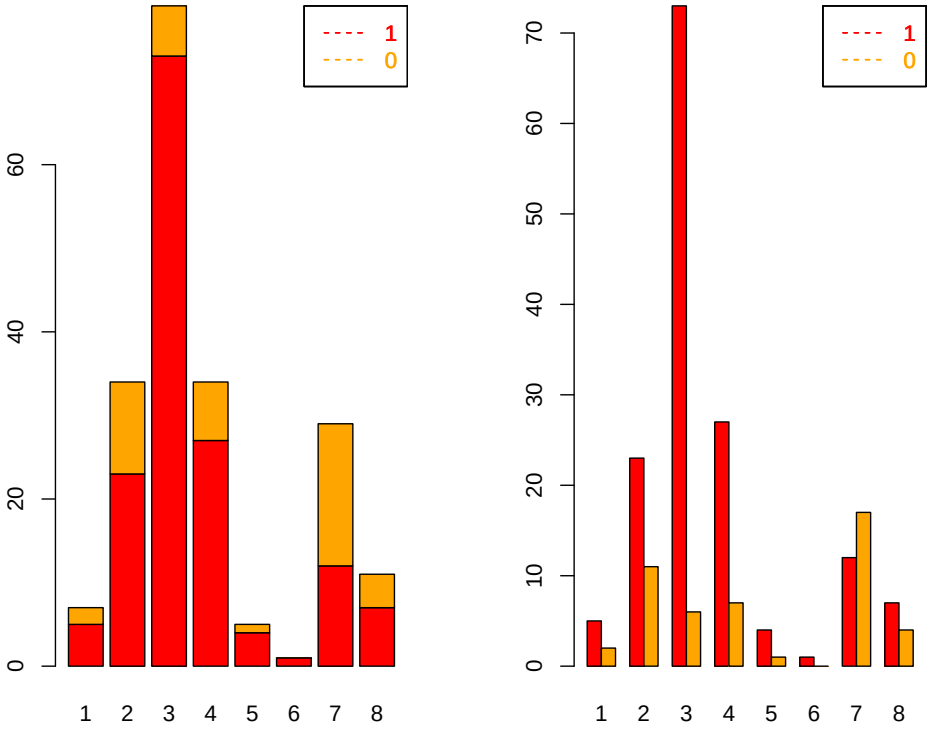
crossR \ typeC	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
1	*5*	*23*	*73*	*27*	*4*	*1*	*12*	*7*
0	*2*	*11*	*6*	*7*	*1*	*0*	*17*	*4*

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

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

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

```
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',
```



Слика 3: Тракасти дијаграми типова несрећа на раскрсницама и ван њих (лево-наслагани дијаграм, десно-груписани дијаграм)

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 расејање.

За прављење 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. Нека је препорука 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 где ће елемент припасти, за почетак првог

интервала не треба узети минималну вредност узорка већ се померити мало лево, као и да величине категорија буду на једну децималу више него што су дати подаци. Величину категорије одређујемо на основу *распона* узорка $R = x_{(n)} - x_{(1)}$. Са $x_{(k)}$ се означили k-ти по реду елемент сортираног узорка. Тако добијен низ се назива *варијациони низ*. Затим, уколико имамо k категорија, њихова приближна величина је $\frac{R}{k}$. Категорије не морају да буду једнаке величине али се тако обично ради. Хистограм управо представља графички приказ учесталости по категоријама. Уколико су на y-оси фреквенције говоримо о *хистограму апсолутних фреквенција*, уколико је приказа тај број подељен са величином узорка говоримо о *хистограму релативних фреквенција*, док ако је то још подељено и са величином категорија, говоримо о *хистограму густине*.

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

Са хистограма можемо да видимо које од познатих расподела долазе у обзир за моделирање посматраног обележја.

Пример 4. *Посматрајмо месечне плате 195 случајно изабраних просветних радника у Србији.*

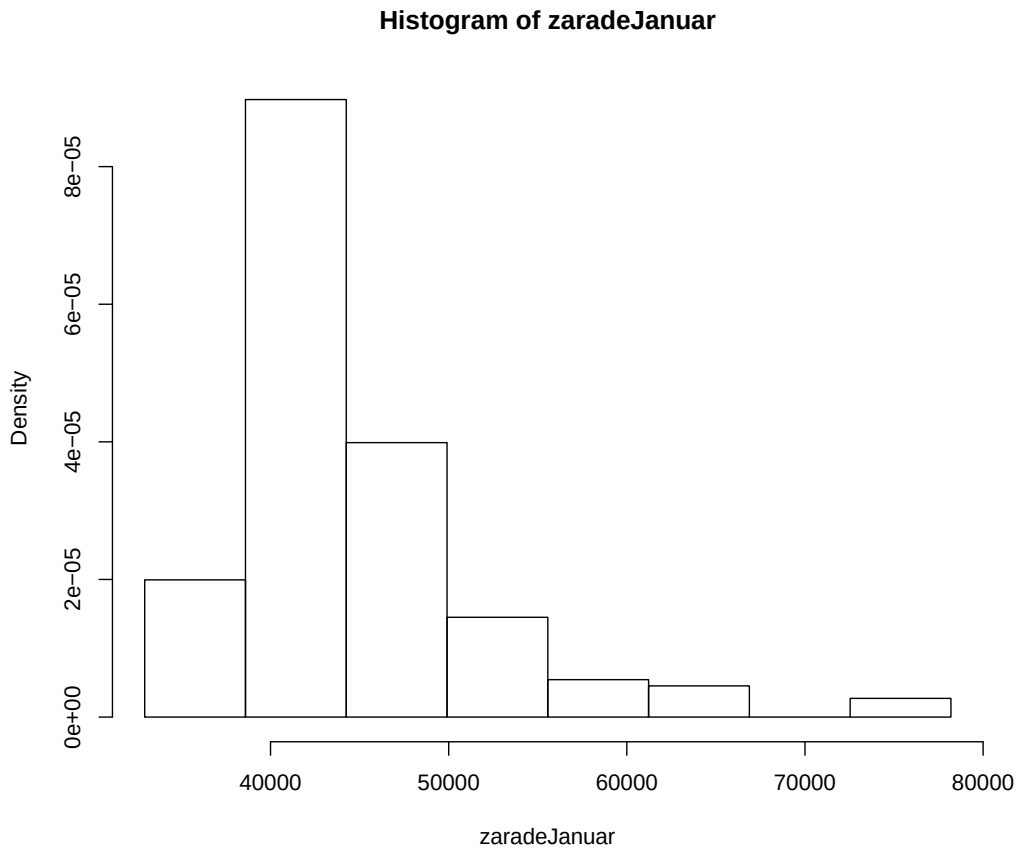
50048	56252	51988	44744	38930	41330	39694	41371	38331	42323	42578	38525	43416
54469	66816	40586	39641	45327	44024	46376	44667	38702	48222	40054	40650	52196
60933	42324	44349	45383	46939	43450	44787	43621	36581	40310	40113	45716	47261
60933	74314	41837	42914	44033	43673	61068	45366	43537	39958	38568	40114	41658
43518	59550	50722	39628	48517	37357	47186	37181	40477	40334	53538	45110	39270
63109	44603	41789	43056	44292	40704	40522	50125	37675	44743	51382	41426	39203
78193	47289	50370	44968	40205	48373	39594	43581	41332	49214	64458	35044	50190
43500	43131	38465	43207	42950	41016	46510	41208	39541	53625	42750	37447	47116
62943	42622	47819	43338	44142	42343	42392	44833	43281	46478	43275	37415	41890
55647	41396	41443	46666	44994	42732	39570	40720	38423	49064	44721	36619	50242
66056	41542	43371	41519	40008	39093	38323	40576	41221	42991	38208	32936	42329
42460	44661	42871	46092	43972	42476	38007	41179	40681	48148	43500	46861	41562
75423	48511	40274	40773	43222	48087	37114	40031	46525	39702	39964	38294	39594
50655	42635	54328	44677	46009	48106	40768	40841	45289	40265	38242	43899	38683
55463	41515	54867	43102	43803	47474	44532	41702	39376	37221	36184	41035	40910

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

Добијамо да је минимална плата у узорку 32 936 динара а максимална 78193. Дакле, $R = 45257$. *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]	(, 44250.3]	(, 49907.5]	(, 55564.7]	(, 61221.9]	(,66879.1]	(,72536.3]	(,78193.5]
22	99	44	16	6	5	0	3

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



Слика 4: Хистограм густине месечних зарада у просвети

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

На слици 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?*

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.

Са графика 4 видимо да је расподела зарада померена удесно.

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

- очекивана вредност;
- медијана;
- мода;

Природна оцена за очекивану (средњу вредност) је средња вредност елемената узорка која се назива *узорачка средина* и означава са $\bar{X} = \frac{X_1+\cdots+X_n}{n}$, односно њена реализована вредност $\bar{x} = \frac{x_1+\cdots+x_n}{n}$. Од сада нећемо нагалшавати да се ради о реализованој вредонсти неке статистике [3].

Медијана расподеле је онај параметар μ за који је истовремено

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

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

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

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 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 од расподела које сте учили имају јединствену моду?

Поред наведених параметара расподеле важне су и такозване мере расејања:

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

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

[4] има једну моду

На основу узорка распон 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 одсупање 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*

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

па се σ оцењује са $\bar{S}$.

Из разлога који ћемо убрзо видети, умести (2) се користи *поправљена узорачка дисперзија* дата са

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

На основу узорка одређујемо q_1 и q_3 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 подниза 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 5. *У случају посматраних зарада из претходног примера добија се*

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

Питање: На основу претходне анализе података шта закључујете о расподели зарада?

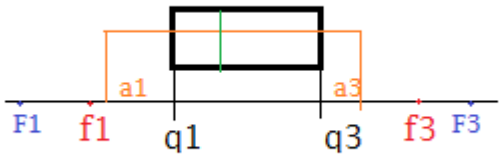
0.0.4 Идентификација аутлајера

Овај појам 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.

Један начин да се представе подаци је такозвани кутијасти дијаграм.[5] На слици 5 је приказан један овакав дијаграм. Ознаке на графику су следеће:

- q_1, q_3 су први и трећи 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 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.

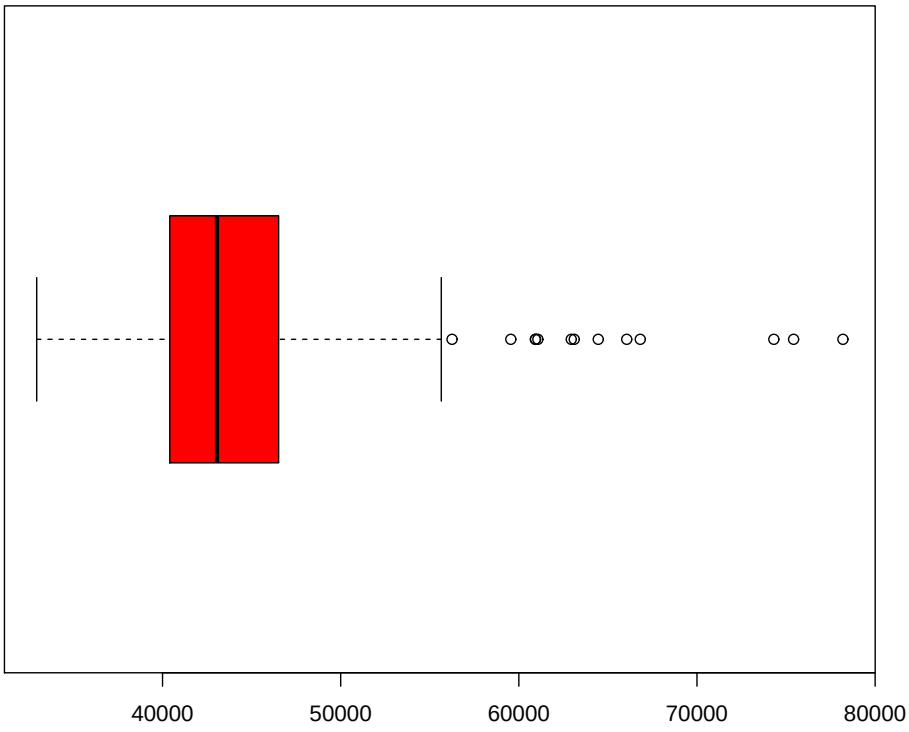
Пример 6. *Кутијасти дијаграм за зараде је приказан на слици 6.* $F_3 = 64853.5$, $f_3 = 55685.5$.

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

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

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

[5] енгл. бош плот



Слика 6: Бокс плот дијаграм

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.

0.1 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 $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Очекивана вредност те

оцене је

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

Последња једнакост важи јер претпостављамо да се ради о п.с.у. 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 кавдратно одступање те оцене од средње вредности видимо из

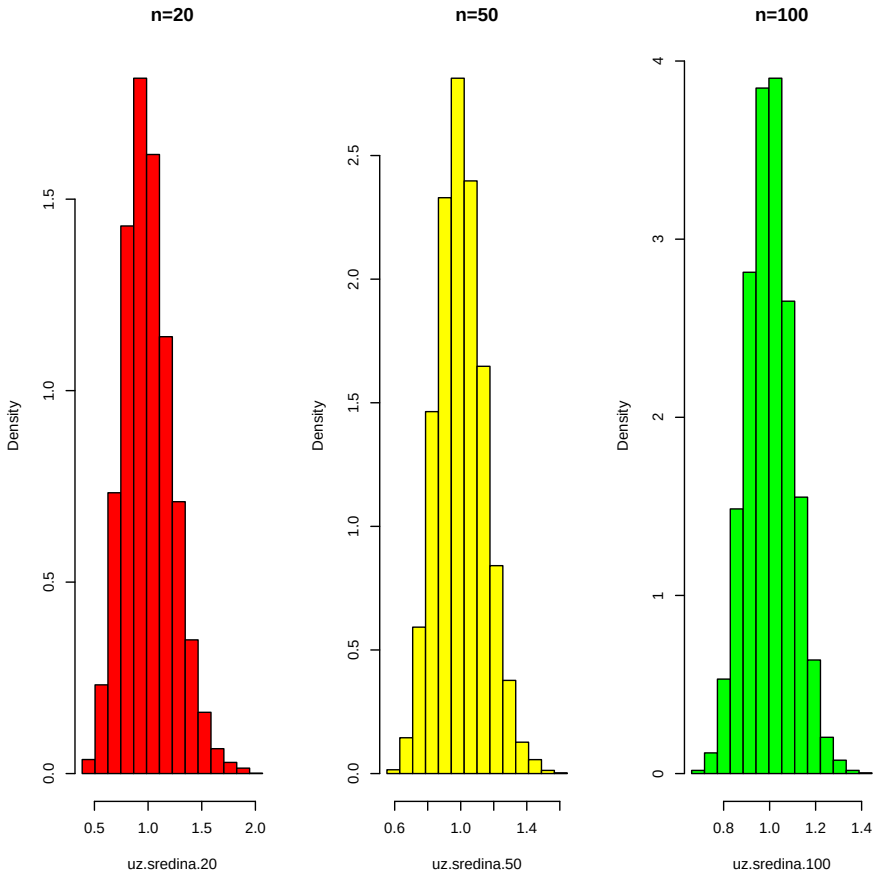
$$D(\bar{X}) = \frac{1}{n^2}\sum_{i=1}^{n} D(X_i) = \frac{D(X)}{n}.$$

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 7. *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 овиме узорка, приказани су на слици 7.*

Јасно се уочава да се са порастом обима узорка смањује одступање оцене од стварне вредности 1. Примећујемо да оцене имају нормалну расподелу. Објашњење тога лежи у Централној граничној теореми.

```
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,
```



Слика 7: Хистограм оцена средње вредности

```
main='n=50')
hist(uz.sredina.100,breaks=0.661+0.056*0:14,col='green',prob=TRUE,
main='n=100')
```

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

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

Увођењем фактора корекције добијамо да је очекивана вредност поправљене оцене баш једнака дисперзији обележја, односно параметру који се оцењује.

Задатак: Одредити $D(n\bar{S}^2)$, и нацртајте хистограме оцене за различите обиме узорка. Како се мења прецизност оцене са порастом обима узорка?

0.2 Емпиријска функција расподеле

Нека је $X_1, X_2, ..., X_n$ п.с.у. из популације на којој посматрамо обележје X са функцијом расподеле F. 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 па има Биномну $\mathcal{B}(n, F(x))$ расподелу (јер је $p = P\{X \leq x\} = F(x)$), 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 тако се и емпиријска функција расподеле "приближава" правој функцији расподеле F. Ово запажање садржано је у следећој теореми која је позната још и као централа теорема статистике.

Теорема 0.2.1 (Гливенко-Кантелијева теорема)**.** *Нека је $X_1, X_2, ..., X_n$ п.с.у. из популације са обележјем X са функцијом расподеле $F(x)$. Даље, нека је $F_n(x)$ одговарајућа емпиријска функција расподеле. Тада*

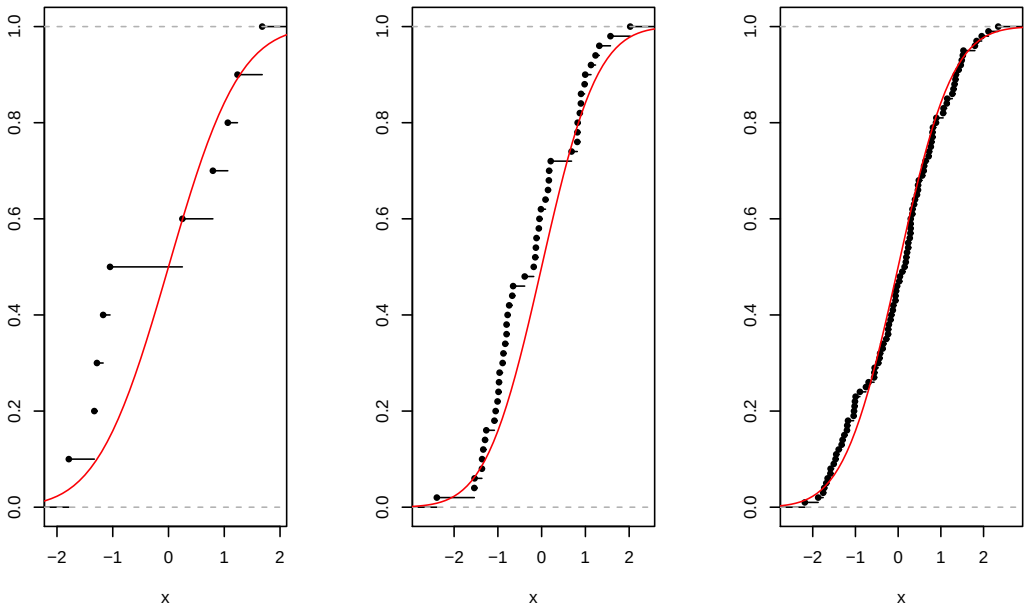
$$P\{\sup_x |F_n(x) - F(x)| \to 0, \text{ кад } n \to \infty\} = 1.$$

На графику 8 приказане су емпиријске функције расподеле узорка из обележја са нормалном $\mathcal{N}(0, 1)$ расподелом. Црвеном линијом приказана је одговарајућа функција расподеле.

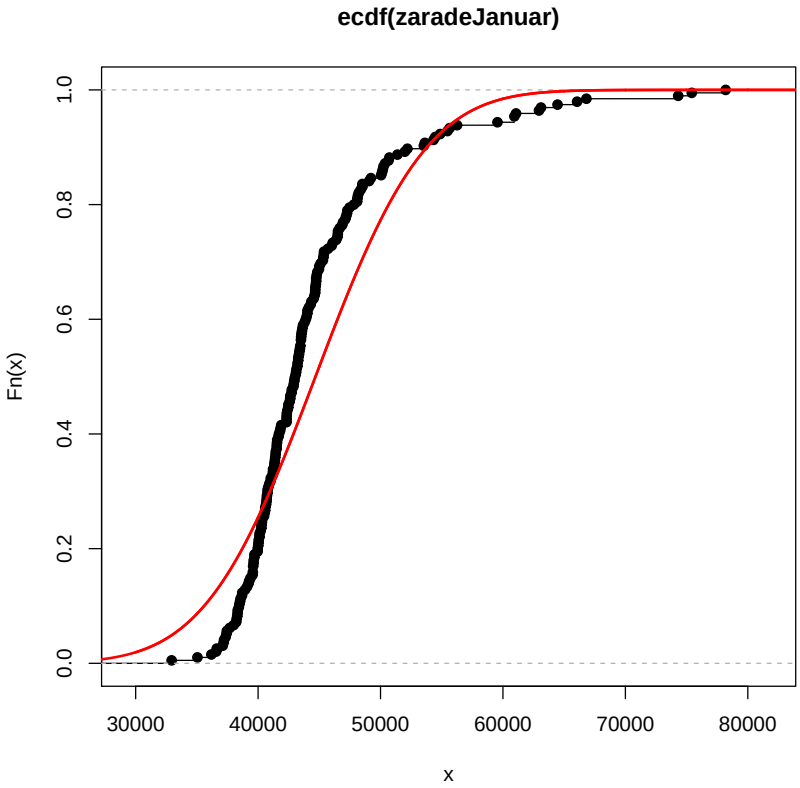
Емпиријска функција расподеле је доста погодна за "имитирање" расподеле посматраног обележја, али из њеног графичког приказа се не закључује о расподели лако као са хистограма, па се у ту сврху најчешће не користи. С друге стране уколико имамо већ неку претпоставку о расподели, лакше ће одбацити исту на основу емпиријске функције расподеле него на основу хистограма.

Пример 8. *Посматрајмо податке из примера 4. Желимо да видимо да ли има смисла претпоставити да је расподела нормална. Параметре расподеле проценићемо са узорачком средином и узорачком дисперзијом. На истом графику (види 9) приказаћемо емпиријску функцију расподеле и претпостављену функцију расподеле.*

```
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)
```



Слика 8: Емпиријска функција расподеле на основу узорка обима $n = 10$ (лево), $n = 50$ (у средини) и $n = 100$ (десно)


ecdf(zaradeJanuar)
Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
30000
40000
50000
60000
70000
80000
x

Слика 9:

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

Напомена: Емпиријска функција расподеле један је од основних појмова у непараметарској статистици[6].

[6]параметарски приступ подразумева да знамо расподелу обележја до на непозна параметар

Поглавље 1

Оцењивање непознатих параметара расподела

До сада смо се упознали са неким непараметарским оценама (хистограм и емпиријска функција расподеле као примери функционалних оцена, оценама параметара централне тенденције и расејања). Тада нисмо водили рачуна из које је расподеле обележје X. Сада претпостављамо да је статистички модел одређен до на непознате параметре расподеле. То значи да већ имамо претпоставку о расподели обележја и остаје нам само да оценимо параметре те расподеле.

Нека је статистички модел одређен до на непознат параметар θ за који знамо да припада скупу Θ, при чему 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. Приказаћемо два основна метода за оцењивање.

1.0.1 Метод момената

Оцене непознатих параметара се добијају као решење система једначина који се добије кад се изједначе теоријски моменти са одговарајућим узорачким моментима (в. табелу 1.1). У свим наредним примерима претпостављамо да на располагању имамо п.с.у. $X_1, X_2, ..., X_n$.

Пример 9. *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,*

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

Табела 1.1: Теоријски и одговарајући узорачки моменти

и то, имајући у виду шта представљају непознати параметри за нормалну расподелу, очекивање и дисперзију. Тако добијамо

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

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

Пример 10. *Нека* X *има експоненцијалну* $\mathcal{E}(\lambda)$ *расподелу*[1]. *Сада нам је потребна само једна једначина па опет узимамо најједнставнију.*

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

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

Оцену смо могли да добијемо и нпр. из једначине

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

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 и расподелу обележја исправна.*

Пример 11. *Нека* X *има* $\mathcal{U}[0, \theta]$ *расподелу. Оцену добијамо из једначине*

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

Одавде је $\hat{\theta} = 2\bar{X}$.

[1] функција густине је $f(x;\lambda) = \lambda e^{-\lambda x}$.

Пример 12. *Нека је X индикатор са вероватноћом успеха p. Сада је*

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

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

Пример 13. *Нека X има Пуасонову $\mathcal{P}(\lambda)$ расподелу. Сада је*

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

па је $\hat{\lambda} = \bar{X}$.

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 14. *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 $\hat{\lambda} = \frac{\log 2}{m_e}$.

1.0.2 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.

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

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

Веома често је лакше маскимизирати неку монотону трансформацију функције веродостојности. Најчешће се максимизира $\log L(\theta)$.

Пример 15. *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

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

Пример 16. *Нека X има експоненцијалну $\mathcal{E}(\lambda)$ расподелу.*

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

Како је $\frac{\partial^2 l(\lambda)}{\partial \lambda^2}|_{\lambda=\hat{\lambda}} < 0$ *па* $\hat{\lambda}$ *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*

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

Оцена добијена овом методом не мора бити јединствена. То можемо видети у наредном примеру.

Пример 17. *Нека* X *има униформну* $U[\theta-1, \theta+1]$ *расподелу. Функција веродостојности је*

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

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

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 18. *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 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* $\hat{\theta} = X_{(n)}$*.*

Пример 19. *Оцена максималне веродостојности за дисперзију индикатора је* $\hat{p}(1-\hat{p})$*, где је* $\hat{p} = \bar{X}$ *оцена максималне веродостојности за* p*.*

Задатак: Нека X има Биномну $\mathcal{B}(N, p)$ расподелу, при чему је N познато и $p > 0.5$. Одредити оцену за p методом максималне веродостојности.

1.1 Особине оцена

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 п.с.у. $X_1, X_2, ..., X_n$.

Дефиниција 1.1.1. *Уколико је $E(\hat{\theta}_n) = \theta$ за оцену $\hat{\theta}_n$ кажемо да је непристрасна оцена параметра θ.*

Дефиниција 1.1.2. *Уколико је $\lim_{n\to\infty} E(\hat{\theta}_n) = \theta$ оцена $\hat{\theta}_n$ је асимптотски непристрасна оцена параметра θ.*

Дефиниција 1.1.3. *Уколико је $\hat{\theta}_n \xrightarrow{P} \theta,\ n \to \infty$, оцена $\hat{\theta}_n$ је постојана оцена параметра θ.*

Својство 1.1.3 нам заправо каже да са за довољно велики узорак можемо наћи у произвољно малој околини стварне вредности параметра θ са великом вероватноћом, односно да

$$\forall \varepsilon > 0,\ P\{|\hat{\theta}_n - \theta| > \varepsilon\} \to 0.$$

Довољан услов да ово важи је да

$$E(\hat{\theta}_n - \theta)^2 \to 0, \quad n \to \infty. \tag{1.1}$$

У случају да је оцена непристрасна услов 1.2 је еквивалентан са

$$D(\hat{\theta}_n) \to 0, \quad n \to \infty.$$

У претходном поглављу видели смо два метода за добијање оцена (који нису једини познати). Самим тим се намеће питање коју од добијених оцена одабрати. Да бисмо на то одговорили потребно је да уведемо неку меру квалитета оцена. То можемо на много начина. Једна могућност је да посматрамо средње квадратно одсупање оцене од праве вредности параметра.

Дефиниција 1.1.4. *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*

$$E(\hat{\theta}_n - \theta)^2 < E(\hat{\hat{\theta}}_n - \theta). \tag{1.2}$$

Врло често, када није могуће наћи расподелу оцена, квалитет испитује Монте Карло методом (оцехкивања која је потребно одредити се оцењују на основу великог броја понављања експеримента у коме се оцењује θ). Алгоритам којим бисмо добили оцене је следећи:

1. Генеришемо узорак $\mathbf{x}$ обима n из расподеле $F(\theta)$;
2. На основу $\mathbf{x}$ одредимо $\hat{\theta}_n(\mathbf{x})$;
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 $(\hat{\theta} - \theta)^2$ оцењујемо са

$$\frac{1}{N}\sum_{i=1}^{N}((\hat{\theta}_n^{(i)} - \theta)^2). \tag{1.3}$$

Оцена средње квадратног одступања 1.3 има све "лепе" особине узорачке средине.

Пример 20. *У примеру 9 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 $E(\hat{m}) = m$ закључујемо да је $\hat{m}$ 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 21. *Испитајмо непристраснос оцене $\hat{\lambda}$ из примера 16, за $n > 3$.*

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

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 $\frac{1}{EY}$.

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

Одавде је

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

Оцена није непристрасна али јесте асимптотски непристрасна.

За испитивање постојаности одредићемо

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

Даље је

$$E(\hat{\lambda}^2) = n^2 E\Big(\frac{1}{Y^2}\Big) = n^2 \int_0^\infty \frac{x^{n-3}\lambda^n e^{-\lambda x}}{\Gamma(n)} dx$$
$$n^2 \int_0^\infty \frac{x^{n-3}\lambda^{n-2} e^{-\lambda x}}{\Gamma(n-2)} dx \cdot \frac{\lambda^2\Gamma(n-2)}{\Gamma(n)} = 1 \cdot \frac{\lambda^2 n^2}{(n-1)(n-2)},$$

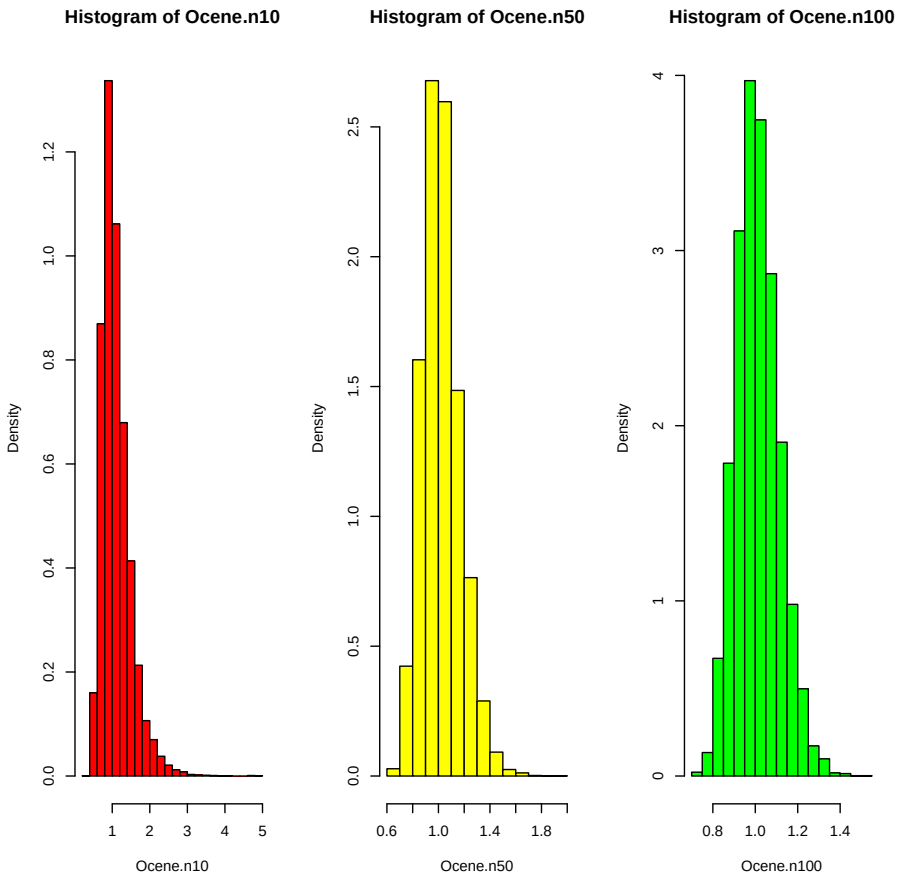
па је

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

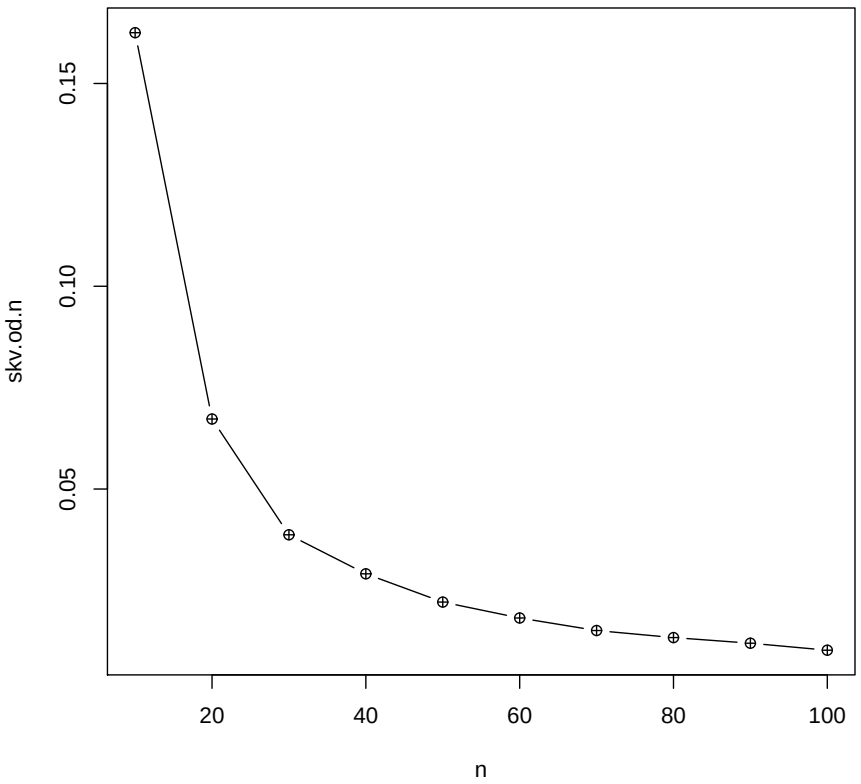
Одавде закључујемо да је $\hat{\lambda}$ постојана оцена за λ.

У случају да се нисмо присетили да Y има $\gamma(n,\lambda)$ расподелу, о истој бисмо могли ипак нешто закључити примењујући кораке 1-3 алгоритма 1.1. На слици 1.1 су приказани хистограми густина оцена које се добијају када је права вредност параметра $\lambda = 1$. Ту можемо да уочимо својство асмптотске непристрасности и постојаности. Како n расте вредности које се добијају су "све ближе" правој вредности λ. То се још боље види на слици 1.2

```
OceneExpLambda<-function(n,lambda,N)
{
ocene=rep(0,N)
for(i in 1: N)
{
uzorak=rexp(n,lambda)
ocene[i]=1/mean(uzorak)

    }
```



Слика 1.1: Хистограм оцена за λ



Слика 1.2: Средње квадратно одступање $\hat{\lambda}$ од λ за узорке различитих обима

```
return(ocene)

    }

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

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)


    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 22. *Нека X има $\mathcal{U}[0,\theta]$ расподелу. У примерима 11 и 18 смо*

видели да су оцене добијене методом момената и методом маскималне веродостоности редом

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

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.

$$E(\hat{\theta} - \theta) = D(\hat{\theta}) = D(\bar{X}_n) = \frac{\theta^2}{12n} \tag{1.4}$$

Да бисмо одредили друго средњеквадратно одступање потребно је да нађемо расподелу за $X_{(n)}$. Имамо да је

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

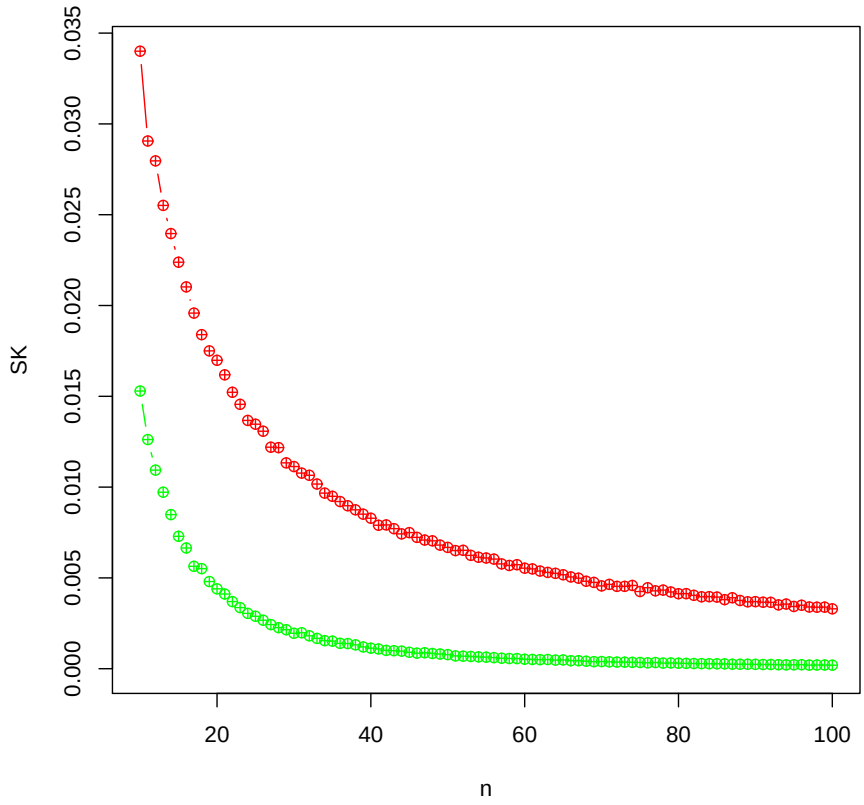
Одавде је

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

Одавде је

$$\begin{aligned} E(\hat{\hat{\theta}} - \theta)^2 &= EX_{(n)}^2 - 2\theta EX_{(n)} + \theta^2 \\ &= \frac{n\theta^2}{n+2} - 2\frac{n\theta^2}{n+1} + \theta^2 = \frac{\theta^2}{(n+2)(n+1)}. \end{aligned} \tag{1.5}$$

Из (1.4) *и* (1.5) *видимо да је оцена добијена методом максималне веродостојности боља у средње квадратном смислу. То се види и на слици 1.3.*



Слика 1.3: Средње квадратна одступања $\hat{\theta}$ (црвено) и $\hat{\hat{\theta}}$ (зелено) од θ за узорке различитих обима

```
n=10:100
N=10000
SK1=c()
SK2=c()
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,tlab='SK')
lines(n,SK2,type='b',col='green',pch=10)
```

Задатак: 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 средњекваратног одступања оцена од стварне вредности параметра $\theta = 0.2$, од обима узорка. Урадити то за обе посматране оцене. Шта закључујете са тог графика?

1.2 Интервалне оцене параметара

Нека је θ непознат параметер. Нека су L_n и U_n статистике за које је $P\{L_n \leq \theta \leq U_n\} = \beta$. Интервал (L_n, U_n) се назива *β% двострани интервал поверења за параметар θ*, а β је *ниво поверења*. Аналогно се дефинишу једнострани доњи и једнстрани 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 β% 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 (стожерну величину). За параметре неких расподела већ постоји устаљена процедура које функције од узорка треба користити и коју расподелу имају. Илустроваћемо неке од њих.

1.2.1 Закључивање у моделу са нормалном расподелом

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

Теорема 1.2.1. *Нека је $X_1, X_2, ..., X_n$ н.с.у. из $\mathcal{N}(m, \sigma^2)$ расподеле. Тада*

- $\frac{\bar{X}_n - m}{\frac{\sigma}{\sqrt{n}}}$ *има $\mathcal{N}(0,1)$ расподелу;*
- $\frac{(n-1)\widetilde{S}_n^2}{\sigma^2} = \frac{n\bar{S}_n^2}{\sigma^2}$ *има χ^2_{n-1} расподелу;*
- $\frac{\sqrt{n}(\bar{X}_n - m)}{\widetilde{S}_n}$ *има t_{n-1} расподелу.*

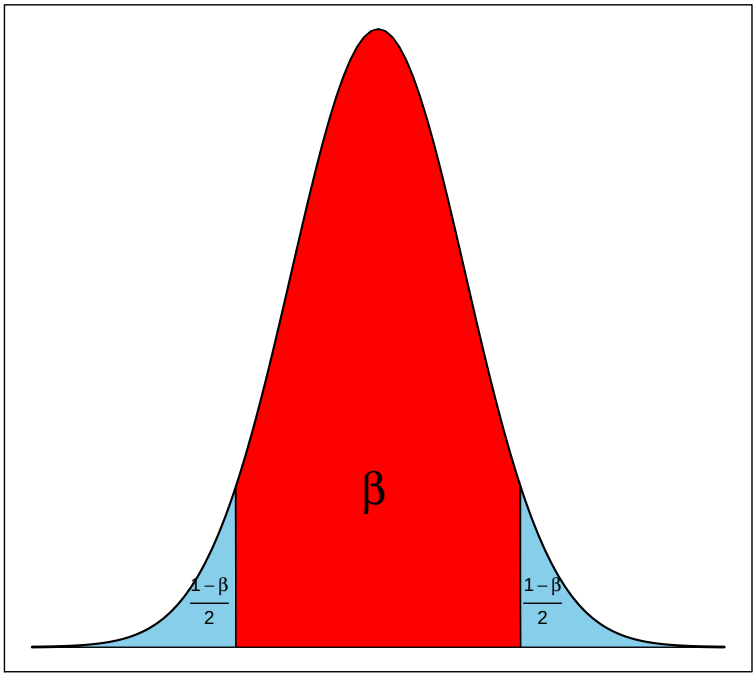
Сада можемо одредити β% интервал поверења за поменуте параметре.

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 1.2.1, је

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

за коју знамо да има $\mathcal{N}(0,1)$ расподелу. Зато можемо одредити константу C тако да је $P\{|T| \leq C\} = \beta$ (види слику 1.4). Због симетричности нормалне расподеле $C = F^{-1}(\frac{1+\beta}{2})$. Даље, неједнакост $|T| \leq C$ је еквивалентна са $\bar{X}_n - C\frac{\sigma}{\sqrt{n}} \leq m \leq \bar{X}_n + C\frac{\sigma}{\sqrt{n}}$. Одавде видимо шта су статистике L_n и U_n које смо тражили.



Слика 1.4: Конструкција $\beta\%$ интервала поверења

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

Пример 23. *Овим примером ћемо илустровати суштину интервала поверења како не би долазило до грешака у његовој интерпретацији. Генерисаћемо узорке из* $\mathcal{N}(0,1)$ *(стварна вредност параметра* $m = 0$*) и правићемо* 95% *интервале поверења за сваки од 10000 узорака. Добили смо да ју у 9515 случајева интервал поверења садржао стварну вредност параметра* m*.*

```
interval.poverenja.m1 <- function(n,m,sigma,beta) {
xsr=mean(rnorm(n,m, sigma))
L=xsr - qnorm((1+beta)/2)/ sqrt(n)
U =xsr + qnorm((1+beta)/2)/ 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
```

Интервал поверења за m када је σ^2 непознато

Сада се за стожерну величину може узети

$$T = \frac{\bar{X}_n - m}{\frac{\widetilde{S}_n}{\sqrt{n}}},$$

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

Пример 24. *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* $\bar{X} = 14.5$ *и* $s = 3.5$. *На основу неких претходних истраживања је установљено да се може сматрати да цена књиге има нормалну расподелу, па можемо применити резултат из овог поглавља да нађемо одговарајући интервал поверења.*

Прво одређујемо C. Имамо да је $C = C = F_{t_{24}}^{-1}(\frac{1+0.9}{2}) = 1.64$, па је доња граница интервала $14.5 - \frac{3.5 \cdot 1.64}{\sqrt{25}} = 13.352$. Слично се добија да је горња граница 13.352. Сада издавачка кућа може да искористи ту информацију за формирање цене.

Интервал поверења за σ^2

Једна од могућности за стожерну величину у овом случају је $T = \frac{(n-1)\widetilde{S}^2}{\sigma^2}$ која има χ^2_{n-1} расподелу.

Као што знамо, ова расподела није симетрична, иако се за велико n може апроксимирати нормалном расподелом, па не 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 одредићемо 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 $C_1 = F^{-1}(\frac{1-\beta}{2})$ и $C_2 = F^{-1}(\frac{1+\beta}{2})$. Неједнакост $C_1 \leq T \leq C_2$ је еквивалентна са

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

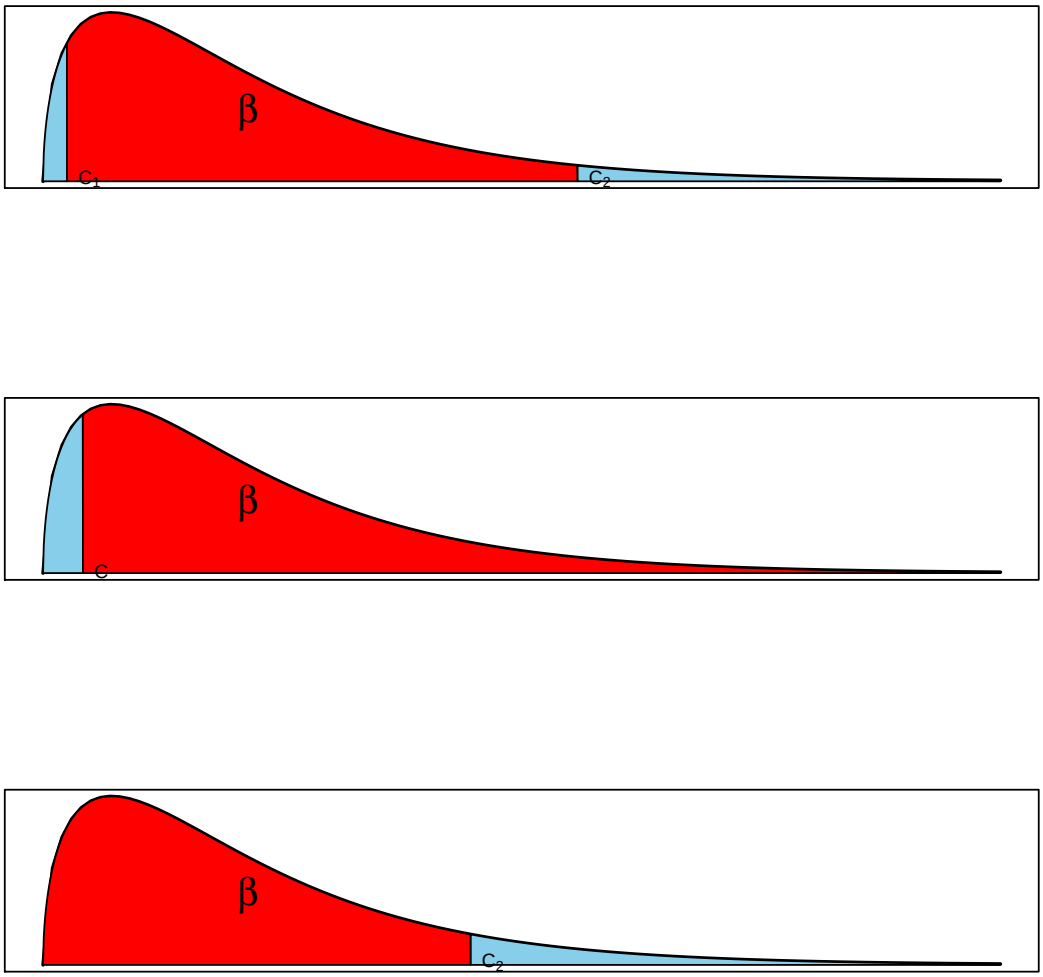
односно добили смо двострани интервал поверења за σ^2. Често нам је битно да нађемо горњу, односно доњу границу за σ^2, тј. да нађемо једностране интервале поверења $[L_n, \infty]$ или $(0, U_n)$. За конструкцију $[L_n, \infty)$ потребно је да одредимо C тако да је $P\{T < C\} = \beta$, док за $(0, U_n)$ је потребно да одредимо C тако да је $P\{T > C\} = \beta$.

Илустрација како се одређују константе C неопходне за конструкцију интервала, приказана је на слици 1.5.

1.2.2 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 $\mathcal{N}(m_1, \sigma_1^2)$ и $\mathcal{N}(m_2, \sigma_2^2)$. 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:

Теорема 1.2.2. *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:*



Слика 1.5: Конструкција $\beta\%$ интервала поверења

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

$$\nu = \frac{\left(\frac{\tilde{S}_{n_1}^2}{n_1}+\frac{\tilde{S}_{n_2}^2}{n_2}\right)^2}{\frac{\left(\frac{\tilde{S}_{n_1}^2}{n_1}\right)^2}{n_1-1}+\frac{\left(\frac{\tilde{S}_{n_2}^2}{n_2}\right)^2}{n_2-1}}. \tag{1.6}$$

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

Сада, уколико су нам потребни двострани интервали поверења за разлику очекивања, одређујемо C тако да је $P\{|T| < C\} = \beta$, при чему користимо резултате из теореме 1.2.2 у зависности од тога да ли су нам дисперзије обележја познате или не, и ако нису, да ли знамо да су једнаке или не. 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 $Q = \frac{\frac{\tilde{S}_{n_1}^2}{\sigma_1^2}}{\frac{\tilde{S}_{n_2}^2}{\sigma_2^2}}$ за коју знамо да има F_{n_1-1,n_2-1}. 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 $P\{Q < C_1\} = \frac{1-\beta}{2}$ и $P\{Q > C_2\} = \frac{1-\beta}{2}$, па се интервал поверења добија из интервала $C_1 < Q < C_2$.

Пример 25. *Желимо да нађемо* 95% *интервал поверења за разлику прошечне 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*
$\tilde{S}$ mg	*5*	*6*

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

Користећи (2.9) *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* $(80-77)+2.08\cdot\sqrt{\frac{5^2}{15}+\frac{6^2}{12}}=7.49$.

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

Претпостављамо да је X индикатор и да је вероватноћа успеха p. Као и до сада, треба да нађемо неку функцију чију расподелу знамо. Најчешће се користи

$$T=\frac{\bar{X}-p}{\frac{p(1-p)}{n}}.$$

За велико n гранична расподела од T 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 $P\{|T|\leq C\}=\beta$. Приметимо да је неједнакост $|T|<C$ еквивалента са $T^2<C$, односно

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

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

$$p^2\left(\frac{C^2}{n}+1\right)-p\left(\frac{C^2}{n}+2\bar{X}_n\right)+\bar{X}_n^2\leq 0. \tag{1.7}$$

Неједнакост (1.7) 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

$$T=\frac{\bar{X}_n-p}{\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}} \tag{1.8}$$

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

$$\left(\bar{X}_n - C\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}, \bar{X}_n + C\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}\right). \tag{1.9}$$

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 $\hat{p}$. Имамамо 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 узмемо (под претпоставком да можемо) прво узорак неке, унапред одабране, величине, нпр. 20, и да на основу тог узорка оценимо $\hat{p}$ а онда у скалду са тим, одредимо n тако да је $d < C$, и онда извучемо нови узорак величине n на основу кога ћемо одредити тражени интервал поверења.

Пример 26. *Компанија за производњу играчака жели да пласира нови производ на тржиште. Направљен је пилот пројекат у коме је направљено 10000 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:*

- $\hat{p} = \frac{800}{1000} = 0.8$*, и* $n(1-\hat{p}) = 200$ *што је заиста велики број, па може да се користи* (1.8)*;*
- $C = \Phi^{-1}\left(\frac{1+0.99}{2}\right) = 2.58$;
- *на основу* (1.9) *интервал поверења је* $(0.8 - 2.58\sqrt{\frac{0.2\cdot 0.8}{1000}}, 0.8 + 2.58\sqrt{\frac{0.2\cdot 0.8}{1000}}) = (0.76, 0.83)$.

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 из које добијамо да

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

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

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

Као и у случају Биномног модела, основа за заклучивање у Пуасоновом моделу је Централна гранична теорема. Уколико су $X_1, ..., X_n$ независне и једнако расподељене случајне променљиве са $\mathcal{P}(\lambda)$ онда

$$T = \frac{\bar{X}_n - \lambda}{\sqrt{\frac{\lambda}{n}}} \tag{1.10}$$

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

$$\beta = P\{|T| \leq C\} = P\{T^2 \leq C^2\} = P\{(\bar{X}_n - \lambda)^2 \leq C^2 \frac{\lambda}{n}\} =$$
$$= P\{\lambda^2 - \lambda(\frac{C^2}{n} + 2\bar{X}_n) + \bar{X}_n^2 \leq 0\}.$$

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

$$T = \frac{\bar{X}_n - \lambda}{\sqrt{\frac{\bar{X}_n}{n}}}, \tag{1.11}$$

и тада је интервал поверења

$$\left(\bar{X}_n - C\sqrt{\frac{\bar{X}_n}{n}}, \bar{X}_n + C\sqrt{\frac{\bar{X}_n}{n}}\right). \tag{1.12}$$

1.2.5 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

$$T = \frac{\bar{X}_{n_1} - \bar{Y}_{n_2} - (\lambda_1 - \lambda_2)}{\sqrt{\frac{\bar{X}_{n_1}}{n_1} + \frac{\bar{Y}_{n_2}}{n_2}}}, \tag{1.13}$$

има нормалну $\mathcal{N}(0,1)$ расподелу, па интервал поверења за $\lambda_1 - \lambda_2$ добијамо из услова $P\{|T| < C\} = \beta$.

1.2.6 Интервал поврења 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 фигурише разлика $\bar{X}_n - m$.

На основу Централне граничне теореме знамо да ако је $DX = \sigma^2 < \infty$ онда, за велико n

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

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

Поглавље 2

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).

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 вахзно да "знамо шта хоћемо" и то правилно искажемо;

- 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.

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 једном добијем негативан резултат тестирања, док да бисмо потврдили хипотезу потребно је то урадити са доста

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

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

Табела 2.1: Резултат статистичког тестирања

тестова. Тако да се тестирање заправо врши да би се одбацила нулта хипотеза у корист прихватања алтернативне хипотезе.

Пример 27. *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 закључка могуће је направити грешке. То је илустровано у табели 2.1.

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 назива ниво значајности теста и означава са α. Вероватноћа грешке друге врсте се означава са β. $1-\beta$ представља моћ теста. Тест је моћнији уколико боље одбацује нетачне хипотезе.

Мера теста је α за које је $\sup_{H_0} P\{\text{грешка } I \text{ врсте}\} = \alpha$.

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.

Још треба напоменути да хипотезе могу бити просте и сложене. Просте су оне за које је скуп допустивих вредности параметара једночлан.

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 критична област. Јасно је да је за то потребна расподела тест статистике под нултом хипотезом. Међутим то није увек једноставно

одредити, а некада чак није ни могуће. Зато се често расподела оцењује Монте Карло методама, слично као кад смо испитивали квалитет оцена. Генерише се N узорака из расподеле која је одређена нултом хипотезом и за сваки од њих одреди вредност тест статистике. На тај начин добијамо низ вредности тест статистике $T_n^{(1)}, ..., T_n^{(N)}$ на основу ког се може оценити расподела статистике T_n, а након тога и одредити критична област. На пример, ако је критична област $W = \{T > C\}$ и ако знамо да је $P_{H_0}\{T > C\} = \alpha$, 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 уколико нам је то потребно. Међутим, како је моћ теста $P_{H_1}\{T \in W\} = E_{H_1}(I\{T \in W\})$, 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 $p < \alpha$ онда одбацујемо хипотезу, у супротном је прихватамо. Овај број је стандардни излаз за већину тестова у разним алатима тако да се статистичко закључивање у великом броју случаја на основу њега дешава. У наредним поглављима на конретним примерима ћемо илустровати како га мозехмо израчунати.

Постоји много начина да класификујемо статистичке тестове. Најприроднији је по томе да ли је расподела узорка позната до на непознат параметар или не (параметарски и непараметарски тестови). 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.

Сваку од поменутих типова тестова ћемо детаљније описати.

2.1 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.

2.1.1 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. Дакле, потребно је тестирати $H_0 : m = m_0$ против неке алтернативе и 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 и 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 се упадне у критичну област већа од α, и да са што већом вероватноћом одбаци нетачна нулта хипотеза (има велику моћ), као и да моћ достиже вредност 1 за довољно велики обим узорка.

Сада ћемо приказати тест статистике које се најчешће користе у овом случају.

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

$$T = \frac{\bar{X}_n - m_0}{\frac{\sigma}{\sqrt{n}}},$$

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

$$T = \frac{\bar{X}_n - m_0}{\frac{\tilde{S}_n}{\sqrt{n}}}, \tag{2.1}$$

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

Најчешће три алтернативе су:

- $$H_1 : \ m \neq m_0; \tag{2.2}$$

- $$H_1 : \ m < m_0; \tag{2.3}$$

- $$H_1 : \ m > m_0. \tag{2.4}$$

Ово нам је јако битно за одређивање критичне области.

$\bar{X}_n$ је оцена за m тако да ако H_0 није тачно T_n неће бити "довољно блиско нули". У случају (2.2) природно је да је критична област облика $W = \{|T_n| > C\}$ јер је расподела тест статистике под нултом хипотезом симетрична, а много мале и много велике вредности тест статистике упућују на алтернативну хипотезу. Константу C одређујемо из услова $P_{H_0}\{|T_n| > C\} = \alpha$. 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 (2.3) $W = \{T_n < C\}$ па је $C = \Phi^{-1}(\alpha)$, док је за алтернативну хипотезу (2.4) $W = \{T_n > C\}$ и $C = \Phi^{-1}(1 - \alpha)$.

Напомена: Приметимо да се критична област може добити инвертовањем интервала поверења за ниво $1 - \alpha$. Свакако важи и обрнут закључак, да се интервал поверења може добити инвертовањем критичне области. На пример, двостани $(1 - \alpha)\%$ 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

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

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

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

$$
\begin{aligned}
M(\theta) &= P_\theta\left\{\Big|\frac{\bar{X}_n - m_0}{\frac{\sigma}{\sqrt{n}}}\Big| > C\right\} = 1 - P_\theta\left\{\Big|\frac{\bar{X}_n - m_0}{\frac{\sigma}{n}}\Big| \le C\right\} \\
&= 1 - P_\theta\left\{-C \le \frac{\bar{X}_n - m_0}{\frac{\sigma}{\sqrt{n}}} \le C\right\} \\
&= 1 - P_\theta\left\{m_0 - \frac{C\sigma}{\sqrt{n}} \le \bar{X}_n \le m_0 + \frac{C\sigma}{\sqrt{n}}\right\} \\
&= 1 - P_\theta\left\{\frac{m_0 - \frac{C\sigma}{n} - \theta}{\frac{\sigma}{\sqrt{n}}} \le \frac{\bar{X}_n - \theta}{\frac{\sigma}{\sqrt{n}}} \le \frac{m_0 + \frac{C\sigma}{\sqrt{n}} - \theta}{\frac{\sigma}{\sqrt{n}}}\right\} \\
&= 1 - P_\theta\left\{-C + \frac{m_0 - \theta}{\frac{\sigma}{\sqrt{n}}} \le \frac{\bar{X}_n - \theta}{\frac{\sigma}{\sqrt{n}}} \le C + \frac{m_0 - \theta}{\frac{\sigma}{\sqrt{n}}}\right\} \\
&= 1 - \Phi\Big(C + \frac{m_0 - \theta}{\frac{\sigma}{\sqrt{n}}}\Big) + \Phi\Big(-C + \frac{m_0 - \theta}{\frac{\sigma}{\sqrt{n}}}\Big). \qquad (2.5)
\end{aligned}
$$

Уколико је $\theta = m_0$ онда је $M(\theta) = \alpha$ што и треба да важи јер "смо тада у нултој хипотези". Уколико је $\theta > m_0$ онда једнакост (2.5) 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 (2.5) можемо написати у облику

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

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

Из једнакости (2.5) 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 2.1).

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

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

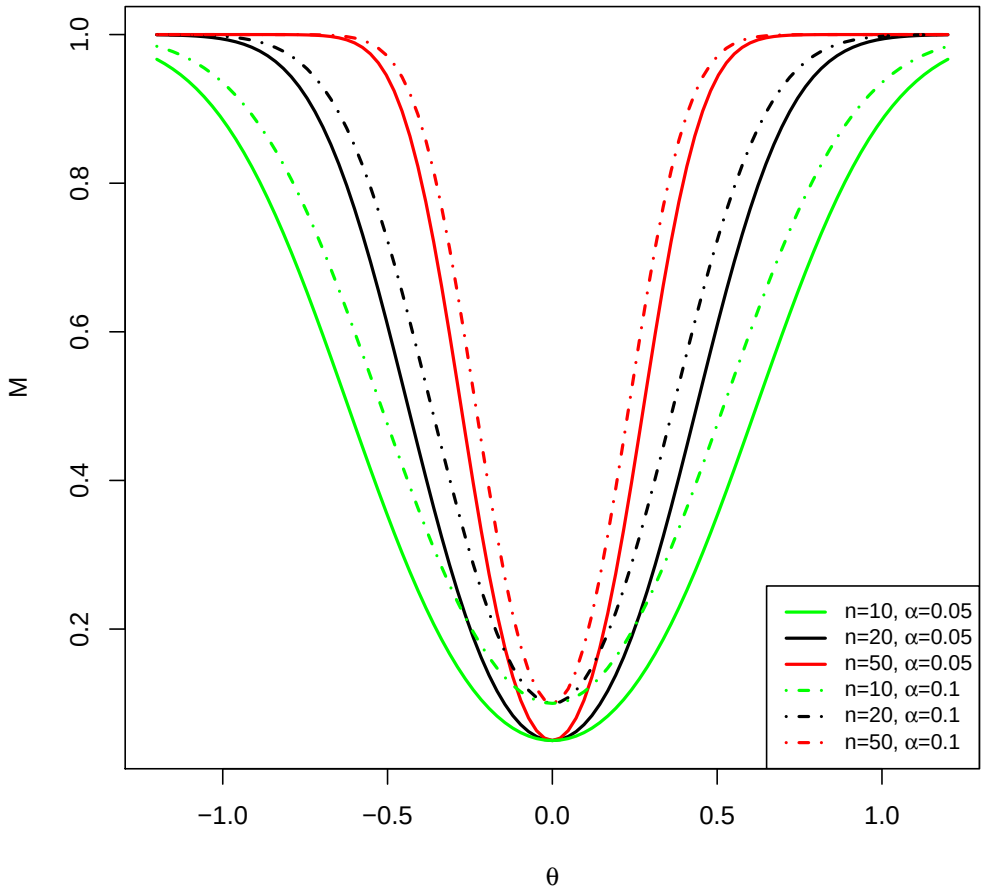
за коју знамо да, уколико је хипотеза тачна, има χ^2_{n-1} расподелу.

Најчешће алтернативе су

- $$H_1 : \ \sigma^2 \neq \sigma_0^2; \tag{2.6}$$
- $$H_1 : \ \sigma^2 < \sigma_0^2; \tag{2.7}$$
- $$H_1 : \ \sigma^2 > \sigma_0^2. \tag{2.8}$$

Посмтрајмо алтернативу (2.6). 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 $W = \{T < C_1\} \cup \{T > C_2\}$. Константе C_1 и C_2 ћемо бирати тако да је $\frac{\alpha}{2} = P_{H_0}\{T < C_1\} = P_{H_0}\{T > C_2\}$. 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 (2.7), или (2.8), критичне области су редом облика $\{T < C\}$, односно $\{T > C\}$.



Слика 2.1: Моћ двостраног теста за $H_0 : m = 0$, кад је $\sigma = 1$ познато

Пример 28. *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* $\{T > C\}$. *при чему је* $c = F^{-1}_{\chi^2_9}(0.95) = 16.92$. *Реализована вредност*

статистике је $\hat{T} = \frac{9 \cdot 1.2^2}{0.9^2} = 16$*, па не упада у критичну област. Закључак је да не одбацујемо тврђу произвођача.*

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

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 $\frac{\sigma_1^2}{\sigma_2^2}$. За прву поменуту статистику разликујемо три случаја:

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

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

   Уколико је нулта хипотеза тачна T 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

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

   где је $S^2 = \frac{(n_1-1)\widetilde{S}_{n_1}^2 + (n_2-1)\widetilde{S}_{n_2}^2}{n_1+n_2-2}$, за коју знамо 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 статистику

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

која, ако је H_0 рачна, има t_ν расподелу, где је

$$\nu = \frac{\left(\frac{\tilde{S}^2_{n_1}}{n_1} + \frac{\tilde{S}^2_{n_2}}{n_2}\right)^2}{\frac{\left(\frac{\tilde{S}^2_{n_1}}{n_1}\right)^2}{n_1-1} + \frac{\left(\frac{\tilde{S}^2_{n_2}}{n_2}\right)^2}{n_2-1}}. \tag{2.9}$$

У сва три случају, ако је $H_1 : m_1 - m_2 \neq m_0$ критична област је $W = \{|T| > C\}$, ако је $H_1 : m_1 - m_2 < m_0$ критична област је $W = \{T < C\}$, а ако је $H_1 : m_1 - m_2 > m_0$ онда је критична област $W = \{T > C\}$. Константе C се одређују у складу са расподелама уколико важи H_0.

Да бисмо одлучили да ли да користимо статистику за случај кад су дисперзије једнаке или за случај кад су различите треба да тестирамо $H_0 : \ \sigma_1^2 = \sigma_2^2$. При чему, грешка друге врсте у овом тестирању је 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 $H_0 : \ \frac{\sigma_1^2}{\sigma_2^2} = 1$ користимо статистику

$$T = \frac{\tilde{S}^2_{n_1}}{\tilde{S}^2_{n_2}},$$

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 $W = \{T < C_1\} \cup \{T > C_2\}$ а константе C_1 и C_2 се одређују из услова $P_{H_0}\{T < C_1\} = P\{T > C_2\} = \frac{\alpha}{2}$. Уколико је $H_1 : \ \sigma_1^2 > \sigma_2^2$ онда је $W = \{T > C\}$, а ако је $H_1 : \ \sigma_1^2 < \sigma_2^2$ онда је $W = \{T < C\}$.

Спарени тест

Могуће је да се деси да посматрана обележја нису независна,и да имамо п.с. узорак парова $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$. Желимо да тестирамо $H_0 : m_1 - m - 2 = 0$ против неке од стандардних алтернатива. Тада можемо направити помоћни узорак $D_1, ..., D_n$, где је $D_i = X_i - Y_i$ и свести случај на тест на основу једног узорка дат изразом (2.1), при чему оцену за σ рачунамо на основу помоћног узорка (јер је σ^2 дисперзија обележја $D = X - Y$).

Тада можемо направти помоћни узорак $D_1 = X_1 -$

2.1.2 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

- $$H_1 \ : p \neq p_0; \tag{2.10}$$
- $$H_1 \ : p < p_0; \tag{2.11}$$
- $$H_1 \ : p > p_0. \tag{2.12}$$

Уколико имамо велики узорак можемо искористити статистику

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

и чињеницу да, уколико је H_0 тачна T има нормалну $\mathcal{N}(0,1)$. Имајући у виду да је $\bar{X}_n$ тачкаста оцена за p, у случају алтернативне хипотезе (2.10) критична област је облика $W = \{|T| > C\}$, у случају алтернативне хипотезе (2.11) $W = \{T < C\}$, док је у случају алтернативне хипотезе (2.12).

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 (2.10) критна област ће бити облика $W = \{T \leq C_1\} \cup \{T \geq C_2\}$ 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.

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

$$T = \frac{\bar{X}_{n_1} - \bar{Y}_{n_2} - p_0}{\sqrt{\frac{\bar{X}_{n_1}(1-\bar{X}_{n_1})}{n_1} + \frac{\bar{Y}_{n_2}(1-\bar{Y}_{n_2})}{n_2}}} \tag{2.13}$$

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

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

У принципу, у овом специјалном случају би требало ову оцену за p_1 на основу обједињеног узорка, али нећемо много погрешити и ако користимо (2.13).

Када је $H_0 : p_1 - p_2 = p_0$ може се наћи адекватнија оцена за p_1 и p_2 тако да овај услов буде задовољен, али о томе нећемо говорити на овом курсу.

Пример 29. *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 испитанику из центра града њих 47 је пушило, док је од 153 испитаника са периферије, њих 52 пушило. Какав је закључак на основу одговрајућег 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 $2P\{T > 0.626\}) = 0.533$ *па не одбацујемо* H_0*, односно закључујемо да нема разлика у центру града и на периферији. Да смо користили статистику* (2.13) *добили бисмо да је* $\hat{T} = 0.624861$*, што се незнатно разликује од претходног и резултат би свакако био исти.*

2.1.3 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

$$T = \frac{\bar{X}_n - \lambda_0}{\sqrt{\frac{\lambda_0}{n}}},$$

која, уколико је нулта хипотеза тачна, има $\mathcal{N}(0,1)$ расподелу, а крртицичне области формирамо у зависности алтернативе. За $H_1 : \lambda \neq \lambda_0$ критична област је облика $W = \{|T| > C\}$,у случају $H_1 : \lambda < \lambda_0$ имамо да је $W = \{T < C\}$, док је у случају $H_1 : \lambda > \lambda_0$ критична област облика $W = \{T > C\}$.

Када је узорак мали можемо искористити статистику $S_n = X_1 + \cdots + X_n$ за коју знамо да, уколико је нулта хипотеза тачна, има Пуасонову $\mathcal{P}(n\lambda_0)$ расподелу.

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 $H_0 : \lambda_1 - \lambda_2 = \lambda_0$. Уколико имамо велике узорке можемо користити статистику

$$T = \frac{\bar{X}_{n_1} - \bar{Y}_{n_2} - \lambda_0}{\sqrt{\frac{\bar{X}_{n_1}}{n_1} + \frac{\bar{Y}_{n_2}}{n_2}}}.$$

2.2 Непараметарски тестови

2.2.1 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

$$T = \sum_{i=1}^{n} I\{X_i > m_{e0}\}. \tag{2.14}$$

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 је приближно $\frac{n}{2}$.

Можемо користити и "центрирану" верзију

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

Уколико је H_0 тачна T има $\mathcal{B}(n, \frac{1}{2})$. За велико n се може користити нормална апроксимација, односно да статистика

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

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

Критичну област формирамо као и до сада, у зависности од алтернативне хипотезе. Нпр. уколико је $H_1 : m_e \neq m_{e0}$ критична област за тестирање $W = \{|T^*| \geq C\}$ или $W = \{|T^c| \geq C\}$.

Приметимо, да ако је расподела обележја X симетрична онда се H_0 своди на H_0 $m = m_0$, где је $m = EX$.

Пример 30. *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 оценом од 1 до 5. Добијене су следеће оцене: 2,5,3,4,1,4,5. Да ли на основу ових резултата, са нивоом значајности* 0.05*, мозехмо 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 $T = \sum_{i=1}^{7} I\{X_i > 3\} = 4$. *Критична област је облика* $W = \{|T^c| \geq C\}$ *па је p-вредност* $P\{T\} =$ 0.45 *и одбацујемо* H_0.

Напомена: Овај тест се често користи и за тестирање симетрије око нуле.

2.2.2 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 су расподеле за X и Y симетричне, и 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.

Пример 31. *У старачком дому група испитаника је била изложена новом леку за који се сматра да успорава деменцију. Мерена је вредност једног мождданог параметра пре 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 $P\{X > Y\} > 1/2$ на

ће критична област бити облика $\{T \geq C\}$ *где је* T *број случајева када је* $X_i > Y_i$*, односно* $T = \sum_{i=1}^{11} I\{X_I > Y_i\}$*. Из студије видимо да је било укупно 9 таквих случајева, односно* $\hat{T} = 9$*. P- вредност теста је* $P\{T \geq 9\} = 0.03$*, при чему смо користили да, уколико је* H_0 *тачна,* T *има Биномну* $\mathcal{B}(11, 0.5)$ *расподелу.*

Да смо користили нормалну апроксимацију добили бисмо да је $\hat{T}^* = 2.11$*, и да је p-вредност 0.02.*

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

2.2.3 Вилкоксонов тест заснован на ранговима и знаковима

Претпостављамо да је расподела обележја X симетрична и желимо да тестирамо $H_0 : m = m_0$.

Уколико је нулта хипотеза тачна $X - m_0$ има исту расподелу као $m_0 - X$.

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

$$T = \sum_{i=1}^{n} r_i I\{X_i - m_0 \geq 0\}$$

може се показати да је $ET = \frac{n(n+1)}{4}$ и да је $DT = \frac{n(n+1)(2n+1)}{24}$. Егзактна расподела под нултом хипотезом се може наћи и за њу постоје таблице. Уколико је $n > 12$, за одређивање критичне области може се корстити нормална апроксимација, тј. да $\frac{T-ET}{\sqrt{DT}}$ има стандардну нормалну расподелу.

Приметимо да T 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.

Случај два узорка

Вилкоксонов тест се може лако адаптирати на случај два спарена, као и независна узорка. 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 $n > 12$.

У случају независних узорака, морају се увести додатне претпоставке, као што су да расподеле за X и Y имају исту расподелу до на константу, тј. да је $X = Y + c$ за неко c. Тада се овај тест може користити као

непараметарска алтернатива t-тесту за једнакост очекивања обележја са нормалним расподелама са истим дисперзијама.

Претпоставимо да на располагању имамо узорке $X_1, ..., X_n$ и $Y_{n+1}, ..., Y_{n+m}$. Ова два узорка се обједине у један узорак (зато смо их тако и означили).

Статистика, у овом случају је

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

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

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

где је $z_i = 1$ ако је i-ти елемент из првог узорка, у супротном $z_i = 0$.

Показаћемо да, ако је H_0 тачна, $ET = \frac{n(n+m+1)}{2}$ и да је $D(T) = \frac{nm(n+m+1)}{12}$,а може се показати и да се за $n, m > 10$ може користи нормална апроксимација.

$$E(T) = \sum_{i=1}^{n} E(r_i) = n \cdot E(r_1) = n \cdot \sum_{i=1}^{n+m} \frac{i}{n+m} = \frac{n(n+m+1)}{2}. \quad (2.15)$$

Дисперзију рачунамо на основу формуле $D(T) = ET^2 - (ET)^2$.

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

$$E(r_1^2) = \sum_{i=1}^{n+m} \frac{i^2}{n+m} = \frac{(m+m+1)(2n+2m+3)}{6}$$

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

Даље је

$$E(T^2) = \frac{n(n+m+1)(2n+2m+3)}{6} - \frac{n(n-1)(m+n+1)(3m+3n+2)}{12}. \quad (2.16)$$

Замењујући (2.15) и (2.16) у израз за дисперзију, добијамо да је $D(T) = \frac{nm(n+m+1)}{12}$.

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

$$U^\star = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} I\{X_i > Y_{n+j}\} = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} I\{X_{(i)} > Y_{(n+j)}\}$$
$$= \frac{1}{nm}\sum_{i=1}^{n}(r_i - i) = \frac{1}{nm}T - \frac{(n+1)}{2m}.$$

Статистика $U = mnU^\star$ је позната и као Ман-Витнијева статистика.

Уколико је $X > Y$ онда ће вредност статистике T бити већа него што очекујемо јер ће елементи узорка $X_1, .., X_n$ имати углавном већи ранг од оних из узорка $Y_{n+1}, ..., Y_{n+m}$.

Пример 32. *За израду каблова коришћене су две различите технике и како би се упоредио квалитет производа, одређене су њихове максималне силе издрзљивости (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 издрзљивости 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 $\hat{T} = 123$, *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, приликом нормализације тест статистике врши се корекција непрекидности јер су вредности које узима T *целобројне, односно рачуна се*

$$\frac{T - ET - 0.5}{\sqrt{DT}}.$$

И управо та вредност је имплементирана у R-у. Поред тога, одузета је и минимална сума рангова која износи $\frac{n(n+1)}{2}$.

2.2.4 Тестови сагласности са расподелом

Претпоставимо да обележје X им функцију расподеле F. У овом поглављу бавићемо се тестовима који се односе на тестирање $H_0 : F = F_0$, при чему F_0 може зависити од непознатих параметара. Приказаћемо неколико 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. За сада претпостављамо да је F_0 апсолутно непрекидна функција расподеле. Такође, претпостављамо да F_0 не зависи од непозатих параметара.

- Тест Колмогоров-Смирнова

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

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

  За мале вредности обима узорка може се егзактно исвести расподела, док је за велико n нађена асимптотска расподеластатистике $\sqrt{n}|F_n(x)-F_0(x)|$. Ова статистика има једно јако лепо својство, а то је да, ако је H_0 тачна, расподела T не зависи од F_0. То следи из познате особина да ако X има функцију расподеле F_0 онда $F_0(X)$ има униформну $\mathcal{U}[0,1]$ расподелу.

- Тест Крамер-фон Мизеса

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

  Вредности подинтегралне функције ће бити блиске нули уколико је H_0 тачна, док ако није, вредности ће бити веће од нуле.

  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 $goftests$.

У случају да желимо да тестирамо $H_0 : F = F_0(\theta)$, где је θ непознат параметар, можемо адаптирати претходно описане тестове и то на следећи начин: оценимо θ методом максималне веродостојности, а затим

применимо претходне тестове са $F_0 = F_0(\hat{\theta})$. У неким (честим ситуацијама) кад је θ параметар скалирања или локације, може се опет показати да је 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).

Пример 33. *Тестираћемо да ли зараде просветних радника из примера 4 имају нормалну расподелу. За то ћемо применити класичне тестове, али, пошто немамо претпоставку о параметрима, мораћемо да искористимо њихове модификације са оцењеним параметрима, и прво 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)
n=195

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
}
(Cks=quantile(Tks0,1-alfa))
(Ccvm=quantile(Tcvm0,1-alfa))
#реалзиована вредности тест статистика
    m=mean(zaradeJanuar)
so=sd(zaradeJanuar)
(Tks=ks.test(zaradeJanuar,'pnorm',m,so)$stat)
(Tcvm=cvm.test(zaradeJanuar,'pnorm',m,so)
```

Добили смо да су сва три теста одбацила нулту хипотезу јер су добијене p-вредности тестова биле веће од 0.05.

χ^2-тест

У овом случају X не мора бити апсолутно непрекидна случајна величина.

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

Домен обележја X се подели у k дисјунктних категорија а затим се преброји број чланова из узорка у свакој од категорија и упореди са очекиваним бројем. Нека је M_j број елемената у j-тој категорији. Тада M_j има биномну $\mathcal{B}(n, p_j)$ где је $p_j = P_{H_0}\{X$ је у j-тој категорији$\}$. Одавде је $EM_j = np_j$. Сада можемо конструисати тест статистику

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

Уколко је H_0 тачно T има χ^2_{k-1} расподелу. Критична област је природно облика $W = \{T > C\}$, 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.

Пример 34. *Желимо да проверимо да ли је коцкица за игру заиста хомогеа, 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* $H_0 : P\{X = k\} = \frac{1}{6},\ k = 1, 2..., 6$. *Алтернатива хипотеза је да* H_0 *не важи.*

Бацили смо коцкицу 60 пута и добили следећи резултат:

Реализована вредност тест статистике је

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

Како је критична област облика $W = \{T > C\}$, *и* T, *ако је* H_0 *тачна има* χ^2_5 *расподелу, добијамо да је p-вредност теста* $p = P\{T > C\} = 0.98$, *па прихватамо* H_0.

За тестирање можемо користити и уграђену функцију chisq.test *у R-у.*

```
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 расподела, уколико важи H_0, $\chi^2_{k-1-\text{број оцењених параметара}}$.

Пример 35. *Желимо да тестирамо да ли узорак зарада из примера 4 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*

$$\hat{m} = \bar{x}_n = 44687.38, \quad \hat{\sigma}^2 = \bar{s}_n^2 = 50149822.$$

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.

	$[-\infty, 38593.1]$	$(, 44250.3]$	$(, 49907.5]$	$(,55564.7]$	$(, 61221.9]$	$(,66879.1]$	$(,\infty]$
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

$$p_1 = P\{X \leq 38593.1\} = \Phi(\frac{38593.1 - \hat{m}}{\hat{\sigma}}) = 0.1947366$$

$$p_2 = P\{38593.1 < X \leq 44250.3\} = \Phi(\frac{44250.3 - \hat{m}}{\hat{\sigma}}) - \Phi(\frac{38593.1 - \hat{m}}{\hat{\sigma}})$$
$$= 0.2806563$$

$$p_3 = P\{44250.3 < X \leq 49907.5\} = \Phi(\frac{49907.5 - \hat{m}}{\hat{\sigma}}) - \Phi(\frac{44250.3 - \hat{m}}{\hat{\sigma}})$$
$$= 0.2940863$$

$$p_4 = P\{49907.5 < X \leq 55564.7\} = \Phi(\frac{55564.7 - \hat{m}}{\hat{\sigma}}) - \Phi(\frac{49907.5 - \hat{m}}{\hat{\sigma}})$$
$$= 0.1682498$$

$$p_5 = P\{55564.7 < X \leq 61221.9\} = \Phi(\frac{61221.9 - \hat{m}}{\hat{\sigma}}) - \Phi(\frac{55564.7 - \hat{m}}{\hat{\sigma}})$$
$$= 0.05249501$$

$$p_6 = P\{61221.9 < X \leq 66879.1\} = \Phi(\frac{66879.1 - \hat{m}}{\hat{\sigma}}) - \Phi(\frac{61221.9 - \hat{m}}{\hat{\sigma}})$$
$$= 0.008912812$$

$$p_6 = P\{X \geq 66879.1\} = 1 - \Phi(\frac{66879.1 - \hat{m}}{\hat{\sigma}}) = 0.0008631106.$$

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

На основу табеле 35 закључујемо да последње три категорије треба спојити. Тада добијамо:

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

2.2.5 Тестови о једнакој расподељености два узорка

Један од тестова из ове категорије је и већ описани Вилкоксонов тест. Поред њега можемо направити аналогне тестове класичним тестовима сагласности, само за два узорка. Нека су F и G функције расподела обележја X и Y. Тада се 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:

- Колмогоров-Смирнов $T = \sup_{x\in\mathbb{R}} |F_{n_1}(x) - G_{n_2}(x)|$. За $\sqrt{\frac{n_1 n_2}{N}}T$ је нађена гранична расподела под нултом хипотезом.
- Крамер-фон Мизесов $T = \int_{-\infty}^{\infty}(F_{n_1}(x) - G_{n_2}(x))^2 dH_N(x)$, где је $H_N(x)$ емпиријска функција расподеле обједињеног узорка. За $\frac{n_1 n_2}{N}T$ 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 из пакета RVAideMemoire,

Пример 36. *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 величина 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.

```
library("RVAideMemoire")
ks.test(xx,yy)
CvM.test(xx,yy)
```

2.2.6 Тестови независности

χ^2 тест

Желимо да тестирамо H_0 да су обележја X и Y независна. Подсетимо се то значи да је за 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 за вредности X и 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.

Тада сличним резоновањем као у χ^2-тесту сагласности са расподелом, формирамо тест статистику

$$T = \sum_{ij} \frac{(M_{ij} - n\hat{p}_{ij})^2}{n\hat{p}_{ij}},$$

при чему је $\hat{p}_{ij}$, оцењено на основу узорка под претпоставком да важи H_0, једнако $\hat{p}_{i,\cdot}\hat{p}_{\cdot,j} = \frac{\sum_j M_{ij}}{n} \cdot \frac{\sum_i M_{ij}}{n}$. Тест статистика, уколико важи H_0, има $\chi^2_{(K-1)(L-1)}$ расподелу.

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

Пример 37. *Желимо да испитамо да ли постоји веза између учсталости физичке активности и пушења. Разликоваћемо 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 237 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 1. и 2. обележја X. Сада добијамо:

$Y \setminus X$	*1-2.*	*3.*	$\sum$
1.	*6.43*	*4.57*	*11*
2.	*110.52*	*78.48*	*189*
3.	*11.11*	*7.89*	*19*
4.	*9.92*	*7.06*	*17*
$\sum$	*138*	*98*	*236*

Добијамо да је реализована вредност тест статистике

$$\hat{T} = \frac{(8-6.43)^2}{6.43} + \frac{(3-4.57)^2}{4.57} + \frac{(105-110.52)^2}{110.52} + \frac{(84-78.48)^2}{78.48} + \frac{(15-11.11)^2}{11.11} + \frac{(4-7.89)^2}{7.89} + \frac{(10-9.92)^2}{9.92} + \frac{(7-7.06)^2}{7.06} = 4.86.$$

Како T, под H_0, има χ^2_3 расподелу, добијамо да је p-вредност $P\{T > \hat{T}\} = 0.182$ 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.

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

Пирсонов и Спирманов тест некорелисаности

Многе статистичке процедуре су засноване на претпоставци о некорелисаности обележја (слабији услов од независности).[2] Зато се је често довољно проверити некорелисаност обележја.

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

$$\rho = \frac{\mathrm{cov(X, Y)}}{\sqrt{DX} \cdot = 0 \sqrt{DY}}.$$

Оцена методом замене за ρ[3] је

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

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

[2] Уколико су два обележја независна онда су и некорелисана. Обрнуто не важи.

[3] оцена је познатија под називом Пирсонов коефицијент корелације

што расподела ове тест стастистике, под нултом хипотезом, евидентно зависи од расподеле обележја X и Y. Познато је да, уколико H_0 важи и X и Y имају нормалну расподелу, онда $T = \hat{\rho}_n \sqrt{\frac{n-2}{1-\hat{\rho}_n^2}}$ има t_{n-2} расподелу.

Како бисмо се "ослободили" о претпоставке о расподели обележја можемо одредити емпиријски коефицијент корелације за статистике ранга, 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*.

Поглавље 3

Регресиони модели

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

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*. Y је зависна променљива а X независна или предиктор.

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. Регресиона функција је права линија акко 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. Регресиони 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 функцијом а да читава случајност Y долази од случајних грешака. Формално овако постављен проблем називамо *контролисана регресија*.

Дакле, наш главни задатак је да моделирамо 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.

3.1 Проста линеарна регресија

Претпоставићемо да на располагању имамо узорак $(X_1, Y_1), \ (X_n, Y_n)$ и да је модел који желимо да искористимо

$$E(Y|X) = \beta_0 + \beta_1 X, \ \ Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,$$

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

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

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

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

Ово је последица тога чсто је $E(Y|X)$ функција $f(X)$ која минимизира растојање $E(Y - f(X))^2$.

Добијамо да су тражени $\hat{a}$ и $\hat{b}$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} Y_i X_i - n\bar{x}\bar{Y}}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$
$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}.$$

Приметимо да полазни модел можемо написати у центрираном облику $Y_i = \beta_1(X_i - \bar{X}) + \beta_0 + \beta_1 \bar{X} + \varepsilon_i$. Испоставља се да је овај облик погоднији за прогнозирање јер $\hat{Y}_i = \hat{\beta}_1(X_i - \bar{X}) + \bar{Y}$.

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:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n Y_i(X_i - \bar{X}) - \bar{Y}\sum_{i=1}^n (X_i - \bar{X})}{n\bar{S}_X^2} = \sum_{i=1}^n Y_i \cdot \frac{(X_i - \bar{X})}{n\bar{S}_X^2}. \qquad (3.1)$$

Сада је

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

Облик (3.1) нам је посебно важан јер из претпоставке да су ε_i и ε_j међусобно некорелисани добијамо да су и Y_i међусобно некорелисани. Одавде је

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

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 $\hat{\beta}_0$.

Уколико се уведе додатна претпоставка да грешке модела $\{\varepsilon_i\}$ представљају низ независних случајних величина са $\mathcal{N}(0, \sigma^2)$ расподелом онда имамо још једно лепо својство добијених оцена. Прво, уколико ε_i има $\mathcal{N}(0, \sigma^2)$ онда Y_i има $\mathcal{N}(\beta_0 + \beta_1 X_i)$ расподелу па $\hat{\beta}_0$ и $\hat{\beta}_1$, као линеарне комбинације нормално расподељених случајних величина, 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

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

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 $H_0 : \quad \beta_1 = 0$. За то, на основу претходног, можемо искористити статистику

$$T = \frac{\hat{\beta}_1}{\frac{\sigma}{\sqrt{n}S_X}}, \tag{3.2}$$

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

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

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

Тада $\frac{\hat{\beta}_1 - \beta_1}{\frac{\hat{\sigma}}{\sqrt{n}S_X}}$ има Студентову t_{n-2} расподелу (доказ изостављамо). Зато ћемо, за тестирање $H_0: \beta_1 = 0$, уместо (3.2) користити статистику

$$T = \frac{\hat{\beta}_1}{\frac{\hat{\sigma}}{\sqrt{n}S_X}}. \tag{3.3}$$

Слично, $\frac{\hat{\beta}_0 - \beta_0}{\frac{\hat{\sigma}}{\sqrt{n}}\sqrt{1+\frac{\bar{X}}{S_X^2}}}$ 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_i када предиктор узима вредност X_i је $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$. Како су и $\hat{\beta}_0$ и $\hat{\beta}_1$ линеарне комбинације нормално расподељених случајних величина, и $\hat{Y}_i$ ће то бити. Заиста,

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

Сада је

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

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

$$\frac{\hat{Y}_0 - EY_0}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{nS_x^2}}} \sim t_{n-2}$$

Одавде можемо лако направити интервал поверења за **средњу вредност** зависне променљиве уколико је предиктор једнак X_0. Као што смо и очекивали, интервал је најужи за $X_0 = \bar{X}$. Како је $Y_0 = \beta_0 + \beta_1 X_0 + \varepsilon$, добијамо да важи

$$\frac{\hat{Y}_0 - Y_0}{\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{nS_x^2}}} \sim t_{n-2}.$$

Одавде можемо направити интервал поверења за **вредност** зависне променљиве уколико је предиктор једнак X_0. 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.

Уведимо следеће ознаке:

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

Може се показати да је

$$SSTO = SSR + SSE.$$

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*. Он 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, R^2 је боље рачунати на тест подацима, а не на подацима који се користе за прављење модела (тренинг подаци). Тада је $R^2 \in (-\infty, 1]$ 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 модела закључивање (а не предикција), онда треба проверити претпоставке на основу којих вршимо закључивања, као што су претпостављена нормалност грешака, константна дисперзија и њихова некорелисаност. Треба имати у виду да су грешке модела необзервабилне и да закључивање о овим својствима можемо вршити на основу резидуала $\{e_i\}$.

3.2 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. Примера ради, претпоставимо да 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 саобраајне несрее е дои, у зависности типа возила, типа пута, сигнализације, временских услова.

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 променљиве дефинисане на R. Следећи пример нам може послужити као мотивациони за коришћење ове трансформације.

Пример 38. *Претпоставимо да је Y нека зависна променљива 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 располгању имамо само информацију да ли је Y_i "прешло 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 оцеити вероватноћу да је Y_i веће од неког нивоа c. Тада је

$$\begin{aligned} p_i &= P\{Y_i^c = 1\} = P\{\varepsilon_i > c - \beta_0 - \beta_1 X_i\} = \Phi\Big(\frac{c - \beta_0 - \beta_1 X_i}{\sigma}\Big) \\ &= \Phi\Big(\frac{-c + \beta_0 + \beta_1 X_i}{\sigma}\Big). \end{aligned}$$

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 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.*

Параметре a и b 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

$$l(\beta_0,\beta_1)=\sum_{i=1}^{n}\Big(Y_i\log\frac{p_i}{1-p_i}+\log(1-p_i)\Big)=\sum_{i=1}^{n}\Big(Y_i(\beta_0+\beta_1X_i)+\log\frac{1}{e^{\beta_0+\beta_1X_i}+1}\Big).$$

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. Након што оценимо $\hat{\beta}_0$ и $\hat{\beta}_1$, оцењена логит функција је

$$\hat{\lambda}(X)=\hat{\beta}_0+\hat{\beta}_1X$$

а оцена вероватноће када је предиктор X је

$$\hat{p}(X)=\frac{1}{1+e^{-(\hat{\beta}_0+\hat{\beta}_1X)}},$$

односно

$$\hat{p}_i=\frac{1}{1+e^{-(\hat{\beta}_0+\hat{\beta}_1X_i)}}.$$

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 Валдове статистике Валдова статистика

$$Z=\frac{\hat{\beta}_i}{SE(\hat{\beta}_i)},$$

која ако је нулта хипотеза $H_0:\ \beta_i=0$ тачна, има асимптотски нормалну расподелу.

Важан корак у анализи 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, тада $Y_i\sim\mathcal{B}(1,\theta_i)$). 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). Важи и општије тврђење, ако су D_1 и D_2 девијације два 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.

Поред овога може се дефинисати и уопштени коефицијент детерминације

$$R^2 = 1 - \frac{D}{D_0}.$$

Из дефиниције видимо да је R^2 блиско јединици уколико је модел баш добар, док уколико је блиско нули, предиктори не доприносе бољем квалитету модела.

Пример 39. *БИЋЕ УБАЧЕН КАСНИЈЕ.*

На основу оцењеног модела можемо вршити и класификацију, и она, такође, може послужити за одређивање квалитета модела. Једна могућност је да ако је $\hat{p}_i > 0.5$ онда је $\hat{Y}_i = 1$, у супротном је 0. И онда можемо видети проценат добро класификованих података. Друга могућност је следеће:

$$y_i = \begin{cases} 0, & \hat{p}_i < C \\ 1, & \hat{p}_i \geq C \end{cases}$$

Сада се поставља питање како да одаберемо адекватан праг C. Резултат предвиђања се може приказати следећом таблицом (матрица конфузије).

Ст \ Пр	0	1
0	a	b
1	c	d

Тачност класификације је $A = \frac{a+d}{a+b+c+d}$. Ова мера није адекватна уколико 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 $TPR = \frac{d}{c+d}$ (енгл. true positive rates)
- специфичност $TNR = \frac{a}{a+b}$ (енгл. true negative rates)

- прецизност $PPV = \frac{d}{b+d}$ (енгл. positive predictive value)
- $FPR = 1 - TNR$
- 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.

Поред наведених мера, за одређивање прага се може одредити и ROC крива која представља зависност између TPR и FPR, односно на y-оси је сензитивност, а на x-оси специфичности. Како нам је "тачка" горњи леви угао графика (што се у пракси никада не постиже), бирамо тачку са криве која је најближа тачки $(0, 1)$.

Површина испод ROC криве (AUC) је индекс прецизности. то је веа површина боља је предиктивна мо модела. Заправо, AUC оцењује вероватноћуу да при случајном избору две 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 (Вилкоксонова статистика).

Пример 40. *БИЋЕ УБАЧЕН НАКНАДНО.*

Додатак

У овом поглављу наводимо важне познате расподеле које користимо током курса.

Дискретне расподеле

- X има Бернулијеву расподелу (познату још и као индикатор), уколико је закон расподеле:

$$X : \begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix}, \quad p \in (0,1). \tag{3.4}$$

  Најприроднија интерпретација ове променљиве је да њоме моделирамо да ли се неки опит успешно реализовао или не. Лако се показује да је $EX = p$ и $DX = p(1-p)$.

- X има Биномну $\mathcal{B}(n,p)$ уколико је њен закон расподеле је

$$P\{X = k\} = \binom{n}{n} p^k (1-p)^{n-k}, \quad k = 0, 1.., n. \tag{3.5}$$

  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 (3.4). Одавде је $EX = np$ и $DX = np(1-p)$.

- X има геометријску $\mathcal{G}(p)$ расподелу уколико је њен закон расподеле

$$P\{X = k\} = p(1-p)^{k-1}, \quad k = 1, 2... \tag{3.6}$$

  X представља број понављања опита до првог успеха (укључујући и последње понављање). Нумеричке карактеристике су $EX = \frac{1}{p}$ и $DX = \frac{1-p}{p^2.}$

- X има Пуасонову $\mathcal{P}(\lambda)$, $\lambda > 0$, расподелу уколико је њен закон расподеле

$$P\{X = k\} = \frac{\lambda^k e^{-\lambda}}{k!} \quad k = 0, 1, 2... \tag{3.7}$$

X се може интерпретирати као број неких догађаја у фиксном временском интервалу (нпр. број позива Хитној помоћи у току ноћи). 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.

Апсолутно непрекидне случајне величине

- X са нормалном $\mathcal{N}(m, \sigma^2)$ расподелом има функцију густине

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-m)^2}{2\sigma^2}}, \quad m \in \mathbb{R},\ \sigma^2 > 0. \tag{3.8}$$

Важне особине:

  - $EX = m$ и $DX = \sigma^2$;
  - Функција расподеле $F(x) = \int_{-\infty}^{x} f(u)du$ нема аналитичи облик;
  - $\frac{X-m}{\sigma}$ има $\mathcal{N}(0, 1)$;
  - функција густине је симетрична око m;
  - функција густине достиже максимум за $x = m$;
  - $F(x) = 1 - F(-x)$;
  - Нека 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 нека су $a_1, ..., a_k \in \mathbf{R}$ тако да је $\sum_{i=1}^{k} a_i^2 > 0$. Тада

$$a_1X_1 + \cdots + a_kX_k \sim \mathcal{N}(a_1m_1 + \cdots + a_km_k, a_1^2\sigma_1^2 + \cdots + a_k^2\sigma_k^2). \tag{3.9}$$

- X са χ_n^2 расподелом има функцију густине

$$f(x) = \frac{x^{\frac{n}{2}-1}}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} e^{-\frac{x}{2}}, \quad x > 0. \tag{3.10}$$

Много чешће X се дефинише као збир n квадрата независних случајних величина са стандардном $\mathcal{N}(0, 1)$ расподелом. Зато је $EX = n$ и $DX = 2n$.

- X са Студентовом t расподелом са n степени слободе има функцију густине

$$f(x) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}},\ n \in \mathbf{R}^+,\ x > 0.$$

Много чеш ће 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 $EX = 0$ (за $n > 1$, иначе не постоји) и $DX = \frac{n}{n-2}$ (за $n > 2$, иначе не постоји). Ова расподела је веома слична стандардној нормалној расподели, с напоменом да су репови ове расподеле тежи. 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.

- X са експоненцијалном $\mathcal{E}(\lambda)$, $\lambda > 0$, расподелом има функцију густине

$$f(x) = \lambda e^{-\lambda x},\ \ x > 0. \tag{3.11}$$

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 је *одсуство памћења*, односно

$$P\{X > s + t | X > s\} = P\{X > t\}.$$

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.

- X са $\gamma(\alpha, \beta)$ расподелом има функцију густине

$$f(x) = \frac{x^{\alpha-1}\beta^\alpha}{\Gamma(\alpha)}e^{-\beta x},\ \ x > 0. \tag{3.12}$$

Њене важне особине су:

  - $EX = \frac{\alpha}{\beta}$ и $DX = \frac{\alpha}{\beta^2}$;
  - Када је $\alpha = 1$ онда се ради о $\mathcal{E}(\beta)$ расподели;

  - Збир n независних случајних променљивих са $\mathcal{E}(\beta)$ расподелом има $\Gamma(n, \beta)$ расподелу.

- X са Фишеровом F_{n_1,n_2} расподелом има компликовану функцију густине па је овом приликом изостављамо. Њена важна особина је да се може дефинисати и као

$$X = \frac{\frac{Y_1}{n_1}}{\frac{Y_2}{n_2}}, \tag{3.13}$$

  где су Y_1 и Y_2 независне случајне променљиве са $\chi^2_{n_1}$ и $\chi^2_{n_2}$ расподелама.