Mirta Benšić

Nenad Šuvak

Primijenjena statistika

Sveučilište J. J. Strossmayera u Osijeku

Odjel za matematiku

Osijek, 2013.

M. Benšić, N. Šuvak – Primijenjena statistika.

Izdavač: Sveučilište J.J. Strossmayera, Odjel za matematiku

Recenzenti: Prof.dr.sc. Bojan Basrak
Prof.dr.sc. Anamarija Jazbec

Lektor: Davor Tanocki, prof.

Tehnička obrada: Prof.dr.sc. Mirta Benšić, Doc.dr.sc. Nenad Šuvak

CIP zapis dostupan u računalnom katalogu Gradske i sveučilišne knjižnice Osijek pod brojem ...

ISBN 978-953-6931-59-0

Udžbenik se objavljuje uz suglasnost Senata Sveučilišta J. J. Strossmayera u Osijeku pod brojem 11/13.

Tisak: Grafika d.o.o., Osijek

Predgovor

Ova knjiga nastala je s namjerom da pomogne studentima preddiplomskih i diplomskih studija prilikom svladavanja gradiva iz statističkih kolegija primijenjenog karaktera. Za razumijevanje gradiva prezentiranog u knjizi nije nužno matematičko predznanje veće od sadržaja matematike općih gimnazijskih programa u Republici Hrvatskoj.

Knjiga je podijeljena u sedam poglavlja: Uvod, Prikupljanje i organizacija podataka, Deskriptivna statistika, Slučajna varijabla, Statističko zaključivanje - jedna varijabla, Statističko zaključivanje - dvije varijable i Zadaci za vježbu. U cijeloj je knjizi teorijski dio ilustriran mnoštvom primjera i zadataka temeljenih na podacima koji su ili simulirani ili potječu iz stvarnih istraživanja i korišteni su uz odobrenje voditelja istraživanja. Baze podataka korištene u primjerima i zadacima dostupne su na mrežnim stranicama autora knjige (http://www.mathos.unios.hr/∼mirta/, http://www.mathos.unios.hr/∼nsuvak/) u formatu prikladnom za upotrebu računalnih programa. Kroz treće poglavlje u primjerima je ilustriran način korištenja programskog paketa *Statistica* (StatSoft, inačica 10) za deskriptivnu statistiku. Statističke procedure navedene u knjizi dostupne su u svim standardnim statističkim programima (R, Statistica, SPSS, SAS, itd.).
Zadnje poglavlje Zadaci za vježbu sadrži zadatke s kolokvija i pismenih ispita održanih tijekom nekoliko posljednjih akademskih godina na Odjelu za matematiku, Građevinskom fakultetu, Prehrambeno-tehnološkom fakultetu i Učiteljskom fakultetu Sveučilišta J.J. Strossmayera u Osijeku.

Zahvaljujemo svima koji su pomogli da se ova knjiga tiska i bude što bolja. To se posebno odnosi na recenzente koji su pažljivo pročitali rukopis te svojim primjedbama i sugestijama utjecali na poboljšanje mnogih dijelova teksta, kao i na kolege Natašu Šarliju, Andreu Krajinu, Slobodana Jelića, Mariju Miloloža-Pandur i Ivonu Puljić jer su svojim sugestijama doprinijeli kvaliteti primjera i zadataka.

Autori će biti zahvalni svim čitateljima na primjedbama vezanima uz eventualne pogreške, nepreciznosti ili nedostatke.

U Osijeku, lipanj 2013. Mirta Benšić i Nenad Šuvak

Sadržaj

Poglavlje 1

Uvod

Uporaba riječi **statistika** u svakodnevnom životu najčešće je povezana s brojčanim vrijednostima kojima pokušavamo opisati bitne karakteristike nekog skupa podataka. Na službenim mrežnim stranicama Državnog zavoda za statistiku Republike Hrvatske možemo pročitati (http://www.dzs.hr, 5. rujna 2012.):

Prosječna mjesečna isplaćena neto plaća po zaposlenome u pravnim osobama Republike Hrvatske za lipanj 2012. iznosila je 5492 *kune.*

Minimalna plaća za razdoblje od 1. lipnja 2012. do 31. svibnja 2012. u Republici Hrvatskoj iznosila je 2814 *kuna.*

Stopa registrirane nezaposlenosti za srpanj 2012. iznosila je 17.5%.

Udio aktivnog stanovništva u radno sposobnom stanovništvu (stopa aktivnosti) za siječanj, veljaču i ožujak 2012. iznosila je 51.7%, *istovremeno* 42.9% *radno sposobnih osoba je zaposleno (stopa zaposlenosti), a* 17% *radne snage je nezaposleno (stopa nezaposlenosti).*

Temelj statistike kao znanstvene discipline, kao i svih istraživanja koja se koriste statističkim metodama, čine skupovi podataka.

Statistika kao znanstvena disciplina bavi se razvojem metoda prikupljanja, opisivanja i analiziranja podataka te primjenom tih metoda u procesu donošenja zaključaka na temelju prikupljenih podataka.

Statističko istraživanje fokusirano je na skup **objekata**, tj. **jedinki** (ljudi, životinja, biljaka, stvari, država, gradova, poduzeća, itd.) i skup odabranih veličina koje

se na njima promatraju. Veličine koje se promatraju zovemo **varijablama**. Sve jedinke koje se žele obuhvatiti istraživanjem, tj. o kojima se želi zaključivati, čine **populaciju**.

Primjer 1.1. *Bavimo se istraživanjem uspjeha studenata jedne generacije na ispitu iz kolegija Statistika na nekom sveučilištu (tablica 1.1).*

Jedinke	osobe, imenom i prezimenom ili nekom šifrom
Varijabla	ocjena iz Statistike

Tablica 1.1: Primjer jedinki i varijabli obuhvaćenih opisanim istraživanjem.

U tom primjeru navedena je samo jedna varijabla koja se analizira na jedinkama populacije, tj. uspjeh iz kolegija Statistika. Međutim, često nas zanima nekoliko varijabli i/ili veze među njima. Primjerice, želimo li ispitati ovisi li uspjeh iz kolegija u prethodnom primjeru o spolu studenta, potrebno je u istraživanju populacije za svaku jedinku zabilježiti i vrijednost varijable spol (M ili Ž), a želimo li ispitati ovisi li uspjeh o pripadnosti pojedinoj grupi vježbi, potrebno je za svaku jedinku zabilježiti koju je grupu vježbi pohađala. Zbog preglednosti prikupljene podatke prikazujemo tablično tako da jedan redak odgovara točno jednoj jedinki, a stupac točno jednoj varijabli.

Primjer 1.2. *Bavimo se istraživanjem uspjeha studenata jedne generacije na ispitu iz kolegija Statistika na nekom sveučilištu u ovisnosti o spolu ispitanika i grupi vježbi koju je student pohađao. U ovom slučaju istraživanje se temelji na jedinkama i varijablama prikazanima u tablici 1.2.*

Jedinke	studenti, identificirani svojim matičnim brojem
Varijable	ocjena iz Statistike, spol, grupa vježbi

Tablica 1.2: Istraživanje uspjeha studenata - jedinke i varijable.

Tablicu za bilježenje prikupljenih podataka treba organizirati na način prikazan tablicom 1.3.

Matični broj studenta	Ocjena iz Statistike	Spol	Grupa vježbi
1206	5	Ž	A
1326	2	Ž	B
942	4	Ž	C
⋮	⋮	⋮	⋮

Tablica 1.3: Istraživanje uspjeha studenata - tablica prikupljenih podataka.

U prethodnim primjerima možemo lako istražiti cijelu populaciju s obzirom da generacija koju proučavamo broji konačno mnogo studenata (npr. 83 studenta). Međutim, istražujemo li prije izbora za predsjednika neke države preferencije građana prema nekom od kandidata, ne možemo ispitati sve osobe populacije (tj. sve državljane koji imaju pravo glasa) jer bi to bilo provođenje izbora. Kada nije moguće istražiti veličine koje nas zanimaju na svim jedinkama populacije, potrebno je iz populacije izdvojiti **uzorak** na kojemu će biti prikupljeni podaci. S obzirom da se o cijeloj populaciji želi zaključivati na temelju podataka prikupljenih na uzorku, za istraživanje je vrlo važno znati kako kreirati kvalitetan uzorak.

Primjena statistike u istraživanju podrazumijeva da se u pripremi istraživanja izabranog problema poštuju sljedeća pravila:

Populaciju koja je predmet istraživanja i ciljeve potrebno je jasno odrediti (detaljno proučiti populaciju, zabilježiti njene osnovne karakteristike i ciljeve istraživanja).

Kreirati kvalitetan uzorak i odabrati metodu za prikupljanje podataka.

Izabrati prikladne metode za opis skupa prikupljenih podataka (deskriptivna statistika).

Izabrati prikladne statističke metode za zaključivanje o populaciji na temelju prikupljenih podataka na uzorku.

U skladu s tim u ovom ćemo se kolegiju baviti nekim **metodama prikupljanja podataka i kreiranja uzorka**, **metodama deskriptivne statistike** i **metodama statističkog zaključivanja**. S obzirom da se metode kojima se kreira uzorak i metode statističkog zaključivanja temelje na poznavanju osnovnih pojmova teorije vjerojatnosti, u kolegiju ćemo također navesti temeljne pojmove i zakone teorije vjerojatnosti potrebne za razumijevanje osnovnog statističkog aparata.

Poglavlje 2

Prikupljanje i organizacija podataka

2.1 Populacija i uzorak

Statističko istraživanje usmjereno je na skup jedinki koje zadovoljavaju neka svojstva bitna za obilježje koje se istražuje, tj. **populaciju**. Dakle, **populaciju čine sve jedinke koje su predmet istraživanja.**

Primjer 2.1. *Istražujemo razlike u prehrambenim navikama između stanovnika Slavonije i Baranje i stanovnika Dalmacije. Populaciju čine svi stanovnici Slavonije, Baranje i Dalmacije. Međutim, ako nas zanimaju samo prehrambene navike studenata iz tih područja, onda populaciju čine samo studenti iz Slavonije, Baranje i Dalmacije.*

Populacija može sadržavati vrlo velik broj jedinki i stoga je često teško, ili čak nemoguće, istraživanje provesti na svim jedinkama populacije. Rješenje tog problema sastoji se u odabiru jednog podskupa populacije, koji nazivamo **uzorak**, na kojemu je osigurano kvalitetno provođenje istraživanja.

Da bi zaključci prilikom istraživanja o populaciji na temelju podataka iz uzorka bili ispravni, nužno je da uzorak bude **reprezentativan**, tj. u njemu moraju biti zastupljene tipične karakteristike populacije bitne za istraživanje.

Primjer 2.2. *U prethodnom primjeru, ako populaciju čine svi stanovnici Slavonije, Baranje i Dalmacije, istraživanje ne možemo provesti samo na uzorku djece koja pohađaju srednju školu. To bi možda bilo praktično, ali takav uzorak nije reprezentativan za zaključivanje o cijeloj populaciji.*

Jedan od načina izbora jedinki iz populacije u uzorak jest formiranje takozvanog **slučajnog uzorka** , uz poštivanje zahtjeva da svaka jedinka populacije ima jednaku vjerojatnost (šansu) ući u uzorak.

S obzirom da se u gornjoj definiciji pojavljuje pojam **vjerojatnost**, metodu formiranja slučajnog uzorka ostavljamo za sljedeća poglavlja, nakon što pojasnimo pojam vjerojatnosti.

2.2 Izvori podataka

Način prikupljanja podataka ovisi o karakteristikama obilježja koje je predmet proučavanja. Najčešće korišteni načini prikupljanja podataka jesu sljedeći:

Podaci iz javnih izvora (knjige, časopisi, novine, Internet).

Podaci iz dizajniranog eksperimenta (istraživač raspoređuje eksperimentalne jedinke u skupine s kojima provodi eksperimente te bilježi podatke za varijable koje ga zanimaju).

Podaci iz ankete (istraživač sastavlja anketni upitnik, izabire skupinu ljudi koju anketira i na osnovi njihovih odgovora prikuplja podatke).

Podaci prikupljeni promatranjem (istraživač promatra eksperimentalne jedinke u njihovu prirodnom okruženju i bilježi podatke za varijable od interesa).

Primjer 2.3. *Jedno medicinsko istraživanje proučava snagu nekog lijeka u prevenciji moždanog udara. Ljude s kojima će se provesti istraživanje istraživač dijeli na dvije skupine: tretiranu i kontrolnu. Ljudima u tretiranoj skupini daje se lijek, dok se ljudima u kontrolnoj skupini daje placebo (nadomjestak koji izgleda isto kao lijek, ali zapravo nije ništa što može imati bilo kakav utjecaj na organizam). To istraživanje primjer je dizajniranog eksperimenta kojim se mogu prikupiti određeni podaci o ispitanicima.*

2.3 Tipovi varijabli

U statističkim istraživanjima razlikujemo nekoliko osnovnih tipova varijabli koje se međusobno razlikuju po svojstvima vrijednosti koje mogu poprimiti.

2.3.1 Kvalitativne varijable

Karakteristika je kvalitativnih varijabli da njihove vrijednosti nisu, po svojim svojstvima korištenim u istraživanju, realni brojevi. Tipičan je primjer takve varijable

spol osobe. Vrijednosti kvalitativne varijable uobičajeno svrstavamo u kategorije. Kategorije kvalitativnih varijabli mogu biti definirane u skladu s potrebama statističkog istraživanja.

Primjer 2.4. *Sljedeće su varijable kvalitativnog tipa:*

- *radna mjesta u školi (spremačica, domar, tajnik, nastavnik, pedagog, ravnatelj)*
- *opisne ocjene (ništa, malo, srednje, puno)*
- *boja očiju (plava, smeđa, zelena)*
- *krvne grupe (A, B, AB, 0)*
- *spol (m ili ž).*

2.3.2 Numeričke varijable

Numeričke varijable prirodno primaju vrijednosti iz skupa realnih brojeva. Tipičan primjeri numeričkih varijabli jesu tjelesna masa i visina osobe. Međutim, treba naglasiti da se i kategorije kvalitativnih varijabli mogu izražavati brojevima, što ih ne čini numeričkim varijablama. Primjerice, spol osobe je jedna kvalitativna varijabla. Kategoriju "ženski spol" možemo označiti npr. oznakom "1", a kategoriju "muški spol" npr. oznakom "2", što može biti korisno prilikom unošenja podataka u bazu. Time smo kategorijama kvalitativne varijable pridružili numeričke vrijednosti, ali samu varijablu nismo učinili numeričkom po njenim svojstvima.

Primjer 2.5. *Sljedeće su varijable numeričkog tipa:*

- *postotak prolaznosti na pojedinim ispitima tijekom jedne akademske godine*
- *broj bodova na državnoj maturi iz matematike*
- *broj ulovljenih komaraca u klopku*
- *temperatura mora*
- *koncentracija soli u morskoj vodi.*

Među numeričkim varijablama razlikujemo **diskretne** i **neprekidne** varijable.

Diskretne numeričke varijable mogu poprimiti samo konačno ili prebrojivo mnogo vrijednosti, dok je skup mogućih vrijednosti neprekidnih numeričkih varijabli cijeli skup realnih brojeva ili neki interval.

Primjer 2.6. *Sljedeće su numeričke varijable diskretne:*

- *broj bodova na državnoj maturi iz matematike*
- *broj ulovljenih komaraca u klopku*
- *broj dana u godini s temperaturom zraka većim od $35^o\,C$.*

Primjer 2.7. *Sljedeće su numeričke varijable neprekidne:*

- *postotak prolaznosti na pojedinim ispitima tijekom jedne akademske godine*
- *temperatura mora*
- *vodostaj neke rijeke.*

Radi prikaza podataka i nekih statističkih analiza vrijednosti numeričke varijable također se mogu svrstati u kategorije. Za razliku od kategorija kvalitativne varijable, među kategorijama numeričke varijable uvijek se može prepoznati prirodan poredak.

Primjer 2.8. (auto-centar.sta)
Svrha ovog primjera je prikazati mogućnost kategorizacije numeričke varijable. Taj se postupak najčešće rješava stvaranjem nove kvalitativne varijable čije su vrijednosti svrstane u kategorije kojih je (znatno) manje nego svih mogućih vrijednosti odgovarajuće diskretne numeričke varijable. Baza podataka auto-centar.sta *sastoji se od sljedećih varijabli:*

automobili - *diskretna numerička varijabla koja sadrži podatke o broju prodanih automobila u jednom danu za sto promatranih dana. Budući da broj prodanih automobila u jednom danu može biti vrlo mali (npr. samo nekoliko osobnih automobila), ali i vrlo velik (npr. narudžbe automobila za vozni park nekog poduzeća), zaključujemo da varijabla* automobili *može poprimiti velik broj različitih vrijednosti iz skupa prirodnih brojeva. Zato je u nekim situacijama korisno kategorizirati vrijednosti ove varijable prema točno određenom kriteriju. Na primjer, kategrizacija prema broju prodanih automobila u jednom danu može se realizirati stvaranjem nove varijable* kategorija.

kategorija - *kvalitativna varijabla koja podatke iz varijable* automobili *svrstava u pet kategorija prema kriteriju prikazanom u tablici 2.8.*

broj prodanih automobila	kategorija
0 - 9	E
10 i 11	D
12 i 13	C
14 i 15	B
16 i više	A

Tablica 2.1: Primjer kategorizacije diskretne numeričke varijable automobili.

2.3.3 Ordinalne varijable

Karakteristika je ordinalnih varijabli da su one po svom karakteru kvalitativne, ali među kategorijama se može uspostaviti prirodan poredak. Tipičan je primjer takve varijable stručna sprema osobe.

Primjer 2.9. (matematika.sta)
Baza podataka matematika.sta *sadrži podatke prikupljene anketiranjem studenata nakon održanih predavanja, vježbi, kolokvija te usmenog ispita iz jednog matematičkog kolegija. Prikupljeni podaci organizirani su na sljedeći način:*

prosjek - *varijabla koja sadrži podatke o prosječnoj ocjeni studiranja za* 49 *anketiranih studenata,*

polozeno - *varijabla koja studente svrstava u dvije kategorije s obzirom na to jesu li položili ispit iz promatranog kolegija prema kriteriju prikazanom u tablici 2.2.*

položen/nepoložen ispit	kategorija
položen ispit	1
nepoložen ispit	0

Tablica 2.2: Kategorizacija studenata prema položenosti ispita.

predavanja, vjezbe - *dvije varijable koje prisutnost studenata na predavanjima/vježbama (p/v) svrstavaju u tri kategorije na način prikazan u tablici 2.3.*

prisutnost studenta na p/v	kategorija
student s p/v nije nikada izostao	1
student je s p/v izostao samo jednom	2
student je s p/v izostao barem dva puta	3

Tablica 2.3: Kategorizacija studenata prema broju izostanaka s predavanja/vježbi.

tezina kolegija, materijali - *dvije varijable koje sadrže subjektivne ocjene (u standardnoj skali od 1 do 5) studenata o težini kolegija i dostatnosti dostupnih materijala za pripremanje ispita iz promatranog kolegija.*

Uočimo da se varijabla prosjek *može promatrati kao neprekidna numerička varijabla, varijabla* položeno *je kvalitativna, dok se varijable* predavanja, vjezbe, tezina kolegija i materijali *mogu svrstati u ordinalne varijable.*

2.4 Organizacija baze podataka

Podaci u bazi podataka mogu biti organizirani na različite načine ovisno o informacijama koje želimo dobiti istraživanjem. Za ilustraciju navodimo jedan primjer niza podataka koji su organizirani na dva različita načina.

Primjer 2.10. (student.sta, student-grupe.sta)
Svrha je ovog primjera pokazati kako isti podaci u bazi podataka mogu biti organizirani na različite načine. Način organizacije ovisi o informacijama koje iz podataka želimo dobiti statističkom analizom. Baza podataka student.sta *sastoji se od sljedećih varijabli:*

klasicno studiranje - *neprekidna numerička varijabla koja sadrži podatke o godinama starosti studenata koji studiraju na klasičan način (stanuju u gradu u kojem studiraju ili putuju na predavanja)*

e-learning - *neprekidna numerička varijabla koja sadrži podatke o godinama starosti studenata koji studiraju putem interneta (tzv. e-learning).*

Baza podataka student-grupe.sta *sastoji se od sljedećih varijabli:*

dob studenta - *neprekidna numerička varijabla koja sadrži podatke o godinama starosti za sto studenata koji studiraju ili na klasičan način ili putem interneta*

nacin studiranja - *kvalitativna varijabla koja studente, bez obzira na podatke sadržane u varijabli* dob studenta*, svrstava u dvije kategorije prema kriteriju prikazanom u tablici 2.4.*

način studiranja	kategorija
student studira na klasičan način	1
student studira putem interneta	0

Tablica 2.4: Primjer kategorizacije studenata prema načinu studiranja.

Dakle, baze podataka student.sta *i* student-grupe.sta *sadrže iste podatke (godine starosti sto promatranih studenata) i daju informaciju o načinu studiranja za svakog studenta:*

- *u bazi podataka* student.sta *podaci o dobi studenata organizirani su u dvije varijable, ovisno o tome studira li student na klasičan način (*klasicno studiranje*) ili putem interneta (*e-learning*)*
- *u bazi podataka* student-grupe.sta *varijabla* dob studenta *sadrži podatke o dobi studenata, dok binarna varijabla* nacin studiranja *za svakog studenta sadrži informaciju o načinu studiranja (tablica 2.4).*

2.5 Zadaci

Zadatak 2.1. (stanovnistvo.sta)
Pretpostavimo da želite saznati starosnu strukturu (prema godinama starosti) stanovništva u svom gradu te da ste u tu svrhu prikupili podatke koji su dani u bazi stanovnistvo.sta. Navedena baza sadrži četiri varijable:

osnovna škola - varijabla koja sadrži podatke o godinama starosti za pedeset slučajno odabranih učenika jedne osnovne škole u vašem gradu

kafić - varijabla koja sadrži podatke o godinama starosti za pedeset slučajno odabranih gostiju popularnog kafića u vašem gradu

gradska knjižnica - varijabla koja sadrži podatke o godinama starosti za pedeset slučajno odabranih posjetitelja gradske knjižnice u vašem gradu

telefonska anketa - varijabla koja sadrži podatke o godinama starosti za pedeset osoba iz vašeg grada čije ste telefonske brojeve na slučajan način izabrali iz telefonskog imenika.

Nakon kratke analize baze podataka stanovnistvo.sta komentirajte reprezentativnost uzorka. Razmislite o mogućim načinima prikupljanja podataka kojima biste kreirali reprezentativan uzorak za proučavanje starosne strukture populacije.

Zadatak 2.2. (glukoza.sta)
Baza podataka glukoza.sta sastoji se od sljedećih varijabli:

dob - neprekidna numerička varijabla koja sadrži podatke o godinama starosti 102 promatrane osobe.

koncentracija - neprekidna numerička varijabla koja sadrži podatke o koncentraciji glukoze u krvi za svaku od 102 promatrane osobe.

kategorija - kvalitativna varijabla koja podatke iz varijable koncentracija glukoze svrstava u dvije kategorije (svaka je kategorija jedan interval pozitivnih realnih brojeva) na način prikazan u tablici 2.5.

interval koncentracije glukoze	kategorija
koncentracija < 6 mMol/L	N - normalna koncentracija
koncentracija ≥ 6 mMol/L	P - povišena koncenracija

Tablica 2.5: Primjer kategorizacije neprekidne numeričke varijable koncentracija.

Predložite neku drugu kategorizaciju varijable koncentracija i usporedite je s varijablom kategorija koju je u istu svrhu formirao istraživač u pokusu.

Zadatak 2.3. (kolegij.sta)
Baza podataka sastoji se od sljedećih varijabli:

godina upisa - *kvalitativna varijabla koja sadrži podatke o akademskoj godini upisa na studij za sto promatranih studenata*

kategorija - *kvalitativna varijabla koja podatke iz varijable* godina upisa *svrstava u tri kategorije (svaka je kategorija jedan konačan skup) na način prikazan u tablici 2.6.*

godina upisa	kategorija
student upisan prije 1990. godine	1
student upisan 1990., 1991. ili 1992. godine	2
student upisan 1993. ili 1994. godine	3

Tablica 2.6: Primjer kategorizacije kvalitativne varijable godina upisa.

opća kemija, organska kemija, anorganska kemija, mikrobiologija - *četiri ordinalne varijable koje sadrže podatke o postignutim ocjenama na ispitima iz spomenutih kolegija za svakog od sto promatranih studenata*

prosjek - *neprekidna numerička varijabla koja sadrži prosječne ocjene iz četiriju spomenuta kolegija za svakog od sto promatranih studenata*

uspjeh - *kvalitativna varijabla koja vrijednosti varijable* prosjek *svrstava u četiri kategorije prema kriteriju prikazanom u tablici 2.7.*

prosjek	uspjeh	prosjek	uspjeh	prosjek	uspjeh	prosjek	uspjeh
$[2, 2.5>$	dovoljan	$[2.5, 3.5>$	dobar	$[3.5, 4.5>$	vrlo dobar	$[4.5, 5]$	izvrstan

Tablica 2.7: Primjer kategorizacije neprekidne numeričke varijable prosjek.

Predložite drugačije kategorizacije varijabli godina upisa *i* uspjeh *i obrazložite svoj prijedlog kategorizacije.*

Zadatak 2.4. *Na sličan način proanalizirajte i odredite tipove varijabli u sljedećim bazama podataka:*

a) baza podataka komarci.sta *sadrži dio rezultata proučavanja komaraca u jednom močvarnom području (dostupni su podaci za* 210 *mjerenja na istoj lokaciji):*

varijable brojM *i* brojZ *redom sadrže broj muških i ženskih jedinki komaraca*

varijabla mjesec *sadrži mjesečevu mijenu (M - mlađak, U - uštap) za svako mjerenje*

varijabla doba dana *sadrži doba dana u kojem je mjerenje obavljeno (P - predvečerje, N - noć, S - svitanje)*

varijabla svjetlost *sadrži tip osvjetljenja pri mjerenju*

varijabla temperatura *sadrži temperaturu pri kojoj je mjerenje izvršeno*

varijabla rel vlaznost *sadrži relativnu vlažnost zraka za vrijeme mjerenja*

b) u bazi podataka navike.sta *nalaze se rezultati praćenja nekih životnih navika u jednom danu za svakog od* 300 *ispitanika iz uzorka:*

varijabla dnevne novine *sadrži broj prelistanih različitih dnevnih novina*

varijabla tv vijesti *sadrži broj pogledanih televizijskih vijesti na dostupnim televizijskim kanalima*

varijabla kava *sadrži broj ispijenih kava*

varijabla troskovi *sadrzi informaciju o troškovima hrane za promatrani dan*

varijabla vrijeme *sadrži ispitanikov subjektivan doživljaj vremenskih prilika u njegovu mjestu stanovanja (O - oblačno, S - sunčano)*

varijabla raspolozenje *sadrži ispitanikovu subjektivnu ocjenu vlastitog raspoloženja (L - loše, D - dobro, O - odlično)*

c) u bazi podataka posao.sta *nalaze se podaci o udaljenosti mjesta stanovanja od radnog mjesta (varijabla* udaljenost*) i mjesečnim troškovima putovanja do radnog mjesta (varijabla* troskovi*) za* 100 *slučajno odabranih zaposlenih ljudi*

d) baza podataka TV-program.sta *sastoji se od sljedećih varijabli:*

varijabla spol *sadrži informaciju o spolu ispitanika*

varijable P1, P2, P3 *i* P4 *sadrže subjektivne ocjene kvalitete ljetne programske sheme televizijskih programa P1, P2, P3 i P4*

varijabla prosjek *sadrži prosječnu ocjenu kvalitete ljetne programske sheme navedenih televizijskih programa*

e) u bazi podataka zdravlje.sta *nalaze se neki zdravstveni podaci anketiranih ispitanika:*

varijable godine *i* spol *sadrže podatke o starosti u godinama i spolu ispitanika*

vrijednosti varijable zdravlje *su subjektivne ocjene vlastitog zdravstvenog stanja ispitanika*

varijabla broj pregleda *sadrži informacije o ukupnom broju zdravstvenih pregleda svakog ispitanika u tekućoj kalendarskoj godini*

varijabla dodatno zdravstveno *sadrži podatke o dodatnom zdravstvenom osiguranju svakog ispitanika (*1 - *ispitanik je dodatno osiguran;* 0 - *ispitanik nije dodatno osiguran)*

varijabla cijena *sadrži cijenu u kunama najskupljeg zdravstvenog pregleda svakog ispitanika (u tekućoj kalendarskoj godini)*

f) baza podataka djelatnici.sta *sadrži podatke o uzorcima djelatnika dviju konkurentskih tvornica - tvornice A i tvornice B. U tablici s imenom "tvornica A" zabilježene su vrijednosti sljedećih varijabli za djelatnike* **tvornice** *A:*

varijabla spol *sadrži informaciju o spolu (*M - *muški spol,* Z - *ženski spol)*

varijabla odjel *sadrži naziv odjela u kojem je djelatnik zaposlen (*TR - *transport,* P- *pakiranje,* IS - *isporuka)*

varijabla obrazovanje *sadrži stručnu spremu djelatnika (*SSS - *srednja stručna sprema,* VŠSS - *viša stručna sprema,* VSS - *visoka stručna sprema)*

varijabla dob *sadrži starost djelatnika u godinama*

varijabla visina *sadrži visinu djelatnika u centimetrima*

varijabla rukovostvo *sadrži broj godina rada koje je djelatnik proveo na nekoj od rukovodećih pozicija u toj tvornici*

varijabla placa prije *sadrži iznos godišnje plaće djelatnika prije reorganizacije poslovnog sustava*

varijabla placa poslije *sadrži iznos godišnje plaće djelatnika nakon reorganizacije poslovnog sustava.*

U tablici s imenom "tvornica B", u varijabli placa konkurencija*, zabilježeni su iznosi godišnje plaće za svakog djelatnika iz uzorka iz* **tvornice** *B.*

Poglavlje 3

Deskriptivna statistika

3.1 Metode opisivanja kvalitativnih podataka

Kvalitativne varijable primaju vrijednosti koje su razvrstane u kategorije. Pri proučavanju takvih varijabli pažnju usmjeravamo na zastupljenost pojedine kategorije u uzorku na kojem provodimo istraživanje. Primjer 3.1 uvodi nas u problematiku opisivanja kvalitativnih varijabli.

Primjer 3.1. *Svaki čovjek prema spolu pripada jednoj od dviju kategorija (ženskom spolu (Ž) ili muškom spolu (M)), a prema tipu svoje krvne grupe jednoj od četiriju kategorija (A, B, AB ili 0). Tablica 3.1 sadrži podatke o spolu i tipu krvne grupe za deset ispitanika iz nekog medicinskog istraživanja.*

ispitanik	spol	krvna grupa
1	Ž	A
2	Ž	B
3	M	0
4	Ž	0
5	M	AB
6	M	B
7	Ž	B
8	M	A
9	Ž	AB
10	Ž	A

Tablica 3.1: Tablični prikaz podataka o spolu i krvnoj grupi.

Iz tablice 3.1 vidimo da za svakog ispitanika iz promatranog uzorka vrijednost varijable spol *pripada kategoriji M ili kategoriji Ž, a vrijednost varijable* krvna grupa *jednoj od kategorija A, B, AB ili*

0. *Prema tome, varijable* spol *i* krvna grupa *jesu* kvalitativne varijable. *Informacije koje je moguće dobiti iz prethodne tablice vezane su uz zastupljenost pojedine kategorije u promatranom uzorku. Tako je npr. moguće dobiti odgovore na sljedeća i slična pitanja:*

Koliko ispitanika ženskog spola ima u promatranom uzorku?

Koliki je udio ispitanika s krvnom grupom 0 *u promatranom uzorku?*

Koliko ispitanika ženskog spola iz promatranog uzorka ima krvnu grupu A?

Koliki udio ispitanika muškog spola iz promatranog uzorka ima krvnu grupu B ili AB?

Kako izmjeriti zastupljenost pojedine kategorije u uzorku?
Osnovna mjera kojom opisujemo zastupljenost jedne kategorije u uzorku jest **frekvencija** kategorije.
Neka varijabla, koju ćemo označiti s X, ima k kategorija (recimo $k = 4$ znači da varijabla ima 4 kategorije - npr. krvne grupe). **Označimo pojedine kategorije s $x_1, x_2, \ldots, x_k$, odnosno u drugom zapisu $\{x_i : i = 1, \ldots, k\}$. Frekvencija kategorije x_i je broj izmjerenih vrijednosti varijable koje pripadaju kategoriji x_i, $i = 1, \ldots, k$. Frekvenciju kategorije x_i označavamo s**

$$f_i.$$

Frekvencija pojedine kategorije ovisi o broju izvršenih mjeranja, tj. veličini uzorka. Da bismo lakše usporedili i tumačili rezultate raznih istraživanja, u opisu zastupljenosti jedne kategorije u uzorku često koristimo i **relativnu frekvenciju** kategorije. **Relativna frekvencija kategorije x_i je broj izmjerenih vrijednosti varijable koje pripadaju kategoriji x_i podijeljen ukupnim brojem izmjerenih vrijednosti za ispitivanu varijablu, $i = 1, \ldots, k$. Ako je n veličina uzorka, tj. broj svih izmjerenih vrijednosti ispitivane varijable, relativnu frekvenciju kategorije x_i računamo kao**

$$\frac{f_i}{n}.$$

Relativna frekvencija kategorije je mjera zastupljenosti koja daje informaciju o udjelu kategorije u uzorku poznate veličine i često se izražava kao postotak. **Frekvencije i relativne frekvencije pojedinih kategorija prikazujemo tablično i grafički.**

3.1.1 Tablični prikaz frekvencija i relativnih frekvencija

U tabličnom prikazu frekvencija i relativnih frekvencija trebaju biti zastupljene sve kategorije promatrane varijable.

Primjer 3.2. *Frekvencije i relativne frekvencije svih kategorija varijabli* spol *i* krvna grupa *iz primjera 3.1 prikazane su u tablicama 3.2 i 3.3.*

spol	frekvencija	relativna frekvencija
Ž	6	$6/10 = 0.6 = 60\%$
M	4	$4/10 = 0.4 = 40\%$

Tablica 3.2: Tablica frekvencija i relativnih frekvencija svih kategorija varijable spol.

krvna grupa	frekvencija	relativna frekvencija
A	3	$3/10 = 0.3 = 30\%$
B	3	$3/10 = 0.3 = 30\%$
AB	2	$2/10 = 0.2 = 20\%$
0	2	$2/10 = 0.2 = 20\%$

Tablica 3.3: Tablica frekvencija i relativnih frekvencija svih kategorija varijable krvna grupa.

Primjer 3.3. *Od velike su važnosti u mnogim istraživanjima i kategorizirane tablice frekvencija i relativnih frekvencija. Frekvencije i relativne frekvencije za izmjerene vrijednosti varijable* krvna grupa *iz primjera 3.1 kategorizirane prema spolu ispitanika dane su u tablicama 3.4 (za ženski spol) i 3.5 (za muški spol).*

spol = Ž		
krvna grupa	frekvencija	relativna frekvencija
A	2	2/6
B	2	2/6
AB	1	1/6
0	1	1/6

Tablica 3.4: Frekvencije i relativne frekvencije krvnih grupa za ženski spol.

spol = M		
krvna grupa	frekvencija	relativna frekvencija
A	1	$1/4 = 0.25 = 25\%$
B	1	$1/4 = 0.25 = 25\%$
AB	1	$1/4 = 0.25 = 25\%$
0	1	$1/4 = 0.25 = 25\%$

Tablica 3.5: Frekvencije i relativne frekvencije krvnih grupa za muški spol.

Na temelju prethodnih dviju tablica i tablica iz primjera 3.2 možemo redom odgovoriti na pitanja postavljena u primjeru 3.1:

U uzorku ima šest ispitanika ženskog spola (tj. frekvencija žena u uzorku je šest).

U uzorku ima 20% *ispitanika s krvnom grupom* 0 *(tj. relativna frekvenicja krvne grupe nula u uzorku je* 20%*).*

U uzorku ima dvije žene s krvnom grupom A (tj. frekvencija žena s krvnom grupom A u uzorku je dva).

Od svih ispitanika muškog spola njih 50% *ima krvnu grupu B ili AB.*

Primjer 3.4. (krvne-grupe.sta)
U ovom primjeru naučit ćemo kako bazu podataka te tablice frekvencija i relativnih frekvencija napraviti u programskom paketu Statistica. Rezultat postupka u tom programskom paketu prikazan je za varijable krvna grupa *i* spol *iz primjera 3.1, tj. iz baze podataka* krvne-grupe.sta. *Tablične prikaze frekvencija i relativnih frekvencija u programskom paketu Statistica možemo dobiti provodeći sljedeći postupak (koji provodimo slijedeći navedeni niz opcija u izborniku):*

Statistics → Basic Statistics/Tables → Freq. Tables → Variables → Summary.

Rezultat provedbe prethodnog postupka jesu tablice prikazane na slici 3.1.

	Frequency table: krvna_grupa (krvne-grupe.sta)			
Category	Count	Cumulative Count	Percent	Cumulative Percent
0	2	2	20,00	20,00
A	3	5	30,00	50,00
B	3	8	30,00	80,00
AB	2	10	20,00	100,00
Missing	0	10	0,00	100,00

(a) krvna grupa

	Frequency table: spol (krvne-grupe.sta)			
Category	Count	Cumulative Count	Percent	Cumulative Percent
Ž	6	6	60,00	60,00
M	4	10	40,00	100,00
Missing	0	10	0,00	100,00

(b) spol

Slika 3.1: Frekvencije i relativne frekvencija svih kategorija varijabli krvna grupa i spol.

Promatranje vrijednosti varijable spol *kategorizirane prema krvnoj grupi ispitanika omogućuju kategorizirane tablice frekvencija i relativnih frekvencija. Za izradu takvih tablica podatke iz varijabli od interesa moramo profiltrirati, tj. moramo zadati uvjet prema kojemu će u daljnju analizu biti uključena samo uvjetom određena kategorija podataka. Kategorizirane tablice frekvencija i relativnih frekvencija u programskom paketu Statistica možemo dobiti provodeći sljedeći postupak:*

Selection → označiti Enable Selection Conditions → pod Include Cases odabrati opciju "Specific, selected by expression" (u polje za unos teksta upisati krvna grupa="A" ako želimo u obzir uzeti samo ispitanike s krvnom grupom A; analogno se postavlja uvjet krvna grupa="B" za krvnu grupu B, krvna grupa="AB" za krvnu grupu AB, krvna grupa="0" za krvnu grupu 0) → OK.

Rezultat provedbe prethodnog postupka jesu tablice prikazane na slici 3.2.

Category	Frequency table: spol (krvne-grupe.sta) Include condition: krvna_grupa="A"			
	Count	Cumulative Count	Percent	Cumulative Percent
Ž	2	2	66,67	66,67
M	1	3	33,33	100,00
Missing	0	3	0,00	100,00

(a) kategorija: krvna grupa A

Category	Frequency table: spol (krvne-grupe.sta) Include condition: krvna_grupa="B"			
	Count	Cumulative Count	Percent	Cumulative Percent
Ž	2	2	66,67	66,67
M	1	3	33,33	100,00
Missing	0	3	0,00	100,00

(b) kategorija: krvna grupa B

Category	Frequency table: spol (krvne-grupe.sta) Include condition: krvna_grupa="AB"			
	Count	Cumulative Count	Percent	Cumulative Percent
Ž	1	1	50,00	50,00
M	1	2	50,00	100,00
Missing	0	2	0,00	100,00

(c) kategorija: krvna grupa AB

Category	Frequency table: spol (krvne-grupe.sta) Include condition: krvna_grupa=0			
	Count	Cumulative Count	Percent	Cumulative Percent
Ž	1	1	50,00	50,00
M	1	2	50,00	100,00
Missing	0	2	0,00	100,00

(d) kategorija: krvna grupa 0

Slika 3.2: Frekvencije i relativne frekvencije kategorija varijable spol za krvne grupe A, B, AB i 0.

3.1.2 Grafički prikazi frekvencija i relativnih frekvencija

Frekvencije i relativne frekvencije kategorija kvalitativnih varijabli grafički prikazujemo korištenjem **stupčastog dijagrama** (eng. Bar Chart ili Bar Plot) frekvencija i stupčastog dijagrama relativnih frekvencija. U istu svrhu može se koristiti i **kružni dijagram** (eng. Pie Chart) frekvencija i relativnih frekvencija . Popularni naziv za isti grafički prikaz je "pita").

Primjer 3.5. (hormon.sta)
Grafičke prikaze frekvencija i relativnih frekvencija kvalitativnih varijabli prikazat ćemo na primjeru varijable dijagnoza *iz baze podataka* hormon.sta *(koja je opisana u zadatku 3.1). Stupčasti dijagram frekvencija u programskom paketu Statistica možemo dobiti provodeći sljedeći postupak:*

Statistics → Basic Statistics/Tables → Frequency Tables → Choose variables → Histograms.

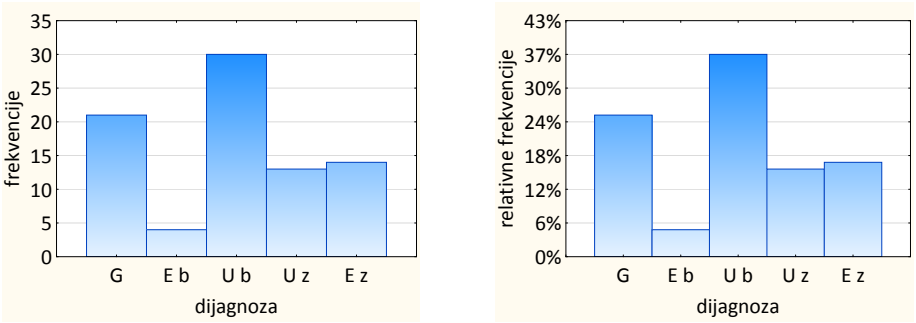
Stupčasti dijagram koji prikazuje i frekvencije i relativne frekvencije u programskom paketu Statistica možemo dobiti provodeći sljedeći postupak:

Graphs → Histograms → Choose variables → Advanced → Pod "Y axis" uključiti "% and N" → OK.

Stupčasti dijagrami frekvencija i relativnih frekvencija svih kategorija varijable dijagnoza *prikazani su na slici 3.3. Drugi način grafičkog prikazivanja mjera zastupljenosti pojedinih kategorija neke kvalitativne varijable u uzorku jesu kružni dijagrami frekvencija i relativnih frekvencija koje u programskom paketu Statistica možemo dobiti provodeći sljedeći postupak:*

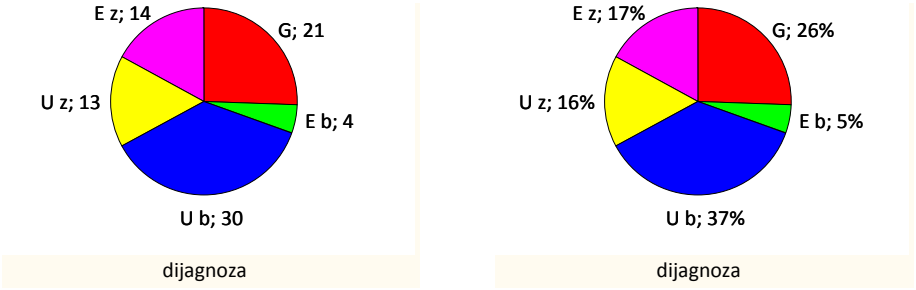
Graphs → 2D Graphs → Graph type (opcija "Pie Chart - Counts") → Choose variables → Advanced → Pie Legend - odabrati opciju "Text and Value" za kružni dijagram frekvencija, a opciju "Text and Percent" za kružni dijagram relativnih frekvencija → OK.

Kružni dijagrami frekvencija i relativnih frekvencija kategorija varijable dijagnoza *prikazani su na slici 3.4.*



(a) frekvencije (b) relativne frekvencije

Slika 3.3: Stupčasti dijagrami frekvencija i relativnih frekvencija svih kategorija varijable dijagnoza.



(a) frekvencije (b) relativne frekvencije

Slika 3.4: Kružni dijagrami frekvencija i relativnih frekvencija svih kategorija varijable dijagnoza.

Primjer 3.6. (djelatnici.sta)
Često se u praksi pokazuje korisnim poznavanje zastupljenosti kategorija jedne varijable za svaku od kategorija neke druge kvalitativne varijable proučavane na istom uzorku. U ovom ćemo primjeru tablično i grafički prikazati frekvencije i relativne frekvencije svih kategorija varijable obrazovanje *iz baze podataka* djelatnici.sta *opisane u primjeru 2.4 posebno za ispitanike ženskog spola, a posebno za ispitanike muškog spola. Tablice tako kategoriziranih frekvencija i relativnih frekvencija varijable* obrazovanje *prikazane su u tablici 3.5.*

Category	Frequency table: obrazovanje (djelatnici.sta) Include condition: spol="Z"			
	Count	Cumulative Count	Percent	Cumulative Percent
SSS	21	21	51.22	51.22
VŠSS	18	39	43.90	95.12
VSS	2	41	4.88	100.00
Missing	0	41	0.00	100.00

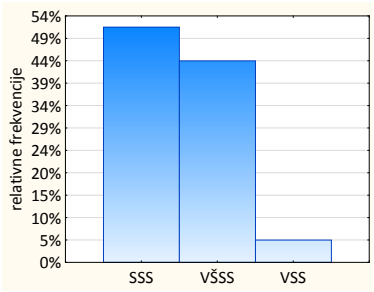
(a) spol = Z

Category	Frequency table: obrazovanje (djelatnici.sta) Include condition: spol="M"			
	Count	Cumulative Count	Percent	Cumulative Percent
SSS	30	30	50.85	50.85
VŠSS	25	55	42.37	93.22
VSS	4	59	6.78	100.00
Missing	0	59	0.00	100.00

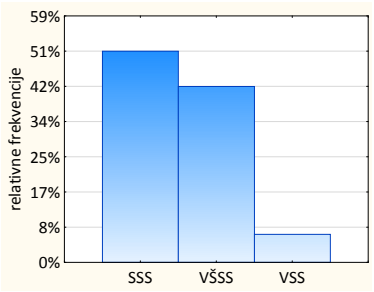
(b) spol = M

Slika 3.5: Tablica frekvencija i relativnih frekvencija svih kategorija varijable obrazovanje posebno za svaku kategoriju varijable spol.

Stupčasti dijagrami frekvencija i relativnih frekvencija svih kategorija varijable obrazovanje *za kategorije* Z *i* M *varijable* spol *prikazani su na slici 3.6, a kružni dijagramovi na slici 3.7.*

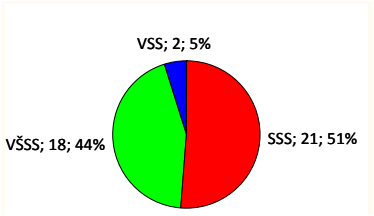


(a) spol=Z

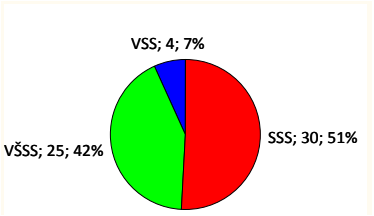


(b) spol=M

Slika 3.6: Stupčasti dijagrami relativnih frekvencija svih kategorija varijable obrazovanje posebno za svaku kategoriju varijable spol.



(a) spol=Z



(b) spol=M

Slika 3.7: Kružni dijagram frekvencija i relativnih frekvencija svih kategorija varijable obrazovanje posebno za svaku kategoriju varijable spol.

3.2 Metode opisivanja numeričkih podataka

Numerički podaci mogu biti prikupljeni promatranjem (mjerenjem) numeričke ili ordinalne varijable. Ordinalne varijable najčešće se zadaju tako da mogu primiti samo nekoliko međusobno različitih vrijednosti, dok kod numeričkih varijabli to vrlo često nije slučaj. Numeričke varijable, po svojoj prirodi, mogu biti diskretne ili neprekidne, kao što je opisano u poglavlju 2.3.2. U oba slučja, a posebno kod neprekidnih varijabli, može se dogoditi da u prikupljenim podacima postoji mnogo međusobno različitih vrijednosti. U takvim slučajevima tablični i grafički prikazi uvedeni za kvalitativne varijable mogu biti nedovoljno informativni. Ilustracija tog problema dana je sljedećim primjerom.

Primjer 3.7. (cijena.sta, hormon.sta, komarci.sta, matematika.sta)

Baza podataka cijena.sta *sadrži informacije o prodajnim mjestima (varijabla* trgovina*) i cijenama nekog proizvoda na tim prodajnim mjestima (varijabla* cijena*). Evidentirane vrijednosti obje varijable jesu brojevi, ali varijabla* trgovina *je, po svojoj prirodi, kvalitativna, a varijabla* cijena *neprekidna. Uočite da su svi prikupljeni podaci za varijablu* cijena *međusobno različiti.*

U bazi podataka komarci.sta *(opisanoj u zadatku 3.1) varijable* brojM *i* brojZ *su diskretne numeričke varijable, a varijable* temperatura *i* rel-vlaznost *neprekidne numeričke varijable. Uočite da se u podacima za sve te varijable pojavljuje mnogo međusobno različitih vrijednosti.*

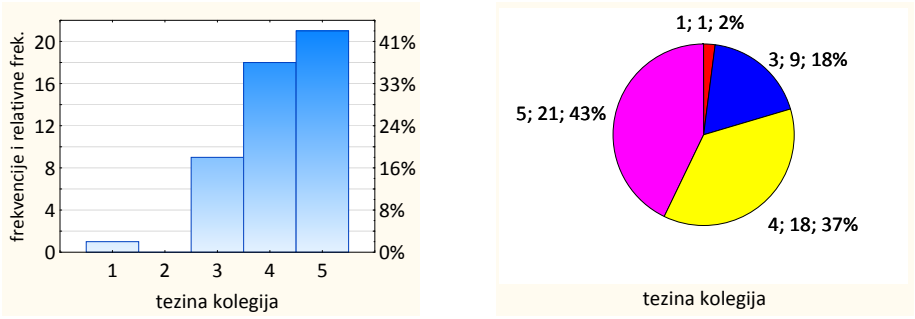
Ako su numeričke varijable diskretne s malo mogućih vrijednosti ili ako su varijable ordinalne, za opis podataka možemo koristiti iste metode kao pri opisivanju kvalitativnih podataka, tj. frekvencije i relativne frekvencije te ih grafički prikazivati stupčastim dijagramima i kružnim dijagramima.

Primjer 3.8. (matematika.sta)

Tablični i grafički prikazi (stupčasti dijagram i kružni dijagram) frekvencija i relativnih frekvencija svih vrijednosti ordinalne varijable tezina-kolegija *prikazani su na slikama 3.8 i 3.9.*

	Frequency table: tezina kolegija (matematika.sta)			
Category	Count	Cumulative Count	Percent	Cumulative Percent
1	1	1	2.04	2.04
3	9	10	18.37	20.41
4	18	28	36.73	57.14
5	21	49	42.86	100.00
Missing	0	49	0.00	100.00

Slika 3.8: Tablica frekvencija i relativnih frekvencija za varijablu tezina-kolegija.



(a) stupčasti dijagram (b) kružni dijagram

Slika 3.9: Grafički prikazi frekvencija i relativnih frekvencija za varijablu tezina-kolegija.

Iz prikazanih opisa varijable tezina-kolegija *možemo dobiti npr. sljedeće informacije:*

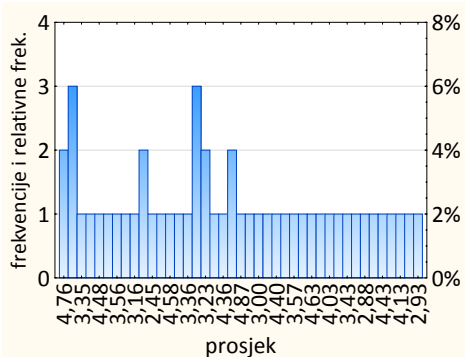
Ocjenom većom od 3 *težinu kolegija ocijenilo je čak* 39 *ispitanika, tj. čak* $39/49 \approx 79.59\%$ *od ukupnog broja ispitanika.*

Ocjenom 3 *težinu kolegija ocijenilo je* 9 *(*$9/49 \approx 18.37\%$*), a ocjenom* 4 *čak* 18 *(*$18/49 \approx 36.73\%$*) ispitanika. Dakle, dvostruko više ispitanika težinu kolegija ocijenilo je ocjenom* 4 *nego ocjenom* 3*.*

U sljedećem primjeru prikazano je šta se događa ako koristimo uobičajeni stupčasti dijagram za prikazivanje numeričkih podataka među kojima ima velik broj različitih vrijednosti.

Primjer 3.9. (matematika.sta)
Stupčasti dijagram za podatke neprekidne numeričke varijable prosjek *iz baze podataka* matematika.sta *(vidi primjer 2.9) prikazan je na slici 3.10. Pri opisivanju ove varijable pretpostavili smo da svi međusobno različiti podaci varijable* prosjek *čine zasebne kategorije. Zbog velikog broja različitih podataka broj kategorija je prevelik i rezultat analize grafičkog prikaza 3.10 ne daje željene informacije.*

Radi dobivanja korisnijih stupčastih i kružnih dijagrama za podatke iz neprekidnih numeričkih varijabli vrijednosti je potrebno **kategorizirati**, tj. razvrstati ih u odabrane kategorije. Pri tome podatke kategoriziramo u disjunktne intervale po kriteriju za koji smatramo da će nam dati željene rezultate. Za potrebe opisivanja skupa podataka obično biramo disjunktne intervale tako da dobivenim tabličnim i grafičkim prikazima možemo ilustrirati karakteristike skupa podataka koje želimo naglasiti.



Slika 3.10: Stupčasti dijagram za podatke varijable prosjek.

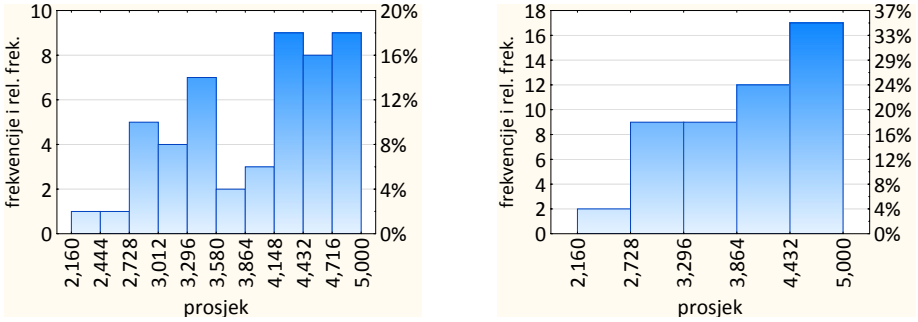
3.2.1 Postupak razvrstavanja numeričkih podataka u kategorije

Razvrstavanje vrijednosti neprekidne numeričke varijable u kategorije moguće je provesti na nekoliko načina, npr. moguće je skup svih podataka (ili nešto veći skup koji sadrži skup svih podataka, ali koji je jednostavnije podijeliti na jednake dijelove) podijeliti na disjunktne intervale jednake duljine. No, nije nužno da su intervali jednake duljine, tj. nema točno definiranog pravila po kojemu bi trebalo definirati duljine intervala niti nihov broj, ali je jasno da ih ne smije biti ni previše ni premalo da bi cijeli postupak imao smisla i služio svrsi (a to je u ovom času prikazivanje skupa podataka).

Za prikaz frekvencija ili relativnih frekvencija tako kategoriziranih podataka možemo koristiti i specifičan stupčasti dijagram koji zovemo **histogram**. Histogram mora imati stupce postavljene u koordinatni sustav nad odgovarajućim intervalima. Širina svakog stupca histograma odgovara duljini odgovarajućeg intervala, a visina frekveniciji, odnosno relativnoj frekvenciji intervala.

Primjer 3.10. (matematika.sta)
Primjerom 3.9 pokazali smo da je teško analizirati varijablu prosjek *iz baze* matematika.sta *ako za kategorije uzmemo sve različite izmjerene vrijednosti te varijable. Stoga ćemo provesti kategorizaciju izmjerenih vrijednosti.*

Dva primjera kategorizacije, tj. podjele izmjerenih vrijednosti u disjunkte intervale, rezultiraju histogramima prikazanim na slici 3.11.

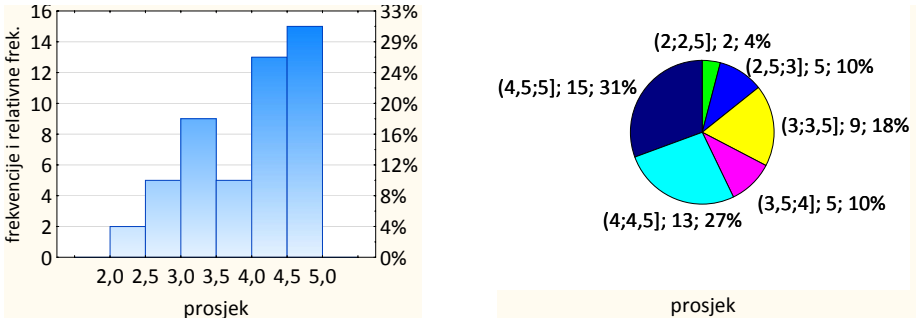


(a) kategorizacija na 10 disjunktnih intervala

(b) kategorizacija na 5 disjunktnih intervala

Slika 3.11: Stupčasti dijagrami za podatke varijable prosjek.

Kriterij kategorizacije treba biti prilagođen zahtjevima istraživanja, tj. treba omogućiti dobivanje odgovora na postavljena pitanja. Npr. ako nas zanima zastupljenost studenata s prosjekom većim od 3.5 *u promatranom uzorku, tada podatke iz varijable* prosjek *možemo kategorizirati u šest disjunktnih intervala duljine* 0.5*, počevši od* 2.0*. Iz grafičkih prikaza sa slike 3.12 očitavamo da je frekvencija takvih studenata* 33*, a relativna frekvencija* $33/49 \approx 67.35\%$.



Slika 3.12: Stupčasti i kružni dijagram za podatke varijable prosjek razvrstane u 6 disjunktnih intervala počevši od ocjene 2.0.

3.2.2 Mjere centralne tendencije i raspršenosti podataka

Karakteristika numeričkih i ordinalnih varijabli jest da među njihovim vrijednostima postoji prirodan uređaj. Na osnovi te činjenice možemo definirati numeričke karakteristike podataka iz tih varijabli koje imaju logičnu interpretaciju i mogu se iskoristiti za prikazivanje skupa podataka. U ovom poglavlju navodimo osnovne numeričke karakteristike skupa podataka te primjerima ilustriramo njihovu inter-

pretaciju u praktičnim problemima.

Aritmetička sredina

Aritmetička sredina (eng. arithmetic mean) niza podataka $x_1, x_2, \ldots, x_n$ iz varijable X definirana je izrazom

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

Aritmetička sredina je numerička karakteristika koja spada u mjere centralne tendencije, tj. ona mjeri "srednju vrijednost" podataka.

Primjer 3.11. *Neka su izmjerene vrijednosti jedne varijable sljedeće:*

$$1.2, 2.1, 3.2, 4.3, 5.4, 6.5, 7.6, 8.7, 9.8.$$

S obzirom da ih ima ukupno devet, aritmetička sredina ovog skupa izmjerenih vrijednosti je

$$\frac{1.2 + 2.1 + 3.2 + 4.3 + 5.4 + 6.5 + 7.6 + 8.7 + 9.8}{9} \approx 5.42.$$

Medijan

Da bismo razumjeli i odredili medijan potrebno je prvo poredati izmjerene vrijednosti $x_1, x_2, \ldots, x_n$ varijable X po veličini (u rastućem poretku, tj. od manjeg prema većem). Medijan je također jedna mjera centralne tendencije kao i aritmetička sredina, a karakterizira ga činjenica da je barem pola podataka manje ili jednako medijanu, a istovremeno je barem pola podataka veće ili jednako od medijana. Način njegova izračuna ovisi o tome imamo li **neparan** ili **paran** broj podataka. Ako imamo **neparan broj** podataka, onda postoji vrijednost koja je na srednjoj poziciji u uređenom skupu podataka pa nju definiramo kao medijan.

Primjer 3.12. *Neka su izmjerene vrijednosti jedne varijable sljedeće:*

$$1, 2, 5, 6, 5, 1, 2, 7, 2, 2, 3.$$

Prvo ove vrijednosti poredamo po veličini:

$$1, 1, 2, 2, 2, \mathbf{2}, 3, 5, 5, 6, 7.$$

S obzirom da ih ima ukupno jedanaest, medijan je vrijednost koja je na šestoj poziciji u tako dobivenom nizu, tj. broj 2.

Ako imamo **paran broj** podataka, onda ne postoji podatak koji je na srednjoj poziciji jer srednju poziciju "zauzimaju" dva podatka. Zapravo, zahtjev na temelju kojega želimo odrediti medijan ispunjavaju svi brojevi iz intervala čije su granice dva srednja podatka. Da bismo jedinstveno odredili medijan podataka, u tom ga slučaju definiramo kao broj na polovini tog intervala, tj. kao aritmetičku sredinu tih dvaju podataka.

Primjer 3.13. *Neka su izmjerene vrijednosti jedne varijable sljedeće:*

$$1, 2, 5, 6, 5, 1, 2, 7, 2, 2, 3, 3.$$

Prvo ove vrijednosti poredamo po veličini:

$$1, 1, 2, 2, 2, \mathbf{2}, \mathbf{3}, 3, 5, 5, 6, 7.$$

S obzirom da ih ima dvanaest, "sredinu" čine šesti i sedmi podatak, tj. brojevi 2 *i* 3*. Medijan ovog skupa podataka je aritmetička sredina ta dva broja, tj. medijan je* $(2 + 3)/2 = 2.5$.

Postotna vrijednost, donji i gornji kvartil

Medijan je karakteriziran činjenicom da je barem pola (50%) podataka manje ili jednko od medijana, dok je istovremeno i barem 50% podataka veće ili jednako njemu. Analognim rezoniranjem karakterizirat ćemo postotnu vrijednost. Postotna vrijednost (eng. percentile value) za neki izabrani broj $p \in \langle 0, 100\rangle$, označimo je s x'_p, definira se poštujući zahtjev da je barem $p\%$ izmjerenih vrijednosti manje ili jednako x'_p, dok je barem $(100 - p)\%$ vrijednosti veće ili jednako x'_p. Dvadeset pet postotna vrijednost zove se donji kvartil (eng. lower quartile), a sedamdeset pet postotna vrijednost zove se gornji kvartil (eng. upper quartile). Donji i gornji kvartil su mjere koje spadaju u grupu mjera raspršenosti podataka.

Analogno kao i kod određivnja medijana, navedena karakterizacija postotne vrijednosti često ne određuje postotnu vrijednost podataka jedinstveno, tj. često postoji cijeli intarval realnih brojeva koji zadovoljava zadani kriterij. Predloženo je nekoliko metoda za određivanje postotne vrijednosti u takvim slučajevima. Programski paket *Statistica* u inačici 10 nudi šest načina računanja postotne vrijednosti čiji opis zainteresirani čitatelj može naći u elektronskom priručniku programskog paketa. Jedan od tih načina navodimo u nastavku teksta.

Postupak računanja postotne vrijednosti

Pretpostavimo da imamo n podataka i da želimo odrediti p-tu postotnu vrijednost x'_p, $p \in \langle 0, 100\rangle$. Prvo je potrebno podatke poredati u rastućem poretku i odrediti "poziciju" j koja je ključna za određivanje zadanog percentila kao $j = np/100$. Ako j nije prirodan broj, onda podatak na poziciji $j + 1$ odgovara p-toj postotnoj vrijednosti. Ako je j prirodan broj onda, se p-ta postotna vrijednost računa kao aritmetička sredina podataka na pozicijama j i $j + 1$.

Primjer 3.14. *Neka su izmjerene vrijednosti jedne varijable sljedeće:*

$$1, 2, 5, 6, 6, 1, 3, 7, 3, 3, 3, 3.$$

Prvo ove vrijednosti poredamo po veličini:

$$1, 1, 2, 3, 3, 3, 3, 3, 5, 6, 6, 7.$$

Želimo li odrediti donji kvartil, potrebno je prvo odrediti četvrtinu podataka (25%). S obzirom da imamo 12 podataka, četvrtinu (25%) čine tri podatka. Treći podatak u gornjem skupu je broj 2, a četvrti 3. Donji kvartil je 2.5. Deveti broj u gornjem skupu podataka je broj 5, a deseti 6 pa je gornji kvartil 5.5.

Najmanja i najveća vrijednost, raspon podataka

Raspon (eng. range) podataka je mjera koja pokazuje koliko su podaci raspršeni, tj. to je jedna od mjera raspršenosti podataka. Definiran je kao razlika najveće i najmanje vrijednosti u skupu mjerenih vrijednosti varijable (tj. razlika maksimalne i minimalne izmjerene vrijednosti varijable). Ako su $x_1, x_2, \ldots, x_n$ izmjerene vrijednosti varijable X, označimo najmanju od njih (minimum) s $x_{\min}$, a najveću s $x_{\max}$.

Primjer 3.15. *Neka su izmjerene vrijednosti jedne varijable sljedeće:*

$$1, 2, 5, 6, 5, 1, 2, 7, 2, 2, 3, 3.$$

Vidimo da je vrijednost 1 najmanja izmjerena vrijednost, a 7 najveća. Prema tome, raspon ovog skupa izmjerenih vrijednosti je $7 - 1 = 6$.

U mnogim primjerima zanimljivo je promatrati **maksimalno odstupanje izmjerenih vrijednosti varijable od "prosjeka", tj. aritmetičke sredine**, izmjerenih vrijednosti. Ta je numerička karakteristika definirana kao veći od brojeva $(\bar{x}_n - x_{\min})$ i $(x_{\max} - \bar{x}_n)$, tj. broj

$$\max\{(\bar{x}_n - x_{\min}), (x_{\max} - \bar{x}_n)\}.$$

Primjer 3.16. *Neka su $1, 2, 5, 6, 5, 1, 2, 7, 2, 2, 3, 3$ izmjerene vrijednosti neke varijable X. Tada je*

$$x_{\min} = 1, \quad x_{\max} = 7, \quad \bar{x}_n = \frac{1+2+5+6+5+1+2+7+2+2+3+3}{12} = 3.25.$$

Maksimalno odstupanje izmjerenih vrijednosti ove varijable od prosjeka izmjerenih vrijednosti je

$$\max\{3.25 - 1, 7 - 3.25\} = \max\{2.25, 3.75\} = 3.75.$$

Varijanca i standardna devijacija

Varijanca i standardna devijacija također spadaju u grupu mjera raspršenosti podataka. One karakteriziraju raspršenost podataka oko aritmetičke sredine. Varijanca niza izmjerenih vrijednosti $x_1, x_2, \ldots, x_n$ varijable X definirana je izrazom:

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$

a standardna devijacija je kvadratni korijen varijance, tj.

$$s_n = \sqrt{s_n^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Primjer 3.17. *Neka su izmjerene vrijednosti jedne varijable sljedeće:*

$$1.2, 2.1, 3.2, 4.3, 5.4, 6.5, 7.6, 8.7, 9.8.$$

Iz primjera 3.11 znamo da je aritmetička sredina ovog skupa podataka približno jednaka 5.42. *Varijanca ovog skupa podataka jest*

$$s_n^2 \approx \frac{1}{9}\sum_{i=1}^{9}(x_i - 5.42)^2 \approx 7.87,$$

a standardna devijacija

$$s_n \approx \sqrt{\frac{1}{9}\sum_{i=1}^{9}(x_i - 5.42)^2} \approx 2.81.$$

Mod

Mod je vrijednost iz niza izmjerenih vrijednosti varijable X kojoj pripada najveća frekvencija, tj. izmjerena je najviše puta. Mod ne mora biti jedinstven.

Primjer 3.18. *Neka su izmjerene vrijednosti jedne varijable sljedeće:*

$$1, 2, 5, 6, 5, 1, 2, 7, 2, 2, 3, 3.$$

Vidimo da je vrijednost 2 *izmjerena najviše puta (četiri puta) pa je* 2 *mod ovog skupa podataka.*

Primjer 3.19. *Neka su izmjerene vrijednosti jedne varijable sljedeće:*

$$1, 2, 5, 6, 5, 3, 1, 2, 7, 2, 2, 3, 3.$$

Vidimo da su najviše puta izmjerene dvije vrijednosi - 2 *i* 3 *su obje izmjerene točno četiri puta. Dakle, mod ovog skupa podataka nije jedinstven. U programskom paketu Statistica za mod ovog skupa izmjerenih vrijednosti pisalo bi* mod = multiple *te bismo u tom slučaju sve vrijednosti moda saznali analizom pripadne tablice frekvencija.*

Korištenjem numeričkih karakteristika podataka skup podataka može se prikazati grafički pomoću **kutijastog dijagrama** (eng. box plot, boxplot ili box-and-whiskers plot).

Kutijastim dijagramom prikazujemo odnos pet numeričkih karakteristika skupa izmjerenih vrijednosti: minimalnu vrijednost, donji kvartil, medijan, gornji kvartil i maksimalnu vrijednost. Na kutijastom dijagramu također se označavaju takozvane stršeće vrijednosti (eng. outliers) ako postoje.

Primjer 3.20. (trgovacki-centri.sta)
Pažljivim proučavanjem kretanja cijena prehrambenih proizvoda analitičar tržišta uočio je da isti proizvodi nemaju jednaku cijenu u različitim trgovačkim centrima. Promatrajući deset trgovačkih centara, zabilježio je cijene proizvoda kod kojega su razlike bile najizraženije (tablica 3.6).

trg. centar	1	2	3	4	5	6	7	8	9	10
cijena	45.52	44.64	39.99	48.95	51.59	46.89	52.02	56.89	50.21	49.99

Tablica 3.6: Cijene jednog proizvoda u deset različitih trgovačkih centara.

Numeričke karakteristike ovog skupa izmjerenih vrijednosti u programskom paketu Statistica možemo izračunati koristeći bazu podataka trgovacki-centri.sta *i provodeći sljedeći postupak:*

Statistics → Basic Statistics/Tables → Descriptive Statistics → Variables → Advanced → označiti mean (aritmetička sredina), mod, range (raspon), variance, standard deviation, median, minimum & maximum i lower & upper quartiles (donji i gornji kvartil) → Summary.

Rezultat ovog postupka (mjere deskriptivne statistike promatranog skupa izmjerenih vrijednosti) jesu tablice prikazane na slici 3.13.

	Descriptive Statistics (trgovacki-centri.sta)						
Variable	Valid N	Mean	Mode	Frequency of Mode	Range	Variance	Std.Dev.
cijena-proizvoda	10	48,66900	Multiple	1	16,90000	21,79821	4,668855

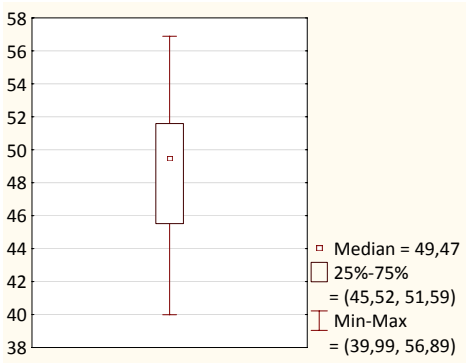
	Descriptive Statistics (trgovacki-centri.sta)						
Variable	Valid N	Median	Minimum	Maximum	Lower Quartile	Upper Quartile	Range
cijena-proizvoda	10	49,47000	39,99000	56,89000	45,52000	51,59000	16,90000

Slika 3.13: Deskriptivna statistika cijena iz tablice 3.6.

Uočimo da mod nije jedinstven - naime sve su izmjerene vrijednosti međusobno različite, tj. svaka je vrijednost izmjerena točno jedanput.

Za analiziranje raspršenosti cijena iz tablice 3.6 korisno je skicirati kutijasti dijagram na bazi medijana (slika 3.14) koji prikazuje odnos numeričkih karakteristika iz donje tablice sa slike 3.13 i koji u programskom paketu Statistica možemo napraviti provodeći sljedeći postupak:

Statistics → Basic Statistics/Tables → Descriptive Statistics → Variables → Options → pod "Options for Box-Whisker Plots" označiti opciju "Median/Quartiles/ Range" → Quick → Box and whisker Plot for all variables.



Slika 3.14: Kutijasti dijagram na bazi medijana za cijene iz tablice 3.6.

3.2.3 Detekcija stršećih vrijednosti

Podatak koji je značajno veći ili manji u odnosu na druge izmjerene vrijednosti jedne varijable nazivamo **stršeća vrijednost** (eng. outlier). Pojavljivanje stršećih vrijednosti najčešće je vezano uz jedan od sljedećih razloga:

- podatak je ili netočno izmjerena ili krivo unesen u bazu podataka
- podatak dolazi iz druge populacije (ne iz populacije koju promatramo u kontekstu problema koji proučavamo) - npr. ako u varijablu čije su izmjerene vrijednosti godišnje plaće 1000 poreznih obveznika u Hrvatskoj upišemo godišnju plaću Microsoftovog managera iz SAD-a, taj će podatak biti stršeća vrijednost
- podatak je točno izmjeren i unesen u bazu, ali predstavlja rijetku pojavu u populaciji - npr. ako se u varijabli čije su izmjerene vrijednosti koncentracije glukoze u krvi za 1000 osoba nađe točno izmjerena vrijednost 46.7, taj ćemo podatak smatrati stršećom vrijednošću jer se radi o vrlo visokoj koncentraciji glukoze koja se rijetko pojavljuje.

Vrlo korisna grafička metoda za detekciju stršećih vrijednosti jest kutijasti dijagram na bazi medijana. U programskom paketu Statistica kutijasti dijagrami osjetljivi na stršeće vrijednosti izrađuju se na sljedeći način:

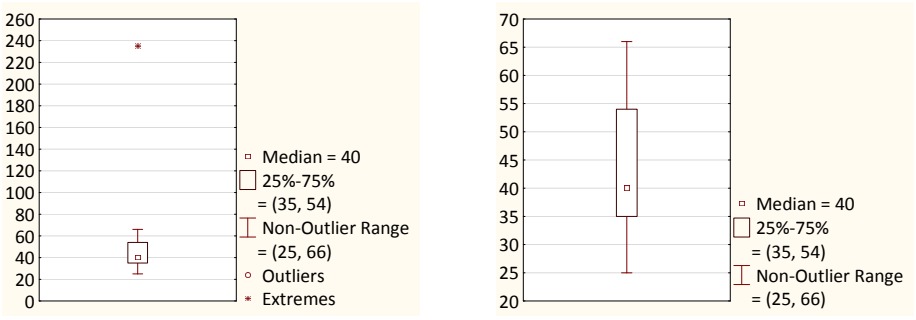
Graphs → 2D Graphs → BoxPlots → Variables → Advanced → pod Whisker odabrati "Non-outlier range" → pod Outliers odabrati "Outl. & Extremes" → OK.

Primjer 3.21. (zdravlje.sta)
Baza podataka zdravlje.sta *sadrži neke zdravstvene podatke za* 51 *ispitanika. Kratkom analizom mjera deskriptivne statistike možemo uočiti da je maksimum skupa izmjerenih vrijednosti* 235*, što u ovom primjeru znači da naš najstariji ispitanik ima* 235 *godina (slika 3.15).*

	Descriptive Statistics (zdravlje.sta)								
Variable	Valid N	Mean	Median	Mode	Frequency of Mode	Minimum	Maximum	Lower Quartile	Upper Quartile
godine	51.00	46.61	40.00	39.00000	7.00	25.00	235.00	35.00	54.00

Slika 3.15: Deskriptivna statistika izmjerenih vrijednosti varijable godine.

Taj je podatak stršeća vrijednost skupa izmjerenih vrijednosti varijable godine*. Međutim, ovaj način analize i detekcije stršećih vrijednosti nije prikladan za velike skupove podataka. Zato za detekciju stršećih vrijednosti često koristimo kutijaste dijagrame. Na slici 3.16 prikazan je kutijasti dijagram za varijablu* godine *sa stršećom vrijednošću te kutijasti dijagram koji dobivamo kad uklonimo stršeće vrijednosti.*



(a) uključena stršeća vrijednost (b) uklonjena stršeća vrijednost

Slika 3.16: Kutijasti dijagrami na bazi medijana za varijablu godine.

Uklanjanjem stršeće vrijednosti mijenjaju se i vrijednosti mjera deskriptivne statistike. Iz tablica sa slike 3.17 vidimo da su se uklanjanjem stršeće vrijednosti aritmetička sredina i gornji kvartil smanjili, dok su mod, medijan i donji kvartil ostali nepromijenjeni. Općenito, uklanjanjem stršećih vrijednosti mod će najčešće ostati nepromijenjen.

	Descriptive Statistics (zdravlje.sta) Include condition: godine < 150								
Variable	Valid N	Mean	Median	Mode	Frequency of Mode	Minimum	Maximum	Lower Quartile	Upper Quartile
godine	50.00	42.84	39.50	39.00000	7.00	25.00	66.00	35.00	53.00

Slika 3.17: Deskriptivna statistika izmjerenih vrijednosti varijable godine nakon uklanjanja stršeće vrijednosti.

3.3 Zadaci

Zadatak 3.1. (hormon.sta, nalaz.sta)
Baza podataka hormon.sta sadrži neke informacije i rezultate nekih medicinskih testova za svakog od 82 ispitanika:

varijabla spol sadrži informaciju o spolu ispitanika (m - ispitanik je muškog spola, z - ispitanik je ženskog spola)

varijable gastrS, somatS i somatZ sadrže izmjerene koncentracije određenih enzima utvrđene prilikom medicinske analize ispitanika

varijable pusenje, alkohol i kava sadrže informaciju o tome konzumira li ispitanik cigarete, alkohol i kavu (0 - ne konzumira, 1 - konzumira)

varijabla CLOtest sadrži rezultate testa na zarazu bakterijom helicobacter pilory (0 - test je negativan, 1 - test je pozitivan)

varijabla dijagnoza sadrži oznake dijagnoze ispitanika.

Baza podataka nalaz.sta sadrži neke informacije i rezultate testova o koncentraciji nekih tvari u krvi za svakog od 102 ispitanika:

varijabla skupina sadrži informaciju o pripadnosti ispitanika jednoj od devet dobnih skupina (g1 - g9)

varijable k1 - k8 sadrže izmjerene koncentracije promatranih tvari u krvi

varijabla stupanj sadrži stupnjevanje rezultata provedenih testova s obzirom na dobnu skupinu kojoj ispitanik pripada (u skali od 1 do 10).

Proučite varijable u prethodno opisanim bazama podataka te pomoću programskog paketa Statistica odredite frekvencije i relativne frekvencije svih kategorija za varijable koje smatrate kvalitativnima. Rezultate prikažite tablično.

Rješenje. *Tablice frekvencija i relativnih frekvencija za kvalitativne varijable s najvećim brojem kategorija - varijable* dijagnoza *iz baze podataka* hormon.sta *i varijable* stupanj *iz baze podataka* nalaz.sta *prikazane su na slici 3.18.*

	Frequency table: dijagnoza (hormon.sta)			
Category	Count	Cumulative Count	Percent	Cumulative Percent
G	21	21	25,61	25,61
E b	4	25	4,88	30,49
U b	30	55	36,59	67,07
U z	13	68	15,85	82,93
E z	14	82	17,07	100,00
Missing	0	82	0,00	100,00

(a) varijabla dijagnoza (hormon.sta)

	Frequency table: stupanj (nalaz.sta)			
Category	Count	Cumulative Count	Percent	Cumulative Percent
1	12	12	11,76	11,76
2	11	23	10,78	22,55
3	12	35	11,76	34,31
4	9	44	8,82	43,14
5	11	55	10,78	53,92
6	10	65	9,80	63,73
7	12	77	11,76	75,49
8	8	85	7,84	83,33
9	8	93	7,84	91,18
10	9	102	8,82	100,00
Missing	0	102	0,00	100,00

(b) varijabla stupanj (nalaz.sta)

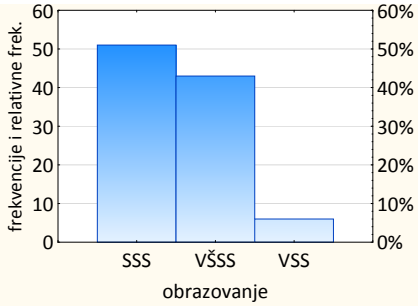
Slika 3.18: Frekvencije i relativne frekvencije svih kategorija varijabli dijagnoza i stupanj.

Zadatak 3.2. (djelatnici.sta)
Baza podataka djelatnici.sta opisana je u primjeru 2.4. Za kvalitativnu varijablu obrazovanje, čije su vrijednosti svrstane u tri kategorije: SSS - srednja stručna sprema, VŠSS - viša stručna sprema, VSS - visoka stručna sprema, odredite zastupljenost tih kategorija u promatranom uzorku od 100 djelatnika.

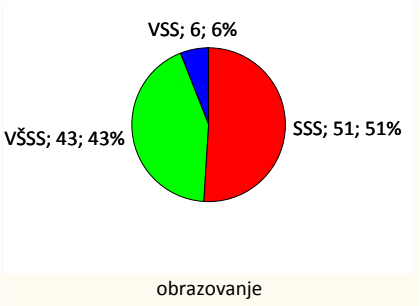
Rješenje. *Zastupljenost kategorija opisana je tablicom frekvencija i relativnih frekvencija 3.19 te stupčastim dijagramom i kružnim dijagramom frekvencija i relativnih frekvencija koji su prikazani na slici 3.20.*

	Frequency table: obrazovanje (djelatnici.sta)			
Category	Count	Cumulative Count	Percent	Cumulative Percent
SSS	51	51	51.00	51.00
VŠSS	43	94	43.00	94.00
VSS	6	100	6.00	100.00
Missing	0	100	0.00	100.00

Slika 3.19: Frekvencije i relativne frekvencije svih kategorija varijabli obrazovanje.



(a) stupčasti dijagram



(b) kružni dijagram

Slika 3.20: Grafički prikazi podataka varijable obrazovanje.

Zadatak 3.3. (nalaz.sta)
U bazi podataka nalaz.sta (opisanoj u zadatku 3.1) odredite frekvencije i relativne frekvencije svih kategorija za varijable koje smatrate kvalitativnima.

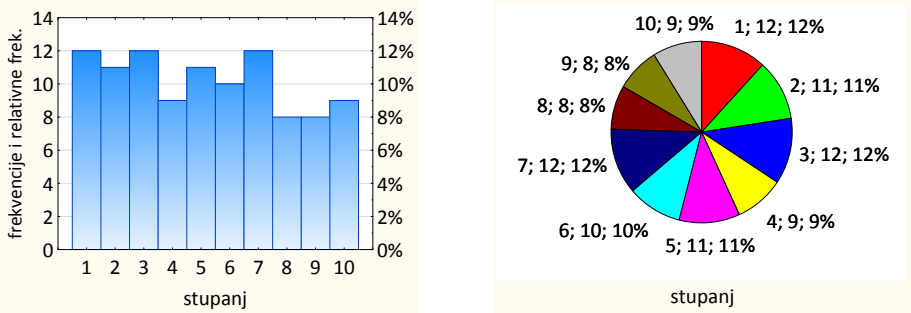
a) Rezultate prikažite grafički koristeći programski paket Statistica.

b) Za koliko je ispitanika vrijednost varijable stupanj manja od tri, za koliko je vrijednost barem četiri, ali manja od sedam, a za koliko je vrijednost barem osam?

c) Za frekvencije iz zadatka b) odredite pripadne relativne frekvencije.

Rješenje.

a) Grafički prikazi frekvencija i relativnih frekvencija kategorija kvalitativne varijable stupanj *prikazani su na slici 3.21.*



Slika 3.21: Grafički prikazi frekvencija i relativnih frekvencija svih kategorija varijable stupanj.

b) Frekvencija ispitanika za koje je vrijednost varijable stupanj *manja od tri je* 23*, frekvencija ispitanika za koje je vrijednost barem četiri, ali manja od sedam je* 30*, a frekvencija ispitanika za koje je vrijednost barem osam je* 25*.*

c) Pripadne relativne frekvencije su redom $23/102 \approx 22.55\%$*,* $30/102 \approx 29.41\%$ *i* $25/102 \approx 24.51\%$*.*

Zadatak 3.4. (djeca.sta)

U bazi podataka djeca.sta nalazi se dio podataka o nekim ocjenama novorođenčeta, načinu poroda i majci iz istraživanja koje je provedeno u jednoj bolnici:

varijabla spol sadrži spol novorođenčeta

varijabla nacin-poroda informaciju o načinu poroda

varijable RM, apgar1 i apgar5 izmjerene vrijednosti nekih obilježja novorođenčeta

varijabla majka-dob godine starosti majke

varijabla majka-bolest informaciju o bolesti majke tijekom trudnoće (N - nije bila bolesna, D - bila je bolesna)

varijabla komplikacije stupanj komplikacija za vrijeme trudnoće (u skali od 0, što označava da komplikacija nije bilo, do 7)

varijabla konvulzije informaciju o konvulzijama kod novorođenčeta (N - konvulzija nije bilo, D - konvulzije su bile prisutne)

varijabla uzv jednu ocjenu ultrazvučnog pregleda mozga novorođenčeta (u skali od 1 do 4).

Odredite frekvencije i relativne frekvencije svih kategorija za varijable koje smatrate kvalitativnima.

a) Rezultate prikažite tablično i grafički koristeći programski paket Statistica.

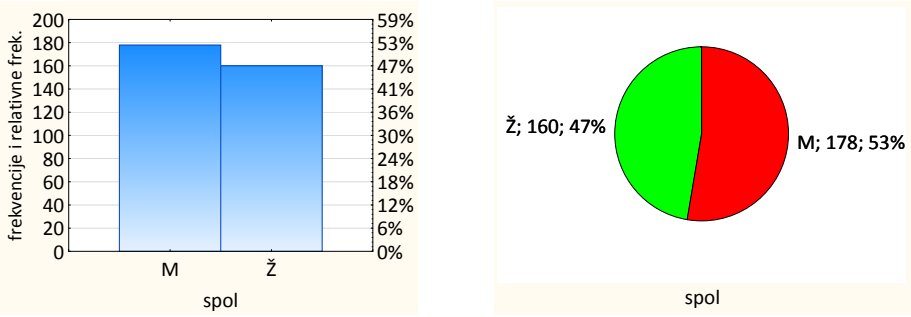
b) Broji li ovaj uzorak više djevojčica ili dječaka? Koliki je udio majki starijih od 35 godina?

Rješenje.

a) Tablični i grafički prikazi frekvencija i relativnih frekvencija svih kategorija varijable spol *prikazani su na slikama 3.22 i 3.23.*

	Frequency table: spol (djeca.sta)			
Category	Count	Cumulative Count	Percent	Cumulative Percent
M	178	178	52,66	52,66
Ž	160	338	47,34	100,00
Missing	0	338	0,00	100,00

Slika 3.22: Tablica frekvencija i relativnih frekvencija svih kategorija varijable spol.



(a) stupčasti dijagram (b) kružni dijagram

Slika 3.23: Grafički prikazi frekvencija i relativnih frekvencija svih kategorija varijable spol.

b) Uzorkom je obuhvaćeno 338 *novorođenčadi* - 160 *djevojčica i* 178 *dječaka. Dakle, u uzorku ima više dječaka. Majki starijih od* 35 *godina ima* $29/338 \approx 8.58\%$.

Zadatak 3.5. (navike.sta)
U bazi podataka navike.sta (opisanoj u zadatku 2.4) odredite frekvencije i relativne frekvencije svih kategorija za varijable koje smatrate kvalitativnima.

a) Rezultate prikažite tablično i grafički koristeći programski paket Statistica.

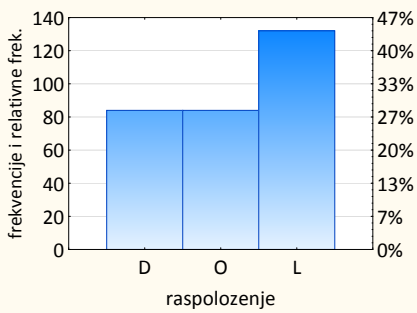
b) Koliko je ispitanika dobro raspoloženo? Je li više ispitanika raspoloženo dobro ili osrednje ili ih je najviše lošeg raspoloženja?

Rješenje.

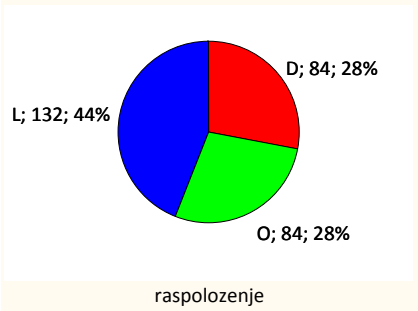
a) Tablični i grafički prikazi frekvencija i relativnih frekvencija svih kategorija varijable raspolozenje *prikazani su na slikama 3.24 i 3.25.*

Category	Frequency table: raspolozenje (navike.sta)			
	Count	Cumulative Count	Percent	Cumulative Percent
D	84	84	28,00	28,00
O	84	168	28,00	56,00
L	132	300	44,00	100,00
Missing	0	300	0,00	100,00

Slika 3.24: Tablica frekvencija i relativnih frekvencija svih kategorija varijable raspolozenje.



(a) stupčasti dijagram



(b) kružni dijagram

Slika 3.25: Grafički prikazi frekvencija i relativnih frekvencija svih kategorija varijable raspolozenje.

b) Uzorkom je obuhvaćeno 300 *ispitanika. Dobro je raspoloženo njih* 84*, što čini* $84/300 = 28\%$ *od ukupnog broja ispitanika. Osrednje je raspoloženo također* 84 *(*28%*) ispitanika, a loše njih* 132 *(*44%*). Dakle, više je ispitanika koji su raspoloženi dobro ili osrednje - u te dvije kategorije spada* 168 *(56 %) ispitanika.*

Zadatak 3.6. (zdravlje.sta)
Često ima smisla analizirati frekvencije i relativne frekvencije numeričkih ili ordinalnih varijabli za pojedine kategorije zadane kvalitativne varijable. Na primjer, korisno je analizirati određene zdravstvene karakteristike posebno za osobe ženskog, a posebno za osobe muškog spola. Analizirajte ordinalnu varijablu zdravlje po kvalitativnoj varijabli spol iz baze podataka zdravlje.sta koja je opisana u zadatku 2.4.

Rješenje. *Prvo ćemo tablično i grafički prikazati frekvencije i relativne frekvencije za podatke sadržane u varijablama* zdravlje *i* spol *(slike 3.26, 3.27 i 3.28).*

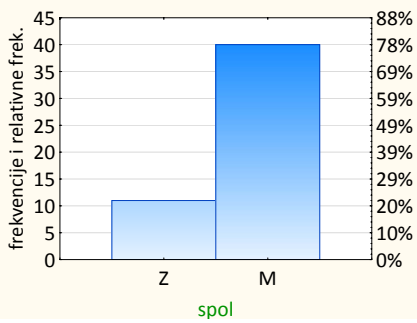
Category	Frequency table: spol (zdravlje.sta) Count	Cumulative Count	Percent	Cumulative Percent
Z: žena	11	11	21,57	21,57
M: muškarac	40	51	78,43	100,00
Missing	0	51	0,00	100,00

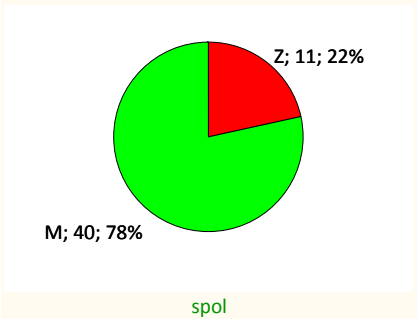
(a) varijabla spol

Category	Frequency table: zdravlje (zdravlje.sta) Count	Cumulative Count	Percent	Cumulative Percent
1	4	4	7,84	7,84
2	8	12	15,69	23,53
3	18	30	35,29	58,82
4	12	42	23,53	82,35
5	9	51	17,65	100,00
Missing	0	51	0,00	100,00

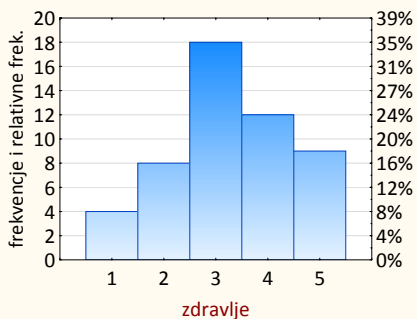
(b) varijabla zdravlje

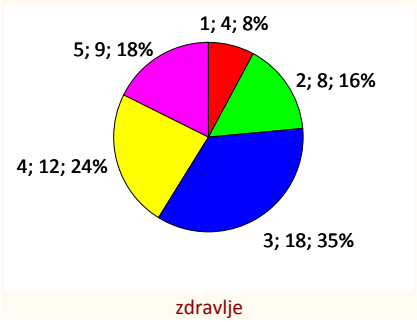
Slika 3.26: Tablice frekvencija i relativnih frekvencija svih podataka varijabli spol i zdravlje.





Slika 3.27: Grafički prikazi frekvencija i relativnih frekvencija svih podataka varijable spol.





Slika 3.28: Grafički prikazi frekvencija i relativnih frekvencija svih podataka varijable zdravlje.

Tablični i grafički prikazi podataka sadržanih u varijabli zdravlje *posebno za kategoriju ispitanika ženskog spola, a posebno za kategoriju ispitanika muškog spola prikazani su na slikama 3.29, 3.30 i 3.31. Kružne dijagrame relativnih frekvencija sa slike 3.31 u programskom paketu Statistica možemo dobiti provodeći sljedeći postupak:*

Graphs → Categorized Graphs → Pie Charts → Graph Type: Pie Chart - Counts → Variables (Vars - zdravlje, X-Category - spol) → Advanced → Pie Legend (Text and Value za kružne dijagrame

frekvencija, Text and Percent za kružne dijagrame relativnih frekvencija).

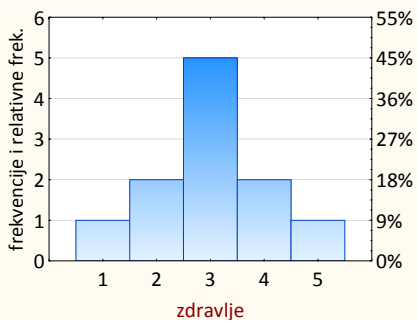
Category	Frequency table: zdravlje (zdravlje.sta) Include condition: spol="Z"			
	Count	Cumulative Count	Percent	Cumulative Percent
1	1	1	9,09	9,09
2	2	3	18,18	27,27
3	5	8	45,45	72,73
4	2	10	18,18	90,91
5	1	11	9,09	100,00
Missing	0	11	0,00	100,00

(a) žene (spol=Z)

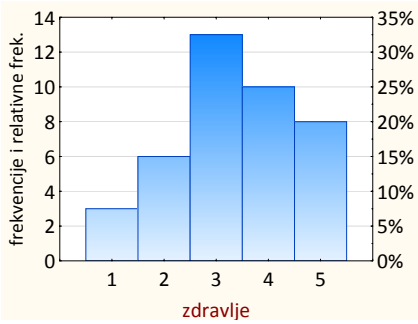
Category	Frequency table: zdravlje (zdravlje.sta) Include condition: spol="M"			
	Count	Cumulative Count	Percent	Cumulative Percent
1	3	3	7,50	7,50
2	6	9	15,00	22,50
3	13	22	32,50	55,00
4	10	32	25,00	80,00
5	8	40	20,00	100,00
Missing	0	40	0,00	100,00

(b) muškarci (spol=M)

Slika 3.29: Tablični prikaz podataka za varijablu zdravlje kategoriziranih prema spolu ispitanika.

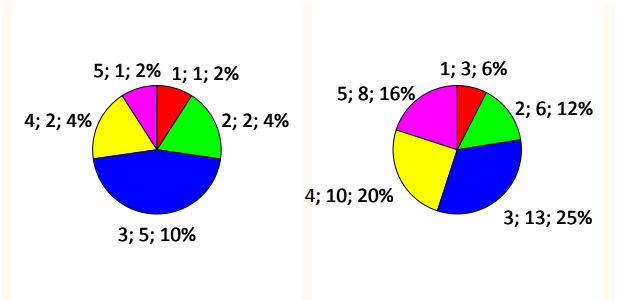


(a) žene (spol=Z)



(b) muškarci (spol=M)

Slika 3.30: Stupčasti dijagrami podataka varijable zdravlje kategoriziranih prema spolu ispitanika.



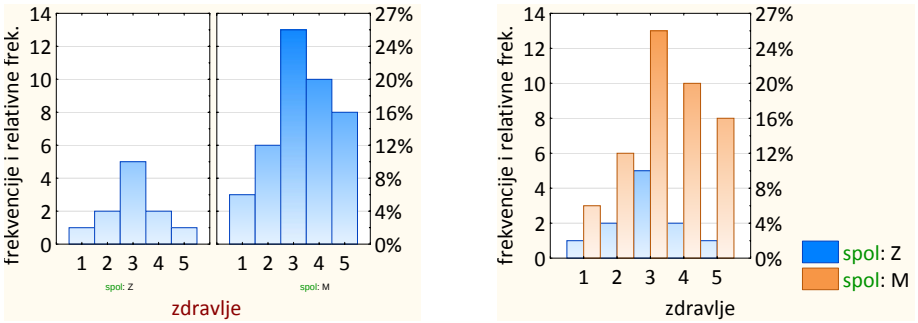
žene (spol=Z) muškarci (spol=M)

Slika 3.31: Kružni dijagrami podataka varijable zdravlje kategoriziranih prema spolu ispitanika.

Radi uspoređivanja rezultata po spolu korisno je stupčaste dijagrame frekvencija i relativnih frek-

vencija podataka sadržanih u varijabli zdravlje *kategoriziranih prema spolu ispitanika prikazati na jednoj slici, tj. grafu (slika 3.32). Objedinjene dijagramske prikaze frekvencija i relativnih frekvencija neke varijable čije su vrijednosti kategorizirane po nekom kriteriju možemo dobiti u programskom paketu Statistica provodeći sljedeći postupak:*

Graphs → Categorized Graphs → Histograms → Variables (Variable - zdravlje, X-Category - spol) → Layout (Separate - za odvojene stupčaste dijagrame kategorija varijable zdravlje kategoriziranih s obzirom na vrijednosti varijable spol; Overlaid - za prikaz frekvencija kategorija varijable zdravlje kategoriziranih s obzirom na vrijednosti varijable spol na istom stupčastom dijagramu)



Slika 3.32: Stupčasti dijagrami podataka varijable zdravlje kategoriziranih prema spolu ispitanika.

Zadatak 3.7. (TV-program.sta)
Za varijable iz baze podataka TV-program.sta napravite sljedeće tablične i grafičke prikaze:

a) napravite tablice i nacrtajte stupčaste dijagrame frekvencija i relativnih frekvencija za podatke sadržane u varijablama spol i P1,

b) napravite tablice i nacrtajte stupčaste dijagrame frekvencija i relativnih frekvencija za podatke sadržane u varijabli P1 posebno za kategoriju ispitanika ženskog spola, a posebno za kategoriju ispitanika muškog spola,

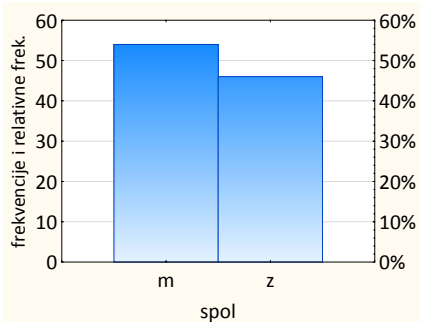
c) nacrtajte kružne dijagrame frekvencija i relativnih frekvencija za podatke sadržane u varijablama spol i P3,

d) nacrtajte kružne dijagrame frekvencija i relativnih frekvencija za podatke sadržane u varijabli P3 posebno za kategoriju ispitanika ženskog spola, a posebno za kategoriju ispitanika muškog spola.

Rješenje.

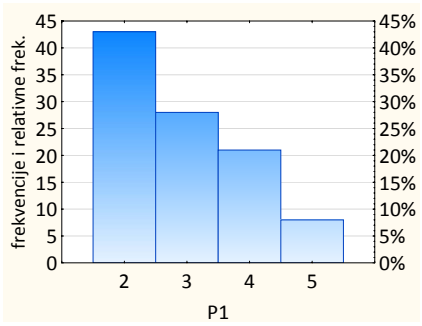
a) Tablični i grafički prikazi frekvencija i relativnih frekvencija svih kategorije varijable spol *i svih različitih vrijednosti varijable* P1 *prikazani su na slikama 3.33 i 3.34.*

Category	Frequency table: spol (TV-program.sta)			
	Count	Cumulative Count	Percent	Cumulative Percent
m	54	54	54,00	54,00
z	46	100	46,00	100,00
Missing	0	100	0,00	100,00



Slika 3.33: Tablica i stupčasti dijagram za podatke varijable spol.

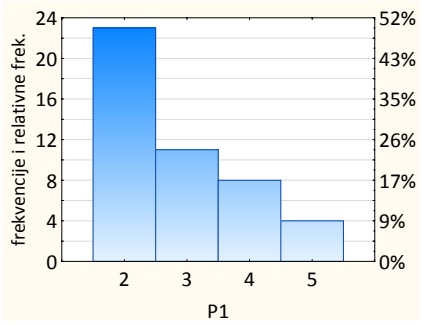
Category	Frequency table: P1 (TV-program.sta)			
	Count	Cumulative Count	Percent	Cumulative Percent
2	43	43	43,00	43,00
3	28	71	28,00	71,00
4	21	92	21,00	92,00
5	8	100	8,00	100,00
Missing	0	100	0,00	100,00



Slika 3.34: Tablica i stupčasti dijagram za podatke varijable P1.

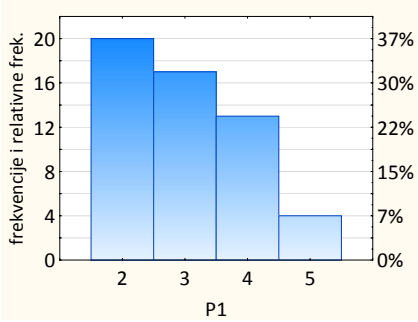
b) Tablični i grafički prikazi frekvencija i relativnih frekvencija svih kategorija varijable P1 *kategoriziranih prema spolu ispitanika prikazani su na slikama 3.35, 3.36 i 3.37.*

Category	Frequency table: P1 (TV-program.sta) Include condition: spol="z"			
	Count	Cumulative Count	Percent	Cumulative Percent
2	23	23	50,00	50,00
3	11	34	23,91	73,91
4	8	42	17,39	91,30
5	4	46	8,70	100,00
Missing	0	46	0,00	100,00

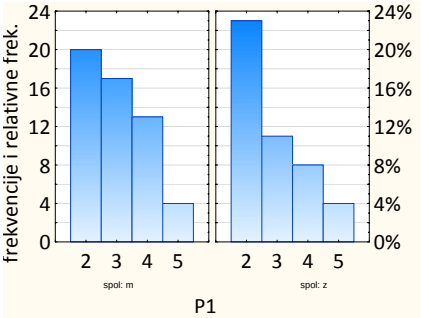


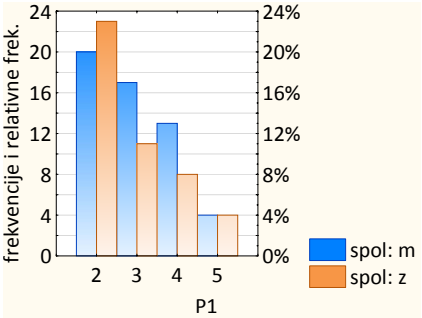
Slika 3.35: Tablica i stupčasti dijagram za podatke varijable P1 za ženski spol.

Category	Frequency table: P1 (TV-program.sta) Include condition: spol="m"			
	Count	Cumulative Count	Percent	Cumulative Percent
2	20	20	37,04	37,04
3	17	37	31,48	68,52
4	13	50	24,07	92,59
5	4	54	7,41	100,00
Missing	0	54	0,00	100,00



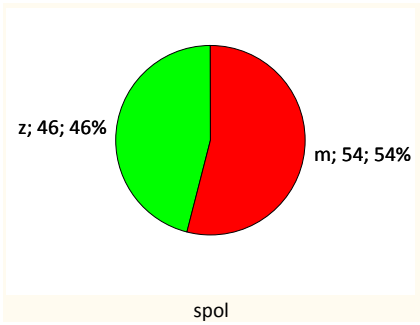
Slika 3.36: Tablica i stupčasti dijagram za podatke varijable P1 za muški spol.



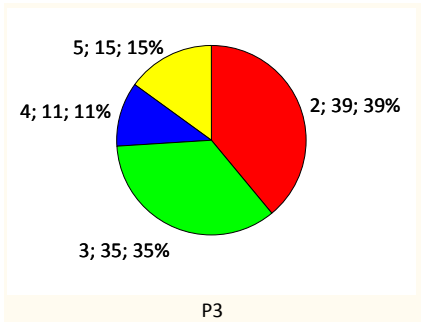


Slika 3.37: Stupčasti dijagrami za podatke varijable P1 kategorizirane prema spolu ispitanika.

d) Kružni dijagrami frekvencija i relativnih frekvencija svih kategorija varijable spol *i svih različitih vrijednosti varijable* P3 *prikazani su na slici 3.38.*



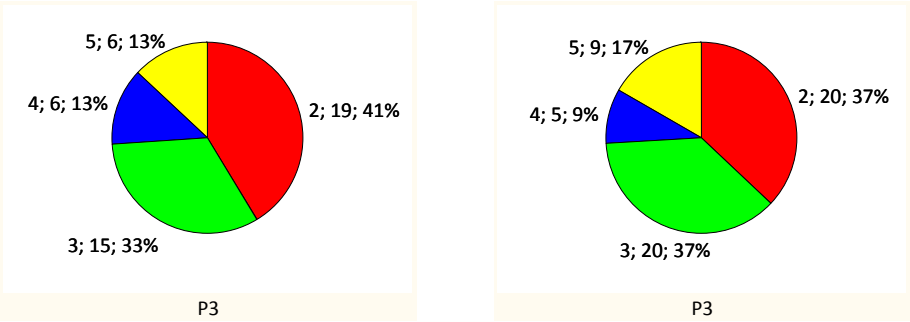
(a) varijabla spol



(b) varijabla P3

Slika 3.38: Kružni dijagrami za podatke varijabli spol i P3.

e) Kružni dijagrami relativnih frekvencija za podatke iz varijable P3 *kategorizirane prema spolu ispitanika prikazani su na slici 3.39.*



(a) spol = z (b) spol = m

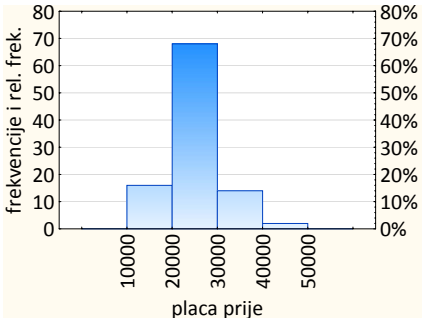
Slika 3.39: Kružni dijagrami za podatke varijable P3 kategorizirane prema spolu ispitanika.

Zadatak 3.8. (djelatnici.sta)
Promotrite varijablu placa prije iz baze podataka djelatnici.sta opisane u primjeru 2.4. Razvrstajte vrijednosti u disjunktne intervale duljine 10000 počevši od nule te prikažite podatke tablično i histogramom.

Rješenje. *Tablični prikaz frekvencija i relativnih frekvencija dan je tablicom 3.7, a pripadni histogram slikom 3.40. Ovakakv histogram jasno ilustrira činjenicu da najviše djelatnika u uzorku ima godišnju plaću od* 20000 *do* 30000 *novčanih jedinica, dok je plaća iz intervala* 40000 *do* 50000 *rijetkost. Intervale za kategorizaciju u ovakvim i sličnim slučajevima obično radimo tako da bismo zadovoljili potrebe za prezentiranjem informacija koje želimo istaknuti.*

iznos plaće	frekvencija	relativna frekvencija
$[0, 10000\rangle$	0	0
$[10000, 20000\rangle$	15	0.15
$[20000, 30000\rangle$	69	0.69
$[30000, 40000\rangle$	14	0.14
$[40000, 50000\rangle$	2	0.02

Tablica 3.7: Tablica frekvencija i relativnih frekvencija kategoriziranih podataka varijable placa prije.



Slika 3.40: Histogram frekvencija i relativnih frekvencija kategoriziranih podataka varijable placa prije.

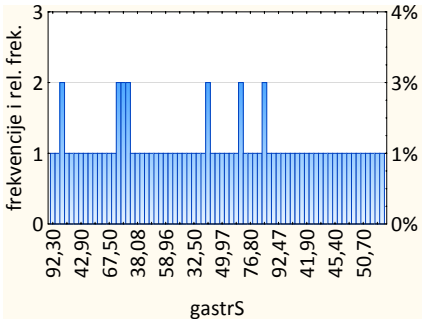
Zadatak 3.9. (hormon.sta)

a) Odredite tablicu frekvencija i stupčasti dijagram za neprekidnu numeričku varijablu gastrS iz baze podataka hormon.sta (koja je opisana u zadatku 3.1) tako da za kategorije uzmete sve međusobno različite izmjerene vrijednosti.

b) Iskoristite izmjerene vrijednosti varijable gastrS, kategorizirajte podatke i prikažite ih histogramom. Mijenjajte broj intervala na koji dijelite skup vrijednosti. Proučavajte što se događa i pribilježite svoj zaključak.

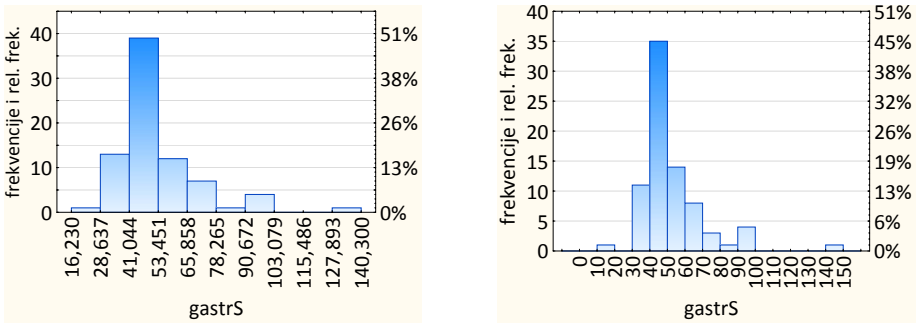
Rješenje.

a) Stupčasti dijagram frekvencija i relativnih frekvencija te kružni dijagram izmjerenih vrijednosti varijable gastrS *u kojima su kao kategorije uzete sve različite izmjerene vrijednosti prikazani su na slici 3.41.*



Slika 3.41: Stupčasti dijagram svih izmjerenih vrijednosti varijable gastr S.

b) Kategorizacija izmjerenih vrijednosti varijable gastrS *na disjunktne intervale daje preglednije grafičke prikaze iz kojih je lakše analizirati izmjerene vrijednosti i donijeti neke zaključke. Grafički prikazi frekvencija i relativnih fekvencija izmjerenih vrijednosti varijable* gastrS *razvrstanih u* 10 *i* 15 *disjunktnih intervala prikazani su na slici 3.42.*



Slika 3.42: Histogram za podatke varijable gastrS.

Zadatak 3.10. (djelatnici.sta)
Odredite numeričke karakteristike skupa izmjerenih vrijednosti varijable placa prije iz baze podataka djelatnici.sta opisane u primjeru 2.4.

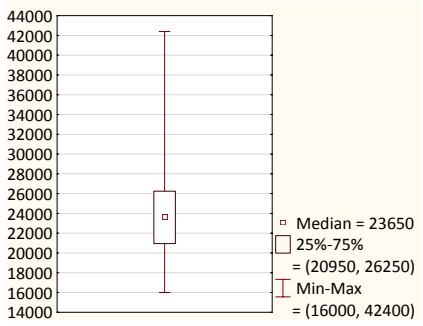
Rješenje. *Numeričke karakteristike prikazane su u tablicama na slici 3.43.*

Variable	Descriptive Statistics (djelatnici.sta)					
	Valid N	Mean	Mode	Frequency of Mode	Variance	Std.Dev.
placa prije	100.00	24522.00	24600.00	4.00	26069208.08	5105.80

Variable	Descriptive Statistics (djelatnici.sta)					
	Median	Minimum	Maximum	Lower Quartile	Upper Quartile	Range
placa prije	23650.00	16000.00	42400.00	20950.00	26250.00	26400.00

Slika 3.43: Deskriptivna statistika izmjerenih vrijednosti varijable placa prije.

Odnos minimuma, donjeg kvartila, medijana, gornjeg kvartila i maksimuma izmjerenih vrijednosti varijable placa prije *prikazani su kutijastim dijagramom 3.44.*



Slika 3.44: Kutijasti dijagram na bazi medijana za varijablu placa prije.

Iz tablice 3.43 i kutijastog dijagrama 3.44 možemo izvesti sljedeće i slične zaključke:

- *najniža godišnja plaća u uzorku iznosi* 16000, *a najviša* 42400
- *bar* 25% *ispitanika iz uzorka ima plaću manju ili jednaku* 20950
- *bar* 25% *ispitanika iz uzorka ima plaću veću ili jednaku* 26250
- *bar* 50% *ispitanika iz uzorka ima plaću manju ili jednaku medijanu, tj.* 23650
- *bar* 50% *ispitanika iz uzorka ima plaću veću ili jednaku* 23650.

Zadatak 3.11. (nastava.sta)
Baza podataka nastava.sta sadrži ocjene u skali od 0 (najniža ocjena) do 10 (najviša ocjena) različitih komponenti probnog nastavnog sata za 65 studenata (budućih nastavnika):

varijabla znanje sadrži ocjene znanja studenta o temi nastavnog sata

varijabla literatura sadrži ocjene primjerenosti korištene literature za pripremu nastavnog sata

varijabla predavac sadrži ocjene predavačeva stava i nastupa pred razredom

varijabla atmosfera sadrži ocjene radne atmosfere na nastavnom satu

varijabla govor sadrži ocjene studentova izražavanja tijekom nastavnog sata

varijabla interes sadrži ocjene pobuđenosti interesa kod učenika za temu nastavnog sata

varijabla bitan sadrzaj sadrži ocjene naglašenosti bitnih sadržaja tijekom nastavnog sata

varijabla primjeri sadrži ocjene odabira i primjerenosti primjera prezentiranih tijekom nastavnog sata

varijabla ukupno sadrži ocjene koje odražavaju ukupan ocjenjivačev dojam o održanom nastavnom satu.

Ako želimo donijeti opći zaključak o uspješnosti budućih nastavnika u stvarnoj nastavnoj situaciji, logično je pažnju usmjeriti na analizu varijable ukupno. Odredite numeričke karakteristike te varijable i kutijasti dijagram na bazi medijana. Diskutirajte o rezultatima.

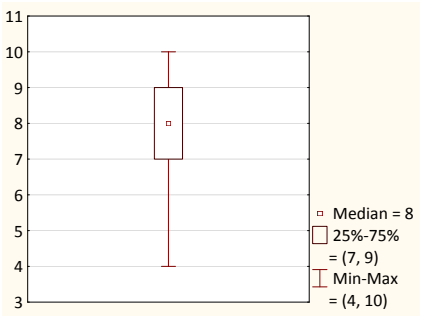
Rješenje. *Numeričke karakteristike te varijable prikazane su u tablici 3.45.*

	Descriptive Statistics (nastava.sta)					
Variable	Valid N	Mean	Mode	Frequency of Mode	Variance	Std.Dev.
ukupno	65.00	8.11	Multiple	19.00	2.16	1.47

	Descriptive Statistics (nastava.sta)					
Variable	Median	Minimum	Maximum	Lower Quartile	Upper Quartile	Range
ukupno	8.00	4.00	10.00	7.00	9.00	6.00

Slika 3.45: Deskriptivna statistika podataka za varijablu ukupno.

Iz tablice frekvencija za varijablu ukupno *lako se vidi da skup podataka te varijable ima dva moda - to su ocjene* 8 *i* 9. *Dakle, probno je predavanje za čak* 19 *studenata ocijenjeno visokom ocjenom* 8 *te za isto toliko ocjenom* 9, *dok je prosječna ocjena ukupnog dojma probnog nastavnog sata* 8.11. *Analizu raspršenosti ocjena napravit ćemo pomoću kutijastog dijagrama (slika 3.46).*



Slika 3.46: Kutijasti dijagram na bazi medijana za podatke varijable ukupno.

Analiza kutijastog dijagrama sugerira sljedeće zaključke: nitko od ispitanika predavanje nije ocijenio ocjenom nižom od četiri, barem 25% *ispitanika predavanje je ocijenilo ocjenama* 4, 5, 6 *ili* 7, *barem* 25% *ocjenama* 7 *ili* 8, *barem* 25% *ocjenama* 8 *ili* 9 *te barem* 25% *ocjenama* 9 *ili* 10. *Zanimljivo je uočiti da je barem* 75% *ispitanika predavanje ocijenilo ocjenom* 7 *i više.*

Zadatak 3.12. (matematika.sta)
Baza podataka matematika.sta (opisana u primjeru 2.9) sadrži rezultate ankete o kvaliteti izvođenja nekog matematičkog kolegija. Ukoliko nas zanima prilagođenost težine sadržaja kolegija predznanju studenata, analizirat ćemo varijablu tezina kolegija. Odredite numeričke karakteristike podataka te varijable i prikažite ih kutijastim dijagramom.

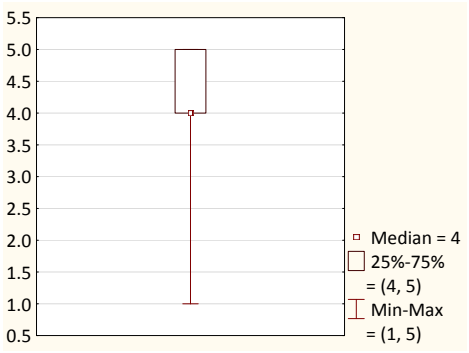
Rješenje. *Mjere deskriptivne statistike varijable* tezina kolegija *prikazane su u tablici na slici 3.47.*

	Descriptive Statistics (matematika.sta)					
Variable	Valid N	Mean	Mode	Frequency of Mode	Variance	Std.Dev.
tezina kolegija	49.00	4.18	5.00	21.00	0.78	0.88

	Descriptive Statistics (matematika.sta)					
Variable	Median	Minimum	Maximum	Lower Quartile	Upper Quartile	Range
tezina kolegija	4.00	1.00	5.00	4.00	5.00	4.00

Slika 3.47: Deskriptivna statistika podataka varijable tezina kolegija.

Uočimo da je čak 21 *ispitanik prilagođenost težine kolegija predznanju studenata ocijenio ocjenom* 5 *(ocjena* 5 *je mod ovog skupa podataka) te da je prosječna ocjena* 4.18. *Za analizu raspršenosti ocjena koristimo kutijasti dijagram prikazan na slici 3.48.*



Slika 3.48: Kutijasti dijagram na bazi medijana za varijablu tezina kolegija.

Analizom kutijastog dijagrama donosimo sljedeći zaključak: barem 25% *ispitanika težinu kolegija ocijenilo je ocjenama* 1, 2, 3 *ili* 4, *barem* 50% *ocjenom* 4 *te barem* 25% *ocjenama* 4 *ili* 5. *Zanimljivo je uočiti da je barem* 75% *ispitanika težinu kolegija ocijenilo ocjenam* 4 *ili* 5.

Zadatak 3.13. (djelatnici.sta)
Varijabla dob iz baze podataka djelatnici.sta opisane u primjeru 2.4 za svakog ispitanika iz uzorka djelatnika promatranog poduzeća sadrži informaciju o dobi u godinama. Odredite numeričke karakteristike podataka iz te varijable, analizirajte postojanje stršećih vrijednosti, prikažite podatke kutijastim dijagramom i diskutirajte o rezultatima.

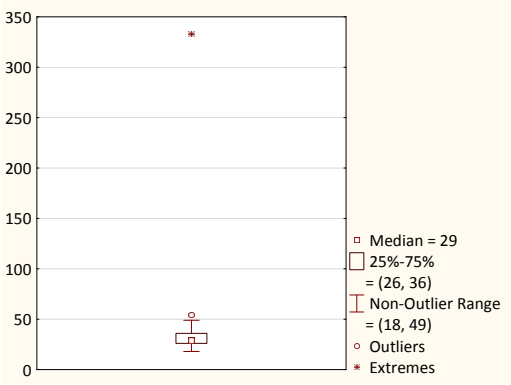
Rješenje. *Iz deskriptivne statistike varijable* dob *(tablica 3.49) vidimo da je maksimalna podatak za dob 333 godine pa je očigledno da postoji stršeći podatak koji je pogrešno upisan u bazu podataka.*

	Descriptive Statistics (djelatnici.sta)					
Variable	Valid N	Mean	Mode	Frequency of Mode	Variance	Std.Dev.
dob	100.00	33.83	28.00000	12.00	964.28	31.05

	Descriptive Statistics (djelatnici.sta)					
Variable	Median	Minimum	Maximum	Lower Quartile	Upper Quartile	Range
dob	29.00	18.00	333.00	26.00	36.00	315.00

Slika 3.49: Deskriptivna statistika podataka varijable dob.

Osim iz tablice 3.49, stršeće vrijednosti među podacima varijable dob *mogli smo detektirati i pomoću kutijastog dijagrama na bazi medijana.*



Slika 3.50: Kutijasti dijagram na bazi medijana s prikazom stršećih vrijednosti varijable dob.

Kao što vidimo iz kutijastog dijagrama 3.50, i dob od 54 *godine prepoznata je kao stršeća vrijednost. Budući da je sasvim razumljivo da promatrano poduzeće može imati djelatnika starog* 54 *godine, taj podatak smatramo točnim, no radi se o dobi koja se rijetko pojavljuje u populaciji djelatnika tog poduzeća.*

Zadatak 3.14. (glukoza.sta)
Varijabla dob baze podataka glukoza.sta sadrži godine starosti, a varijabla koncentracija izmjerene vrijednosti koncentracije glukoze u krvi za 102 ispitanika. Korištenjem programskog paketa Statistica riješite sljedeće zadatke:

a) Napravite deskriptivnu statistiku podataka sadržanih u varijabli koncentracija. Grafičkom metodom odredite stršeću vrijednost u ovom skupu podataka. Možete li se složiti s tvrdnjom da je identificirani podatak moguća izmjerena vrijednost ili ipak sumnjate u dobiveni rezultat? Obrazložite svoj odgovor.

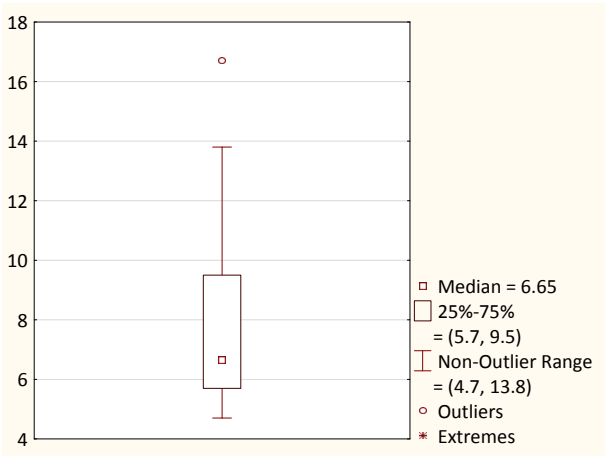
b) Grafičkom metodom identificirajte stršeće vrijednosti među podacima u varijabli dob. Što se događa s numeričkim karakteristikama podataka nakon uklanjanja stršeće vrijednosti?

Rješenje.

a) Deskriptivna statistika i kutijasti dijagram s označenim stršećim vrijednostima skupa izmjerenih vrijednosti varijable koncentracija *prikazani su na slikama 3.51 i 3.52.*

	Descriptive Statistics (glukoza.sta)							
Variable	Valid N	Mean	Median	Mode	Frequency of Mode	Minimum	Maximum	Lower Quartile
koncentracija	102.00	7.70	6.65	5.500000	14.00	4.70	16.70	5.70

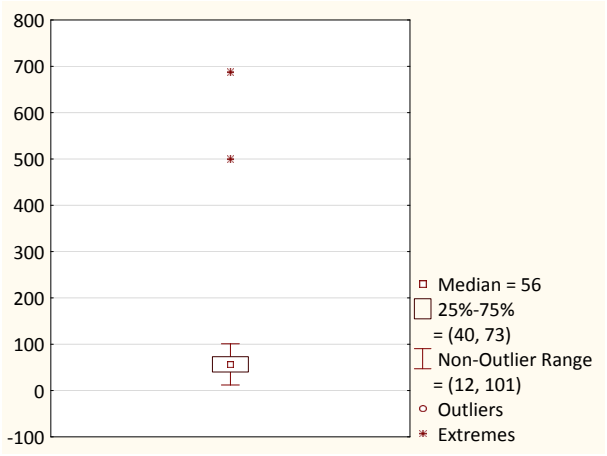
Slika 3.51: Deskriptivna statistika izmjerenih vrijednosti varijable koncentracija.



Slika 3.52: Kutijasti dijagram na bazi medijana s prikazom stršećih vrijednosti varijable dob.

Statistica je kao stršeću vrijednost detektirala podatak 16.7. *Kako se ta koncentracija glukoze u krvi može zaista pojaviti pri mjerenjima, taj podatak nećemo tretirati kao stršeću vrijednost.*

b) Kutijasti dijagram s označenim stršećim vrijednostima i deskriptivna statistika skupa izmjerenih vrijednosti varijable dob *prikazani su na slikama 3.53 i 3.54.*



Slika 3.53: Kutijasti dijagram na bazi medijana s prikazom stršećih vrijednosti varijable dob.

Variable	Descriptive Statistics (glukoza.sta)								
	Valid N	Mean	Median	Mode	Frequency of Mode	Minimum	Maximum	Lower Quartile	Upper Quartile
dob	102	66.73	56.00	Multiple	4.00	12.00	688.00	40.00	73.00

(a) uključene stršeće vrijednosti

Variable	Descriptive Statistics (glukoza.sta)								
	Valid N	Mean	Median	Mode	Frequency of Mode	Minimum	Maximum	Lower Quartile	Upper Quartile
dob	100	56.18	55.50	Multiple	4.00	12.00	101.00	40.00	71.50

(b) uklonjene stršeće vrijednosti

Slika 3.54: Deskriptivna statistika izmjerenih vrijednosti varijable dob.

Statistica je kao stršeće vrijednosti među izmjerenim vrijednostima varijable dob *detektirala podatke* 500 *i* 688. *Zaključujemo da uklanjanjem tih vrijednosti dolazi do smanjenja aritmetičke sredine i medijana izmjerenih vrijednosti.*

Zadatak 3.15. (komarci.sta)
Proučite bazu podataka komarci.sta koja je opisana u zadatku 2.4. Odredite tablicu i histogram frekvencija i relativnih frekvencija varijable brojM tako da za kategorije uzmete sve međusobno različite izmjerene vrijednosti te varijable. Zatim podijelite skup izmjerenih vrijednosti na određen broj disjunktnih intervala i ponovno odredite frekvencije i relativne frekvencije pojedinih kategorija (tj. intervala). Mijenjajte broj intervala, proučavajte što se događa i pribilježite svoj zaključak.

Zadatak 3.16. Koristeći javne izvore podataka ili podatke koje ste prikupljali u sklopu nekog istraživanja formirajte jednu bazu podataka koja će sadržavati najmanje dvije kvalitativne varijable, najmanje jednu diskretnu numeričku varijablu i jednu neprekidnu numeričku varijablu. Opišite o kakvom se istraživanju radi i zašto se mjere vrijednosti navedenih varijabli. Vodite računa da baza sadrži što više jedinki. Navedite točan izvor podataka. Iskoristite prethodno opisane postupke i pojmove te opišite svoju bazu podataka.

Poglavlje 4

Slučajna varijabla

4.1 Uvod

U prethodnom poglavlju naučili smo da su predmet istraživanja, u kojemu želimo napraviti statističku analizu, varijable čije vrijednosti mjerimo na jedinkama.

Primjer 4.1. *Pretpostavimo da je građanima iz reprezentativnog uzorka stanovnika grada Osijeka jednog dana u podne izmjerena koncentracija glukoze u krvi. Rezultat tog istraživanje je podatak o koncentraciji glukoze u krvi za svaku osobu iz uzorka - te izmjerene vrijednosti radi statističke analize podataka organiziramo u varijablu* koncentracija glukoze*. U tablici 4.1 prikazano je samo nekoliko izmjerenih vrijednosti te varijable.*

osoba	koncentracija glukoze (mmol/L)
1	5.635
2	12.560
3	19.817
⋮	⋮

Tablica 4.1: Izmjerene vrijednosti varijable koncentracija glukoze.

Međutim, jasno je da su ove izmjerene vrijednosti samo neke od svih vrijednosti koje koncentracija glukoze u krvi može poprimiti. Medicinska istraživanja pokazuju da koncentracija glukoze u krvi čovjeka može biti bilo koji realan broj iz intervala $(0, 131]$*. Dakle, izmjerena vrijednost varijable* koncentracija glukoze *za svaku osobu iz ovog uzorka je jedna vrijednost iz skupa svih mogućih vrijednosti koje koncentracija glukoze u krvi čovjeka može poprimiti.*

Primjer 4.2. *Na nekoj mjernoj postaji svakog se sata mjeri vodostaj rijeke Drave. Nekoliko zadnjih izmjerenih vodostaja prikazao je u tablici 4.2.*

dan i sat	vodostaj (cm)
17.11.2010. - 9:00	174
17.11.2010. - 8:00	161
17.11.2010. - 7:00	152
⋮	⋮

Tablica 4.2: Izmjerene vrijednosti varijable vodostaj.

Prema povijesnim podacima najniži izmjereni vodostaj Drave na ovoj mjernoj postaji bio je 105 *cm (1978.), a najviši čak* 511 *cm (1972.). Ove činjenice opravdavaju visok stupanj vjerovanja da vodostaj rijeke Drave na promatranoj mjernoj postaji može biti bilo koji realan broj iz intervala* $[105, 511]$. *Prema tome, svaka izmjerena vrijednost varijable* vodostaj *iz gornje tablice jedna je vrijednost iz skupa svih mogućih vrijednosti koje vodostaj Drave može poprimiti na toj mjernoj postaji. Podaci su preuzeti sa http://www.voda.hr.*

Varijable koje su navedene u prethodnim primjerima (koncentracija glukoze u krvi ili vodostaj rijeke Drave) želimo opisati matematičkim modelom. Pri tome smo svjesni da prije samog mjerenja i tijekom mjerenja istraživač ne zna koji će rezultat mjerenja dobiti, ali zna iz kojeg skupa izmjerena vrijednost te varijable može biti (iz $(0, 131]$ za varijablu koncentracija glukoze te iz $[105, 511]$ za varijablu vodostaj). Da bismo napravili model na osnovi kojega možemo raditi statističko zaključivanje, varijable ćemo modelirati kao **slučajne varijable**. Zašto ove varijable treba nazvati slučajnima? Razlog je taj što one mogu primiti mnogo različitih vrijednosti, a mi u trenutku njihova proučavanja ne možemo sa sigurnošću znati koja će se od tih vrijednosti realizirati. Zapravo, mjerenje varijable provodimo, između ostalog, zato da ocijenimo stupanj izvjesnosti da varijabla u određenim uvjetima primi dane vrijednosti.
Slučajna varijabla i način kako je opisujemo predmet su ovog poglavlja. Slučajne varijable označavat ćemo velikim slovima, recimo X, Y, Z. Podsjetimo da se u matematici varijable obično označavaju malim slovima x, y, z. Biranjem velikog slova za oznaku varijable naglašavamo da se ovdje radi o slučajnoj varijabli.

Varijablu nazivamo slučajnom varijablom ako su njene moguće realizacije (ishodi) realni brojevi, ali vrijednost koja će se realizirati u pojedinom eksperimentu nije jednoznačno određena uvjetima koje možemo sagledati prilikom istraživanja.

Već iz primjera 4.1 i 4.2 možemo vidjeti da je osnovni objekt koji služi za modeliranje slučajne varijable **skup svih mogućih realizacija slučajne varijable** (u matematici taj skup zovemo slika slučajne varijable). Skup svih mogućih realizacija slučajne varijable X označit ćemo s $\mathfrak{R}(X)$.

Primjer 4.3. *Bacamo novčić i smatramo uspjehom ako je palo pismo. Realizacije ovog pokusa možemo modelirati slučajnom varijablom. Recimo, kažemo da slučajna varijabla X prima vrijednost* 1 *ako je palo pismo, a* 0 *ako nije palo pismo (tj. ako je pala glava). Na taj način dolazimo do skupa mogućih realizacija te slučajne varijable:* $\mathfrak{R}(X) = \{0, 1\} \subset \mathbb{R}$.

Primjer 4.4. *Bacamo igraću kockicu. Broj koji se okrene prilikom jednog bacanja na gornjoj strani kockice je realizacija jedne slučajne varijable, označimo je s X. Prirodno, skup svih mogućih realizacija te slučajne varijable je* $\mathfrak{R}(X) = \{1, 2, 3, 4, 5, 6\} \subset \mathbb{R}$.

Primjer 4.5. *Bacamo igraću kockicu dva puta. Zbroj brojeva koji se okrenu prilikom tih dvaju bacanja je realizacija jedne slučajne varijable X. Skup svih mogućih realizacija te slučajne varijable je* $\mathfrak{R}(X) = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\} \subset \mathbb{R}$.

Primjer 4.6. *Broj ulovljenih komaraca u jednu klopku u Osijeku, u jednom danu lipnja 2012. godine, možemo modelirati kao slučajnu varijablu. Naime, jedan dan u klopku smo uhvatili, npr.* 20 *komaraca, drugi dan* 25*, treći dan* 45*, četvrti dan opet* 20*. Koliko ćemo ih uhvatiti sutra, prekosutra? Ne znamo kako će se ta varijabla realizirati sutra i prekosutra, ali znamo da će to svakako biti neki prirodan broj ili nula. Osim toga, ako smo postavili dvije identične klopke jednu pored druge, može se dogoditi da je u istom danu na jednu klopku uhvaćeno* 20 *komaraca, a na drugu* 23*. Dakle, prirodno je smatrati tu varijablu slučajnom varijablom (označimo je s X) jer, u uvjetima koje mi možemo sagledati, ne možemo sa sigurnošću znati kako će se realizirati. Skup svih mogućih realizacija ove slučajne varijable je skup* $\mathfrak{R}(X) = \{0, 1, 2, \ldots, n\}$*, gdje je* $n \in \mathbb{N}$ *najveći broj komaraca koji mogu biti ulovljeni korištenom klopkom.*

4.2 Vjerojatnost

Promatrajući skup vrijednosti koji može primiti slučajna varijabla X može se dogoditi da je naše uvjerenje u realizaciju nekog podskupa $A \subseteq \mathfrak{R}(X)$ veće od uvjerenja da će se realizirati poskup $B \subseteq \mathfrak{R}(X)$. U tom slučaju uobičajeno kažemo da skup A ima veće **šanse** za realizaciju nego podskup B.

Primjer 4.7. *U primjeru bacanja igraće kockice ishod jednog bacanja modelirali smo slučajnom varijablom X čiji je skup svih mogućih realizacija $\Re(X) = \{1, 2, 3, 4, 5, 6\}$. Pretpostavite da je igraća kockica pravilno izrađena. Razmislite i odgovorite na sljedeća pitanja:*

Kojim biste realnim brojem iskazali šanse za realizaciju skupa $\{3\}$?

Očekujete li da se šanse za realizaciju skupa $\{3\}$ razlikuje od šansi za realizaciju skupa $\{5\}$?

Kojim biste realnim brojem iskazali šanse za realizaciju parnog broja pri bacanju ove kockice (tj. da se na kockici okrene paran broj)?

Ima smisla također govoriti i o šansama za realizaciju nekog podskupa skupa čiji elementi nisu realni brojevi, nego proizvoljni objekti (npr. slova, neki specijalni znakovi, razne kategorije). Sljedeći primjer ilustrira jedan takav slučaj.

Primjer 4.8. *Promotrimo skup $\Omega = \{\clubsuit, \diamondsuit, \heartsuit, \spadesuit\}$. Elementi ovog skupa su oznake za crne karte (tref i pik) i crvene karte (karo i herc) u standardnom svežnju angloameričkih igraćih karata. Dakle skup Ω ima četiri elementa i možemo reći da njegovi elementi opisuju ishod pokusa koji se sastoji od izvlačenja jedne karte iz svežnja pri čemu nas za izvučenu kartu zanima samo boja (crvena ili crna) i tip (tref, pik, karo ili herc). Razmislite i odgovorite na sljedeća pitanja:*

Kojim biste realnim brojem iskazali šanse za realizaciju skupa $\{\clubsuit\} \subset \Omega$?

Očekujete li da se šanse za realizaciju skupa $\{\clubsuit, \diamondsuit\} \subset \Omega$ razlikuje od šansi za realizaciju skupa $\{\heartsuit, \spadesuit\}$?

Kojim biste realnim brojem iskazali šanse za realizaciju skupa $\{\clubsuit, \diamondsuit, \heartsuit\} \subset \Omega$, a kojim šanse za realizaciju skupa $\{\heartsuit\}$?

Broj kojim izražavamo šanse za realizaciju nekog podskupa možemo definirati općenito za neprazan skup Ω, bez obzira jesu li njegovi elementi brojevi ili neki drugi objekti. Mjeru koja modelira šansu da će se realizirati neki podskup promatranog nepraznog skupa Ω zvat ćemo vjerojatnost. Podskupove skupa Ω zvat ćemo događajima. U ovom poglavlju navodimo definiciju vjerojatnosti, načine kako u konkretnim primjerima možemo modelirati vjerojatnost te neka osnovna svojstva vjerojatnosti. **Neka je Ω neprazan skup te neka familija skupova $\mathfrak{F}$ sadrži određene podskupove od Ω (tj. određene događaje). Vjerojatnost (oznaka P) je funkcija koja svakom događaju $A \in \mathfrak{F}$ pridružuje realan broj iz intervala $[0, 1]$ (tj. $0 \leq P(A) \leq 1$) tako da vrijede sljedeći zahtjevi:**

V1. $P(\Omega) = 1$,

V2. **ako su A_1 i A_2 događaji iz $\mathfrak{F}$ koji nemaju zajedničkih elemenata, tj. $A_1, A_2 \in \mathfrak{F}$ i $A_1 \cap A_2 = \emptyset$, tada vrijedi**

$$P(A_1 \cup A_2) = P(A_1) + P(A_2),$$

tj. vjerojatnost unije događaja A_1 i A_2 jednaka je zbroju vjerojatnosti $P(A_1)$ i $P(A_2)$.[1]

Vidimo da je na ovaj način definirana vjerojatnost na familiji podskupova proizvoljnog nepraznog skupa Ω. Uzmemo li da je $\Omega = \mathfrak{R}(X)$, dobivamo definiciju vjerojatnosti na familiji podskupova skupa svih mogućih realizacija (slike) slučajne varijable X.

Uobičajene oznake i nazivi

Neka je $\mathfrak{R}(X)$ skup svih mogućih realizacija slučajne varijable X i $\mathfrak{F}$ familija podskupova od $\mathfrak{R}(X)$ na kojoj je definirana vjerojatnost P. Familiju $\mathfrak{F}$ obično zovemo **familija događaja**.
Zbog lakšeg razumijevanja i opisivanja događaja koje razmatramo, tj. podskupova od $\mathfrak{R}(X)$, skup $C \subseteq \mathfrak{R}(X)$ označavat ćemo oznakom $\{X \in C\}$. Naime, skup C će se dogoditi (realizirati) ako slučajna varijabla X primi vrijednosti (realizacije) iz skupa C. Na taj način lakše povezujemo događaje sa slučajnom varijablom na koju se odnose.

Primjer 4.9. *Skup* $\{X \in [2,3]\}$ *definira događaj koji se dogodi ako se slučajna varijabla realizira nekom vrijednošću iz intervala* $[2,3]$. *Uočimo da isti događaj možemo zapisati i na sljedeći način:*

$$\{2 \leq X \leq 3\}.$$

Skup $\{4 < X \leq 7\}$ *definira događaj koji se dogodi ako se slučajna varijabla realizira brojem koji je veći od* 4, *ali manji ili jednak* 7.

Slučajnu varijablu X definirali smo ako smo definirali $\mathfrak{R}(X)$ i vjerojatnost P na familiji podskupova $\mathfrak{F}$. Tada kažemo da smo zadali **razdiobu (distribuciju) slučajne varijable** X.

Definiranje vjerojatnosti za pojedine primjere temelji se na dosadašnjem iskustvu u istraživanju danog slučajnog pokusa i može biti vrlo složen postupak. U nastavku opisujemo metodu određivanja vjerojatnosti na konačnom skupu **pod uvjetom da su svi ishodi jednako mogući**. Takav pristup temelji se na intuitivnoj ideji

[1] Ukoliko familija $\mathfrak{F}$ sadrži beskonačno mnogo događaja, ovaj zahtjev mora se pojačati. Tada se traži da za proizvoljan niz događaja $(A_n, n \in \mathbb{N})$ koji nemaju zajedničkih točaka, tj. $A_i \bigcap A_j = \emptyset$, za sve $i \neq j$, vrijedi:

$$P(\bigcup_{i \in \mathbb{N}} A_i) = \sum_{i \in \mathbb{N}} P(A_i).$$

koju je formulirao jedan od osnivača teorije vjerojantnsoti James Bernoulli (1654.—1705.), a možemo je prevesti kao *"Vjerojatnost se prema sigurnosti odnosi kao dio prema cjelini".*

4.2.1 Jednako mogući ishodi

Pretpostavimo da prilikom izvođenje pokusa vrijede vrijede sljedeći uvjeti:

(1) skup $\Omega \neq \emptyset$ ima konačno mnogo elemenata, tj. Ω je oblika

$$\Omega = \{\omega_1, \ldots, \omega_n\}, \quad n \in \mathbb{N},$$

(2) svi jednočlani podskupovi skupa Ω su jednako vjerojatni, tj.

$$P(\{\omega_i\}) = P(\{\omega_j\}), \qquad \text{za sve } i, j \in \{1, \ldots, n\}.$$

Tada vjerojatnost skupa (događaja) $A \subseteq \Omega$ definiramo na sljedeći način:

$$P(A) = \frac{\text{broj elemenata od } A}{\text{broj elemenata od } \Omega} = \frac{k(A)}{k(\Omega)},$$

gdje je $k(\cdot)$ oznaka za broj elemenata skupa (tj. $k(A)$ je oznaka za broj elemenata skupa A, a $k(\Omega)$ za broj elemenata skupa Ω).
Taj pristup modeliranju vjerojatnosti temelji se na ideji da vjerojatnost predstavlja mjeru dijela u odnosu na cjelinu. Problemi u primjeni ovog pristupa odnose se na provjeru pretpostavki. Npr. kako možemo biti sigurni da su svi jednočlani podskupovi skupa Ω jednako vjerojatni?
Na potpuno isti način možemo definirati vjerojatnost na familiji podskupova skupa svih mogućih realizacija slučajne varijable X, tj. skupu $\mathfrak{R}(X)$, pod uvjetom da $\mathfrak{R}(X)$ ima konačno mnogo jednako vjerojatnih elemenata. Dakle, ako je $\Omega = \mathfrak{R}(X)$, tada vjerojatnost skupa $B \subseteq \mathfrak{R}(X)$ definiramo na sljedeći način:

$$P(B) = \frac{k(B)}{k(\mathfrak{R}(X))}.$$

Primjer 4.10. *Iz svežnja koji se sastoji od 32 karte*[2] *izvlačimo jednu kartu. Odredimo:*

vjerojatnost da je izvučena karta as

vjerojatnost da izvučena karta nije as

vjerojatnost da je izvučena karta as ili kralj.

[2]Svežanj od 32 karte koji se spominje u ovoj knjizi podrazumijeva karte dolaze u četiri "boje" (crvena, zelena, žir i bundeva) i osam tipova (sedmica, osmica, devetka, desetka, dečko, dama, kralj i as)

Uočimo da ovakav svežanj možemo podijeliti na osam skupina karata koje se sastoje od po četiri karte istog tipa (četiri sedmice, četiri asa, četiri kralja, četiri dame, ...). Prema tome, tipove karata u svežnju možemo označiti brojevima $1, \ldots, 8$. *U skladu s ovim označavanjem zaključujemo da se izvlačenjem jedne karte zapravo realizira jedan od brojeva* $1, \ldots, 8$. *Time smo zapravo definirali slučajnu varijablu* X *koja svakoj karti iz svežnja (koji možemo shvatiti kao skup* Ω*) pridružuje točno jedan od brojeva* $1, 2, 3, 4, 5, 6, 7, 8$. *Dakle, skup svih mogućih realizacija slučajne varijable* X *je* $\mathfrak{R}(X) = \{1, 2, 3, 4, 5, 6, 7, 8\}$. *To je skup koji ima* 8 *elemenata koji su, zbog jednakobrojnosti svih osam skupina karata, svi jednako vjerojatni. Prema tome, odgovori na prethodno postavljena pitanja su:*

vjerojatnost da izvučemo asa je $1/8$,

vjerojatnost da ne izvučemo asa je $7/8$,

iz zahtjeva V2. iz definicije vjerojatnosti slijedi da je vjerojatnost da izvučemo asa ili kralja $1/8 + 1/8 = 1/4$.

Primjer 4.11. *Pri bacanju pravilno izrađene igraće kockice može pasti bilo koji od brojeva* $1, \ldots, 6$, *tj. skup svih mogućih ishoda ovog pokusa je* $\Omega = \{1, 2, 3, 4, 5, 6\}$. *Pretpostavimo da se ovo bacanje kockice vrši u sklopu igre u kojoj zarađujemo jednu kunu ako se na kockici okrene paran broj, a gubimo jednu kunu ako se okrene neparan broj.*

Kolika je vjerojatnost zarade jedne kune?

Budući da jednu kunu zarađujemo ako se okrene 2 *ili* 4 *ili* 6, *slijedi da je skup svih za nas povoljnih ishoda skup* $A = \{2, 4, 6\} \subset \Omega$. *Slijedi da je vjerojatnost zarade jedne kune*

$$P(A) = \frac{k(A)}{k(\Omega)} = \frac{3}{6} = \frac{1}{2}.$$

Drugi način rješavanja ovog problema uključuje definiranje slučajne varijable X *čija je realizacija* 1 *ako se pri bacanju kockice okrene paran broj, a* (-1) *ako se pri bacanju kockice okrene neparan broj. Dakle,* $\mathfrak{R}(X) = \{-1, 1\}$. *Povoljan događaj u ovom kontekstu je događaj* $\{1\} \subset \mathfrak{R}(X)$, *pa je vjerojatnost zarade jedne kune*

$$P\{X = 1\} = \frac{k(\{1\})}{k(\mathfrak{R}(X))} = \frac{1}{2}.$$

Primjer 4.12. *Bacamo jednom dvije pravilno izrađene igraće kockice. Budući da se pri bacanju na svakoj od kockica realizira neki od brojeva iz skupa* $\{1, 2, 3, 4, 5, 6\}$, *zaključujemo da je jedna realizacija bacanja dviju kockica uređeni par brojeva. Dakle, skup svih mogućih ishoda ovog pokusa je skup* $\Omega = \{(i, j) : i, j \in \{1, 2, 3, 4, 5, 6\}\}$ *koji se sastoji od* 36 *elemenata. Pitamo se:*

Kolika je vjerojatnost da je suma brojeva koji su pali na obje kockice jednaka 6*?*

Kolika je vjerojatnost da je suma brojeva koji su pali na obje kockice manja od 6*?*

Neka je A *skup koji sadrži one uređene parove iz* Ω *za koje je suma prve i druge komponente jednaka* 6, *tj.*

$$A = \{(1, 5), (2, 4), (3, 3), (4, 2), (5, 1)\},$$

a B *skup koji sadrži one uređene parove iz* Ω *za koje je suma prve i druge komponente manja od* 6, *tj.*

$$B = \{(1, 1), (1, 2), (1, 3), (1, 4), (2, 1), (2, 2), (2, 3), (3, 1), (3, 2), (4, 1)\}.$$

Slijedi:

$$P(A) = \frac{k(A)}{k(\Omega)} = \frac{5}{36}, \qquad P(B) = \frac{k(B)}{k(\Omega)} = \frac{10}{36} = \frac{5}{18}.$$

Drugi način rješavanja istih problema uključuje definiranje slučajne varijable X čija je realizacija zbroj brojeva koji su pali pri bacanju dviju kockica, dakle $\Re(X) = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$. Vjerojatnosti skupova $\{X = 2\}, \ldots, \{X = 12\}$ mogu se pregledno prikazati tablicom 4.3.

k	2	3	4	5	6	7	8	9	10	11	12
$P\{X = k\}$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Tablica 4.3: Tablični prikaz vjerojatnosti skupova $\{X = 2\}, \ldots, \{X = 12\}$.

4.2.2 Statistička interpretacija vjerojatnosti

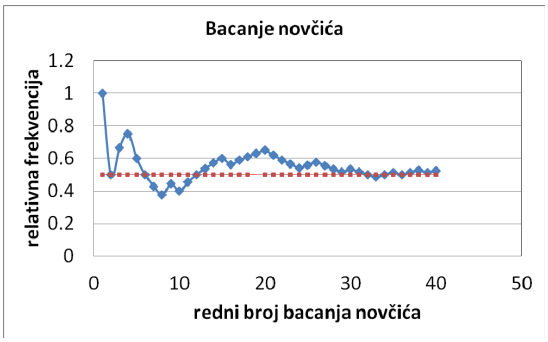
Prethodno opisan način određivanja vjerojatnosti može se primijeniti pod pretpostavkom da je broj jednako mogućih ishoda slučajnog pokusa konačan. Mnogo je pokusa koji ne zadovoljavaju te pretpostavke. Kako tada pridružiti vjerojatnost skupu? U ovom poglavlju ilustrirat ćemo statističku interpretaciju vjerojatnosti koja može biti od pomoći ako želimo odrediti vjerojatnost događaja u pokusu koji možemo puno puta nezavisno[3] izvoditi. Za ilustraciju zakonitosti o kojoj će biti riječi izvedite pokus tako da bacite novčić 40 puta. Bilježite realizacije pisma (oznaka 1) ili glave (oznaka 0) kao što je to prikazano u tablici 4.4.

Redni broj bacanja	Realizacija
1	0
2	1
3	0
4	0
⋮	⋮

Tablica 4.4: Realizacije bacanja novčića.

Izračunajte relativne frekvencije pojavljivanja pisma u prvih n bacanja za svaki $n = 1, \ldots, 40$. Grafički prikaz relativnih frekvencija pojavljivanja pisma za 40 bacanja novčića zabilježenih u Excel dokumentu novcic.xls prikazan je na slici 4.1. Usporedite svoje rezultate s navedenim grafom!

[3] Smatramo da se pokusi izvode nezavisno ako činjenica da se dogodio neki događaj prilikom izvođenja jednog od njih ne mijenja šanse za realizaciju bilo kojeg događaja drugog pokusa. Npr. bacanje igrače kocke dva puta čini dva nezavisna pokusa, ali izvlačenje drugog broja u igri loto pokus je koji nije nezavisan od izvlačenja prvog broja u toj igri.



Slika 4.1: Grafički prikaz relativnih frekvencija pojavljivanja pisma za 40 bacanja novčića.

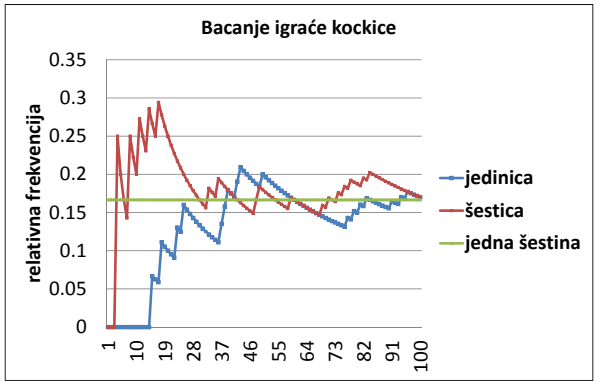
Ako ste imali pravilan novčić (tj. novčić kod kojeg su realizacije pisma i glave jednako mogući ishodi), možete uočiti dvije sličnosti vašeg grafa s grafom 4.4: za velike n relativna frekvencija stabilizira se i to blizu 0.5. Uočite da je u svakom pojedinom bacanju novčića vjerojatnost pojavljivanja pisma ista jer bacamo isti novčić u istim uvjetima. Osim toga, tu vjerojatnost možemo izračunati na temelju pretpostavke jednako mogućih ishoda i ona iznosi točno 0.5.
Ovaj primjer ilustrira zakonitost o kojoj će biti riječi u poglavlju 5, a može se sažeti u sljedeću formulaciju:

Ako je pokus takav da ga možemo nezavisno ponavljati mnogo puta, relativna frekvencija pojavljivanja događaja A će se s povećanjem broja ponavljanja pokusa stabilizirati oko broja koji predstavlja vjerojatnost pojavljivanja događaja A.

Primjer 4.13. (kockica.xls)
Pokus koji možemo nezavisno ponavljati mnogo puta je i bacanje igraće kockice. Znamo da se pri jednom bacanju igraće kockice realizira broj iz skupa $\{1, 2, 3, 4, 5, 6\}$ te da su, uz pretpostavku da je kockica pravilno izrađena, svi ishodi jednako mogući, tj.

$$P(\{1\}) = P(\{2\}) = P(\{3\}) = P(\{4\}) = P(\{5\}) = P(\{6\}) = \frac{1}{6}.$$

Očekujemo da će se s povećanjem broja bacanja igraće kockice relativne frekvencije mogućih realizacija stabilizirati oko 1/6. Baza podataka kockica.xls sadrži ishode za 100 bacanja igraće kockice zajedno s pripadnim frekvencijama i relativnim frekvencijama. Relativne frekvencije realizacija jedinice i šestice u ovisnosti o broju bacanja grafički su prikazane na slici 4.2 - vidimo da se relativne frekvencije stabiliziraju oko $1/6 \approx 0.1667$.



Slika 4.2: Grafički prikaz relativnih frekvencija pojavljivanja 1 i 6 za 100 bacanja igraće kockice.

4.2.3 Neka svojstva vjerojatnosti

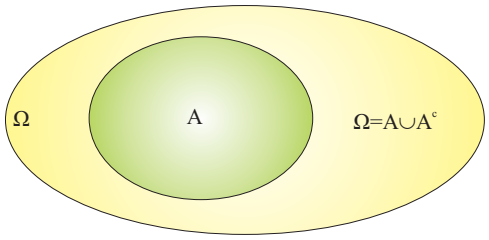
Da bismo lakše računali vjerojatnosti događaja za razne podskupove konkretnog skupa Ω, u ovom poglavlju navest ćemo osnovna svojstva vjerojatnosti.

S1. **Vjerojatnost suprotnog događaja**
Ako je $A \in \mathcal{F}$, tada je

$$P(A^c) = 1 - P(A),$$

gdje je $A^c = \Omega \setminus A$ komplement skupa A.

Dokaz. Zahtjev V1. iz definicije vjerojatnosti glasi $P(\Omega) = 1$. Cijeli skup Ω možemo prikazati kao uniju skupova A i A^c (slika 4.3).



Slika 4.3: Događaj A (zeleno područje) i njegov komplemet A^c (žuto područje).

Ti skupovi nemaju zajedničkih elemenata, tj. $A \cap A^c = \emptyset$. Sada prema zahtjevu V2. iz definicije vjerojatnosti slijedi

$$1 = P(\Omega) = P(A \cup A^c) = P(A) + P(A^c) \;\Rightarrow\; P(A^c) = 1 - P(A).$$

S2. **Vjerojatnost nemogućeg događaja**

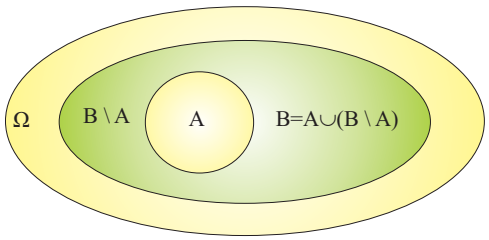
$$P(\emptyset) = 0.$$

Dokaz.
S obzirom da je $\emptyset = \Omega^c$, primjenom zahtjeva V1. iz definicije vjerojatnosti i prethodno dokazanog svojstva S1. slijedi da je

$$P(\emptyset) = P(\Omega^c) = 1 - P(\Omega) = 0.$$

S3. **Monotonost vjerojatnosti**
Ako su A i B skupovi iz $\mathcal{F}$ takvi da je $A \subseteq B$, tada je $P(A) \leq P(B)$. Osim toga vrijedi i da je $P(B \setminus A) = P(B) - P(A)$.

Dokaz. Prikažimo skup B kao uniju skupova koji nemaju zajedničkih elemenata: $B = A \cup (B \setminus A)$, $A \cap (B \setminus A) = \emptyset$.



Slika 4.4: Skup B kao unija skupova A (manje žuto područje) i $(B \setminus A)$ (zeleno područje).

Sada prema zahtjevu V2. iz definicije vjerojatnosti slijedi da je

$$P(B) = P(A \cup (B \setminus A)) = P(A) + P(B \setminus A) \geq P(A),$$

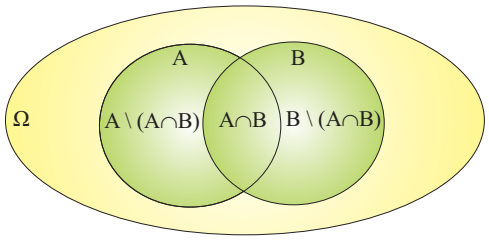
jer je zbog nenegativnosti vjerojatnosti $P(B \setminus A) \geq 0$. Slijedi da je u tom slučaju $P(B) \geq P(A)$, tj. $P(A) \leq P(B)$. Primjenom istog pristupa kao u dokazu prethodne tvrdnje također slijedi da je $P(B) = P(A \cup (B \setminus A)) = P(A) + P(B \setminus A)$, tj.

$$P(B \setminus A) = P(B) - P(A).$$

(S4) **Vjerojatnost unije**
Ako su $A, B \in \mathcal{F}$ proizvoljni događaji (koji ne moraju biti disjunktni), tada je

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Dokaz. Sa slike 4.5 vidimo da se skup $A \cup B$ može prikazati kao unija triju skupova koji nemaju zajedničkih elemenata.



Slika 4.5: Skup $A \cup B$ (zeleno područje) kao unija skupova $A \setminus (A \cap B)$, $(A \cap B)$ i $B \setminus (A \cap B)$.

Dakle,

$$A \cup B = (A \setminus B) \cup (A \cap B) \cup (B \setminus A) = (A \setminus (A \cap B)) \cup (A \cap B) \cup (B \setminus (A \cap B)),$$

gdje je $A \cap B \subseteq A$ i $A \cap B \subseteq B$. Sada prema zahtjevu V2. iz definicije vjerojatnosti slijedi

$$P(A \cup B) = P(A \setminus (A \cap B)) + P(A \cap B) + P(B \setminus (A \cap B)) =$$

$$P(A) - P(A \cap B) + P(A \cap B) + P(B) - P(A \cap B) = P(A) + P(B) - P(A \cap B).$$

Primjer 4.14. Računalo slučajno generira posljednju znamenku telefonskog broja. Skup svih mogućih ishoda generiranja zadnje znamenke je

$$\Omega = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}, \quad k(\Omega) = 10.$$

Korištenjem svojstava S1., S2. S3. i S4. možemo odredite vjerojatnost sljedećih i sličnih događaja:

a) vjerojatnost da je slučajno generirana znamenka jednaka 8 je

$$P(\{8\}) = 1/10$$

b) vjerojatnost da je slučajno generirana znamenka jednaka 8 ili 9 je

$$P(\{8\} \cup \{9\}) = P(\{8, 9\}) = 2/10 = 1/5$$

c) vjerojatnost da je slučajno generirana znamenka neparna ili 2 je

$$P(\{1\} \cup \{3\} \cup \{5\} \cup \{7\} \cup \{9\} \cup \{2\}) = P(\{1, 2, 3, 5, 7, 9\}) = 6/10 = 3/5$$

d) vjerojatnost da je slučajno generirana znamenka parna ili 2 je

$$P(\{0\} \cup \{2\} \cup \{4\} \cup \{6\} \cup \{8\} \cup \{2\}) = P(\{0, 2, 4, 6, 8\}) = 5/10 = 1/2$$

e) vjerojatnost da je slučajno generirana znamenka neparna, ali nije 3, je

$$P(\{1, 5, 7, 9\}) = P(\{1, 3, 5, 7, 9\} \setminus \{3\}) = (5/10) - (1/10) = 4/10 = 2/5.$$

4.3 Diskretna slučajna varijabla

Kao što smo opisali u poglavlju o tipovima varijabli koje su predmet statističkog opisivanja i istraživanja, bitna je razlika u opisu numeričkih varijabli koje su diskretnog tipa i onih koje su neprekidnog tipa. Te razlike vidljive su i u načinu koji koristimo kada opisujemo vjerojatnosna svojstva slučajnih varijabli kojima modeliramo varijable u istraživanju. Mi ćemo razlikovati dva tipa slučajnih varijabli: **diskretne slučajne varijable** i **neprekidne slučajne varijable**

Ako je $\mathfrak{R}(X)$ konačan skup ili ga možemo prikazati kao $\mathfrak{R}(X) = \{x_i \in \mathbb{R}; i \in \mathbb{N}\}$ (prebrojiv skup), kažemo da je slučajna varijabla X diskretna.

U tom slučaju skup svih mogućih realizacija označit ćemo s $\mathfrak{R}(X) = \{x_1, x_2, x_3, \ldots, x_n\}$ ako je on konačan, odnosno, s $\mathfrak{R}(X) = \{x_1, x_2, x_3, \ldots\}$ ako je beskonačan. Vjerojatnosti događaja vezanog uz realizaciju diskretne slučajne varijable možemo najjednostavnije računati koristeći vjerojatnosti da se dogode pojedinačne realizacije. Zato, uz skup svih mogućih realizacija diskretne slučajne varijable X, u njezinu opisu ključnu ulogu ima i pridruženi niz pozitivnih realnih brojeva $(p_1, p_2, p_3, \ldots, p_n)$ (odnosno $(p_i, i \in \mathbb{N})$, ako je $\mathfrak{R}(X)$ beskonačan) kojim su zadane vjerojatnosti da se dogode pojedinačne realizacije iz $\mathfrak{R}(X)$. Preciznije to možemo iskazati na sljedeći način.

Neka je X diskretna slučajna varijabla s konačnim skupom svih mogućih realizacija $\mathfrak{R}(X) = \{x_1, x_2, x_3, \ldots, x_n\}$ (odnosno prebrojivim skupom svih mogućih realizacija $\mathfrak{R}(X) = \{x_1, x_2, x_3, \ldots\}$). Za svaku pojedinu realizaciju x_i definiramo realan broj

$$p_i = P\{X = x_i\}.$$

Distribucija (razdioba) diskretne slučajne varijable X u potpunosti je zadana skupom $\mathfrak{R}(X)$ i pripadnim nizom $(p_i, i = 1, \ldots, n)$ (odnosno nizom $(p_i, i \in \mathbb{N})$ ako je $\mathfrak{R}(X)$ prebrojiv skup).

Uočimo da za ovako definiran niz realnih brojeva $(p_i, i = 1, \ldots, n)$, odnosno $(p_i, i \in \mathbb{N})$, moraju vrijediti sljedeća dva bitna svojstva kako bi on definirao vjerojatnost na $\mathfrak{R}(X)$:

1. $p_i \geq 0$ za sve pripadne $x_i \in \mathfrak{R}(X)$,
2. $\sum_{x_i \in \mathfrak{R}(X)} p_i = 1.$

Također, korištenjem zahtjeva V2 iz definicije vjerojatnosti izvodimo način računanja vjerojatnosti da diskretna slučajna varijabla primi vrijednosti iz nekog skupa $A \subseteq \mathfrak{R}(X)$. Naime, vrijedi:

$$P\{X \in A\} = \sum_{x_i \in A} p_i.$$

Zaista, svaki skup $A \subseteq \mathfrak{R}(X)$ možemo prikazati kao uniju jednočlanih podskupova $\{x_i\}$ od $\mathfrak{R}(X)$ gdje je $i \in \mathrm{I}_A$, tj.

$$A = \bigcup_{i \in \mathrm{I}_A} \{x_i\}.$$

Odavde korištenjem poopćenja svojstva V2. iz definicije vjerojatnosti slijedi:

$$P\{X \in A\} = P\left\{X \in \bigcup_{i \in \mathrm{I}_A} \{x_i\}\right\} = \sum_{i \in \mathrm{I}_A} P\{X = x_i\} = \sum_{x_i \in A} p_i.$$

Korištenjem ovih rezultata, diskretna slučajna varijabla se često prikazuje pomoću ta dva bitna niza na sljedeći način:

$$X \sim \begin{pmatrix} x_1 & x_2 & \dots & x_n \\ p_1 & p_2 & \dots & p_n \end{pmatrix}, \quad \text{odnosno} \quad X \sim \begin{pmatrix} x_1 & x_2 & x_3 & \dots \\ p_1 & p_2 & p_3 & \dots \end{pmatrix},$$

pri čemu se prvom tablicom zadaje diskretna slučajna varijabla karakterizirana konačnim skupom $\mathfrak{R}(X)$, a drugom tablicom diskretna slučajna varijabla karakterizirana prebrojivim skupom $\mathfrak{R}(X)$. Ovakvu tablicu zovemo **tablica distribucije** diskretne slučajne varijable. Tablice distribucije možemo prikazivati i u klasičnom tabličnom obliku (tablica 4.5) (usporedite s tabicom iz primjera 4.12).

vrijednosti od X	x_1	x_2	$\dots$	x_n
vjerojatnosti $P\{X = x_i\}$	p_1	p_2	$\dots$	p_n

Tablica 4.5: Tablica distribucije diskretne slučajne varijable karakterizirane konačnim skupom $\mathfrak{R}(X)$.

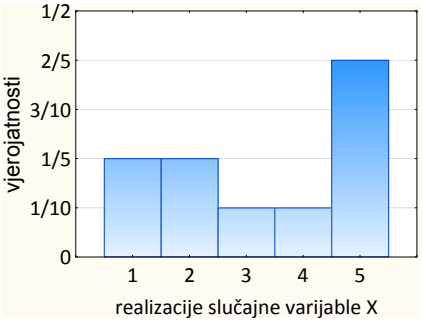
Distribuciju diskretne slučajne varijable X možemo slikovito prikazati **stupčastim dijagramom** u kojem svaki stupić odgovara jednoj vrijednosti x_i koju ta slučajna varijabla može poprimiti (tj. jednom elementu iz $\mathfrak{R}(X)$), a visina stupića jednaka je vjerojatnosti $p_i = P\{X = x_i\}$.

Primjer 4.15. *Diskretna slučajna varijabla X koja može poprimiti vrijednosti $1, 2, 3, 4, 5$ zadana je tablicom distribucije 4.6.*

vrijednosti	1	2	3	4	5
vjerojatnosti	1/5	1/5	1/10	1/10	2/5

Tablica 4.6: Tablica distribucije slučajne varijable sa slikom $\{1, 2, 3, 4, 5\}$.

Stupčasti dijagram distribucije slučajne varijable zadane tablicom distribucije 4.6 prikazan je na slici 4.6.



Slika 4.6: Grafički prikaz distribucije slučajne varijable X zadane tablicom distribucije 4.6.

Pomoću tablice distribucije slučajne varijable X možemo odrediti vjerojatnosti za podskupove od $\Re(X)$. Npr.

$$P\{X = 5\} = \frac{2}{5}, \qquad P\{X \in \{2, 3\}\} = P\{X = 2\} + P\{X = 3\} = \frac{1}{5} + \frac{1}{10} = \frac{3}{10}.$$

Primjer 4.16. Procjenjuje se učinak investicije na jednom području izražen u obliku dobiti, odnosno gubitka. Neka je X diskretna slučajna varijabla čije su realizacije iznosi dobitka (odnosno gubitka) u tisućama kuna. Distribucija vjerojatnosti učinka investicije zadana je tablicom 4.7.

dobit (gubitak) u tisućama kn	vjerojatnost
-400	0.05
-200	0.15
-100	0.3
0	0.1
100	0.3
200	0.03
300	0.04
400	0.03

Tablica 4.7: Tablica distribucije vjerojatnosti učinka investicije.

Prema tablici 4.7 je

$$\Re(X) = \{-400, -200, -100, 0, 100, 200, 300, 400\},$$

a pripadne vjerojatnosti su sljedeće:

$$P\{X=-400\}=0.05,\ P\{X=-200\}=0.15,\ P\{X=-100\}=P\{X=100\}=0.3,$$
$$P\{X=0\}=0.1,\ P\{X=200\}=P\{X=400\}=0.03,\ P\{X=300\}=0.04.$$

Dakle, tablicom 4.7 zadana je distribucija diskretne slučajne varijable X te pomoću nje možemo odrediti vjerojatnosti sljedećih događaja:

Investicija rezultira gubitkom ako slučajna varijabla X primi neku od vrijednosti iz skupa

$$\{-400,-200,-100\}\subset\mathfrak{R}(X),$$

pa je vjerojatnost da će investicija rezultirati gubitkom

$$P\{X\in\{-400,-200,-100\}\}=0.05+0.15+0.3=0.5.$$

Investicija neće rezultirati dobitkom ako slučajna varijabla X primi neku od vrijednosti iz skupa

$$\{-400,-200,-100,0\}\subset\mathfrak{R}(X),$$

pa je vjerojatnost da investicija neće rezultirati dobitkom

$$P\{X\in\{-400,-200,-100,0\}\}=0.05+0.15+0.3+0.1=0.6.$$

Vjerojatnost da će dobit biti barem 100000, ali manje od 300000 kuna je

$$P\{X\in\{100,200\}\}=0.3+0.03=0.33.$$

4.4 Neprekidna slučajna varijabla

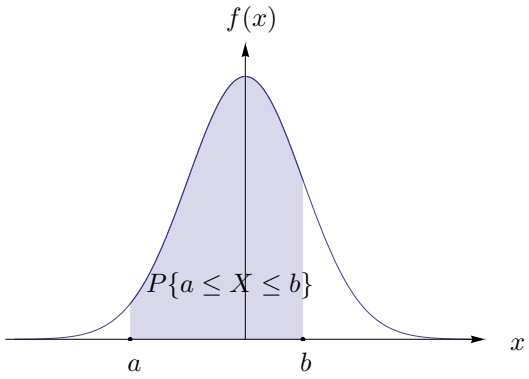
Diskretne slučajne varijable poslužit će kao matematički model kojim opisujemo diskretne numeričke varijable u istraživanju. Za modeliranje neprekidnih numeričkih varijabli u istraživanjima trebat će nam model slučajne varijable čiji je skup svih mogućih realizacija $\mathfrak{R}(X)$ interval ili segment realnih brojeva ili je $\mathfrak{R}(X)=\mathbb{R}$. Za takve slučajne varijable bit će karakteristično da se, općenito, njihova vjerojatnosna svojstva ne mogu modelirati korištenjem niza vjerojatnosti pojedinačnih realizacija kao u diskretnom slučaju. Kao što smo već uočili kod neprekidnih varijabli, naglasak pri njihovu opisivanju stavljen je na interval vrijednosti koje takva varijabla prima, a ne na pojedinačne realizacije.

Za slučajnu varijablu X kažemo da je neprekidna slučajna varijabla ako postoji nenegativna realna funkcija f, definirana na skupu realnih brojeva, takva da je za $a,b\in\mathbb{R}$ $(a\le b)$ vjerojatnost

$$P\{a\le X\le b\}=P\{a<X\le b\}=\int_a^b f(x)\,dx.$$

Takvu funkciju f zovemo funkcija gustoće neprekidne slučajne varijable X.

Uočimo da vjerojatnost $P\{a\le X\le b\}=P\{a<X<b\}$ zapravo predstavlja površinu između osi x i grafa funkcije f na intervalu $[a,b]$ (slika 4.7).



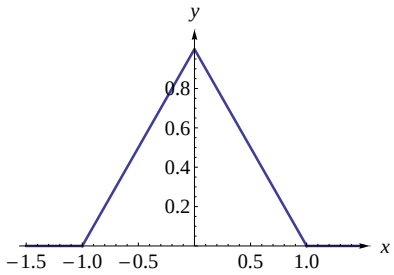
Slika 4.7: Vjerojatnost kao površina između osi x i grafa funkcije f na intervalu $[a, b]$

Na taj način lako vizualiziramo značenje vjerojatnosti da neprekidna slučajna varijabla primi vrijednost iz nekog podskupa skupa $\mathbb{R}$. Dakle, prilikom računanja vjerojatnosti za neprekidnu slučajnu varijablu treba prvo skicirati graf njene funkcije gustoće i koristiti ga prilikom analiziranja slučajne varijable i računanja vjerojatnosti da ona primi vrijednost iz nekog skupa.
Neprekidna slučajna varijabla zadana je ako je poznata njena funkcija gustoće. Tada kažemo da poznajemo **razdiobu** ili **distribuciju** neprekidne slučajne varijable.

Primjer 4.17. *Funkcija gustoće neprekidne slučajne varijable X dana je izrazom*

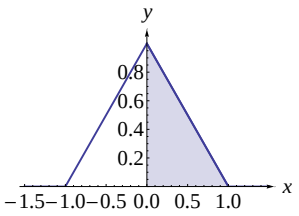
$$f(x) = \begin{cases} -|x| + 1 & , \ x \in [-1, 1] \\ 0 & , \ x \notin [-1, 1] \end{cases}.$$

Graf funkcije f prikazan je slikom 4.8.

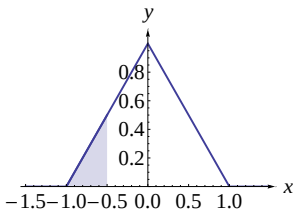


Slika 4.8: Graf funkcije gustoće f.

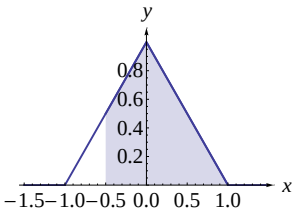
Računanjem površine ispod grafa funkcije f možemo odrediti vjerojatnost da se X realizira realnim brojem iz nekog intervala. Na primjer:



(a) $P\{X \in (0,1)\} = \frac{1}{2}$



(b) $P\{X \in (-1,-\frac{1}{2})\} = \frac{1}{8}$



(c) $P\{X \in (-\frac{1}{2},1)\} = \frac{7}{8}$

4.5 Mjere centralne tendencije i raspršenosti slučajne varijable

Kao što smo kod podataka prikupljenih mjerenjem numeričkih varijabli koristili mjere centralne tendencije i raspršenosti da bismo lakše opisali skup podataka, tako i kod slučajnih varijabli možemo koristiti analogne mjere za lakši opis svojstava slučajnih varijabli. Međutim, teorijska analiza takvih mjera precizno definiranih na osnovi tablice distribucije kod diskretnih, odnosno funkcije gustoće kod neprekidnih, slučajnih varijabli donosi i neke njihove bitne značajke koje se koriste u statističkom zaključivanju. U ovom poglavlju definirat ćemo mjere centralne tendencije i raspršenosti slučajne varijable posebno za diskretne, a posebno za neprekidne slučajne varijable.

Neka je X diskretna slučajna varijabla zadana tablicom distribucije

$$X \sim \begin{pmatrix} x_1 & x_2 & \dots & x_n \\ p_1 & p_2 & \dots & p_n \end{pmatrix}, \quad \text{odnosno} \quad X \sim \begin{pmatrix} x_1 & x_2 & x_3 & \dots \\ p_1 & p_2 & p_3 & \dots \end{pmatrix}.$$

Ako red $\sum_{x_i \in \mathfrak{R}(X)} |x_i| p_i$ konvergira, možemo definirati očekivanje slučajne varijable X kao realan broj

$$\mu = EX = \sum_{x_i \in \mathfrak{R}(X)} x_i p_i.$$

Ako i red $\sum_{x_i \in \mathfrak{R}(X)} x_i^2 p_i$ konvergira, možemo definirati **varijancu** kao realan broj

$$\sigma^2 = \mathrm{Var}X = \sum_{x_i \in \mathfrak{R}(X)} (x_i - \mu)^2 p_i.$$

Primjer 4.18. *Promotrimo bacanje pravilno izrađene igraće kockice. Znamo da će se pri jednom bacanju te kockice okrenuti jedan broj iz skupa* $\{1,2,3,4,5,6\}$*, no ne znamo točno koji. Kako je kockica pravilo izrađena, znamo da se svaki od brojeva iz tog skupa realizira s vjerojatnošću* $1/6$*. Dakle, ishod jednog bacanja ovakve kockice modeliramo diskretnom slučajnom varijablom* X *s tablicom distribucije*

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \end{pmatrix}.$$

Očekivanje ove slučajne varijable je broj

$$EX = \frac{1}{6}(1+2+3+4+5+6) = 3.5,$$

a njezina varijanca broj

$$\mathrm{Var}X = \sum_{k=1}^{6} \frac{1}{6}(k-3.5)^2 \approx 2.92.$$

Neka je X neprekidna slučajna varijabla s funkcijom gustoće f. Ako postoji $\int_{-\infty}^{\infty} |x|\, f(x)\, dx$, onda definiramo **očekivanje** ove slučajne varijable kao realan broj

$$\mu = EX = \int_{-\infty}^{\infty} x\, f(x)\, dx.$$

Ako postoji i $\int_{-\infty}^{\infty} x^2\, f(x)\, dx$, definiramo **varijancu** kao realan broj

$$\sigma^2 = \mathrm{Var}X = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx.$$

Primjer 4.19. *Neprekidna slučajna varjabla iz zadatka 4.14 zadana je funkcijom gustoće*

$$f(x) = \begin{cases} 1/2 \ , & x \in [-1,1] \\ 0 \quad , & x \notin [-1,1] \end{cases}.$$

Izračunajmo očekivanje i varijancu ove neprekidne slučajne varijable:

$$EX = \int_{-\infty}^{\infty} x\, f(x)\, dx = \int_{-1}^{1} \frac{x}{2}\, dx = 0,$$

$$\mathrm{Var}X = \int_{-\infty}^{\infty} (x-EX)^2\, f(x)\, dx = \int_{-1}^{1} \frac{x^2}{2}\, dx = \frac{1}{3}.$$

Drugi korijen iz varijance zovemo **standardna devijacija** slučajne varijable i označavamo ga sa σ.

Očekivanje je jedna od mjera centralne tendencije, a varijanca i standardna devijacija mjere raspršenja oko očekivanja. Tu činjenicu potkrijepljuju mnogi rezultati teorije vjerojatnosti, a jedan od njih je i takozvana **Čebiševljeva nejednakost**.

Čebiševljeva nejednakost:
Neka je X slučajna varijabla koja ima varijancu. Neka je σ standardna devijacija te slučajne varijable, a μ njeno očekivanje. Tada za svaki prirodan broj k vrijedi:

$$P\{|X-\mu| \geq k\sigma\} \leq \frac{1}{k^2}, \quad k \in \mathbb{N}.$$

Primjenom svojstva vjerojatnosti suprotnog događaja slijedi da je

$$P\{|X-\mu| < k\sigma\} \geq 1 - \frac{1}{k^2}.$$

Interpretacije:

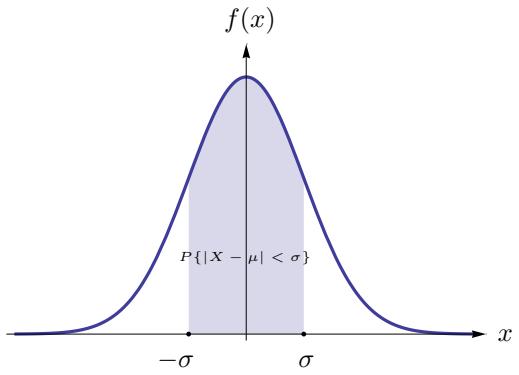
Vjerojatnost da se slučajna varijabla realizira vrijednostima koje su od očekivanja μ udaljene više ili jednako $k\sigma$ manja je ili jednaka $1/k^2$.

Vjerojatnost da se slučajna varijabla realizira vrijednostima koje su od očekivanja μ udaljene manje od $k\sigma$ veća je od $1 - 1/k^2$.

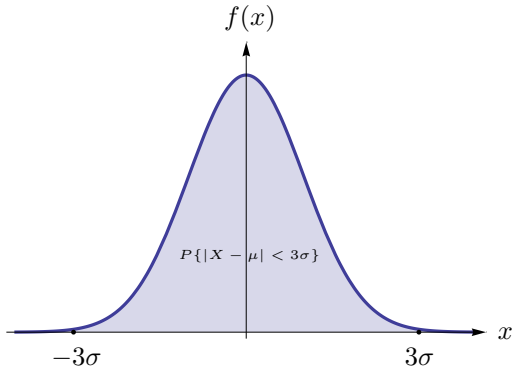
Uvrštavajući $k = 3$, vidimo da realizacija slučajne varijable pada u interval $(\mu - 3\sigma, \mu+3\sigma)$ s vjerojatnošću većom od 0.88 (≈ 0.9). Ta činjenica praktično znači da barem 88% realizacija slučajne varijable X padne u interval $(\mu-3\sigma, \mu+3\sigma)$.

Ove tvrdnje vrijede za sve slučajne varijable koje imaju varijancu pa je za očekivati da tako dobivena ocjena nije jako precizna, ali ona svakako svjedoči o standardnoj devijaciji kao jednoj mjeri raspršenosti realizacija slučajne varijable oko njenog očekivanja.

Na slikama 4.9 i 4.10 prikazana je vjerojatnost $P\{|X-\mu| < k\sigma\}$ za $k = 1$ i $k = 3$ za normalnu slučajnu varijablu X s parametrima $\mu = 0$ i $\sigma = 1$.



Slika 4.9: Prema Čebiševljevoj nejednakosti je $P\{|X-\mu|<\sigma\} = P\{X \in (\mu-\sigma, \mu+\sigma)\} \geq 0$.



Slika 4.10: Prema Čebiševljevoj nejednakosti je $P\{|X-\mu|<3\sigma\} = P\{X \in (\mu-3\sigma, \mu+3\sigma)\} \geq \frac{8}{9}$.

Medijan slučajne varijable X je realan broj m za koji vrijedi da je

$$P\{X \geq m\} \geq \frac{1}{2} \quad \text{i} \quad P\{X \leq m\} \geq \frac{1}{2}.$$

Medijan je također jedna mjera centralne tendencije, ali ne mora nužno biti jedinstven.

Primjer 4.20. *Kockar sudjeluje u igri u kojoj dobiva kada se pri bacanju igraće kockice okrene šestica. No, odlučio je varati i u tu je svrhu nabavio nepravilno izrađenu igraću kockicu za koju je*

$$P(\{k\}) = \frac{1}{15}, \quad k \in \{1,2,3,4,5\}$$

i $P(\{6\}) = 2/3$. *Dakle, bacanje te kockice modeliramo slučajnom varijablom* X *čija je distribucija dana tablicom*

$$X = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 1/15 & 1/15 & 1/15 & 1/15 & 1/15 & 2/3 \end{pmatrix}.$$

Uočimo da je

$$P\{X \leq 6\} = 1 \quad i \quad P\{X \geq 6\} = \frac{2}{3},$$

pa je 6 *medijan slučajne varijable* X*. Također uočimo da ova slučajna varijabla ima jedinstven medijan.*

Primjer 4.21. *a) U primjeru 4.18 definirali smo diskretnu slučajnu varijablu* X *kojom modeliramo bacanje pravilno izrađene igraće kockice i čija je distribucija dana tablicom*

$$X = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \end{pmatrix}.$$

Uočimo da je

$$P\{X \leq 3\} = \frac{1}{2} \quad i \quad P\{X \geq 3\} = \frac{2}{3}$$

te da je

$$P\{X \leq 4\} = \frac{2}{3} \quad i \quad P\{X \geq 4\} = \frac{1}{2},$$

pa slijedi da je svaki realan broj iz intervala $[3, 4]$ *medijan ove slučajne varijable.*

b) U primjeru 4.16 definirali smo diskretnu slučajnu varijablu X *čije su realizacije iznosi dobitka, odnosno gubitka, od neke investicije izraženi u tisućama kuna i čija je distribucija dana tablicom 4.7 koju možemo zapisati i na sljedeći način:*

$$X = \begin{pmatrix} -400 & -200 & -100 & 0 & 100 & 200 & 300 & 400 \\ 0.05 & 0.15 & 0.3 & 0.1 & 0.3 & 0.03 & 0.04 & 0.03 \end{pmatrix}.$$

Uočimo da je

$$P\{X \leq 0\} = P\{X \in \{-400, -200, -100, 0\}\} = 0.6 \quad i \quad P\{X \geq 0\} = P\{X \in \{0, 100, 200, 300, 400\}\} = 0.5,$$

te da je

$$P\{X \leq -100\} = P\{X \in \{-400, -200, -100\}\} = 0.5 \quad i \quad P\{X \geq -100\} = P\{X \in \{-100, 0, 100, 200, 300, 400\}\} = 0.8,$$

pa slijedi da je svaki realan broj iz intervala $[-100, 0]$ *medijan ove slučajne varijable.*

Primjer 4.22. *Promotrimo normalnu slučajnu varijablu s parametrima* 0 *i* 1*. Iz kalkulatora vjerojatnosti u programskom paketu Statistica možemo pročitati da je* $\mu = EX = 0$*,* $\sigma^2 = \mathrm{Var}X = 1$*,* $\sigma = 1$ *i medijan* $= 0$*. Iz distribucije ove slučajne varijable slijedi:*

$$P\{|X - \mu| < 3\sigma\} = P\{|X| < 3\} = P\{-3 < X < 3\} = P\{X < 3\} - P\{X \leq -3\} =$$

$$= \int_{-3}^{3} f(x)\,dx = \int_{-\infty}^{3} f(x)\,dx - \int_{-\infty}^{-3} f(x)\,dx = 0.998650 - 0.001350 = 0.9973.$$

Ocjena ove vjerojatnosti dobivena pomoću Čebiševljeve nejednakosti je (pogledajte sliku 4.10)

$$P\{|X - \mu| < 3\sigma\} = P\{|X| < 3\} \geq 1 - \frac{1}{9} = \frac{8}{9} \approx 0.888.$$

Uočimo da je ocjena dobivena pomoću Čebiševljeve nejednakosti realno gruba.

4.6 Važni primjeri diskretnih i neprekidnih slučajnih varijabli

4.6.1 Bernoullijeva slučajna varijabala

Ako varijabla koju istražujemo može primiti samo dvije vrijednosti (npr. 0 ili 1), možemo je modelirati korištenjem Bernoullijeve slučajne varijable.

Bernoullijeva slučajna varijabla s parametrom $p \in (0,1)$ je svaka slučajna varijabla koja ima tablicu distribucije sljedećeg oblika:

$$X = \begin{pmatrix} 0 & 1 \\ q & p \end{pmatrix}, \quad p \in (0,1), \quad q = 1 - p.$$

Parametar $p \in (0,1)$ ima značenje vjerojatnosti da slučajna varijabla X primi vrijednost 1.

Primjer 4.23. *Igramo kockarsku igru u kojoj ostvarujemo dobitak ako se na igraćoj kocki okrene šestica.*

$$X = \begin{pmatrix} 0 & 1 \\ 5/6 & 1/6 \end{pmatrix}.$$

Dakle, realizaciju šestice možemo modelirati Bernoullijevom slučajnom varijablom: ako se pri bacanju kockice realizira šestica, Bernoullijeva slučajna varijabla X poprima vrijednost 1*, a inače poprima vrijednost* 0*. Uočite da su vjerojatnosti u tablici distribucije slučajne varijable X određene na temelju pretpostavke jednko mogućih ishoda.*

Primjer 4.24. *Izvlačimo jedan proizvod iz velike pošiljke u kojoj je* 2% *loših proizvoda (oznake:* 0 *- loš proizvod,* 1 *- dobar proizvod). Rezultat izvlačenja modeliramo Bernoullijevom slučajnom varijablom s tablicom distribucije*

$$X = \begin{pmatrix} 0 & 1 \\ 0.02 & 0.98 \end{pmatrix}.$$

Očekivanje Bernoulijeve slučajne varijable s parametrom p je

$$EX = 1 \cdot p + 0 \cdot q = p,$$

a **varijanca**

$$\mathrm{Var}X = pq.$$

4.6.2 Binomna slučajna varijabla

Binomna slučajna varijabla vezana je uz n nezavisnih ponavljanja pokusa koji ima samo dva moguća ishoda - uspjeh i neuspjeh (oznake: 1 - uspjeh; 0 - neuspjeh). Pri tome se u svakom izvođenju pokusa uspjeh realizira s vjerojatnošću $p \in (0,1)$. Svako ponavljanje takvog pokusa opisano je Bernoullijevom slučajnom varijablom. Binomna slučajna varijabla s parametrima $n \in \mathbb{N}$ i $p \in (0,1)$ (oznaka $X \sim \mathcal{B}(n,p)$) broji uspjehe u tih n nezavisnih ponavljanja pokusa. Njena distribucija zadana je sljedećom tablicom:

$$X = \begin{pmatrix} 0 & 1 & 2 & \dots & n \\ q^n & \binom{n}{1} pq^{n-1} & \binom{n}{2} p^2 q^{n-2} & \dots & p^n \end{pmatrix}, \quad q = 1-p.$$

Objašnjenje: pokus čijim se jednim izvođenjem može realizirati ili uspjeh (1) ili neuspjeh (0) ponavljamo nezavisno n puta. Zanima nas kolika je vjerojatnost da se pojavi točno k uspjeha (tj. točno k jedinica), $k = 0, 1, \dots, n$. Prema tablici distribucije binomne slučajne varijable slijedi da je

$$P\{X = k\} = \binom{n}{k} p^k q^{n-k}$$

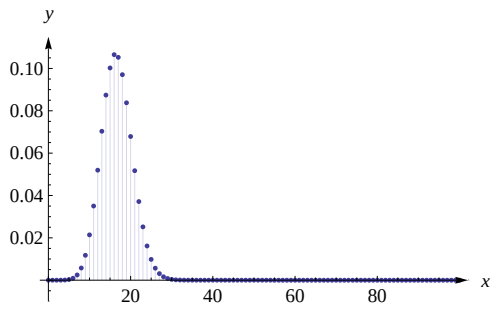
jer se u n nezavisnih ponavljanja pokusa točno k puta (svaki puta s vjerojatnošću p) pojavila realizacija koju nazivamo uspjeh i točno $(n-k)$ puta realizacija koju nazivamo neuspjeh (svaki puta s vjerojatnošću q).

Značenje parametara binomne distribucije:

n - broj nezavisnih ponavljanja pokusa sa samo dva moguća ishoda,

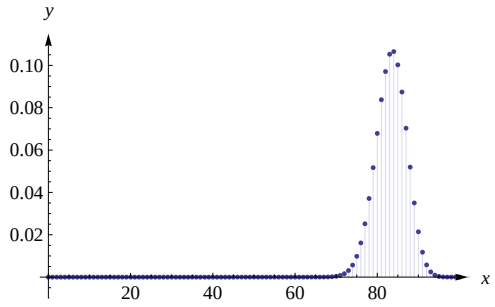
p - vjerojatnost realizacije uspjeha (označenog brojem 1) u jednom izvođenju tog pokusa.

Primjer 4.25. *Broj realizacija šestice pri n nezavisnih bacanja pravilno izrađene igraće kockice možemo modelirati binomnom slučajnom varijablom X s parametrima n i $p = 1/6$. Neka je $n = 100$, dakle $X \sim \mathcal{B}(100, 1/6)$. Stupčasti dijagram koji prikazuje distribuciju (tj. vjerojatnosti iz tablice distribucije) slučajne varijable X prikazan je slikom 4.11, pri čemu je u koordinatnom sustavu na x-osi prikazan broj bacanja kockice, a na y-osi vjerojatnost realizacije šestice u tom broju bacanja.*



Slika 4.11: Graf binomne distribucije s parametrima $n = 100$ i $p = 1/6$.

Nadalje, jedna realizacija slučajne varijabe $Y \sim \mathcal{B}(100, 5/6)$ *u ovom kontekstu je broj koji nam kaže koliko se puta šestica nije pojavila u* 100 *nezavisnih bacanja ove igraće kockice. Stupčasti dijagram koji prikazuje distribuciju slučajne varijable* Y *prikazan je slikom 4.12.*



Slika 4.12: Graf binomne distribucije s parametrima $n = 100$ i $p = 5/6$.

Očekivanje binomne slučajne varijable s parametrom p je

$$EX = np,$$

a **varijanca**

$$\mathrm{Var}X = npq.$$

Primjer 4.26. *Neka je* X *binomna slučajna varijabla s parametrima* $n = 10$ *i* $p = 0.1$*, tj.* $X \sim \mathcal{B}(10, 0.1)$*. Prema tome očekivanje, varijanca i standardna devijacija slučajne varijable* X *su*

$$EX = 1, \quad Var(X) = 0.9, \quad \sigma = \sqrt{0.9} \approx 0.95.$$

Vjerojatnost da realizacija slučajne varijable X *padne u interval* $(EX - \sigma, EX + \sigma) = (0.05, 1.95)$ *je*

$$P\{|X - 1| < 0.95\} = P\{X \in (0.05, 1.95)\} = P\{X = 1\} = 0.38742.$$

Nadalje, vjerojatnost da realizacija slučajne varijable X padne u interval $(EX - 3\sigma, EX + 3\sigma) = (-1.85, 3.85)$ je

$$P\{|X - 1| < 3 \cdot 0.95\} = P\{X \in (-1.85, 3.85)\} =$$
$$= P\{X = 0\} + P\{X = 1\} + P\{X = 2\} + P\{X = 3\} =$$
$$= \sum_{k=0}^{3} \binom{10}{k} 0.1^k 0.9^{10-k} \approx 0.987205.$$

4.6.3 Normalna slučajna varijabala

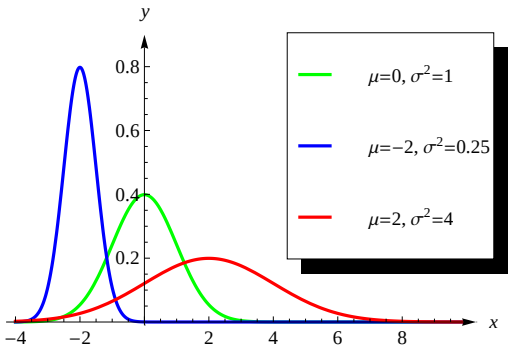
Normalna slučajna varijabla najvažnija je neprekidna slučajna varijabla. Njena važnost posljedica je činjenice da se **suma mnogo nezavisnih i jednako distribuiranih slučajnih varijabli koje imaju konačnu varijancu može dobro aproksimirati slučajnom varijablom s normalnom distribucijom**. Analogna tvrdnja često vrijedi i ako sve slučajne varijable u sumi nisu jednako distribuirane, a također i u nekim slučajevima kada nisu nezavisne.

Normalna slučajna varijabla (oznaka $X \sim \mathcal{N}(\mu, \sigma^2)$) je neprekidna slučajna varijabla za koju je $\Re(X) = \mathbb{R}$, a funkcija gustoće vjerojatnosti definirana je izrazom

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R},$$

gdje je μ bilo koji realan broj, a $\sigma > 0$.

Graf funkcije gustoće normalne slučajne varijable ovisi o izboru parametara μ i σ^2. Na slici 4.13 prikazani su grafovi funkcije gustoće normalne distribucije za različite vrijednosti parametara μ i σ^2.



Slika 4.13: Graf funkcije gustoće normalne distribucije za različite μ and σ^2.

Specijalno, ako je $\mu = 0$, $\sigma^2 = 1$, normalnu slučajnu varijablu zovemo **standardna normalna slučajna varijabla**.

Značenje parametara normalne distribucije:

$$\mu = EX, \quad \sigma^2 = Var\, X.$$

Uočimo:

- funkcija gustoće normalne slučajne varijable ima maksimum za $x = \mu$
- funkcija gustoće normalne slučajne varijable simetrična je u odnosu na pravac koji prolazi maksimumom krivulje i paralelan je y osi
- standardna devijacija je pozitivan broj i ona određuje koliko je funkcija gustoće "široka".

Postupak standardizacije.

Neka je X normalna slučajna varijabla $X \sim \mathcal{N}(\mu, \sigma^2)$. Tada je slučajna varijabla

$$Z = \frac{X - \mu}{\sigma}$$

standardna normalna slučajna varijabla (tj. normalna slučajna varijabla s očekivanjem 0 i varijancom 1).

Primjer 4.27. Pokažite da za $X \sim \mathcal{N}(\mu, \sigma^2)$ vrijede sljedeće tvrdnje:

- Vjerojatnost da realizacija od X padne u interval $[\mu - \sigma, \mu + \sigma]$ iznosi 0.68.
- Vjerojatnost da realizacija od X padne u interval $[\mu - 2\sigma, \mu + 2\sigma]$ iznosi 0.95.
- Vjerojatnost da realizacija od X padne u interval $[\mu - 3\sigma, \mu + 3\sigma]$ iznosi 0.9972.

(Koristite postupak standardizacije i neki kalkulator površine ispod grafa funkcije gustoće standardne normalne slučajne varijable, npr. kalkulator vjerojatnosti iz programskog paketa Statistica.)

4.7 Empirijska distribucija

Pretpostavimo da u statističkom ispitivanju bilježimo realizacije jedne diskretne numeričke varijable u M promatranja. Uvjereni smo da svi ti podaci predstavljaju nezavisne realizacije iste diskretne slučajne varijable X. Za sada pretpostavljamo

da ta diskretna slučajna varijabla X može primiti samo konačno mnogo vrijednosti $x_1, \ldots, x_n$. Tada je X zadana tablicom distribucije

$$X \sim \begin{pmatrix} x_1 & x_2 & \ldots & x_n \\ p_1 & p_2 & \ldots & p_n \end{pmatrix},$$

ali pripadni niz vjerojatnosti p_i, $i = 1, \ldots, n$, ne znamo i želimo ga odrediti na temelju prikupljenih podataka. U tu svrhu prisjetit ćemo se statističke interpretacije vjerojatnosti po kojoj se relativna frekvencija pojavljivanja realizacije x_i u prikupljenim podacima može dovesti u vezu s p_i ako je broj mjerenja dovoljno velik. **Empirijska distribucija** diskretne slučajne varijable X koristi upravo ovu logiku i definira p_i točno kao relativnu frekvenciju pojavljivanja x_i u M ponavljanja mjerenja.
Dakle, ako s f_i označimo frekvenciju pojavljivanja realizacije x_i u podacima, onda je empirijska distribucija ove slučajne varijable zadana tablicom

$$\begin{pmatrix} x_1 & x_2 & \ldots & x_n \\ \frac{f_1}{M} & \frac{f_2}{M} & \ldots & \frac{f_n}{M} \end{pmatrix}, \qquad f_1 + f_2 + \cdots f_n = M.$$

Rezultati koji su dokazani u okviru matematičke statistike pokazuju da će empirijska distribucija to bolje oslikavati stvarnu distribuciju slučajne varijable što je broj promatranja (tj. izmjerenih vrijednosti varijable od interesa) veći.

Primjer 4.28. *U jednoj trgovini uveden je novi proizvod. Nakon nekog vremena vlasnika zanima sviđa li se kupcima taj proizvod ili ne pa je provedeno ispitivanje slučajnog uzorka kupaca. Pri tome je provedeno sljedeće kodiranje odgovora:*

odgovor "ne sviđa mi se" označen je s -1

odgovor "niti mi se sviđa niti mi se ne sviđa" označen je s 0

odgovor "sviđa mi se" označen je s 1.

Bilježenjem odgovora na ovaj način, ispitivanjem 50 *kupaca dobiven je niz nula, jedinica i minus jedinica koji preglednije prikazujemo tablicom frekvencija 4.8.*

x_i	-1	0	1
n_i	24	11	15

Tablica 4.8: Tablica frekvencija odgovora kupaca.

Iz tablice 4.8 možemo odrediti empirijsku distribuciju slučajne varijable X kojom modeliramo odgovor na postavljeno pitanje slučajno odabranog kupca. Ta slučajna varijabla može primiti vrijednosti $-1, 0, 1$, no pripadne vjerojatnosti $P\{X = -1\}$, $P\{X = 0\}$ i $P\{X = 1\}$ nisu nam poznate. Dakle, distribuciju slučajne varijable X ne znamo. Međutim, pomoću tablice frekvencija 4.8 možemo odrediti empirijsku distribuciju slučajne varijable X (tablica 4.9).

$$\begin{pmatrix} -1 & 0 & 1 \\ 0.48 & 0.22 & 0.3 \end{pmatrix}.$$

Tablica 4.9: Empirijska distribucija sl. var. kojom modeliramo odnos kupca prema proizvodu.

Ako pretpostavimo da empirijska distribucija odgovara stvarnoj distribuciji varijable X, možemo donijeti npr. sljedeće zaključke:

vjerojatnost da se slučajno odabanom kupcu iz populacije sviđa novi proizvod je $P\{X = 1\} = 0.3$

ako u trgovinu dođe 200 *kupaca iz pripadne populacije, među njima će biti približno* $200 \cdot 0.3 = 60$ *kupaca kojima se ovaj proizvod sviđa.*

Varijable o kojima želimo zaključivati ne moraju biti uvijek diskretnog tipa s konačnim skupom vrijednosti. Da bismo bili u stanju koristiti prikupljene podatke za aproksimativno računanje vjerojatnosti vezane uz slučajnu varijablu i kod ostalih tipova varijabli, definirat ćemo **empirijsku distribuciju** dobivenu korištenjem podataka $v_1, \ldots, v_M$ koji predstavljaju nezavisne realizacije slučajne varijable X. Prije svega, uočimo da je broj prikupljenih podataka mjerenjem vrijednosti slučajne varijable uvijek konačan. Među izmjerenim podacima može biti i jednakih pa pretpostavimo da se u nizu $v_1, \ldots, v_M$ pojavljuju različite vrijednosti $x_1, \ldots, x_n$ s odgovarajućim frekvencijma $f_1, \ldots f_n$. Na temelju dobivenih podataka možemo definirati empirijsku distribuciju tablicom

$$\begin{pmatrix} x_1 & x_2 & \ldots & x_n \\ \frac{f_1}{M} & \frac{f_2}{M} & \cdots & \frac{f_n}{M} \end{pmatrix}, \qquad f_1 + f_2 + \cdots f_n = M.$$

Neovisno o stvarnom tipu distribucije slučajne varijable iz koje dolaze navedeni podaci, ovako definiranu empirijsku distribuciju možemo koristiti za aproksimativno računanje vjerojatnosti realiziranja varijable X u nekom skupu ako je M velik broj. Tada, npr. vrijedi:

$$P\{X \in [a, b]\} \approx \text{relativna frekvencija pojavljivanja realizacije iz intervala } [a, b].$$

Treba također uočiti da očekivanje empirijske distribucije odgovara aritmetičkoj sredini podataka, a varijanca empirijske distribucije varijanci podataka, tj. ako je S slučajna varijabla definirana empirijskom tablicom distribucije gore opisanih podataka

$$S \sim \begin{pmatrix} x_1 & x_2 & \ldots & x_n \\ \frac{f_1}{M} & \frac{f_2}{M} & \cdots & \frac{f_n}{M} \end{pmatrix}, \qquad f_1 + f_2 + \cdots f_n = M,$$

onda je

$$ES = \frac{1}{n} \sum_i x_i = \bar{x}_n, \quad \text{Var}S = \frac{1}{n} \sum_i (x_i - \bar{x}_n)^2 = s_n.$$

Upitno je koliko je opravdano empirijsku distribuciju podataka prikupljenih na osnovi nezavisnih realizacija slučajne varijable smatrati njezinom pravom distribucijom. Kod varijabli koje su po karakteru neprekidne i želimo ih modelirati kao neprekidne slučajne varijable, očigledno je da računanje vjerojatnosti korištenjem empirijske distribucije može biti samo aproksimacija stvarnih vjerojatnosti (vidi definiciju neprekidne slučajne varijable).
Zapravo, empirijska distribucija podataka prikupljenih na osnovi nezavisnih realizacija slučajne varijable X samo je **procjena** za njenu stvarnu distribuciju dok su aritmetička sredina, varijanca, standardna devijacija i medijan tih podataka **procjene** za očekivanje, varijancu, standardnu devijaciju i medijan slučajne varijable, ali to je tema sljedećih poglavlja.

Primjer 4.29. (gradjevina.sta)
U bazi podataka gradjevina.sta *u varijabli* placa2009 *nalaze se iznosi u eurima prosječnih mjesečnih plaća zaposlenika u* 2009. *godini za* 100 *građevinskih poduzeća srednje veličine u nekoj zemlji. Prirodno je tu varijablu modelirati neprekidnom slučajnom varijablom X koja prima vrijednosti iz intervala $[0, x]$, gdje je x broj koji je veći ili jednak najvišoj ikad zabilježenoj plaći u građevinskom poduzeću srednje veličine u toj zemlji. Za računanje vjerojatnosti vezanih uz realizacije slučajne varijable X trebali bismo poznavati njezinu distribuciju, tj. funkciju gustoće vjerojatnosti. To ovdje, kao i u većini praktičnih problema, nije slučaj. Međutim, raspolažemo sa* 100 *izmjerenih vrijednosti (realizacija) neprekidne slučajne varijable X. Iz tih realizacija možemo odrediti empirijsku distribuciju od X (određujemo ju iz tablice relativnih frekvencija):*

$$\begin{pmatrix} 121 & \dots & 479 & \dots & 1559 \\ 1/100 & \dots & 2/100 & \dots & 1/100 \end{pmatrix}.$$

Uz pretpostavku da empirijska distribucija zadana gornjom tablicom dobro aproksimira stvarnu (nepoznatu) distribuciju neprekidne slučajne varijable X, možemo ju iskoristiti za određivanje približnih vrijednosti vjerojatnosti vezanih uz realizacije od X. Tako je npr. vjerojatnost da je prosječna mjesečna plaća u slučajno odabranom građevinskom poduzeću srednje veličine u toj zemlji veća od 500 *eura približno jednaka* 0.66*, tj.*

$$P\{X > 500\} \approx 0.66,$$

dok je vjerojatnost da je prosječna mjesečna plaća barem 300 *eura, ali manja od* 500 *eura približno jednaka* 0.32*, tj.*

$$P\{300 \leq X < 500\} \approx 0.32.$$

Očekivanje od X procjenjujemo aritmetičkom sredinom 100 *dostupnih realizacija, tj. brojem*

$$\overline{x}_n = 600.13,$$

a standardnu devijaciju procjenjujemo standardnom devijacijom tih podataka, tj. brojem

$$s_n = 194.63.$$

4.8 Zadaci

Zadatak 4.1. Ako imamo jako preciznu vagu i mjerimo neto masu šećera koji je pakiran u vrećice deklarirane mase 1 kg, hoćemo li dobiti točno 1 kg? Ako uzmemo drugo pakiranje istog tipa, koliko vam se čini izvjesno da će neto težina biti ista kao u prethodno vaganom pakiranju? Očekujete li velika odstupanja? Ako neto masu šećera u toj seriji pakiranja modeliramo slučajnom varijablom X, koji biste skup svih mogućih realizacija Vi definirali za tu slučajnu varijablu?

Zadatak 4.2. Iz svežnja koji se sastoji od 32 karte izvlačimo dvije karte za redom. Kolika je vjerojatnost da su obje izvučene karte asovi?

Rješenje. *Budući da iz svežnja izvlačimo dvije karte jednu za drugom, skup Ω ovdje se sastoji od svih parova različitih karata iz svežnja. Zanima nas koliko elemenata ima skup Ω. Odgovor nam daje sljedeće razmatranje:*

- *svežanj se sastoji od* 32 *karte i prva izvučena karta (koja se nakon izvlačenja ne vraća u svežanj) može biti bilo koja karta iz svežnja*
- *prvu izvučenu kartu možemo spariti sa svakom od preostale* 31 *karte u svežnju*
- *takvih parova karata ima* $32 \cdot 31 = 992$*, tj.* $k(\Omega) = 992$.

Na sličan način određujemo broj elemenata skupa A koji se sastoji od svih parova različitih asova. Budući da u svežnju ima četiri različita asa, za svakog prvog izvučenog asa drugog asa biramo od preostala tri asa pa takvih parova ima $4 \cdot 3 = 12$, tj. $k(A) = 12$. Prema tome vrijedi:

$$P(A) = \frac{k(A)}{k(\Omega)} = \frac{12}{992} = \frac{3}{248}.$$

Zadatak 4.3. Pravilno izrađena igraća kockica baca se dva puta. Zanimaju nas vjerojatnosti sljedećih događaja:

a) A - pali su jednaki brojevi

b) B - suma brojeva koji su pali je 8

c) C - produkt brojeva koji su pali je 8.

Rješenje.

a) $A = \{(i,j) \in \Omega : i = j\}$, $P(A) = 6/36 = 1/6$

b) $B = \{(i,j) \in \Omega : i + j = 8\}$, $P(B) = 5/36$

c) $C = \{(i,j) \in \Omega : i \cdot j = 8\}$, $P(C) = 2/36 = 1/18$.

Zadatak 4.4. U kutiji se nalazi 100 papirića numeriranih brojevima $1, 2, \ldots, 100$. Realizacija slučajne varijable X je broj na jednom slučajno izvučenom papiriću. Odredite vjerojatnosti sljedećih događaja:

a) A - izvučeni je broj jednoznamenkast

b) B - izvučeni je broj dvoznamenkast

c) C - izvučeni je broj manji ili jednak 57

d) D - izvučeni je broj strogo veći od 57.

Rješenje.

a) $A = \{1, \ldots, 9\}$, $P(A) = 9/100$

b) $B = \{10, \ldots, 99\}$, $P(B) = 9/10$

c) $C = \{1, \ldots, 57\}$, $P(C) = 57/100$

d) $D = \{58, \ldots, 100\}$, $P(D) = 1 - (57/100) = 43/100$.

Zadatak 4.5. Ako ispunite listić s 12 kombinacija u igri *LOTO* 6 *od* 45, kolika je vjerojatnost da osvojite dobitak na pogođenih svih šest brojeva, a kolika je vjerojatnost da osvojite dobitak na pet pogođenih brojeva?

Zadatak 4.6. Pravilno izrađena igraća kockica baca se dva puta. Zanimaju nas vjerojatnosti sljedećih događaja:

a) A - barem se jednom okrenuo broj 2

b) B - suma brojeva koji su pali je 7

c) C - produkt brojeva koji su pali je 4.

Zadatak 4.7. Na raspolaganju nam je kutija u kojoj se nalazi 150 papirića numeriranih brojevima $1, 2, \ldots, 150$. Realizacija slučajne varijable je broj na jednom slučajno izvučenom papiriću. Odredite vjerojatnosti sljedećih događaja:

a) A - izvučeni je broj djeljiv s tri

b) B - izvučeni je broj troznamenkast

c) C - izvučeni je broj manji ili jednak od 99

d) D - izvučeni je broj strogo veći od 99.

Zadatak 4.8. Iz svežnja od 52 karte na slučajan način biramo 5 karata. Izračunajte vjerojatnost da su izvučene točno tri dame ili točno dva asa.

Zadatak 4.9. Između 100 istovrsnih objekata označenih različitim brojevima od 1 do 100, na slučajan način izabiremo jedan objekt. Odredite vjerojatnosti sljedećih događaja:

a) izabran je objekt označen brojem većim ili jednakom 30 (Rješenje: $71/100$)

b) izabran je objekt označen brojem većim od 30 ili manjim od 10 (Rješenje: $79/100$)

c) izabran je objekt označen parnim brojem (Rješenje: $1/2$)

d) izabran je objekt označen parnim brojem ili brojem većim od 30 (Rješenje: $17/20$)

e) izabran je objekt označen brojem čija je zadnja znamenka 8 (Rješenje: $1/10$)

f) izabran je objekt označen brojem čija zadnja znamenka nije 8 (Rješenje: $9/10$)

g) izabran je objekt označen parnim brojem čija zadnja znamenka nije 8 (Rješenje: $41/100$).

Zadatak 4.10. Pretpostavimo da jednom bacamo pravilno izrađen novčić:

- ako pri bacanju novčić padne na glavu (G), tada jednom gađamo metu jednostavnog karaktera (što znači da su jedine moguće realizacije gađanja mete promašaj (0) ili pogodak(1))
- ako pri bacanju novčić padne na pismo (P), tada jednom bacamo pravilno izrađenu igraću kockicu.

Odredite skup elementarnih događaja tog slučajnog pokusa te korištenjem elementarnih događaja odredite sljedeće događaje:

a) A - pismo je palo točno jednom

b) B - glava je pala točno jednom.

Primjenom skupovne operacije prikažite događaj A pomoću događaja B. Ako je zadana vjerojatnost događaja A, tj. ako je $P(A) = 0.5$, odredite vjerojatnost događaja B.

Zadatak 4.11. Diskretna slučajna varijabla koja može primiti vrijednosti $2, 3, 8$ i 10 zadana je tablicom distribucije 4.10.

vrijednosti	2	3	8	10
vjerojatnosti	0.15	0.10	0.25	0.5

Tablica 4.10: Tablica distribucije slučajne varijable sa slikom $\{2, 3, 8, 10\}$.

Odredite vjerojatnost da ova slučajna varijabla primi vrijednosti manje ili jednake 8.

Rješenje. *Vjerojatnost da ova slučajna varijabla primi vrijednosti manje ili jednake* 8 *je*

$$P\{X \leq 8\} = P\{X \in \{2, 3, 8\}\} = P\{X = 2\} + P\{X = 3\} + P\{X = 8\} = 0.5.$$

Zadatak 4.12. Promotrimo tablicu 4.11.

vrijednosti	2	3	8	10
vjerojatnosti	0.15	0	0.21	0.2

Tablica 4.11: Tablica kojom nije zadana distribucija slučajne varijable sa slikom $\{2, 3, 8, 10\}$.

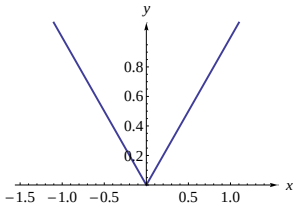
Može li ovom tablicom biti zadana distribucija jedne slučajne varijable?

Rješenje. *Zanima nas je li ovom tablicom zadana distribucija slučajne varijable X sa slikom $\mathfrak{R}(X) = \{2, 3, 8, 10\}$. Vidimo da su brojevi u drugom retku tablice nenegativni (tj. ≥ 0) i manji od jedan, ali u sumi daju* 0.56 *što nije u skladu s drugim navedenim svojstvom distribucije diskretne slučajne varijable. Dakle, konačan niz brojeva* $0.15, 0, 0.21, 0.2$ *ne definira vjerojatnost na skupu* $\{2, 3, 8, 10\}$.

Zadatak 4.13. Funkcija gustoće neprekidne slučajne varijable X dana je izrazom

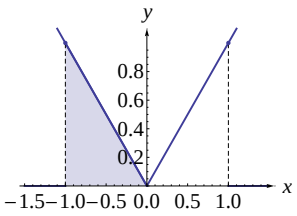
$$f(x) = \begin{cases} |x| \ , \ x \in [-1, 1] \\ 0 \quad , \ x \notin [-1, 1] \end{cases} .$$

Graf funkcije gustoće prikazan je slikom 4.14. Odredite vjerojatnosti $P\{X \in (0, 1)\}$, $P\{X \in (-1, -1/2)\}$ i $P\{X \in (-1/2, 1)\}$.

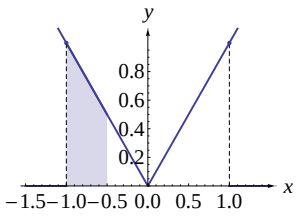


Slika 4.14: Graf funkcije gustoće slučajne varijable iz zadatka 4.13.

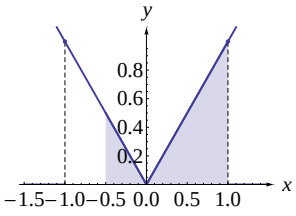
Rješenje. *Analognim postupkom kao u primjeru 4.17 možemo odrediti vjerojatnost da se X realizira realnim brojem iz nekog intervala realnih brojeva.*



(a) $P\{X \in (0, 1)\} = \frac{1}{2}$



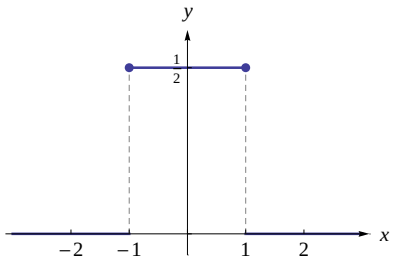
(b) $P\{X \in (-1, -1/2)\} = \frac{3}{8}$



(c) $P\{X \in (-1/2, 1)\} = \frac{5}{8}$

Zadatak 4.14. Funkcija gustoće neprekidne slučajne varijable dana je izrazom

$$f(x) = \begin{cases} 1/2 \ , \ x \in [-1,1] \\ 0 \quad\ , \ x \notin [-1,1] \end{cases} .$$



Slika 4.15: Graf funkcije gustoće slučajne varijable iz zadatka 4.14.

Odredite vjerojatnosti sljedećih događaja:

a) $P\{X \in (0,1)\}$ (Rješenje: $1/2$),

b) $P\{X \in (-1,-1/2)\}$ (Rješenje: $1/4$),

c) $P\{X \in (-1/2,1)\}$ (Rješenje: $3/4$),

d) $P\{X \in (-3/2,1/2)\}$ (Rješenje: $3/4$),

e) $P\{X \in (-2,2)\}$ (Rješenje: 1).

Zadatak 4.15. Pokrenite programski paket Statistica te izaberite:

Statistics → Probability Calculator[4] → Distributions.

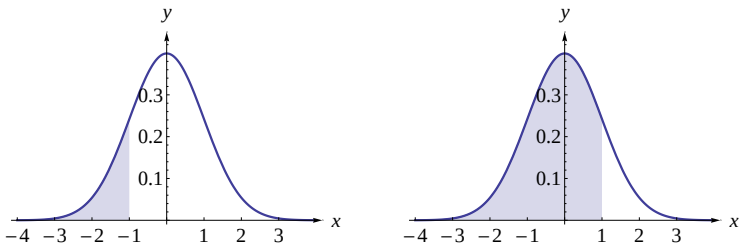
Pogledajte grafove nekih funkcija gustoća vjerojatnosti koje se koriste u primjenama. Diskutirajte o razlikama u grafovima. Odaberite jednu slučajnu varijablu koja prima brojeve bliske nuli s velikom vjerojatnošću.

Zadatak 4.16. U programskom paketu Statistica pod opcijom Distributions u kalkulatoru vjerojatnosti (probability calculator) proučite grafove funkcija gustoća normalne slučajne varijable. Uočite da se u izborniku nalaze i imana drugih neprekidnih slučajnih varijabli koje nismo spominjali. Potražite u dodatnoj literaturi opis Studentove, Fisherove, eksponencijalne i χ^2 slučajne varijable i za svaku od njih, korištenjem programskog paketa Statistica, odredite $P\{X \leq -1\}$, $P\{X \leq 1\}$, $P\{X \geq 1\}$ i $P\{0 < X \leq 1\}$. Pri tome koristite vrijednosti parametara koji su zadani u programskom paketu.

[4] kalkulator vjerojatnosti

Rješenje.

1. *Normalna distribucija s parametrima* $\mu = 0$ *i* $\sigma = 1$ *(mean=0, st.dev.=1):*



(a) $P\{X \leq -1\}$ (b) $P\{X \leq 1\}$

(c) $P\{X \geq 1\}$ (d) $P\{0 < X \leq 1\}$

Slika 4.16: Normalna distribucija - geometrijski prikaz vjerojatnosti.

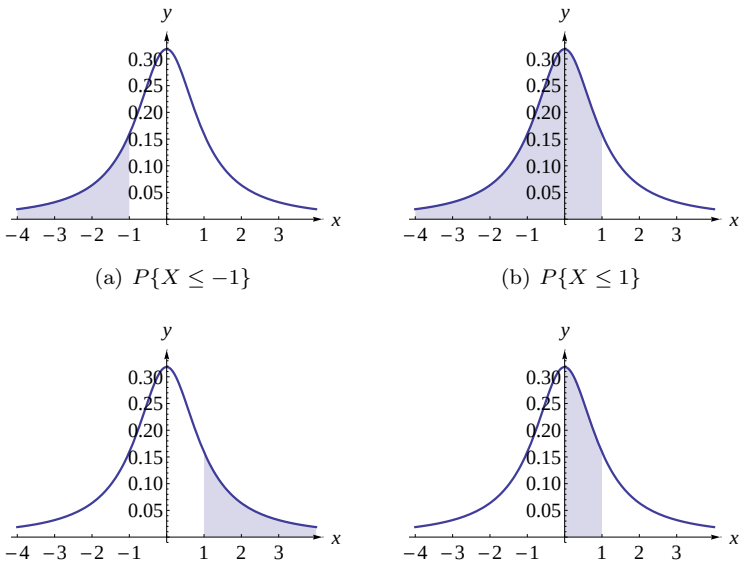
$$P\{X \leq -1\} = \int_{-\infty}^{-1} f(x)\,dx = 0.158655$$

$$P\{X \leq 1\} = \int_{-\infty}^{1} f(x)\,dx = 0.841345$$

$$P\{X \geq 1\} = 1 - \int_{-\infty}^{1} f(x)\,dx = 1 - 0.841345 = 0.158655$$

$$P\{0 < X \leq 1\} = \int_{-\infty}^{1} f(x)\,dx - \int_{-\infty}^{0} f(x)\,dx = 0.841345 - 0.341345 = 0.341345.$$

2. Studentova distribucija s jednim stupnjem slobode (df=1):



(a) $P\{X \leq -1\}$ (b) $P\{X \leq 1\}$

(c) $P\{X \geq 1\}$ (d) $P\{0 < X \leq 1\}$

Slika 4.17: Studentova distribucija - geometrijski prikaz vjerojatnosti.

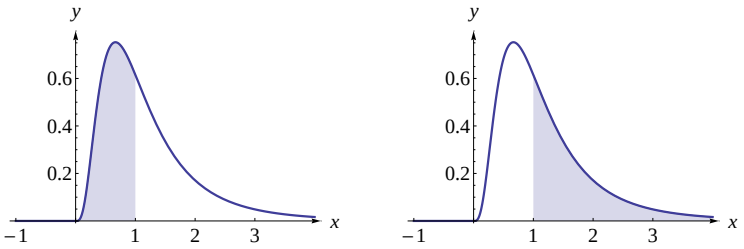
$$P\{X \leq -1\} = \int_{-\infty}^{-1} f(x)\,dx = 0.25$$

$$P\{X \leq 1\} = \int_{-\infty}^{1} f(x)\,dx = 0.75$$

$$P\{X \geq 1\} = 1 - \int_{-\infty}^{1} f(x)\,dx = 1 - 0.75 = 0.25$$

$$P\{0 < X \leq 1\} = \int_{-\infty}^{1} f(x)\,dx - \int_{-\infty}^{0} f(x)\,dx = 0.75 - 0.25 = 0.25.$$

3. Fisherova distribucija sa stupnjevima slobode $m = 10$ i $n = 10$ (df1=10, df2=10):



(a) $P\{X \leq 1\} = P\{0 < X \leq 1\}$ (b) $P\{X \geq 1\}$

Slika 4.18: Fisherova distribucija - geometrijski prikaz vjerojatnosti.

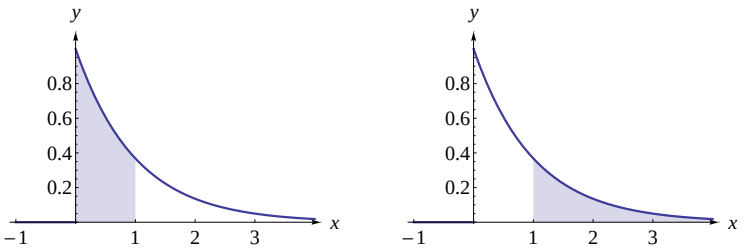
$$P\{X \leq -1\} = \int_{-\infty}^{-1} f(x)\,dx = 0$$

$$P\{X \leq 1\} = \int_{-\infty}^{1} f(x)\,dx = 0.5$$

$$P\{X \geq 1\} = 1 - \int_{-\infty}^{1} f(x)\,dx = 1 - 0.5 = 0.5$$

$$P\{0 < X \leq 1\} = \int_{-\infty}^{1} f(x)\,dx - \int_{-\infty}^{0} f(x)\,dx = 0.5 - 0 = 0.5.$$

4. Eksponencijalna distribucija s parametrom $\lambda = 1$:



(a) $P\{X \leq 1\} = P\{0 < X \leq 1\}$ (b) $P\{X \geq 1\}$

Slika 4.19: Eksponencijalna distribucija - geometrijski prikaz vjerojatnosti.

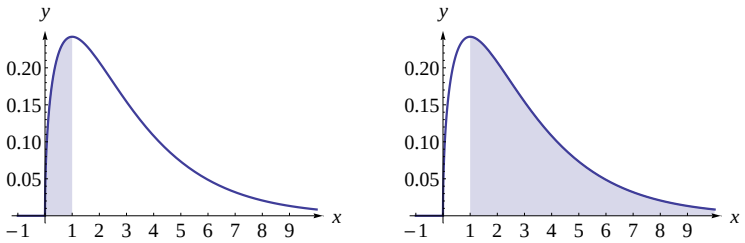
$$P\{X \le -1\} = \int_{-\infty}^{-1} f(x)\,dx = 0$$

$$P\{X \le 1\} = \int_{-\infty}^{1} f(x)\,dx = 0.632121$$

$$P\{X \ge 1\} = 1 - \int_{-\infty}^{1} f(x)\,dx = 1 - 0.632121 = 0.367879$$

$$P\{0 < X \le 1\} = \int_{-\infty}^{1} f(x)\,dx - \int_{-\infty}^{0} f(x)\,dx = 0.632121 - 0 = 0.632121.$$

5. χ^2 distribucija s 3 stupnja slobode (df=3):



(a) $P\{X \le 1\} = P\{0 < X \le 1\}$ (b) $P\{X \ge 1\}$

Slika 4.20: χ^2 distribucija - geometrijski prikaz vjerojatnosti.

$$P\{X \le -1\} = \int_{-\infty}^{-1} f(x)\,dx = 0$$

$$P\{X \le 1\} = \int_{-\infty}^{1} f(x)\,dx = 0.198748$$

$$P\{X \ge 1\} = 1 - \int_{-\infty}^{1} f(x)\,dx = 1 - 0.198748 = 0.801252$$

$$P\{0 < X \le 1\} = \int_{-\infty}^{1} f(x)\,dx - \int_{-\infty}^{0} f(x)\,dx = 0.198748 - 0 = 0.198748.$$

Uočavamo da je kod normalne distribucije s parametrima 0 *i* 1 *i Studentove distribucije s parametrom df* $= 1$ *(tj. s jednim stupnjeva slobode)* $P\{X \leq -1\} = P\{X \geq 1\}$*, što ukazuje na simetričnost tih distribucija. Budući da je kod Fisherove, eksponencijalne i* χ^2 *distribucije* $P\{X \leq -1\} = P\{X \leq 0\} = 0$*, zaključujemo da su te tri distribucije nenegativne, tj. da slučajne varijable s tim distribucijama ne poprimaju negativne vrijednosti.*

Uočimo sličnost grafa funkcija gustoća normalne distribucije s parametrima 0 *i* 1 *i Studentove distribucije. Graf funkcije gustoće Studentove distribucije s povećanjem vrijednoste parametra df (tj. s povećanjem broja stupnjeva slobode) sve više nalikuje grafu funkcije gustoće normalne distribucije s parametrima* 0 *i* 1*. Što je broj stupnjeva slobode veći, to je vjerojatnost da slučajna varijabla sa Studentovom distribucijom poprimi vrijednosti iz nekog intervala realnih brojeva bliža vjerojatnosti da slučajna varijabla s normalnom distribucijom poprimi vrijednosti iz tog istog intervala.*

Kada vrijednost parametra mean normalne distribucije nije 0 *nego npr.* 1*, uočimo da je* $P\{X \leq -1\} = 0.022750$*, a* $P\{X \geq 1\} = 0.5$*. No u ovom je slučaju* $P\{X \leq 1\} = P\{X \geq 1\} = 0.5$ *i također* $P\{X \leq 0\} = P\{X \geq 2\} = 0.158655$*. Zaključujemo da je normalna distribucija simetrična s obzirom na vrijednost parametra mean.*

Zadatak 4.17. Po uzoru na primjer 4.37 odredite očekivanje, varijancu, standardnu devijaciju i medijan diskretnih slučajnih varijabli kojima modeliramo diskretne numeričke varijable iz primjera poglavlja 4.3 (pretpostavite da se stvarna i empirijska distribucija tih slučajnih varijabli podudaraju). Za svaku od promatranih slučajnih varijabli odredite $P\{|X - \mu| < 3\sigma\}$ korištenjem distribucije te dobiveni rezultat usporedite s ocjenom vjerojatnosti dobivenom pomoću Čebiševljeve nejednakosti.

Zadatak 4.18. Neka je distribucija slučajne varijable X dana tablicom:

$$X = \begin{pmatrix} -3 & -2 & -1 & 4 & 5 & 6 \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{12} & \frac{1}{12} & \frac{1}{12} & \frac{5}{12} \end{pmatrix}.$$

a) Odredite vjerojatnost skupova: $\{X < 0\}$, $\{X = -3\}$, $\{X = 0\}$, $\{X > 6\}$, $\{X \geq 5\}$.

b) Odredite njeno očekivanje, varijancu i standardnu devijaciju.

c) Odredite $P\{|X - \mu| \leq 2\sigma\}$, gdje je μ očekivanje a σ^2 varijanaca.

Zadatak 4.19. Poznato je da je u velikom skladištu trgovine informatičkom opremom vjerojatnost pojavljivanja prijenosnog računala s greškom nastalom u proizvodnji jednaka 0.02. Pretpostavimo da iz tog skladišta biramo 10 prijenosnih računala. Odredite sljedeće vjerojatnosti:

a) vjerojatnost da je točno 5 prijenosnih računala s greškom (Rješenje: $7.28922 \cdot 10^{-7}$)

b) vjerojatnost da su s greškom najviše 3 prijenosna računala (Rješenje: 0.999969)

c) vjerojatnost da je s greškom barem 6 prijenosnih računala (Rješenje: $1.25423 \cdot 10^{-8}$).

Zadatak 4.20. Jedno je istraživanje pokazalo da se 5% Amerikanaca boje biti sami u kući tijekom noći. Ako na reprezentativan način odaberemo uzorak od 20 Amerikanaca, odredite sljedeće vjerojatnosti:

a) ima točno pet ljudi u uzorku koji se boje biti sami noću (Rješenje: 0.00224465)

b) ima najviše tri osobe u uzorku koje se boje biti same noću (Rješenje: 0.984098)

c) ima barem tri osobe u uzorku koje se boje biti sami noću (Rješenje: 0.0754837).

Zadatak 4.21. Računovodstvena služba nekog poduzeća utvrdila je da 40% kupaca ne plaća račune na vrijeme. Iz skupa svih kupaca koji su nešto kupili od tog poduzeća na slučajan način odabire se 6 kupaca.

a) Kolika je vjerojatnost da su svi odabrani kupci podmirili račune na vrijeme?
(Rješenje: 0.046656)

b) Kolika je vjerojatnost da je preko 3/4 odabranih kupaca podmirilo račune?
(Rješenje: 0.23328)

c) Kolika je vjerojatnost da 50% odabranih kupaca nije platilo račune na vrijeme?
(Rješenje: 0.27648)

Zadatak 4.22. Vjerojatnost da izvještaj o povratu poreza neke osobe bude ponovo pregledan iznosi 1.5% za prihod manji od 100000 dolara, a 3% ako je prihod jednak 100000 dolara i veći (izvor: Statistical Abstract of the USA, 1998).

a) Kolika je vjerojatnost da poreznom obvezniku, čiji je prihod manji od 100000 $, porezna kartica bude ponovno pregledana, a kolika za onoga čiji je prihod jednak ili veći od 100000 $?
(Rješenje: 0.015, 0.03.)

b) Ako se odabere pet poreznih obveznika s prihodom manjim od 100000 $, kolika je vjerojatnost da će biti pregledana samo jedna porezna prijava, a kolika da će ih biti pregledano više od jedne? (Rješenje: 0.0706002, 0.00218326.)

c) Isto izračunajte za pet poreznih obveznika s prihodom većim od 100000 $.
(Rješenje: 0.132794, 0.00847205.)

d) Koje ste pretpostavke morali postaviti da biste riješili prethodne zadatke upotrebom binomne distribucije? (Rješenje: pretpostavljamo da se radi o malom uzorku (pet osoba) iz velike populacije, što aproksimativno odgovara modelu u kojem pet puta nezavisno ponavljamo isti Bernoullijev pokus. Ta pretpostavka ovdje omogućuje upotrebu binomne distribucije.)

Zadatak 4.23. U pošiljci od 100 čokolada iz neke tvornice nalazi se samo 5% čokolada s lješnjacima, a sve su ostale obične mliječne čokolade. Pretpostavimo da želimo kušati čokoladu s lješnjacima:

- prvo na slučajan način iz pošiljke odaberemo jednu čokoladu i bez obzira na to je li sa lješnjacima ili ne, pojedemo ju
- nakon toga od preostalih čokolada u pošiljci odaberemo još jednu čokoladu.

Kolika je vjerojatnost da je druga odabrana čokolada s lješnjacima, ako znamo da je:

a) prva odabrana čokolada bila obična mliječna čokolada

b) prva odabrana čokolada bila čokolada s lješnjacima.

Zadatak 4.24. Neka je Z standardna normalna slučajna varijabla, tj $Z \sim \mathcal{N}(0,1)$. Odredite sljedeće vjerojatnosti:

a) $P\{-0.5 \leq Z \leq 1.1\}$ (Rješenje: 0.555796)

b) $P\{-0.38 \leq Z \leq 1.72\}$ (Rješenje: 0.605311)

c) $P\{Z \geq 1.6\}$ (Rješenje: 0.054799)

d) $P\{Z \leq -1.8\}$ (Rješenje: 0.035930).

Zadatak 4.25. Prinos usjeva određenog gospodarstva mjeri se količinom proizvoda koji se proizvede po hektaru. Poznato je da se normalna slučajna varijabla može upotrijebiti za opis prinosa kroz vrijeme (izvor: American Journal of Agricultural Economics, 1999). Povijesni podaci pokazuju da prinos pamuka za iduću godinu može biti opisan normalnom distribucijom s očekivanjem 1500 funti po hektaru i standardnom devijacijom 250. Poljoprivredno gospodarstvo koje promatramo bit će profitabilno ako proizvede barem 1600 funti po hektru.

a) Kolika je vjerojatnost da će to gospodarstvo izgubiti novac sljedeće godine? (Rješenje: 0.655422.)

b) Kolika je vjerojatnost da sljedeće godine prinos padne unutar dvije standardne devijacije oko 1500? (Rješenje: 0.9545.)

Zadatak 4.26. Količina novca koji aviokompanije troše na hranu po jednom putniku normalno je distribuirana s očekivanjem 64 kn i standardnom devijacijom 16. Korištenjem statističke interpretacije vjerojatnosti odgovorite na pitanja:

a) Koliki postotak aviokompanija troši više od 100 kn po putniku? (Rješenje: 0.012224.)

b) Koliki postotak aviokompanija troši između 48 i 80 kn po putniku? (Rješenje: 0.68269.)

Zadatak 4.27. Dnevna zarada nekog kafića može se opisati slučajnom varijablom koja ima normalnu distribuciju s očekivanjem 2000 i standardnom devijacijom 250. Korištenjem programskog paketa Statistica odredite vjerojatnost da dnevna zarada tog kafića padne unutar dvije standardne devijacije oko očekivanja, tj. u interval $(\mu - 2\sigma, \mu + 2\sigma)$?

Zadatak 4.28. Odredite vjerojatnosti skupova $\{X \leq 1\}$, $\{X \geq 5\}$, $\{1 < X < 3\}$, ako je X normalna slučajna varijabla s očekivanjem 2 i varijancom 4.

Zadatak 4.29. (kafic.sta)
Broj gostiju koji dnevno dolaze na kavu u jedan kafić nalazi se u bazi podataka kafic.sta.

a) Kojim tipom slučajne varijable možemo modelirati broj gostiju koji dnevno dolaze na kavu u promatrani kafić? Odredite njezinu empirijsku distribuciju.

b) Pretpostavimo da empirijska distribucija odgovara stvarnoj distribuciji te slučajne varijable. Tada možemo odrediti vjerojatnosti vezane uz broj gostiju, što vlasniku kafića može pomoći pri donošenju poslovnih odluka. Na primjer, pretpostavimo da je prije bilježenja broja gostiju vlasnik odlučio da će zaposliti još jednog konobara ako vjerojatnost da će dnevno biti više od 55 gostiju iznosi više od 0.5. Pomoću empirijske distribucije odredite tu vjerojatnost te odgovorite hoće li vlasnik kafića zaposliti još jednog konobara ili ne.

c) Pomoću empirijske distribucije (koja prema pretpostavci odgovara teorijskoj diistribuciji) odredite vjerojatnost da će u jednom danu kafić posjetiti izmedu 50 i 54 gosta.

Rješenje.

a) Empirijska distribucija diskretne slučajne varijable kojom modeliramo broj gostiju koji u jednom danu posjete promatrani kafić dana je tablicom 4.12.

$$\begin{pmatrix} 45 & 46 & \dots & 67 \\ 0.057 & 0.0143 & \dots & 0.0143 \end{pmatrix}.$$

Tablica 4.12: Empirijska distribucija varijable broj-gostiju.

b) $P\{X \geq 56\} = 0.53$*, pa vlasnik kafića ima osnovu zaposliti još jednog konobara.*

c) $P\{50 < X < 54\} = 0.086$.

Zadatak 4.30. (zdravlje.sta)
Varijabla zdravlje baze podataka zdravlje.sta (baza podataka opisana je u primjeru 2.4) sadrži subjektivne ocjene u standardnoj skali od jedan do pet osobnog zdravstvenog stanja za svakog ispitanika. Subjektivnu ocjenu zdravstvenog stanja možemo modelirati slučajnom varijablom X koja može primati vrijednosti iz skupa $\{1, 2, 3, 4, 5\}$.

a) Pomoću zabilježenih vrijednosti varijable zdravlje odredite empirijsku distribuciju te slučajne varijable X i prikažite je stupčastim dijagramom.

b) Uz pretpostavku da empirijska distribucija odgovara stvarnoj distribuciji slučajne varijable X, odredite vjerojatnost da slučajno odabrani ispitanik svoje zdravstveno stanje ocijeni ocjenom većom od 3. (Rješenje: $P\{X > 3\} = 0.4118$.)

c) Uz pretpostavku da empirijska distribucija odgovara stvarnoj distribuciji slučajne varijable X, odredite što je vjerojatnije - da slučajno odabrani ispitanik svoje zdravlje ocijeni kao nedovoljno (ocjena 1) ili da ga ocijeni kao izvrsno (ocjena 5)?
(Rješenje: $P\{X = 1\} = 0.0784$, $P\{X = 5\} = 0.1765$.)

d) Označimo sa Z slučajnu varijablu kojom modeliramo subjektivnu ocjenu zdravstvenog stanja ispitanika ženskog spola, a M slučajnu varijablu kojom modeliramo subjektivnu ocjenu zdravstvenog stanja ispitanika muškog spola. Odredite empirijske distribucije slučajnih varijabli Z i M. Uz pretpostavku o jednakosti empirijskih distribucija stvarnim distribucijama

slučajnih varijabli Z i M odredite je li vjerojatnije da ocjenom izvrstan svoje zdravstveno stanje ocijeni slučajno odabrana žena ili slučajno odabrani muškarac.
(Rješenje: $P\{Z = 5\} = 0.0909$, $P\{M = 5\} = 0.2$.)

Zadatak 4.31. (gradjevina.sta)
Varijable zaposleni2007, zaposleni2008 i zaposleni2009 baze podataka gradjevina.sta sadrže podatke o broju zaposlenika u 100 građevinskih poduzeća srednje veličine u jednoj tranzicijskoj zemlji u 2007., 2008. i 2009. godini. Broj zaposlenika u građevinskim poduzećima srednje veličine možemo modelirati diskretnom slučajnom varijablom X koja prima vrijednosti iz konačnog skupa $\{0, 1, \ldots, n\}$, gdje je $n \in \mathbb{N}$. Koristeći zabilježeni broj zaposlenih u promatranim poduzećima u 2007., 2008. i 2009. godini odredite empirijske distribucije pripadnih slučajnih varijabli (označimo ih s X_{2007}, X_{2008} i X_{2009}). Ako pretpostavimo da dobivene empirijske distribucije odgovaraju stvarnim distribucijama slučajnih varijabli X_{2007}, X_{2008} i X_{2009}, tada nam one mogu biti svojevrsni indikatori kretanja broja zaposlenih u građevinskim poduzećima srednje veličine u promatranom trogodišnjem periodu. Riješite sljedeće zadatke:

a) Kolika je proporcija (relativna frekvencija) srednje velikih građevinskih poduzeća s brojem zaposlenika većim od 50 u 2007., kolika u 2008., a kolika u 2009. godini?
(Rješenje: proporcije su 0.83 za 2007., 0.93 za 2008. te 0.95 za 2009. godinu.)

b) Ako slučajno odaberemo jedno srednje veliko građevinsko poduzeće, kolika je vjerojatnost da ono ima više od 50 zaposlenih u 2007., kolika u 2008., a kolika u 2009. godini?
(Rješenje: $P\{X_{2007} > 50\} = 0.83, P\{X_{2008} > 50\} = 0.93, P\{X_{2009} > 50\} = 0.95$.)

c) Riješite sljedeće zadatke za slučaj da je broj zaposlenika veći od 100 te za slučaj da je broj zaposlenika veći od 200.
(Rješenje: $P\{X_{2007} > 100\} = 0.32, P\{X_{2008} > 100\} = 0.38, P\{X_{2009} > 100\} = 0.36,$
$P\{X_{2007} > 200\} = 0.03, P\{X_{2008} > 200\} = 0.04, P\{X_{2009} > 200\} = 0.03$.)

Zadatak 4.32. (komarci.sta)
Baza podataka komarci.sta sadrži dio rezultata proučavanja komaraca u jednom močvarnom području i detaljnije je opisana u zadatku 2.4. Varijable brojM i brojZ sadrže broj muških i ženskih jedinki komaraca uhvaćenih jednom klopkom za svako od promatranih 210 mjerenja. Ako želimo broj uhvaćenih komaraca tom klopkom modelirati kao slučajnu varijablu, možemo koristiti diskretnu slučajnu varijablu sa skupom vrijednosti $\{0, 1, \ldots, n\}$, pri čemu je $n \in \mathbb{N}$ ograničenje klopke (tj. najveći broj komaraca koji mogu biti ulovljeni korištenom klopkom). Slučajnu varijablu kojom modeliramo broj muških jedinki komaraca označimo s M, a slučajnu varijablu kojom modeliramo broj muških jedinki komaraca označimo sa Z.

a) Koristeći zabilježeni broj muških i ženskih jedinki komaraca u varijablama brojM i brojZ, odredite empirijske distrbucije slučajnih varijabli M i Z te ih prikažite stupčastim dijagramom.

b) Pod pretpostavkom o jednakosti empirijskih i stvarnih distribucija slučajnih varijabli M i Z odgovorite na sljedeće pitanje: što je vjerojatnije - da je u slučajno odabranom mjerenju u klopku uhvaćeno više od 50 muških ili više od 50 ženskih jedinki komaraca?
(Rješenje: $P\{Z > 50\} = 0.1857$, $P\{M > 50\} = 0.0381$.)

Zadatak 4.33. (auto-centar.sta)
Broj dnevno prodanih automobila u jednom autocentru za proteklih 100 dana dan je u varijabli automobili u bazi podataka auto-centar.sta opisanoj u primjeru 2.8.

a) Odredite empirijsku distribuciju slučajne varijable kojom modeliramo broj automobila prodanih u jednom danu u promatranom autocentru.

b) Uz pretpostavku da empirijska distribucija odgovara stvarnoj distribuciji voditelj autocentra može donijeti izvjesne zaključke o dnevoj prodaji što mu može pomoći u donošenju poslovnih odluka. U tom kontekstu odredite vjerojatnost da će u jednom danu biti prodano više od 13 automobila te vjerojatnost da će biti prodano više od 9, ali manje od 12 automobila.

Rješenje.

a) Empirijska distribucija slučajne varijable X kojom modeliramo broj automobila prodanih u jednom danu dana je tablicom 4.13).

$$\begin{pmatrix} 9 & 10 & 11 & 12 & 13 & 14 & 15 & 16 & 17 \\ 0.07 & 0.15 & 0.08 & 0.11 & 0.12 & 0.09 & 0.11 & 0.13 & 0.14 \end{pmatrix}.$$

Tablica 4.13: Empirijska distribucija varijable automobili.

b) $P\{X > 13\} = 0.47$, $P\{9 < X < 12\} = 0.23$.

Zadatak 4.34. (prihod.sta)
Raspolažemo podacima o prihodima za 153 trgovačka poduzeća srednje veličine u jednoj zemlji. Pretpostavimo da prihod takvih poduzeća u promatranoj zemlji možemo modelirati neprekidnom slučajnom varijablom koja prima vrijednosti iz konačnog intervala $[0, r]$, gdje je realan broj r veći ili jednak od ikada zabilježenog prihoda trgovačkog poduzeća srednje veličine u toj zemlji.

a) Koji je najčešći prijavljeni prihod za ispitana poduzeća? (Rješenje: mod je 999999.)

b) Nacrtajte stupčasti dijagram i izračunajte očekivanje i standardnu devijaciju empirijske distribucije. (Rješenje: $\overline{x}_n = 742398.4$, $s_n = 525905.9$.)

c) Uz pretpostavku da empirijska distribucija dobro aproksimira stvarnu distribuciju ove neprekidne slučajne varijable, odredite vjerojatnost da će prihod biti 1200000 i veći te da će prihod biti izmedu 300000 i 700000 eura? (Rješenje: 0.196078, 0.078432.)

Zadatak 4.35. (poduzetnici.sta)
Raspolažemo podacima o dobi 200 poduzetnika u nekoj zemlji. Poznato je da dob poduzetnika u toj zemlji možemo modelirati kontunuiranom slučajnom varijablom X koja prima vrijednosti iz konačnog intervala $[0, s]$, gdje je s starost najstarijeg poduzetnika u toj zemlji. Pretpostavimo da u uvjetima ovog primjera empirijska distribucija dobro aproksimira stvarnu distribuciju ove neprekidne slučajne varijable.

a) Odredite očekivanje i standardnu devijaciju empirijske distribucije.
(Rješenje: $\mu = \overline{x}_n = 42.605$, $s_n = 8.994078$.)

b) Korištenjem empirijske distribucije i statističkog načina računanja vjerojatnosti ocijenite proporciju poduzetnika mlađih od 35 godina u toj zemlji. (Rješenje: $P\{X < 35\} = 0.19$.)

c) Ocijenite kolika je vjerojatnost da slučajno odabrani poduzetnik ima između 46 i 60 godina. (Rješenje: $P\{46 < X < 60\} = 0.275$.)

Zadatak 4.36. Uz pretpostavku o dobroj aproksimiranosti stvarne distribucije empirijskom odredite očekivanje, varijancu i standardnu devijaciju svake neprekidne slučajne varijable iz primjera poglavlja 4.4. Za svaku od navedenih slučajnih varijabli odredite $P\{|X - \mu| < 3\sigma\}$ korištenjem empirijske distribucije. Dobiveni rezultat usporedite s ocjenom vjerojatnosti dobivenom pomoću Čebiševljeve nejednakosti.

Zadatak 4.37. (auto-centar.sta)
Varijablu automobili baze podataka auto-centar možemo modelirati diskretnom slučajnom varijablom koja prima vrijednosti iz konačnog skupa $\{0, 1, \ldots, n\}$, gdje je $n \in \mathbb{N}$ najveći ikada prodani broj automobila u jednom danu u promatranom autocentru. Pretpostavimo da se stvarna i empirijska distribucija (tablica 4.13) ove slučajne varijable podudaraju.

a) Odredite očekivanje, varijancu, standardnu devijaciju i medijan te slučajne varijable.

b) Pomoću empirijske distribucije odredite vjerojatnost da ta slučajna varijabla odstupa od svog očekivanja za manje od tri standardne devijacije te dobiveni rezultat usporedite s Čebiševljevom ocjenom te vjerojatnsoti.

Rješenje.

a) Očekivanje, varijanca, standardna devijacija i medijan slučajne varijable kojom modeliramo broj automobila prodanih u jednom danu dani su u tablici 4.21.

	Descriptive Statistics (auto-centar)				
Variable	Valid N	Mean	Median	Variance	Std.Dev.
automobili	100	13,26	13,00	6,84	2,62

Slika 4.21: Numeričke karakteristike slučajne varijable kojom modeliramo varijablu automobili.

b) Iz empirijske distribucije 4.13 ove slučajne varijable slijedi da je

$$P\{|X - \mu| < 3\sigma\} = P\{\mu - 3\sigma < X < \mu + 3\sigma\} = P\{5.413518 < X < 21.106482\} = 1.$$

Ocjena ove vjerojatnosti dobivena pomoću Čebiševljeve nejednakosti je (pogledajte sliku 4.10)

$$P\{|X - \mu| < 3\sigma\} \geq 1 - \frac{1}{9} = \frac{8}{9} \approx 0.888.$$

Poglavlje 5

Statističko zaključivanje — jedna varijabla

5.1 Procjena distribucije, očekivanja i varijance

U prethodnim poglavljima naučili smo da se veličine promatrane na jedinkama obuhvaćenim nekim istraživanjem nazivaju varijablama te da ih u statistici modeliramo korištenjem slučajnih varijabli. U ovom poglavlju vrijednosti varijable izmjerene na jedinkama iz uzorka (tj. vrijednosti zabilježene u stupac baze podataka) smatramo nezavisnim realizacijama slučajne varijable kojom modeliramo promatranu veličinu. Slučajna varijabla u potpunosti je zadana svojom distribucijom - tablicom distribucije ako se radi o diskretnoj slučajnoj varijabli, odnosno funkcijom gustoće vjerojatnosti ako se radi o neprekidnoj slučajnoj varijabli. Poznavanje distribucije slučajne varijable omogućuje izračunavanje vjerojatnosti vezanih uz njezine realizacije te izračunavanje njezinih numeričkih karakteristika kao što su npr. očekivanje, varijanca i standardna devijacija. Problem se javlja u slučaju kad distribucija slučajne varijable nije poznata jer tada ne možemo točno izračunati vjerojatnosti vezane uz njezine realizacije niti možemo izračunati njezino očekivanje, varijancu i standardnu devijaciju. Problem ovog tipa ilustriran je u primjeru 5.1.

Primjer 5.1. (automobili.sta)
Raspolažemo podacima o realizaciji slučajne varijable X koja opisuje potrošnju goriva novog modela automobila pri brzini od 110 *km/h na autocesti za* 300 *nezavisnih mjerenja. Podaci se nalaze u bazi podataka* automobili.sta*. Često nas zanimaju odgovori na pitanja sljedećeg tipa:*

Kolika je vjerojatnost da je potrošnja goriva tog modela u ovim uvjetima manja od 5.5 *l?*

Kolika je očekivana potrošnja goriva u ovim uvjetima?

Kolika je standardna devijacija slučajne varijable koja opisuje potrošnju goriva u ovim uvjetima?

Kao što je već rečeno, problem pri odgovaranju na ova pitanja jest činjenica da ne znamo stvarnu distribuciju slučajne varijable X koja opisuje potrošnju goriva u danim uvjetima. Ta nam distribucija treba za precizno odgovaranje na postavljena pitanja. Temeljem statističke interpretacije vjerojatnosti znamo da ima smisla koristiti empirijsku distribuciju ovih podataka kao osnovu za računanje vjerojatnosti događaja oblika $P\{X \in [a,b]\}$, $a, b \in \mathbb{R}$, ako imamo velik broj realizacija (tj. mnogo izmjerenih vrijednosti potrošnje goriva u prethodnom primjeru). Što je broj realizacija veći, korištenje empirijske distribucije za računanje ovih vjerojatnosti je opravdanije. Zato kažemo da je **empirijska distribucija podataka** $(x_1, \ldots, x_n)$**, koji predstavljaju nezavisne realizacije jedne slučajne varijable** X**, dobar procjenitelj za distribuciju slučajne varijable** X**.** Što je tih podataka više, empirijska distribucija bit će bliža stvarnoj distribuciji slučajne varijable X.

Ako razmislimo o tipu slučajne varijable koja opisuje potrošnju goriva u gornjem primjeru, prirodno je modelirati je kao neprekidnu slučajnu varijablu. Međutim, empirijska distribucija podataka koju koristimo kao temelj za računanje vjerojatnosti u upravo predloženom postupku je diskretna. Na osnovi poznavanja svojstava varijable koju proučavamo često možemo unaprijed odrediti oblik neprekidne distribucije koju je opravdano koristiti prilikom modeliranja slučajne varijable. Npr., već je spomenuto da suma puno nezavisnih slučajnih utjecaja na varijablu osigurava da se slučajan karakter varijable može opisati korištenjem normalne distribucije s nepoznatim očekivanjem μ i varijancom σ^2. U takvim slučajevima za odrađivanje distribucije iz podataka možemo koristiti pretpostavljeni tip distribucije te procijeniti nepoznate parametre: očekivanje i varijancu.

Za procjenu očekivanja slučajne varijable koristimo aritmetičku sredinu podataka $(x_1, x_2, \ldots, x_n)$ **dobivenih mjerenjem realizacija navedene slučajne varijable u međusobno nezavisnim ponavljanjima pokusa, tj.**

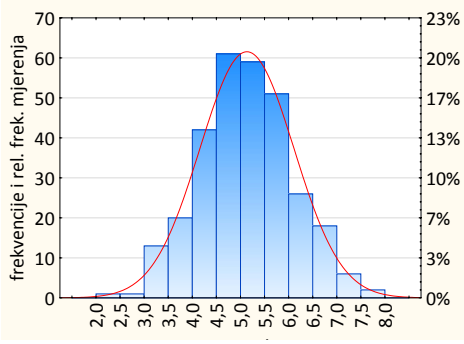
$$\bar{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Za procjenu varijance slučajne varijable koristimo korigiranu varijancu podataka $(x_1, x_2, \ldots, x_n)$ **dobivenih mjerenjem realizacija navedene slučajne varijable u međusobno nezavisnim ponavljanjima pokusa, tj.**

$$s_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2,$$

a za procjenu standardne devijacije koristimo $\sqrt{s_n^2}$.
Korištenjem metoda opisanih u prethodnom razmatranju možemo doći do aproksimativnog odgovora na pitanja koja smo postavili u primjeru 5.1.

Primjer 5.2. (automobili.sta)
Promotrimo podatke o potrošnji goriva iz baze podataka automobili.sta. *Pretpostavimo da su izmjerene vrijednosti potrošnje goriva u primjeru 5.1. realizacije neprekidne slučajne varijable X. Pokušajmo odrediti o kojem se tipu neprekidne slučajne varijable radi tako da kategoriziramo podatke i nacrtamo histogram frekvencija i relativnih frekvencija (slika 5.1).*



Slika 5.1: Histogram izmjerenih vrijednosti potrošnje goriva kategoriziranih u intervale duljine 0.5.

Histogram sa slike 5.1 sugerira da potrošnju goriva u danim uvjetima ima smisla modelirati kao normalnu slučajnu varijablu. Dakle, potrebno je još odrediti očekivanje i varijancu da bi distribucija bila potpuno određena. Stvarno očekivanje i varijancu znamo, no obje ove numeričke karakteristike možemo procijeniti na temelju 300 izmjerenih vrijednosti potrošnje goriva iz baze podataka automobili.sta*:*

$$\begin{aligned} &\textit{procjena očekivanja slučajne varijable } X\textit{:} \quad \overline{x}_{300} = 5.12, \\ &\textit{procjena varijance slučajne varijable } X\textit{:} \quad s^2_{300} = 0.97^2. \end{aligned}$$

Procjene za očekivanje i varijancu možemo iskoristiti kao parametre normalne distribucije kojom vršimo modeliranje, tj. možemo uzeti da je $X \sim \mathcal{N}(5.12, 0.97^2)$. Sada, pomoću ovako određene normalne distribucije, možemo izračunati vjerojatnost da je potrošnja goriva tog modela u navedenim uvjetima manja od 5.5 l. Korištenjem kalkulatora vjerojatnosti u programskom paketu Statistica slijedi da je

$$P\{X < 5.5\} \approx 0.652.$$

Istu vjerojatnost mogli smo izračunati i korištenjem empirijske distribucije slučajne varijable X. Tim pristupom dobivamo da je

$$P\{X < 5.5\} \approx 0.657.$$

U prethodnom primjeru dobili smo dva različita broja kao aproksimacije za $P\{X < 5.5\}$. Logično je da se oni razlikuju jer su to samo procjene za stvarnu vjerojatnost

$P\{X < 5.5\}$ korištenjem različitih metoda. Uočimo da se razlika u ovom primjeru pojavljuje tek na trećoj decimali. Odgovor na pitanje koja metoda daje bolje rezultate nije jednostavan. To je područje kojim se bavi matematička statistika. U ovoj knjizi navest ćemo metode koje je primjereno koristiti pod zadanim pretpostavkama bez detaljnog obrazloženja kriterija na temelju kojih su metode određene.

Za razumijevanje procjene potrebno je uočiti da broj kojim smo aproksimirali $P\{X < 5.5\}$ ne ovisi samo o primijenjenoj metodi, nego i o podacima. Ako se promijene podaci, taj broj više ne mora biti isti niti kod primjene iste metode. Isto se događa i s brojevima kojima smo aproksimirali očekivanje i varijancu. Ilustrirajmo tu činjenicu sljedećim primjerom.

Primjer 5.3. (automobili.sta)
Odredimo procjene za očekivanje i standardnu devijaciju korištenjem samo sto podataka iz baze automobili.sta*.*

Ako koristimo samo prvih 100 podataka (1-100), procjena za očekivanje je 5.17, a za standardnu devijaciju 1.03.

Ako koristimo samo drugih 100 podataka (101-200), procjena za očekivanje je 5.02, a za standardnu devijaciju 0.89.

Ako koristimo samo trećih 100 podataka (201-300), procjena za očekivanje je 5.15, a za standardnu devijaciju 0.10.

(Ponovite procjenu tako da samostalno izaberete 150 *podataka na različite načine.)*

Kako interpretirati dobivene rezultate i što nam zapravo govore izračunate aproksimacije o stvarnim vrijednostima vjerojatnosti, očekivanja i standardne devijacije, lakše ćemo razumjeti nakon što opišemo matematički model **jednostavnog slučajnog uzorka** koji koristimo za modeliranje skupa prikupljenih podataka jedne varijable te pojam **procjenitelj**.

5.1.1 Jednostavni slučajni uzorak i procjenitelj

Do sada smo naučili da varijablu koju istražujemo modeliramo kao slučajnu varijablu, označimo je s X. Podatak x koji smo pri tome dobili mjerenjem (odnosno nekom drugom metodom prikupljanja podataka opisanom u uvodu) jedna je realizacija te slučajne varijable. S obzirom da smo iz te varijable prikupili n podataka, označili smo ih s $x_1, \ldots, x_n$. Pri tome je svaki x_i jedna realizacija slučajne varijable X_i, $i \in \{1, \ldots, n\}$ koja je distribuirana jednako kao slučajna varijabla X. Osim toga, postupak prikupljanja podataka mora biti takav da su mjerenja međusobno nezavisna. Prema tome prirodno je izmjerene podatke $x_1, \ldots, x_n$ smatrati

jednom realizacijom od n slučajnih varijabli $X_1, \ldots, X_n$ koje imaju distribuciju kao X i međusobno su nezavisne. Takav model u statistici zovemo model jednostavnog slučajnog uzorka iz distribucije koja je zadana slučajnom varijablom X.

Jednostavni slučajni uzorak iz distribucije zadane slučajnom varijablom X je uređena n-torka slučajnih varijabli $(X_1, \ldots, X_n)$ od kojih svaka ima istu distribuciju kao X i međusobno su nezavisne.[1]

S obzirom da ćemo u ovom poglavlju koristiti samo model jednostavnog slučajnog uzorka, umjesto ovog dugačkog naziva koristit ćemo termin **uzorak** za taj model, a **realizacija uzorka** za prikupljene podatke.

U trenutku kada radimo procjenu neke numeričke karakteristike slučajne varijable X, primjenjujemo zadanu formulu na jednu realizaciju uzorka (npr. formulu za aritmetičku sredinu jedne realizacije uzorka ako procjenjujemo očekivanje, formulu za korigiranu varijancu realizacije uzorka ako procjenjujemo varijancu, ...). S obzirom da uzorak ima slučajan karakter, u ponovnom prikupljanju podataka dobivamo neku drugu realizaciju koja rezultira drugom vrijednosti za procjenu. Samim tim procjenu ne možemo smatrati determinističkom, već slučajnom veličinom. Dakle, pojedinačna procjena nije ništa drugo do realizacija jedne slučajne varijable, zovemo je **procjenitelj**, slično kao što je jedno mjerenje samo jedna realizacija slučajne varijable koja nas zanima i o kojoj nastojimo nešto zaključiti.

5.1.2 Intervalna procjena

Iako želimo izvršiti procjenu neke numeričke vrijednosti jednim brojem, valja priznati realnost, tj. slučajan karakter procjenitelja te pokušati dobiti što kvalitetniju informaciju iz postupka procjene. U tu svrhu koristimo činjenicu da je procjenitelj slučajna varijabla i vršimo **procjenu intervalom** uz unaprijed izabran broj $\gamma \in (0, 1)$ koji ćemo zvati **pouzdanost intervalne procjene**.
Neka je $\gamma \in (0, 1)$ odabrani broj. Interval pouzdanosti γ (pouzdani interval) za procjenu neke veličine (recimo očekivanja) ustvari nije pravi interval s granicama koje su realni brojevi. To je interval koji ima slučajne varijable kao granice i određen je temeljem zahtjeva da se stvarna

[1] Intuitivno smatramo da su slučajne varijable nezavisne ako činjenica da se dogodio neki događaj prilikom realizacije nekoliko od njih ne mijenja vjerojatnost za pojavu bilo kojeg događaja prilikom realizacije preostalih slučajnih varijabli. Precizniji opis nezavisnosti slučajih varijabli ostavljemo za Poglavlje 6.

vrijednost veličine koju procjenjujemo nalazi u takvom, slučajnom, intervalu s vjerojatnošću barem γ. Svaki puta kad primijenimo formule za određivanje granica intervala pouzdanosti γ na podatke iz uzorka slučajne varijable, dobit ćemo običan interval s realnim brojevima kao granicama. U $100\gamma\,\%$ slučajeva taj izračunati interval realnih brojeva sadržavat će stvarnu vrijednost veličine koju procjenjujemo.

Dakle, interval pouzdanosti γ takozvani je slučajni interval, tj. granice su mu slučajne varijable. Jedna realizacija intervala pouzdanosti γ, određena na osnovi prikupljenog uzorka, običan je interval realnih brojeva. Uobičajeno je u praksi i tu realizaciju pouzdanog intervala također zvati pouzdani interval. Međutim, važno je znati razliku između pouzdanog intervala kao slučajnog intervala i njegove realizacije - običnog intervala realnih brojeva. Pri tome je važno voditi računa o interpretaciji. Ako smo izabrali pouzdanost $95\,\%$, kažemo da smo procijenili danu veličinu intervalom s pouzdanošću $95\,\%$.

5.2 Intervalna procjena očekivanja za velike uzorke

Predmet je ovog poglavlja određivanje intervala izabrane pouzdanosti γ za očekivanje slučajne varijable iz koje smo sakupili velik uzorak.

Neka je $\overline{X}_n$ aritmetička sredina uzorka veličine n iz slučajne varijable X. Pretpostavimo da je očekivanje slučajne varijable X nepoznato i iznosi μ, a varijanca je poznata i iznosi σ^2. Teorija vjerojatnosti pokazuje da aritmetička sredina uzorka, za velike uzorke, ima približno normalnu distribuciju s očekivanjem μ i varijancom $\frac{\sigma^2}{n}$. Korištenjem postupka standardizacije odavde slijedi da slučajna varijabla

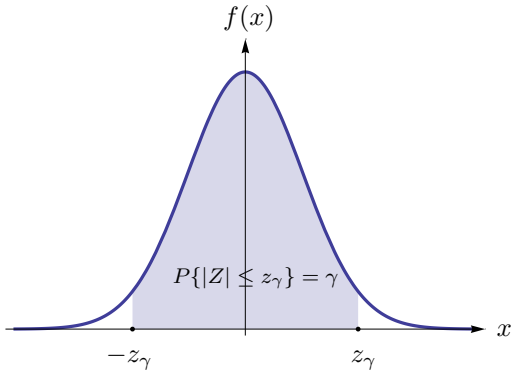
$$Z' = \frac{\overline{X}_n - E\overline{X}_n}{\sqrt{Var(\overline{X}_n)}} = \frac{\overline{X}_n - \mu}{\sigma}\sqrt{n}$$

ima približno standardnu normalnu distribuciju. Označimo sa Z slučajnu varijablu s $\mathcal{N}(0,1)$ distribucijom. Neka je z_γ broj za koji vrijedi

$$P\{|Z| \leq z_\gamma\} = \gamma.$$

Uočimo da vrijednost γ pretstavljaju površinu ispod grafa funkcije gustoće standardne normalne distribucije nad intervalom $[-z_\gamma, z_\gamma]$ (slika 5.2), tj.

$$P\{|Z| \leq z_\gamma\} = \frac{1}{\sqrt{2\pi}} \int_{-z_\gamma}^{z_\gamma} e^{-x^2/2}\,dx = \gamma.$$



Slika 5.2: Vjerojatnost $P\{|Z| \leq z_\gamma\}$.

Uvrštavanjem izraza $Z' = \frac{\overline{X}_n - \mu}{\sigma}\sqrt{n}$ u izraz $P\{|Z| \leq z_\gamma\}$ umjesto Z slijedi:

$$P\{|Z'| \leq z_\gamma\} = P\{-z_\gamma \leq Z' \leq z_\gamma\} =$$

$$= P\left\{-z_\gamma \leq \frac{\overline{X}_n - \mu}{\sigma}\sqrt{n} \leq z_\gamma\right\} =$$

$$= P\left\{\overline{X}_n - z_\gamma \frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X}_n + z_\gamma \frac{\sigma}{\sqrt{n}}\right\}.$$

Dakle, vrijedi:

$$P\left\{\mu \in \left[\overline{X}_n - z_\gamma \frac{\sigma}{\sqrt{n}}, \overline{X}_n + z_\gamma \frac{\sigma}{\sqrt{n}}\right]\right\} \approx \gamma.$$

Ovo razmatranje dovodi do sljedećeg zaključka: ako je $(x_1, \ldots, x_n)$ realizacija uzorka iz slučajne varijable X, $\overline{x}_n$ aritmetička sredina koju smo izračunali iz te realizacije i $\gamma \in (0, 1)$, onda će u približno $100\gamma\,\%$ slučajeva interval izračunat po formuli

$$\left[\overline{x}_n - z_\gamma \frac{\sigma}{\sqrt{n}},\ \overline{x}_n + z_\gamma \frac{\sigma}{\sqrt{n}}\right]$$

$\overline{x}_n$ — aritmetička sredina uzorka
σ — standardna devijacija slučajne varijable X
z_γ — broj za koji vrijedi da je $P\{|Z| \leq z_\gamma\} = \gamma$
Z — standardna normalna slučajna varijabla

sadržavati stvarnu (nepoznatu) vrijednost očekivanja μ slučajne varijable X.

U praksi najčešće ne znamo stvarnu vrijednost standardne devijacije σ. U tom slučaju za velike uzorke procjenjujemo σ korijenom korigirane varijance uzorka, tj.

brojem s_n, i tu procjenu koristimo za izračunavanje realizacije pouzdanog intervala kojim procjenjujemo očekivanje. Isti postupak koristit ćemo i u nastavku, u dijelu koji govori o testiranju statističkih hipoteza.

Primjer 5.4. (automobili.sta)
Za izmjerene vrijednosti potrošnje goriva u uvjetima danim u primjeru 5.1 intervalom pouzdanosti $95\,\%$ *procijenit ćemo očekivanu potrošnju goriva. Mjere deskriptivne statistike potrebne za računanje jedne realizacije intervala pozdanosti* $95\,\%$ *su*

$$n = 300, \quad \overline{x}_{300} = 5.12, \quad s_{300} = 0.97.$$

Vrijednost z_γ *za* $\gamma = 0.95$ *određujemo pomoću kalkulatora vjerojatnosti u Statistici:*

$$z_\gamma = 1.959964 \approx 1.96.$$

Sada uvrštavanjem slijedi:

$$\overline{x}_n - z_\gamma \frac{s_{300}}{\sqrt{n}} = 5.12 - 1.96\,\frac{0.97}{\sqrt{300}} \approx 5.01023,$$
$$\overline{x}_n + z_\gamma \frac{s_{300}}{\sqrt{n}} = 5.12 + 1.96\,\frac{0.97}{\sqrt{300}} \approx 5.22977.$$

Dakle, realizacija intervala pouzdanosti $95\,\%$ *za izmjerene vrijednosti varijable* potrosnja *je*

$$[5.01023, 5.22977].$$

Realizaciju intervala pouzdanosti $95\,\%$ *možemo izračunati i u Statistici provodeći sljedeći postupak:*

Statistics → Basic Statistics/Tables → Descriptive Statistics → Variables → Advanced → označiti "Conf. limits for means interval" i odabrati vrijednost $95\,\%$ → Summary.

Interval pouzdanosti koji kao rješenje daje Statistica je $[5.004597, 5.225560]$. *Razlike u rezultatima posljedica su zaokruživanja vrijednosti mjera deskriptivne statistike u prvom načinu rješavanja.*

Primjer 5.5. (poduzetnici.sta)
Varijabla dob poduzetnika *baze podataka* poduzetnici.sta *sadrži dob u godinama za* 200 *poduzetnika u nekoj zemlji. Procijenimo očekivanje neprekidne slučajne varijable* X *kojom modeliramo dob poduzetnika u toj zemlji intervalima pouzdanosti* $95\,\%$ *i* $97\,\%$ *i usporedimo rezultate. Realizacije intervala pouzdanosti* $95\,\%$ *i* $97\,\%$ *temeljene na godinama starosti* 200 *promatranih poduzetnika jesu*

$$I_{0.95} = [41.35088, 43.85912], \quad I_{0.97} = [41.21490, 43.99510].$$

Uočimo da za ove intervale vrijedi $I_{0.95} \subset I_{0.97}$. *Objašnjenje leži u činjenici da za intervale različitih pouzdanosti* γ_1 *i* γ_2 *takve da je* $\gamma_1 < \gamma_2$ *(npr.* $\gamma_1 = 0.95$, $\gamma_2 = 0.97$*) vrijedi da je*

$$z_{\gamma_1} < z_{\gamma_2},$$

pa za istu realizaciju $(x_1, \ldots, x_n)$ *slučajnog uzorka* $(X_1, \ldots, X_n)$ *vrijedi*

$$\left[\overline{x}_n - z_{\gamma_1}\frac{\sigma}{\sqrt{n}}, \overline{x}_n + z_{\gamma_1}\frac{\sigma}{\sqrt{n}}\right] \subset \left[\overline{x}_n - z_{\gamma_2}\frac{\sigma}{\sqrt{n}}, \overline{x}_n + z_{\gamma_2}\frac{\sigma}{\sqrt{n}}\right].$$

5.3 Intervalan procjena vjerojatnosti događaja za velike uzorke

Vjerojatnost pojavljivanja nekog unaprijed izabranog događaja na osnovi nezavisnih ponavljanja istog pokusa može se dovesti u vezu s pojmom proporcije. To je posljedica interpretacije vjerojatnosti kao odnosa dijela i cjeline, što je ilustrirano u primjeru 5.6.

Primjer 5.6. *Vjerojatnost izvlačenja asa iz svežnja karata odgovara kvocijentu broja asova u svežnju i broja svih karata u svežnju.*

Vjerojatnost pobjede izabrane stranke na izborima odgovara kvocijentu broja osoba koje će glasati za tu stranku i ukupnog broja glasača.

Vjerojatnost izbora pokvarenog proizvoda iz nekog skupa proizvoda odgovara kvocijentu broja pokvarenih proizvoda i broja proizvoda u skupu iz kojeg biramo.

Na primjer, ako želimo procijeniti proporciju loših proizvoda u nekoj velikoj pošiljci možemo se zapitati: "Kolika je vjerojatnost da izvučem loš proizvod iz pošiljke?" Ta vjerojatnost odgovara proporciji loših proizvoda u pošiljci. Dakle, u ovom poglavlju govorimo i o procjeni proporcije i o procjeni vjerojatnosti pojavljivanja izabranog događaja prilikom nezavisnog ponavljanja istog pokusa istovremeno, tj. za oba problema koristimo isti tip statističkog modela.

Statitički model ćemo opisati za problem procjene vjerojatnosti pojavljivanja izabranog događaja, a primjerima ćemo pokazati kako se on koristi u problemu procjene proporcije.

Model za rezultat jednog pokusa u kojem se izabrani događaj dogodi s vjerojatnošću p je Bernoullijeva slučajna varijaba koja je zadana tablicom distribucije

$$X = \begin{pmatrix} 0 & 1 \\ q & p \end{pmatrix} \quad p \in (0,1), \quad q = 1 - p.$$

Pri tome 1 označava realizaciju "uspjeha", a 0 realizaciju "neuspjeha". Dakle, $p = P\{X = 1\}$ je vjerojatnost realizacije "uspjeha".

Nezavisnim ponavljanjem pokusa n puta bilježimo je li se realizirao "uspjeh" (1) ili "neuspjeh" (0). Tako prikupljeni uzorak niz je jedinica i nula (ukupno n njih). Želimo na neki način procijeniti vjerojatnost realizacije "uspjeha", tj. želimo procijeniti parametar p. Međutim, uočimo da je p očekivanje Bernoullijeve slučajne varijable X (potpoglavlje 4.6.1) pa se problem procjene vjerojatnosti p svodi na problem

procjene očekivanja slučajne varijable X. Očekivanje slučajne varijable procjenjujemo aritmetičkom sredinom uzorka. S obzirom da se ovdje uzorak $(x_1, \ldots, x_n)$ sastoji od samih nula i jedinica, aritmetička sredina uzorka odgovara relativnoj frekvenciji jedinica u uzorku.

Za procjenu vjerojatnosti realizacije uspjeha u Bernoullijevoj slučajnoj varijabli, na osnovi n nezavisnih ponavljanja pokusa, koristimo relativnu frekvenciju (proporciju) uspjeha u uzorku, tj. broj

$$\hat{p} = \frac{f_1}{n}.$$

Određivanje pouzdanog intervala za vjerojatnost p možemo ponovo temeljiti na činjenici da, za velike uzorke ($n > 30$), aritmetička sredina uzorka ima približno normalnu distribuciju s očekivanjem koje je jednako očekivanju Bernoullijeve distribucije i varijanci koja je jednaka kvocijentu varijance Bernoullijeve distribucije i veličine uzorka. S obzirom da je, u ovom problemu, očekivanje jednako p, a pq (potpoglavlje 4.6.1), onda slučajna varijabla

$$Z' = \frac{\hat{p} - p}{\sqrt{pq}}\sqrt{n}$$

ima približno standardnu normalnu distribuciju.

Neka je z_γ broj za koji vrijedi da je

$$P\{|Z| \leq z_\gamma\} = \gamma,$$

gdje je $Z \sim \mathcal{N}(0,1)$ (slika 5.2). Uvrštavanjem izraza $Z' = \frac{\hat{p}-p}{\sqrt{pq}}\sqrt{n}$ u izraz $P\{|Z| \leq z_\gamma\} = \gamma$ umjesto Z i analiziranjem nejednakosti $\frac{\hat{p}-p}{\sqrt{pq}}\sqrt{n} \leq z_\gamma$ može se pokazati da vrijedi

$$P\left\{p \in \left[\hat{p} - z_\gamma\sqrt{\frac{\hat{p}\hat{q}}{n}}, \hat{p} + z_\gamma\sqrt{\frac{\hat{p}\hat{q}}{n}}\right]\right\} \approx \gamma.$$

Ovo razmatranje dovodi do sljedećeg zaključka: ako je $\hat{p}$ relativna frekvencija jedinica u n-dimenzionalnom uzorku iz Bernoullijeve distribucije i $\gamma \in (0,1)$, onda će u

približno $100\gamma\,\%$ slučajeva interval izračunat po formuli

$$\left[\hat{p} - z_\gamma \sqrt{\frac{\hat{p}\hat{q}}{n}},\ \hat{p} + z_\gamma \sqrt{\frac{\hat{p}\hat{q}}{n}}\right],$$

$\hat{p}$ — relativna frekvencija jedinice (uspjeha) u uzorku
$\hat{q}$ — relativna frekvencija nula (neuspjeha) u uzorku, $\hat{q} = 1 - \hat{p}$
z_γ — broj za koji vrijedi $P\{|Z| \leq z_\gamma\} = \gamma$
Z — standardna normalna slučajna varijabla

sadržavati pravu (nepoznatu) vrijednost vjerojatnosti p. Također se može pokazati da je broj elemenata u uzorku (n) dovoljno velik za primjenu ovakvog zaključivanja ako interval

$$\left[\hat{p} - 3\sqrt{\frac{\hat{p}\hat{q}}{n}}, \hat{p} + 3\sqrt{\frac{\hat{p}\hat{q}}{n}}\right]$$

ne sadrži ni 0 ni 1. Uočimo da iz ovog razmatranja možemo odrediti veličinu uzorka koja će osigurati zadanu preciznost procjene pouzdanim intervalom, tj. zadanu duljinu intervala.

Primjer 5.7. *Jedna tvornica hrane želi provesti istraživanje tržišta intervjuirajući* 1000 *potrošača kako bi odredila koju marku pahuljica za doručak preferiraju. Prikupljeni podaci pokazali su da* 313 *potrošača odabire upravo marku tvornice koja je provela istraživanje. Na temelju rezultata tog istraživanja možemo odrediti jednu realizaciju intervala pouzdanosti* 95 % *kojim procjenjujemo vjerojatnost da slučajno odabrani potrošač preferira pahuljice tvornice koja je provela istraživanje:*

$$\hat{p} - z_\gamma \sqrt{\frac{\hat{p}\hat{q}}{n}} = 0.313 - 1.96\sqrt{\frac{0.313 \cdot 0.687}{1000}} = 0.284,$$

$$\hat{p} + z_\gamma \sqrt{\frac{\hat{p}\hat{q}}{n}} = 0.313 + 1.96\sqrt{\frac{0.313 \cdot 0.687}{1000}} = 0.342.$$

Dakle, realizacija intervala pouzdanosti 95 % *temeljena na rezultatima istraživanja je interval realnih brojeva* $[0.284, 0.342]$. *Uočimo da taj pouzdani interval možemo interpretirati i kao pouzdani interval proporcije potrošača koji preferiraju danu marku pahuljica za doručak.*

5.4 Testiranje hipoteza

Pretpostavimo da želimo provjeriti je li očekivana vrijednost vremena čekanja u redu studentske menze u vrijeme ručka veća od pet minuta. Naime, ako je veća, onda ćemo u vrijeme ručka pokrenuti još jednu traku u menzi. U tu svrhu od sto slučajno izabranih studenata koji odlaze na ručak u studentsku menzu prikupljamo podatke

o vremenu čekanja za vrijeme ručka. Tako dolazimo do podataka $(x_1, \ldots, x_{100})$ koji su jedna realizacija slučajnog uzorka $(X_1, \ldots, X_{100})$ iz neke, nama nepoznate, distribucije. Da bismo donijeli odluku o pokretanju još jedne trake u menzi, potrebno je testirati hipotezu o iznosu očekivanog vremena čekanja u redu na temelju prikupljenih podataka $(x_1, \ldots, x_{100})$. Takvim i sličnim problemima bavi se teorija testiranja statističkih hipoteza.

Za testiranje hipoteze vezane uz varijablu koja nas zanima koristimo modeliranje varijable kao što je opisano u prethodnim poglavljima, tj. varijable u ispitavanju su slučajne varijable. Slučajna varijabla određena je svojom distribucijom. Kao što je već rečeno, distribucije nam nisu u potpunosti poznate, ali smo naučili kako možemo pribaviti neke informacije o distribuciji na osnovi teorije procjene. **Hipotezu koju želimo testirati korištenjem statističkog testa moramo izraziti u terminima hipoteze koja se odnosi na distribuciju slučajne varijable**. Tako u postupku donošenja odluke o otvaranju nove trake u studentskoj menzi treba testirati jednu hipotezu o vrijednosti očekivanja slučajne varijable koja opisuje vrijeme čekanja u redu studentske menze za vrijeme ručka. Hipotezu koja je formulirana u terminima distribucije slučajne varijable zovemo **statistička hipoteza**.

Postupak testiranja hipoteza uvijek počinje postupkom prevođenja problema koji nas zanima u statističku hipotezu. Primjerice, u uvodnom primjeru u kojem govorimo o mogućnosti otvaranja još jedne trake u studentskoj menzi, u donošenju odluke može nam pomoći testiranje statističke hipoteze da je očekivanje čekanja u redu veće od pet minuta. Statističku hipotezu standardno označavamo s $\mathcal{H}$. **Testirati hipotezu znači donijeti odluku o tome hoćemo li $\mathcal{H}$ odbaciti ili prihvatiti**. Zbog toga često govorimo o testiranju dviju hipoteza u statističkom testu. Jednu od njih zovemo **nul-hipoteza** i označavamo s $\mathcal{H}_0$, a drugu **alternativna hipoteza** i označavamo s $\mathcal{H}_1$. **Alternativna hipoteza je ona koju prihvaćamo u slučaju odbacivanja nul-hipoteze**.
Statistički test koji ćemo koristiti za testiranje statističke hipoteze dizajniran je tako da korištenjem informacija iz prikupljenih podataka o realizacijama slučajne varijable donosimo **odluku o odbacivanju nul-hipoteze** u korist alternativne hipoteze ili **neodbacivanju nul-hipoteze**. Uočimo da nul-hipoteza i alternativna hipoteza u ovoj formulaciji nisu ravnopravne, npr. nigdje nije napisano da prihvaćamo nul-hipotezu. Razlog za ovakvo neobično izražavanje leži u činjenici da se odlučivanje u statističkom testu provodi uz toleranciju malih vjerojatnosti pogrešne odluke. Da bismo bolje razumjeli ovaj koncept, opisat ćemo vrste pogrešaka statističkog testa i mogućnosti koje daje test u odnosu na njihovu kontrolu.

5.4.1 Pogreške statističkog testa

Odluka koja je donesena statističkim testom može biti ili pogrešna ili ispravna. Pri tome se mogu dogoditi dva tipa pogrešne odluke:

pogreška I. tipa: odbaciti $\mathcal{H}_0$ ako je ona istinita
pogreška II. tipa: ne odbaciti $\mathcal{H}_0$ ako je $\mathcal{H}_1$ istinita.

Vjerojatnost pogreške prvog tipa i pogreške drugog tipa ovisi o stvarnoj distribuciji slučajne varijable o kojoj testiramo hipotezu. Htjeli bismo da su te vjerojatnosti pogreške što je moguće manje. Postupak kreiranja statističkog testa, tj. definiranje pravila na osnovi kojih ćemo odlučivati, vodi računa upravo o tom zahtjevu. Statistički test dizajniran je tako da dopušta istraživaču izbor maksimalne vjerojatnosti pogreške prvog tipa koju istraživač želi prihvatiti. Te vrijednosti uglavnom se biraju između brojeva 0.01, 0.05 ili 0.1. Odabrana maksimalna vjerojatnost pogreške prvog tipa zove se **razina značajnosti testa** ili **nivo signifikantnosti testa** i standardno označava s α. Vjerojatnost pogreške drugog tipa određena je dizajnom testa uz izabrani nivo signifikantnosti. Testovi se dizajniraju uz nastojanje da se maksimalna vjerojatnost pogreške drugog tipa učini što manjom i ona se, u pravilu, ne iskazuje u primjeni statističkih testova.

Uzimajući u obzir da ćemo biti u mogućnosti birati maksimalnu vjerojatnost pogreške prilikom odbacivanja nul-hipoteze, to je informacija koju u primjeni testa referiramo. Npr. reći ćemo da **odbacujemo nul-hipotezu na nivou značajnosti α i prihvaćamo hipotezu $\mathcal{H}_1$**, što će značiti da prihvaćamo alternativnu hipotezu uz vjerojatnost najviše α da smo pri tome pogriješili. U suprotnom ćemo reći kako podaci ne podupiru tvrdnju da $\mathcal{H}_0$ treba odbaciti.

Ovakav neravnopravan odnos između nul-hipoteze i alternativne hipoteze prilikom kreiranja statističkog testa upućuje na činjenicu da nije svejedno kako smo izabrali hipoteze i pripadni test. **Ako je moguće, uputno je u primjeni birati statistički test tako da alternativna hipoteza odgovara tvrdnji koju želimo dokazati.**

5.5 Testiranje hipoteza o očekivanju

U ovom poglavlju pokazat ćemo nekoliko statističkih testova koje možemo koristiti prilikom rješavanja problema koji se mogu modelirati analogno kao problem u primjeru o otvaranju nove trake u studentskoj menzi iz prethodnog poglavlja. Način razmišljanja koji treba slijediti u problemima tog tipa objašnjen je u primjeru 5.8.

Primjer 5.8. *Pretpostavimo da želimo provjeriti je li očekivana vrijednost vremena čekanja u redu studentske menze u vrijeme ručka veća od pet minuta. U tu svrhu od sto slučajno izabranih studenata koji odlaze na ručak u studentsku menzu prikupljamo podatke o vremenu čekanja za vrijeme ručka. Tako dolazimo do podataka* $(x_1, \ldots, x_{100})$*. Na osnovi tih podataka aritmetičkom sredinom procijenili smo očekivanje slučajne varijable* X *iz koje potječu ti podaci - procjena je iznosila* 6.5 *minuta. Znajući iz prethodnih proučavanja ove slučajne varijable da je njena varijanca* 25*, ispitajmo je li očekivano vrijeme čekanja u redu za rijeme ručka statistički značajno veće od pet minuta.*

Neka je μ *očekivanje slučajne varijable koja modelira vrijeme čekanja u redu menze za vrijeme ručka. Postavimo hipoteze na sljedeći način:*

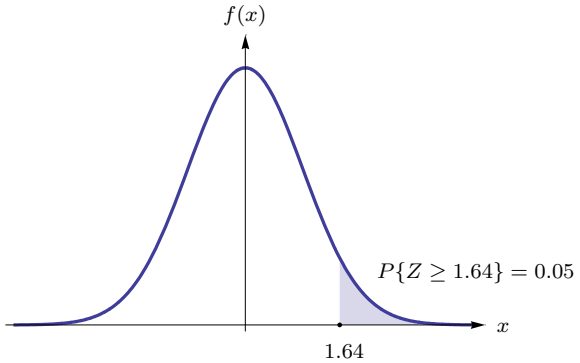
$$\begin{aligned} \mathcal{H}_0\text{:} &\quad \mu = 5 = \mu_0 \\ \mathcal{H}_1\text{:} &\quad \mu > 5 \end{aligned}$$

Ako je $\mathcal{H}_0$ *istinita hipoteza, onda je distribucija aritmetičke sredine uzorka približno normalna s očekivanjem* μ_0 *i varijancom* $\sigma^2/100$*. Dakle, pod pretpostavkom istinitosti nul-hipoteze je distribucija slučajne varijable*

$$Z' = \frac{\overline{X}_{100} - \mu_0}{\sigma}\sqrt{100}$$

približno standardna normalna i velika je vjerojatnost realizacije Z' *blizu nule (slika 5.3). Na primjer, uočimo da se realizacije veće ili jednake* 1.64 *pojavljuju s vjerojatnošću približno* 0.05*, tj. da je*

$$P\{Z' \geq 1.64\} \approx 0.05.$$



Slika 5.3: Vjerojatnost $P\{Z \geq 1.64\}$.

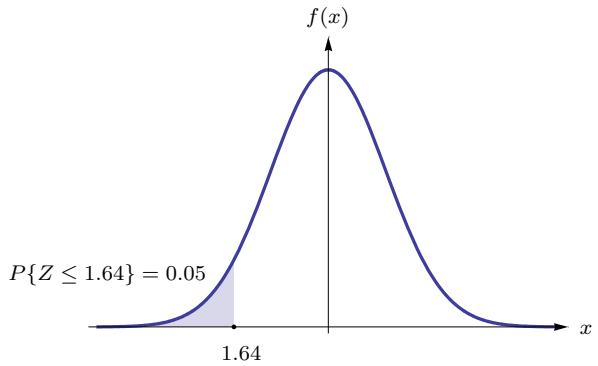
Pretpostavimo da u našem slučaju Z' *realizirala brojem* 3*. Ako je* $\mathcal{H}_0$ *istinita hipoteza, vjerojatnost da se slučajna varijabla* Z' *realizira brojem većim ili jednakim* 3 *iznosi približno* 0.00135*, tj.*

$$P\{Z' \geq 3\} = P\{Z' \in [3, \infty)\} \approx 0.00135.$$

Sada zaključujemo na sljedeći način. Broj 3 *relativno je daleko od nule. Ako je* $\mathcal{H}_0$ *istinita hipoteza, realizacije veće ili jednake* 3 *mogu se pojaviti, ali je vjerojatnost za to tek oko* 0.00135.

Dakle, ako odbacimo nul-hipotezu, vjerojatnost da ćemo time pogriješiti najviše je oko 0.00135*, što je manje od standardno prihvaćenih vrijednosti za maksimalnu vjerojatnost pogreške prvog tipa (tj. nivoa značajnosti testa). To znači da je, na nivou značajnosti* $\alpha = 0.05$*, opravdano odbaciti nul-hipotezu i prihvatiti hipotezu da je očekivanje vremena čekanja u redu studentske menze za vrijeme ručka veće od pet minuta. Za naš problem to znači da treba pokrenuti novu traku u menzi. Izračunatu aproksimaciju maksimalne vjerojatnosti da smo ovom odlukom pogriješili (vjerojatnost koja iznosi* 0.00135*) zovemo* ***p*-vrijednost**.

Na sličan bismo način proveli postupak testiranja na nivou značajnosti $\alpha = 0.05$ u slučaju da je alternativna hipoteza oblika $\mathcal{H}_1 : \mu < \mu_0$. Tada vjerojatnost $p = P\{Z' \leq z\} \approx P\{Z \leq z\}$, $Z \sim \mathcal{N}(0,1)$, uspoređujemo s nivoom značajnosti $\alpha = 0.05$ koji je u ovom slučaju površina ispod grafa funkcije gustoće standardne normalne distribucije nad intervalom $(-\infty, -1.64]$ (slika 5.4).



Slika 5.4: Vjerojatnost $P\{Z \leq 1.64\}$.

U ovim postupcima aritmetičku sredinu uzorka $\overline{x}_n$ koristimo kao procjenu za očekivanje. Objašnjeni postupak općenito zapisujemo na sljedeći način:

Nul-hipoteza:

$$H_0 : \ \mu = \mu_0.$$

Test-statistika:

$$Z' = \frac{\overline{X}_n - \mu_0}{\sigma/\sqrt{n}}.$$

Ovdje je n veličina uzorka, $\overline{X}_n$ aritmetička sredina uzorka, a σ standardna devijacija.

Ako je nul-hipoteza istinita, očekujemo da je na temelju podataka izračunata vrijednost za Z' (označit ćemo je sa $\hat{z}$) blizu 0 jer varijabla Z' ima približno standardnu

normalnu distribuciju. Međutim, ne možemo zanemariti činjenicu da se tako distribuirana slučajna varijablo može realizirati i u intervalu daleko od nule (doduše, vjerojatnost za to je mala, ali ipak je veća od 0).
Ako označimo sa Z slučajnu varijablu s $\mathcal{N}(0,1)$ distribucijom, na osnovi realizacije $\hat{z}$ statistike Z' na podacima možemo odrediti p-vrijednost kao:

- $p = P\{Z \geq \hat{z}\}$ ako je alternativna hipoteza oblika $\mathcal{H}_1 : \mu > \mu_0$
- $p = P\{Z \leq \hat{z}\}$ ako je alternativna hipoteza oblika $\mathcal{H}_1 : \mu < \mu_0$.

Tako izračunatu p-vrijednost uspoređujemo s nivoom značajnosti α. U slučaju da je $p < \alpha$, odbacujemo nul-hipotezu na nivou značajnosti α i prihvaćamo alternativnu hipotezu $\mathcal{H}_1$. Ako je $p > \alpha$, zaključujemo da nemamo dovoljno informacija koje bi poduprle odluku o odbacivanju nul-hipoteze.

Ukoliko pretpostavimo da naš uzorak potječe iz normalne distribucije, analogno testiranje možemo provesti i na malom uzorku.

Nul-hipoteza:

$$H_0 : \quad \mu = \mu_0.$$

Test-statistika:

$$T = \frac{\overline{X}_n - \mu_0}{s_n/\sqrt{n}}.$$

Ovdje je n veličina uzorka, $\overline{X}_n$ aritmetička sredina uzorka, a s_n standardna devijacija uzorka.

Ako je nul-hipoteza istinita, očekujemo da je na temelju podataka izračunata vrijednost za T (označit ćemo je s $\hat{t}$) blizu 0. Zapravo, može se pokazati da, ako je nul-hipoteza istinita, slučajna varijabla T ima Studentovu distribuciju s $(n-1)$ stupnjeva slobode (oznaka za broj stupnjeva slobode u Statistici je df, od eng. degrees of freedom). Na osnovi realizacije $\hat{t}$ na našem uzorku možemo odrediti p-vrijednost kao:

- $p = P\{T \geq \hat{t}\}$ ako je alternativna hipoteza oblika $\mathcal{H}_1 : \mu > \mu_0$
- $p = P\{T \leq \hat{t}\}$ ako je alternativna hipoteza oblika $\mathcal{H}_1 : \mu < \mu_0$.

Na osnovi tako izračunate p-vrijednost zaključujemo o odbacivanju ili neodbacivanju nul-hipoteze kao i do sada.

Primjer 5.9. (televizija.sta)
Godine 1979. *osnovna kablovska televizija u SAD-u u prosjeku je stajala* 7.37 *dolara mjesečno. Godine* 1983. *udruženje kablovskih televizija, koje broji više od* 4000 *kablovskih sustava, zaključilo je da je kablovska televizija poskupjela za samo* 8% *u odnosu na* 1979. *te da ne stoji statistički značajno više od* 8 *dolara mjesečno. No udruženje potrošača sumnja u te izjave pa ćemo ih mi, na temelju* 33 *podatka u bazi* televizija.sta*, provjeriti. U tu svrhu postavljamo sljedeće hipoteze:*

$$\begin{aligned} \mathcal{H}_0: &\quad \mu = \mu_0 = 8, \\ \mathcal{H}_1: &\quad \mu > 8. \end{aligned}$$

Da bismo izračunali vrijednost $\hat{z}$, trebaju nam izračunate vrijednosti $\overline{x}_n$ i s_n:

$$\overline{x}_n = 8.33, \quad s_n = 2.18.$$

Sada slijedi da je

$$\hat{z} = \frac{\overline{x}_n - \mu_0}{s_n/\sqrt{n}} = \frac{8.33 - 8}{2.18 \cdot \sqrt{33}} = 0.87.$$

Korištenjem kalkulatora vjerojatnosti u Statistici slijedi da je u uvjetima istinitosti nul-hipoteze

$$P\{Z' \leq \hat{z}\} \approx p = P\{Z \geq \hat{z}\} = P\{Z \geq 0.87\} = 0.19.$$

Neka je nivo značajnosti testa $\alpha = 0.05$. Budući da je u ovom slučaju $p > \alpha$, na nivou značajnosti $\alpha = 0.05$ ne odbacujemo nul-hipotezu, tj. na nivou značajnosti $\alpha = 0.05$ nemamo argumenata tvrditi da kablovska televizija stoji statistički značajno više od 8 *dolara mjesečno.*

5.6 Testiranje hipoteza o vjerojatnosti događaja za velike uzorke

U ovom poglavlju ponovno se bavimo statističkim zaključivanjem o Bernoullijevoj distribuciji. Neka je slučajan pokus modeliran Bernoullijevom slučajnom varijablom s tablicom distribucije

$$X = \begin{pmatrix} 0 & 1 \\ q & p \end{pmatrix}, \quad p \in (0,1), \quad q = 1 - p.$$

Testirat ćemo hipotezu o vrijednosti parametra p koji ima značenje vjerojatnosti realizacije "uspjeha" u jednom izvođenju pokusa koji se realizira ili "uspjehom" (oznaka 1) ili "neuspjehom" (oznaka 0). U ovom postupku koristimo relativnu frekvenciju (proporciju) realiziranih "uspjeha" (tj. jedinica) kao procjenu za vjerojatnost p.

Nul-hipoteza:

$$H_0: \quad p = p_0.$$

Test-statistika:

$$Z' = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}.$$

n je veličina uzorka, a $\hat{p}$ relativna frekvencija "uspjeha".

Ako je nul-hipoteza istinita, očekujemo da je na temelju podataka izračunata vrijednost za Z' (označit ćemo je sa $\hat{z}$) blizu 0. Može se pokazati da, ako je nul-hipoteza istinita, slučajna varijabla Z' ima standardnu normalnu distribuciju. Označimo li sa Z slučajnu varijablu s $\mathcal{N}(0,1)$ distribucijom, na osnovi realizacije $\hat{z}$ na našem uzorku možemo odrediti p-vrijednost kao:

- $p = P\{Z \geq \hat{z}\}$ ako je alternativna hipoteza oblika $\mathcal{H}_1 : p > p_0$
- $p = P\{Z \leq \hat{z}\}$ ako je alternativna hipoteza oblika $\mathcal{H}_1 : p < p_0$.

Tako izračunatu p-vrijednost uspoređujemo s nivoom značajnosti α. U slučaju da je $p < \alpha$, na nivou značajnosti α odbacujemo nul-hipotezu $\mathcal{H}_0$ i prihvaćamo alternativnu hipotezu $\mathcal{H}_1$. Ako je $p > \alpha$, nemamo dovoljno informacija koje bi poduprle odluku o odbacivanju nul-hipoteze.
Pokazuje se da je uzorak dovoljno velik za provođenje ovog statističkog testa ako interval

$$\left[p_0 - 3\sqrt{\frac{p_0(1-p_0)}{n}}, p_0 + 3\sqrt{\frac{p_0(1-p_0)}{n}}\right]$$

ne sadrži ni 0 ni 1.

Primjer 5.10. (vrtic.sta)
U nekom poduzeću zaposleno je više od 3000 *ljudi. Uprava poduzeća želi ponuditi pomoć svojim zaposlenicima oko organizacije čuvanja djece. Predložene su dvije opcije - otvaranje vrtića u sklopu poduzeća ili plaćanje dijela troškova čuvanja djece koje bi roditelji organizirali sami. Da bi se utvrdilo koja je od ovih dviju mjera popularnija među zaposlenicima, odabran je uzorak od* 60 *roditelja s malom djecom koji su se izjasnili o tome koju opciju preferiraju. Njihovi odgovori označeni su na sljedeći način:*

0 - *radije bih novčanu pomoć za samostalno organiziranje čuvanja djece*
1 - *radije bih da se otvori vrtić u sklopu poduzeća.*

Pretpostavimo da uprava neće organizirati vrtić u sklopu poduzeća ako se pokaže da je proporcija roditelja koji podržavaju tu ideju manja od 0.75. *Da bismo to provjerili, postavljamo sljedeće hipoteze:*

$$\begin{aligned} \mathcal{H}_0: &\quad p = p_0 = 0.75, \\ \mathcal{H}_1: &\quad p < 0.75. \end{aligned}$$

Za izračunavanje vrijednosti $\hat{z}$ treba nam relativna frekvencija (proporcija) roditelja iz uzorka koji podržavaju ideju o organizaciji vrtića u sklopu poduzeća:

$$\hat{p} = 38/60 = 0.63.$$

Sada slijedi da je

$$\hat{z} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.63 - 0.75}{\sqrt{\frac{0.75 \cdot 0.25}{60}}} = -2.15.$$

Korištenjem kalkulatora vjerojatnosti u Statistici slijedi da je, pod pretpostavkom istinitosti nul-hipoteze,

$$P\{Z' \leq \hat{z}\} \approx p = P\{Z \leq \hat{z}\} = P\{Z \leq -2.15\} = 0.016.$$

Neka je nivo značajnosti testa $\alpha = 0.05$. Budući da je u ovom slučaju $p < \alpha$ na nivou značajnosti $\alpha = 0.05$ odbacujemo nul-hipotezu i prihvaćamo alternativnu hipotezu što u ovoj situaciji znači da uprava nema osnovu organizirati vrtić u sklopu poduzeća.

5.7 Testiranje hipoteza o distribuciji općenito

U ovom poglavlju kao procjenu za stvarnu (nepoznatu) distribuciju slučajne varijable koristimo empirijsku distribuciju određenu na temelju podataka koje smo skupili kao nezavisne realizacije slučajne varijable. Želimo testirati ima li naša slučajna varijabla neku pretpostavljenu distribuciju (zovemo je **teorijska distribucija**).

5.7.1 χ^2 test

Neka je teorijska distribucija slučajne varijable zadana tablicom 5.1.

$$\begin{pmatrix} x_1 & x_2 & \dots & x_k \\ p_1 & p_2 & \dots & p_k \end{pmatrix}.$$

Tablica 5.1: Teorijska distribucija diskretne slučajne varijable.

Očito se ovdje radi o distribuciji jedne diskretne slučajne varijable X s konačnom slikom $\mathfrak{R}(X) = \{x_1, \dots, x_k\}$. Testiranje hipoteze da podaci dolaze iz pretpostavljene teorijske distribucije može se provesti tzv. χ^2-testom.

Hipoteze χ^2 testa su:

$\mathcal{H}_0$: distribucija iz koje dolaze podaci jednaka je teorijskoj
$\mathcal{H}_1$: distribucija iz koje dolaze podaci razlikuje se od teorijske.

Neka je n broj prikupljenih podataka. Za testiranje ove hipoteze koristi se test-statistika temeljena na odstupanju stvarnih frekvencija podataka (f_i, $i = 1, \ldots, k$) od teorijskih (np_i, $i = 1, \ldots, k$) definirana izrazom

$$D = \sum_{i=1}^{k} \frac{(np_i - f_i)^2}{np_i}.$$

Pokazuje se da, pod pretpostavkom istinitosti hipoteze $\mathcal{H}_0$, slučajna varijabla D za velike n ima približno χ^2 distribuciju sa stupnjem slobode $(k-1)$ pa se ta statistika može iskoristiti za testiranje $\mathcal{H}_0$ na uobičajeni način.

χ^2 test možemo provesti u većini statističkih programskih paketa. U programskom paketu Statistica potrebno je formirati bazu podataka koja sadrži eksperimentalno dobivene frekvencije i teorijske frekvencije izračunate na bazi teorijske distribucije i veličine uzorka.

S obzirom da je distribucija statistike D približno χ^2 za velike uzorke, potrebno je voditi računa o veličini uzorka prilikom provođenja ovog testa. Može se pokazati da je korištenje χ^2 testa prikladno ako su sve teorijske frekvencije veće od 5, tj. ako je umnožak veličine uzorka n sa svakom vjerojatnošću p_i veći od 5.

χ^2 test može se koristiti također i za diskretne distribucije s prebrojivim skupom stanja kao i za neprekidne teorijske distribucije. Pri tome je potrebno sliku $\mathfrak{R}(X)$ neprekidne slučajne varijable razdvojiti na disjunktne intervale i suprotstaviti teorijske frekvencije tih intervala njihovim uzoračkim frekvencijama. Međutim, pokazuje se da je test jako osjetljiv na izbor podjele slike $\mathfrak{R}(X)$ na disjunktne intervale.

Primjer 5.11. *Tržišni analitičar želi istražiti imaju li potrošači neke posebne sklonosti prema jednom od okusa sokova koji su se pojavili na tržištu. Na uzorku od* 100 *ljudi prikupio je preferencije prema ponuđenih pet okusa. Frekvencije zabilježene tim istraživanjem dane su u tablici 5.2.*

višnja	jagoda	naranča	limun	grejp
32	28	16	14	10

Tablica 5.2: Tablica empirijskih frekvencija za pet ponuđenih okusa sokova.

Ako želimo ispitati postoji li, na nivou značajnosti $\alpha = 0.05$, statistički značajna preferencija potrošača prema nekom od ponuđenih okusa ili je sklonost potrošača jednaka prema svim ponuđenim okusima, možemo provesti χ^2 test, pri čemu je teorijsku distribuciju zadajemo tablicom:

$$\begin{pmatrix} \textit{višnja} & \textit{jagoda} & \textit{naranča} & \textit{limun} & \textit{grejp} \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \end{pmatrix}.$$

Za provođenje χ^2 testa u Statistici trebamo napraviti tablicu s empirijskim i teorijskim frekvencijama svih vrijednosti koje slučajna varijabla s danom distribucijom može primiti. Empirijske frekvencije dane su u tablici 5.2. Teorijske frekvencije određujemo iz poznate teorijske distribucije. U načem slučaju sve su teorijske frekvencije jednake i iznose $100 \cdot (1/5) = 20$. Frekvencije na temelju kojih provodimo χ^2 test dane su u tablici 5.3.

okus	empirijske frekvencije	teorijske frekvencije
višnja	32	20
jagoda	28	20
naranča	16	20
limun	14	20
grejp	10	20

Tablica 5.3: Tablica empirijskih i teorijskih frekvencija za χ^2 test.

p-vrijednost dobivena u Statistici je 0.001234*, što je manje od nivoa značajnosti $\alpha = 0.05$. Dakle, na tom nivou značajnosti odbacujemo nul-hipotezu i prihvaćamo alternativnu hipotezu, tj. možemo tvrditi da postoji statistički značajna preferencija potrošača prema nekim od ponuđenih vrsta sokova.*

5.7.2 Kako saznati dolaze li podaci iz normalne distribucije?

Ako se radi o neprekidnoj slučajnoj varijabli, u ovom kolegiju prvenstveno ćemo se baviti odgovorom na pitanje ima li ona normalnu distribuciju ili ne. Odgovor na ovo pitanje od iznimne je važnosti za točnost statističkih analiza s obzirom da su mnogi statistički testovi kreirani uz pretpostavku normalnosti obilježja.

Za prvi uvid u moguća odstupanja od normalne distribucije možemo koristiti razne mjere deskriptivne statistike i grafičke prikaze (npr. stupčaste dijagrame relativnih frekvencija), no to nije dovoljno za donošenje zaključka o normalnoj distribuiranosti varijable.

Navodimo dva testa koji se mogu koristiti za testiranje hipoteza:

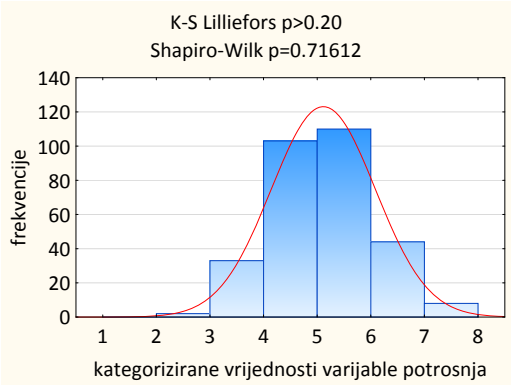
$\mathcal{H}_0$: varijabla ima normalnu distribuciju
$\mathcal{H}_1$: varijabla nema normalnu distribuciju,

a ugrađeni su u većinu statističkih aplikativnih programa. To su:

- Lillieforsova inačica Kolmogorov-Smirnovljeva testa
- Shapiro-Wilk W test.

Važno je naglasiti da su oba testa primjenjiva samo u slučaju velikog broja podataka. Ovdje nećemo opisivati test statistike na osnovi kojih su testovi kreirani. Bit će dovoljno da ih naučimo koristiti i pravilno interpretirati njihove rezultate. U tu svrhu treba voditi računa o tome da nul-hipoteza kod oba testa ide u prilog normalnosti distribucije. Ako je p-vrijednost dobivena provođenjem tih testova na podacima manja od izabranog nivoa značajnosti α, tada odbacujemo nul-hipotezu koja kaže da podaci dolaze iz normalne distribucije.

Primjer 5.12. (automobili.sta)
U bazi podataka automobili.sta *nalaze se rezultati mjerenja prosječne potrošnje novog tipa automobila pri brzini od* 110 *km/h na autocesti za* 300 *nezavisnih mjerenja. Sjetimo se da smo u primjeru 5.2, na temelju stupčastih dijagrama 5.1, zaključili kako ima smisla modelirati ovu varijablu kao normalnu slučajnu varijablu s očekivanjem* $\overline{x}_{300} = 5.12$ *i varijancom* $s^2_{300} = 0.97^2$. *Sada možemo testirati hipotezu o normalnosti distribucije. Zanima nas možemo li, na nivou značajnosti* $\alpha = 0.05$, *tvrditi da je slučajna varijabla kojom modeliramo ovu potrošnju normalno distribuirana.*



Slika 5.5: Stupčasti dijagram izmjerenih vrijednosti potrošnje goriva s p-vrijednostima za Shapiro Wilk test i Lilleforsovu inačicu Kolmogorov-Smirnovljeva testa.

Sa stupčastog dijagrama 5.5 vidimo da su i kod Shapiro Wilk testa i Lilleforsove inačice Kolmogorov-Smirnovljeva testa p-vrijednosti veće od 0.05. *Dakle, na nivou značajnosti* $\alpha = 0.05$ *ne odbacujemo nul-hipotezu da je varijabla normalno distribuirana.*

5.8 Zadaci

Zadatak 5.1. (poduzetnici.sta)
Baza podataka poduzetnici.sta sadrži podatke o godinama starosti za 200 poduzetnika (varijabla dob poduzetnika).

a) Procijenite očekivanje i standardnu devijaciju slučajne varijable X kojom se modelira dob poduzetnika. (Rješenje: $\overline{x}_{200} = 42.61$, $s_{200} = 8.99$.)

b) Kategorizirajte podatke s kojima raspolažete te odlučite ima li smisla modelirati ovu varijablu kao normalnu slučajnu varijablu. Ako ima, korištenjem normalne distribucije s procijenjenim vrijednostima očekivanja i varijance odredite vjerojatnost da je poduzetnik stariji od 30, ali mlađi od 40 godina. Istu vjerojatnost izračunajte i korištenjem empirijske distribucije slučajne varijable X te usporedite rezultate.
Rješenje: *Iz empirijske distribucije slučajne varijable X slijedi da je $P(30 < X < 40) = 0.265$. Ako X modeliramo kao $\mathcal{N}(42.61, 8.99)$ slijedi da je $P\{30 < X < 40\} = 0.31$.*

Zadatak 5.2. (gradjevina.sta)
Baza podataka gradjevina.sta sadrži neke podatke o organizaciji i poslovanju za 100 građevinskih poduzeća srednje veličine u nekoj zemlji (za detaljniji opis pogledajte zadatak 4.31).

a) Procijenite očekivanje i standardnu devijaciju slučajne varijable X kojom se modelira prosječna plaća zaposlenika u građevinskim poduzećima srednje veličine u toj zemlji u 2009. godini. (Rješenje: $\overline{x}_{100} = 600.13$, $s_{100} = 194.63$.)

b) Kategorizirajte podatke s kojima raspolažete te odlučite ima li smisla modelirati ovu varijablu kao normalnu slučajnu varijablu. Ako smatrate da ima, korištenjem normalne distribucije s procijenjenim vrijednostima očekivanja i varijance odredite vjerojatnost da je u 2009. godini u slučajno odabranom poduzeću srednje veličine u toj zemlji prosječna plaća bila viša od 500 eura. Istu vjerojatnost izračunajte i korištenjem empirijske distribucije slučajne varijable X te usporedite rezultate.
Rješenje: *Iz stupčastog dijagrama relativnih frekvencija vidimo da normalna distribucija nije prikladna za modeliranje ovih podataka, a to sugeriraju i izračunate tražene vjerojatnosti: iz empirijske distribucije slučajne varijable X slijedi da je $P(X > 500) = 0.66$, a ako X modeliramo kao $\mathcal{N}(600.13, 194.63^2)$ slijedi da je $P\{X > 500\} = 0.696536$.*

Zadatak 5.3. (farmakologija.sta)
Baza podataka farmakologija.sta sadrži podatke o koncentraciji nekog lijeka u organizmu za 101 mjerenje provedeno od trenutka unosa lijeka u organizam do trenutka njegove eliminacije iz organizma (varijabla koncentracija lijeka).

a) Kategorizirajte izmjerene vrijednosti varijable koncentracija lijeka i nacrtajte stupčasti dijagram frekvencija i relativnih frekvencija. Je li, na temelju nacrtanog stupčastog dijagrama, normalna slučajna varijabla prikladna za modeliranje ovih podataka?

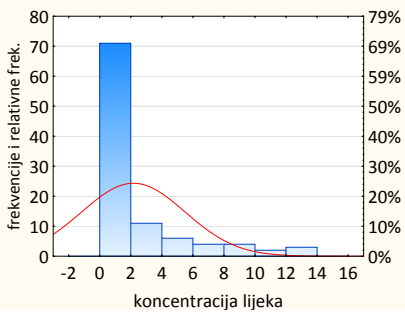
b) Ima li, na temelju nacrtanog stupčastog dijagrama, smisla izmjerene vrijednosti varijable koncentracija lijeka modelirati eksponencijalnom distribucijom? Obrazložite zašto.

c) Procijenite očekivanje i varijancu slučajne varijable X kojom modeliramo koncentraciju promatranog lijeka u organizmu.

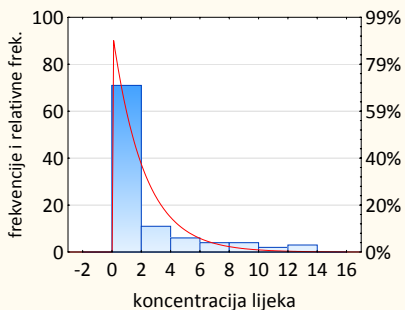
d Pretpostavimo da je slučajna varijabla X eksponencijalna s parametrom $\lambda > 0$. Odredite vjerojatnost da je koncentracija lijeka u krvi u slučajno odabranom trenutku prije njegove eliminacije iz organizma manja od pet te dobiveni rezultat usporedite s rezultatom dobivenim pomoću empirijske distribucije te slučajne varijable.

Rješenje.

a) Ako kategoriziramo izmjerene vrijednosti (tj. podatke iz varijable koncentracija lijeka*) i nacrtamo stupčasti dijagram frekvencija i relativnih frekvencija, uočavamo da normalna slučajna varijabla nije prikladna za modeliranje ovih podataka (uočite crvenu krivulju na slici 5.6 (a)).*



(a) X normalna slučajna varijabla



(b) X eksponencijalna slučajna varijabla

Slika 5.6: Stupčasti dijagram frekvencija i relativnih frekvencija izmjerenih vrijednosti koncentracije lijeka u organizmu.

b) Međutim, moguće je prepoznati drugu neprekidnu distribuciju kojom je opravdano modelirati varijablu koncentracija lijeka*, a to je eksponencijalna distribucija (slika 5.6 (b)). Da biste na slici 5.6 (b) dobili crvenu krivulju koja predstavlja graf funkcije gustoće eksponencijalne distribucije u programskom paketu Statistica slijedite postupak*

Graphs → Histograms → Advanced → Fit type - Exponential.

c) Procjene očekivanja i varijance neprekidne slučajne varijable X kojom modeliramo koncentraciju ovog lijeka u organizmu jesu

$$\overline{x}_{101} = 2.14, \quad s^2_{101} = 13.96.$$

d) Korištenjem empirijske distribucije slijedi da je

$$P(0 \leq X < 5) = 0.84.$$

Poznato je da je parametar eksponencijalne distribucije jednak recipročnoj vrijednosti njezina očekivanja. Tako u ovom primjeru možemo parametar eksponencijalne distribucije

procijeniti s $\lambda = 1/\overline{x}_{101} = 0.47$ *te pomoću kalkulatora vjerojatnosti u Statistici izračunati da je*

$$P\{0 \leq X < 5\} = 0.9.$$

Iako se vjerojatnosti dobivene korištenjem empirijske distribucije i eksponencijalne distribucije s parametrom $\lambda = 0.47$ *razlikuju, stupčasti dijagrami sa slike 5.6 sugeriraju da je u ovom je slučaju eksponencijalna distribucija prikladnija za modeliranja koncentracije lijeka u organizmu od normalne distribucije.*

Zadatak 5.4. (iq.sta)
U nekoj zemlji zakon o diskriminaciji na temelju dobi označava ilegalnim postupak diskriminacije radnika starih 40 godina i više. Oni koji se ne slažu sa zakonom argumentiraju da postoje opravdani razlozi zašto poslodavci nerado zapošljavaju osobe koje su bliže mirovini. Također govore da je radna sposobnost ljudi te dobi upitna. U bazi podataka iq.sta nalaze se rezultati testa inteligencije za dvije generacije ispitanika od kojih je jedna mlađe dobi, a druga starije (varijable iq1, iq2). Procijenite očekivanje slučajne varijable kojom je modeliran rezultat provedenog testa inteligencije intervalom pouzdanosti 95 % za obje dobi. Dajte objašnjenje tih intervala i komentar u kontekstu problema koji je opisan.

Rješenje. *Realizacije intervala pouzdanosti* 95 % *za očekivanje ovih slučajnih varijabli, temeljene na podacima iz baze* iq.sta*, jesu* $[61.98, 71.69]$ *(za stariju generaciju) i* $[41.01, 49.6]$ *(za mlađu generaciju).*

Zadatak 5.5. (gradjevina.sta)
Varijabla placa2009 baze podataka gradjevina.sta (za detaljniji opis pogledajte zadatak 4.31) sadrži prosječne mjesečne plaće zaposlenika u 100 građevinskih poduzeća srednje veličine u nekoj zemlji za 2009. godinu. Intervalom pouzdanosti 95 % procijenite očekivanje slučajne varijable kojom se modelira prosječna mjesečna plaća zaposlenika u 2009. godini u građevinskom poduzeću srednje veličine u toj zemlji.

Rješenje. *Realizacije intervala pouzdanosti* 95 % *za očekivanje ove slučajne varijable, temeljena na podacima iz varijable* placa2009*, jest* $[561.51, 638.75]$.

Zadatak 5.6. (glukoza.sta)
Baza podataka glukoza.sta (za detaljniji opis pogledajte zadatak 2.2) u varijabli dob sadrži podatke o dobi te u varijabli koncentracija koncentraciju glukoze u krvi za 100 ispitanika (podatke za zadnja dva ispitanika ne uključujemo u postupak jer smo u zadatku 3.14 njihove dobi detektirali kao stršeće vrijednosti). Procijenite očekivanje slučajne varijable kojom je modelirana koncentraciju glukoze u krvi intervalom pouzdanosti 95 %. Interpretirajte rezultat.

Rješenje. *Realizacija intervala pouzdanosti za očekivanje ove slučajne varijable, temeljena na izmjerenim koncentracijama, jest* $[7.15, 8.24]$.

Zadatak 5.7. (turizam.sta)
U bazi podataka turizam.sta nalaze se podaci o cijenama sedmodnevnih turističkih aranžmana za dvije osobe u nekim ljetovalištima na obali (varijabla obala) i nekim ljetovalištima na otocima (varijabla otoci).

a) Ima li smisla varijable obala i otoci modelirati normalnim slučajnim varijablama? Ako smatrate da ima, koje ćete vrijednosti parametara normalne slučajne varijable koristiti i zašto?

b) Procijenite očekivanja slučajnih varijabli kojima modeliramo cijene sedmodnevnih turističkih aranžmana na obali i na otocima intervalima pouzdanosti 95 %. Što je veće - očekivana cijena turističkog aranžmana na obali ili očekivana cijena turističkog aranžmana na otocima? Na temelju čega izvodite taj zaključak?

Rješenje.

a) Varijable obala *i* otoci *ima smisla modelirati normalnim slučajnim varijablama. Parametre tih normalnih distribucija procjenjujemo aritmetičkom sredinom i varijancom podataka iz promatranih varijabli:* Obala $\sim \mathcal{N}(1672.49, 245.24)$, Otoci $\sim \mathcal{N}(2349.29, 269.29)$.

b) Realizacije intervala pouzdanosti 95 %*, temeljene na cijenama sadržanima u varijablama* obala *i* otoci*, redom su* $[1588.24, 1756.73]$ *i* $[2256.78, 2441.79]$.

Zadatak 5.8. (vrtic.sta)
Intervalom pouzdanosti 95 % procijenite proporciju zaposlenika iz primjera 5.10 koji preferiraju otvaranje vrtića u okviru poduzeća.

Rješenje. *Realizacija intervala pouzdanosti* 95 % *za ovu vjerojatnost, temeljena na zabilježenim odgovorima* 60 *odabranih roditelja, jest* $[0.51, 0.76]$.

Zadatak 5.9. (gradjevina.sta)
Varijabla placa2009 baze podataka gradjevina.sta (za detaljniji opis pogledajte zadatak 4.31) sadrži prosječne mjesečne plaće zaposlenika u 100 građevinskih poduzeća srednje veličine u nekoj zemlji za 2009. godinu. Intervalom pouzdanosti 95 % procijenite vjerojatnost da je u slučajno odabranom takvom poduzeću prosječna mjesečna plaća zaposlenika viša od procijenjene očekivane plaće u 2009. godini u srednje velikim građevinskim poduzećima u toj zemlji.

Rješenje. *Realizacija intervala pouzdanosti* 95 % *za traženu vjerojatnost, temeljena na podacima iz varijable* placa2009*, jest* $[0.34, 0.54]$.

Zadatak 5.10. Neka agencija provela je istraživanje koje je obuhvatilo 1252 osobe iz populacije osoba koje imaju kreditnu karticu. Njih 180 koristilo je karticu za kupovinu putem interneta.

a) Je li uzorak dovoljno velik za konstruiranje valjanog pouzdanog intervala proporcije korisnika kreditne kartice koji je koriste za kupovinu putem interneta? Obrazložite odgovor.

b) Odredite interval pouzdanosti 98 % za navedenu proporciju. Da ste konstruirali interval pouzdanosti 90 %, bi li on bio uži ili širi?

Rješenje.

a) Budući da interval $\left[\hat{p} - 3\sqrt{\frac{\hat{p}\hat{q}}{n}}, \hat{p} + 3\sqrt{\frac{\hat{p}\hat{q}}{n}}\right] = [0.11, 0.17]$ *ne sadržni ni nulu ni jedinicu, uzorak je dovoljno velik za konstruiranje traženog pouzdanog intervala.*

b) Realizacija intervala pouzdanosti 98 % *jest* $[0.121, 0.167]$. *Realizacija intervala pouzdanosti* 90 % *jest* $[0.127, 0.160]$. *Temeljeno na istim podacima, realizacija intervala pouzdanosti* 90 % *podskup je realizacije intervala pouzdanosti* 98 %.

Zadatak 5.11. (glukoza.sta)
Baza podataka glukoza.sta (za detaljniji opis pogledajte zadatak 2.2) u varijabli dob sadrži podatke o dobi te u varijabli koncentracija koncentraciju glukoze u krvi za 100 ispitanika (podatke za zadnja dva ispitanika ne uključujemo u postupak jer smo u zadatku 3.14 njihove dobi detektirali kao stršeće vrijednosti). Odredite interval pouzdanosti 95 % za vjerojatnost da je koncentracija glukoze za slučajno odabranog ispitanika viša od 4, ali niža od 6 mMol/L. Objasnite rezultat.

Rješenje. $[0.213772, 0.39407]$.

Zadatak 5.12. (kolokvij.sta)
U bazi podataka kolokvij.sta nalaze se rezultati dvaju kolokvija iz nekog kolegija. Varijabla ocjena sadrži prijedloge ocjena s kojima ispitani studenti pristupaju usmenom ispitu, a varijabla stanovanje informaciju o mjestu stanovanja studenta (Osijek - student stanuje u Osijeku; drugo mjesto - student stanuje u nekom drugom mjestu). Intervalom pouzdanosti 95 % procijenite vjerojatnost da slučajno odabrani student usmenom ispitu pristupa s ocjenom većom od 3 za svaku od spomenutih kategorija po mjestu stanovanja. Odredite i interval pouzdanosti 95 % bez obzira na kategorizaciju studenata po mjestu stanovanja.

Rješenje.

Procjena vjerojatnosti intervalom pouzdanosti 95 % *za studente koji žive u Osijeku:*
$[-0.00473237, 0.357732]$.

Procjena vjerojatnosti intervalom pouzdanosti 95 % *za studente koji ne žive u Osijeku:*
$[0.00437127, 0.146629]$.

Procjena vjerojatnosti intervalom pouzdanosti 95 % *bez obzira na mjesto stanovanja:*
$[0.0297206, 0.170279]$.

Zadatak 5.13. (lopta.sta)
Jedan se poduzetnik bavi proizvodnjom loptica za golf. U suradnji s projektantima u poduzeću napravio je preinake na jednom dijelu stroja (ubrizgavalici). Cijeli je proces dizajniran tako da proizvodi loptice prosječne mase 0.25 unci[2]. Kako bi istražio radi li nova ubrizgavalica zadovoljavajuće, odabire 40 loptica i bilježi njihove mase (podaci su dostupni u bazi lopta.sta). Je li na nivou značajnosti $\alpha = 0.05$ očekivana masa loptice statistički značajno veća od 0.25 unci?

Rješenje. $\mathcal{H}_0 : \mu = 0.25$, $\mathcal{H}_1 : \mu > 0.25$, *na nivou značajnosti* $\alpha = 0.05$ *odbacujemo nul-hipotezu.*

[2] 1 unca = 28.35 g

Zadatak 5.14. Kako bi odgovorili na pitanje koji faktori ometaju proces učenja u razredu, istraživači na nekom sveučilištu ispitali su 40 učenika koji su trebali ocjenama od 1 (uopće ne) do 7 (u velikoj mjeri) ocijeniti razinu do koje određeni faktori ometaju proces učenja. Faktor koji je dobio najveću ocjenu je "profesori koji inzistiraju na jednom točnom odgovoru radije nego da evaluiraju cjelokupno razmišljanje i kreativnost". Deskriptivna statistika za ocjenu razine utjecaja ovog faktora je $\overline{x}_{40} = 4.70$, $s_{40} = 1.62$. Premašuje li na nivou značajnosti $\alpha = 0.05$ očekivanje ocjene za navedeni faktor značajno ocjenu 4? Interpretirajte rezultat.

Rješenje. $\mathcal{H}_0 : \mu = 4$, $\mathcal{H}_1 : \mu > 4$, *na nivou značajnosti* $\alpha = 0.05$ *odbacujemo nul-hipotezu.*

Zadatak 5.15. (perec.sta)
Odlučili ste prodavati nove perece u svojoj pekari, no niste sigurni sviđaju li se oni vašim kupcima. O tome ovisi hoćete li nastaviti prodavati te perece ili ne. U bazi podataka perec.sta nalaze se podaci dobiveni iz uzorka od 50 kupaca, pri čemu su njihovi odgovori označeni na sljedeći način:

0 - pereci mi se ne sviđaju
1 - pereci mi se sviđaju
2 - neodlučan sam.

a) Odredite interval pouzdanosti 95 % za proporciju kupaca kojima se sviđaju novi pereci.
Rješenje: $[0.17, 0.43]$.

b) Što ćete učiniti s veličinom uzorka ako želite povećati preciznost procjene?

c) Možete li na nivou značajnosti $\alpha = 0.05$ prihvatiti hipotezu da je proporcija kupaca kojima se ne sviđaju pereci veća od 0.5?
Rješenje: $\mathcal{H}_0 : p = 0.5$, $\mathcal{H}_1 : p > 0.5$, *na nivou značajnosti* $\alpha = 0.05$ *ne odbacujemo nul-hipotezu, tj. na tom nivou značajnosti nemamo dovoljno argumenata tvrditi da je proporcija značajno veća od* 0.5.

Zadatak 5.16. Reputacija mnogih poslova može biti snažno narušena pošiljkom proizvedene robe koja sadrži velik postotak (proporciju) oštećenih proizvoda. Na primjer, proizvođač alkalnih baterija želi biti siguran da je manje od 5% baterija u pošiljci oštećeno. Pretpostavimo da je slučajnim izborom iz vrlo velike pošiljke odabrano 300 baterija od kojih je 10 oštećenih. Je li to dovoljan dokaz proizvođaču da, na nivou značajnosti $\alpha = 0.01$, zaključi da je proporcija neispravnih baterija u pošiljci manja od 0.05?

Rješenje. $\mathcal{H}_0 : p = 0.05$, $\mathcal{H}_1 : p < 0.05$, *na nivou značajnosti* $\alpha = 0.01$ *ne odbacujemo nul-hipotezu. To nije dovoljan dokaz!*

Zadatak 5.17. Savjetnik ekološkog kluba na jednom sveučilištu želi poštovati zahtjev da klub čini 10% brucoša, 20% studenata druge godine, 40% studenata treće godine te 30% apsolvenata. Članstvo ekološkog kluba za ovu godinu brojilo je 14 brucoša, 19 studenata druge godine, 51 studenta treće godine i 16 apslovenata. Provjerite postoji li statistički značajna razlika trenutnog sastava kluba od traženog standarda na nivou značajnosti $\alpha = 0.1$.

Rješenje. *Na nivou značajnosti $\alpha = 0.1$ odbacujemo nul-hipotezu, tj. na tom nivou značajnosti možemo tvrditi da postoji statistički značajno odstupanje sastava kluba od traženog standarda.*

Zadatak 5.18. U studiji temeljenoj na istraživanju razloga povratka umirovljenih ljudi na posao postavljena je sljedeća teorijska distribucija:

38%	ponovo se zaposli u drugom poduzeću
32%	osnuje obrt
23%	rade kao konzultanti
7%	osnuje vlastito poduzeće.

Podudaraju li se, na nivou značajnosti $\alpha = 0.05$, rezultati prikazani u sljedećoj tablici

122	ponovo se zaposlilo u drugom poduzeću
85	osnovalo je obrt
76	rade kao konzultanti
17	osnovalo je vlastito poduzeće.

s prethodno postavljenom teorijskom distribucijom?

Rješenje. *Na nivou značajnosti $\alpha = 0.05$ ne odbacujemo nul-hipotezu, tj. na tom nivou značajnosti nemamo dovoljno argumenata tvrditi da se te dvije distribucije značajno razlikuju.*

Zadatak 5.19. (gradjevina.sta)
Varijabla napredovanje baze podataka gradjevina.sta sadrži ocjene kadrovskih službi 100 građevinskih poduzeća srednje veličine u nekoj zemlji o tome u kolikoj mjeri uspješno obavljanje posla utječe na mogućnost napredovanja na bolje radno mjesto. Zabilježene ocjene interpretiramo na sljedeći način: 1 - uspješnost obavljanja posla uopće ne utječe na mogućnost napredovanja, ..., 5 - napredovanje na bolje radno mjesto isključivo ovisi o uspješnosti u obavljanju posla. Pretpostavimo da bi u idealnom slučaju teorijska distribucija slučajne varijable kojom se modelira ta ocjena bila zadana tablicom

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 1/10 & 1/10 & 1/10 & 1/5 & 1/2 \end{pmatrix}.$$

Postoji li, na nivou značajnosti $\alpha = 0.01$, statistički značajno odstupanje empirijske distribucije te slučajne varijable od pretpostavljene teorijske distribucije?

Rješenje. *Na nivou značajnosti $\alpha = 0.01$ odbacujemo nul-hipotezu.*

Zadatak 5.20. (poduzetnici.sta)
Podaci o dobi 200 poduzetnika u nekoj zemlji nalaze se u bazi podataka poduzetnici.sta. Je li neprekidna slučajna varijabla kojom modeliramo dob poduzetnika u toj zemlji normalno distribuirana? Statističke testove provedite na nivou značajnosti $\alpha = 0.05$ te komentirajte dobiveni rezultat s obzirom na konkretan problem koji proučavate.

Rješenje. *Na nivou značajnosti $\alpha = 0.05$ ne odbacujemo nul-hipotezu da je varijabla normalno distribuirana.*

Zadatak 5.21. (MBA-studij.sta)
Baza podataka MBA-studij.sta sadrži podatke o broju bodova na GMAT (Graduate Management Admission Test) testu za 100 studenata koji žele upisati neki studij. Možemo li na nivou značajnosti $\alpha = 0.05$ tvrditi da je slučajna varijabla kojom modeliramo broj bodova na tom testu normalno distribuirana?

Rješenje. *Na nivou značajnosti $\alpha = 0.05$ ne odbacujemo nul-hipotezu da je varijabla normalno distribuirana.*

Zadatak 5.22. (gradjevina.sta)
Baza podataka gradjevina.sta sadrži neke podatke o organizaciji i poslovanju za 100 građevinskih poduzeća srednje veličine u nekoj zemlji (za detaljniji opis pogledajte zadatak 4.31). Možemo li na nivou značajnosti $\alpha = 0.05$ tvrditi da su slučajne varijable kojima modeliramo prosječnu starost te plaće, troškove i prihode u 2007., 2008. i 2009. godini normalno distribuirane?

Rješenje. *Za sve slučajne varijable na nivou značajnosti $\alpha = 0.05$ odbacujemo nul-hipotezu i prihvaćamo alternativnu hipotezu, tj. zaključujemo da na tom nivou značajnosti spomenute slučajne varijable nisu normalno distribuirane.*

Poglavlje 6

Statističko zaključivanje — dvije varijable

6.1 Razlike u distribuciji između dviju varijabli

U praksi nas često zanima dolazi li do promjene obilježja koje proučavamo zbog provođenja neke aktivnosti, u nekom drugom trenutku ili općenito u nekim drugim uvjetima. Sljedeći primjer ilustrira problematiku tog tipa.

Primjer 6.1. (student.sta)
Neko sveučilište osim klasičnog načina studiranja nudi i studiranje temeljeno na konceptu e-learninga. Povjerenstvo za praćenje kvalitete studiranja želi vidjeti postoji li razlika u dobi između studenata koji studiraju na klasičan način i onih koji studiraju putem e-learninga. Podaci o dobi studenata nalaze se u bazi student.sta *(primjer 2.10). Uvidom u dobnu strukturu tih dvaju uzoraka studenata možemo dobiti procjenu distribucije i numeričkih karakteristika slučajne varijable kojom modeliramo dob studenata koji studiraju klasično i dob studenata koji studiraju putem e-learninga (slike 6.1, 6.2 i 6.3).*

	Frequency table: student.sta			
From To	Count (e-learning)	Percent (e-learning)	Count (klasicno studiranje)	Percent (klasicno studiranje)
18<=x<23	30	60.00000	30	60.00000
23<=x<28	13	26.00000	16	32.00000
28<=x<33	3	6.00000	4	8.00000
33<=x<38	4	8.00000	0	0.00000

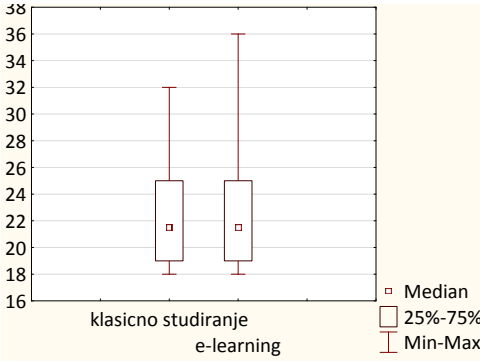
Slika 6.1: Varijable klasicno studiranje i e-learning - tablica kategoriziranih frekvencija i relativnih frekvencija

Variable	Descriptive Statistics (student.sta)					
	Valid N	Mean	Mode	Frequency of Mode	Variance	Std.Dev.
klasicno studiranje	50	22.12	Multiple	9	13.58	3.68
e-learning	50	22.80	19.00000	9	22.94	4.79

Variable	Descriptive Statistics (student.sta)					
	Valid N	Mean	Mode	Frequency of Mode	Variance	Std.Dev.
klasicno studiranje	50	22.12	Multiple	9	13.58	3.68
e-learning	50	22.80	19.00000	9	22.94	4.79

Variable	Descriptive Statistics (student.sta)					
	Median	Minimum	Maximum	Lower Quartile	Upper Quartile	Range
klasicno studiranje	21.5	18	32	19	25	14
e-learning	21.5	18	36	19	25	18

Slika 6.2: Varijable klasicno studiranje i e-learning - deskriptivna statistika.



Slika 6.3: Varijable klasicno studiranje i e-learning - kutijasti dijagrami na bazi medijana.

Budući da se ovdje radi o pručavanju istog obilježja (dobi) na dva uzorka studenata koji nemaju zajedničkih jedinki, kažemo da proučavamo nevezane uzorke.

Primjer 6.2. *Pretpostavimo da želimo usporediti daje li novi tip sjemana kukuruza, razvijen genetičkim metodama, veće prinose nego do sada najčešće korištena sorta kukuruza na ovim područjima. Pokusi moraju biti izvedeni sijanjem ovih sorti na poljima koja osiguravaju iste uvjete za rast. Urod kukuruza po kvadratnom metru parceliranih polja predstavlja bazu podataka na osnovi koje možemo statistički zaključivati o pitanjima razlika. I u ovom se primjeru radi o proučavanju* nevezanih uzoraka.

Primjer 6.3. (igre.sta)
U jednoj je školi napravljeno istraživanje o tome što djeca misle i osjećaju prema sebi. Test se sastojao od toga da na početku testiranja djeca ocjenom od 1 *(ne slažem se) do* 5 *(slažem se)*

ocijene tvrdnju "imam mnogo dobrih osobina". Nakon toga u razdoblju od šest tjedana djeca su igrala četiri igre koje potiču pozitivan stav prema sebi. Poslije tih igara ponovno im je postavljeno isto pitanje koje su na isti način ocijenili. U bazi podataka igre.sta *nalaze se ocjene prije i nakon provođenja igara. Uvidom u utjecaj igara na mišljenje djece o samima sebi možemo dobiti procjenu distribucije i numeričkih karakteristika slučajne varijable kojom modeliramo ocjene prije i nakon tretmana igrama (slike 6.4, 6.5 i 6.6).*

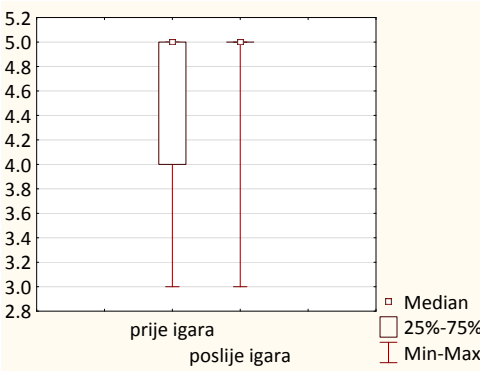
Variable	Descriptive Statistics (igre.sta)					
	Valid N	Mean	Mode	Frequency of Mode	Variance	Std.Dev.
prije igara	33	4.55	5	21	0.44	0.67
poslije igara	33	4.79	5	27	0.23	0.48

Variable	Descriptive Statistics (igre.sta)					
	Median	Minimum	Maximum	Lower Quartile	Upper Quartile	Range
prije igara	5	3	5	4	5	2
poslije igara	5	3	5	5	5	2

Slika 6.4: Varijable prije igara i poslije igara - deskriptivna statistika.

Category	Frequency table: igre.sta			
	Count (prije igara)	Percent (prije igara)	Count (poslije igara)	Percent (poslije igara)
3	3	9.09	1	3.03
4	9	27.27	5	15.15
5	21	63.64	27	81.82

Slika 6.5: Varijable prije igara i poslije igara - tablice kategoriziranih frekvencija i relativnih frekvencija



Slika 6.6: Varijable prije igara i poslije igara - kutijasti dijagrami na bazi medijana.

Budući da se ovdje radi o analizi subjektivnog mišljenja djeteta o samom sebi na istom uzorku djece prije i poslije tretmana igrama, kažemo da proučavamo vezane uzorke.

Na temelju tablica frekvencija i relativnih frekvencija u primjerima 6.1 i 6.3 (tablice 6.1 i 6.5) uočavamo razlike u empirijskoj distribuciji slučajne varijable kojom modeliramo promatrano obilježje na danim vezanim ili nevezanim uzorcima. Preciznije, uočavamo razlike u empirijskoj distribuciji slučajne varijable kojom modeliramo dob studenata koji studiraju na klasičan način i slučajne varijable kojom modeliramo dob studenata koji studiraju putem e-learninga u primjeru 6.1 te razlike u empirijskim distribucijama slučajnih varijabli kojima modeliramo ocjenu mišljenja djece o sebi prije i nakon igara u primjeru 6.3. Također, na temelju deskriptivnih statistika 6.2 i 6.4 uočavamo postojanje razlika u npr. aritmetičkoj sredini (procjeni za očekivanje slučajne varijable) i standardnoj devijaciji podataka (procjeni za standardnu devijaciju slučajne varijable). Postavlja se pitanje jesu li ove razlike uočene na uzorcima statistički značajne ili ne. U tu svrhu u ovom ćemo se poglavlju baviti zaključivanjem o statističkoj značajnosti uočenih razlika u ovim i sličnim primjerima.

Prvi korak u ovakvim analizama uvijek je analiza obilježja koje nas zanima posebno za svaki od dva dana uzorka pa kažemo da analiziramo jedno obilježje u dva **tretmana**. Cilj je utvrditi postoje li razlike u distribuciji obilježja za različite tretmane. S obzirom da ne znamo stvarnu distribuciju promatranog obilježja, o njoj zaključujemo na osnovi prikupljenih podataka. U tu ćemo svrhu usporediti empirijske distribucije obilježja po tretmanima, kao i procijenjene vrijednosti parametara (primjeri 6.1 i 6.3). S obzirom na činjenicu da su procjenitelji koje pri tome koristimo slučajne varijable, prirodno je očekivati da dobivene procijenjene vrijednosti po tretmanima neće biti jednake. Pitanje na koje želimo odgovoriti jest mogu li se razlike koje uočavamo pripisati samo činjenici da su procjenitelji slučajne varijable ili ima razloga vjerovati da su izazvane postojanjem razlika između stvarnih distribucija promatranih slučajnih varijabli (tada kažemo da su razlike statistički značajne). Ukratko, pitanje na koje odgovaramo u ovom poglavlju jest: **"Jesu li uočene razlike po tretmanima statistički značajne?"** Postupak koji ćemo pri tome primjenjivati jest testiranje statističkih hipoteza. Važno je također naglasiti da je prilikom ovakvog analiziranja razlika među distribucijama slučajnih varijabli važno pažljivo pripremanje pokusa tako da se osiguraju dva slučajna uzorka koja se bitno razlikuju samo po tretmanu.

6.1.1 Usporedba očekivanja — nevezani uzorci

Zanima nas postoji li razlika u očekivanju slučajne varijable kojom modeliramo neko obilježje u dva tretmana. U svakom od tretmana biramo jedinke u uzorak na

slučajni način. Uzorci ne sadrže iste jedinke. Neka su n_1, μ_1 i σ_1 veličina uzorka, očekivanje i standardna devijacija slučajne varijable kojom modeliramo obilježje u prvom tretmanu, a n_2, μ_2 i σ_2 veličina uzorka, očekivanje i standardna devijacija slučajne varijable kojom modeliramo obilježje u drugom tretmanu.

Veliki uzorci

U uvjetima kada imamo velike uzorke, možemo testirati hipotezu o jednakosti očekivanja slučajnih varijabli kojima modeliramo promatrano obilježje u dva tretmana **neovisno o distribuciji tih slučajnih varijabli**. Postupak testiranja provodi se na sljedeći način:

Nul-hipoteza:

$$\mathcal{H}_0 : \mu_1 = \mu_2$$

Test-statistika:

$$Z' = \frac{\overline{X}_{n_1} - \overline{X}_{n_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{6.1}$$

Ovdje su n_1 i n_2 veličine uzoraka iz nevezanih tretmana, $\overline{X}_{n_1}$ i $\overline{X}_{n_2}$ su redom aritmetičke sredine, a σ_1 i σ_2 poznate standardne devijacije tih uzoraka, redom. Test-statistika Z', u slučaju velikih uzoraka te ako je nul-hipoteza istinita, ima približno standardnu normalnu distribuciju.

Ako je nul-hipoteza istinita, očekujemo da, na temelju podataka izračunata vrijednost za Z' (označit ćemo je sa $\hat{z}$), nije daleko od 0. Međutim, slučajna varijabla Z' ima približno standardnu normalnu distribuciju pa ne možemo zanemariti mogućnost realizacije te varijable i u intervalu koji je daleko od nule. Ako označimo sa Z slučajnu varijablu s $\mathcal{N}(0,1)$ distribucijom, na osnovi realizacije $\hat{z}$ statistike Z' na podacima možemo odrediti p-vrijednost kao:

- $p = P\{Z \geq \hat{z}\}$ ako je alternativna hipoteza oblika $\mathcal{H}_1 : \mu_1 - \mu_2 > 0$
- $p = P\{Z \leq \hat{z}\}$ ako je alternativna hipoteza oblika $\mathcal{H}_1 : \mu_1 - \mu_2 < 0$.

Tako izračunatu p-vrijednost uspoređujemo s nivoom značajnosti α. U slučaju da je $p < \alpha$ odbacujemo nul-hipotezu na nivou značajnosti α i prihvaćamo alternativnu hipotezu $\mathcal{H}_1$. Ako je $p > \alpha$, zaključujemo da nemamo dovoljno argumenata koji bi poduprli odluku o odbacivanju nul-hipoteze.

U ovim postupcima aritmetičke sredine uzoraka $\overline{X}_{n_1}$ i $\overline{X}_{n_2}$ koristimo kao procjenitelje za očekivanja μ_1 i μ_2 (njihove realizacije za izmjerene vrijednosti u prvom i drugom tretmanu su procjene $\overline{x}_{n_1}$ i $\overline{x}_{n_2}$ očekivanja μ_1 i μ_2). Za primjenu ovog testa potrebno je poznavati i varijancu obilježja (tj. vrijednosti σ_1^2 i σ_2^2), što u primjenama najčešće nije slučaj. Međutim, u slučaju velikih uzoraka možemo iskoristiti korigirane varijance uzoraka $s_{n_1}^2$ i $s_{n_2}^2$ kao procjene nepoznatih varijanci.

Mali uzorci

Ako pretpostavimo da su varijable u tretmanima normalno distribuirane i da imaju jednake varijance, tada možemo primijeniti test koji će biti opisan u ovom odjeljku. Dakle, ako za slučajne varijable X_1 i X_2, kojima modeliramo obilježje u prvom, odnosno drugom tretmanu, vrijede pretpostavke

- $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ i $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$
- $\sigma_1^2 = \sigma_2^2$,

postupak testiranja jednakosti očekivanja slučajnih varijabli X_1 i X_2 možemo provesti i za **male uzorke**. Postupak testiranja provodi se na sljedeći način:

Nul-hipoteza:

$$\mathcal{H}_0 : \ \mu_1 = \mu_2$$

Test-statistika:

$$T' = \frac{\overline{X}_{n_1} - \overline{X}_{n_2}}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{6.2}$$

$$s_p^2 = \frac{(n_1 - 1)s_{n_1}^2 + (n_2 - 1)s_{n_2}^2}{n_1 + n_2 - 2}$$

Ovdje su n_1 i n_2 veličine uzoraka iz nevezanih tretmana, $\overline{X}_{n_1}$ i $\overline{X}_{n_2}$ su redom aritmetičke sredine tih uzoraka, a $s_{n_1}^2$ i $s_{n_2}^2$ korigirane uzoračke varijance za svaki tretman. Ako je nul-hipoteza istinita, test-statistika T' ima Studentovu t-distribuciju s $(n_1 + n_2 - 2)$ stupnjeva slobode.

Ako je nul-hipoteza istinita, očekujemo da je na temelju podataka izračunata vrijednost za T' (označit ćemo je s $\hat{t}$) blizu 0, a vjerojatnost da se T' realizira u intervalu dalekom od nule, koja nam treba za određivanje p-vrijednosti, računamo na temelju

Studentove t-distribuciju s (n_1+n_2-2) stupnjeva slobode. Označimo li s T slučajnu varijablu koja ima t-distribuciju s (n_1+n_2-2) stupnjeva slobode, imamo:

- $p = P\{T \geq \hat{t}\}$ ako je alternativna hipoteza oblika $\mathcal{H}_1 : \mu_1 - \mu_2 > 0$
- $p = P\{T \leq \hat{t}\}$ ako je alternativna hipoteza oblika $\mathcal{H}_1 : \mu_1 - \mu_2 < 0$.

Tako izračunatu p-vrijednost uspoređujemo s nivoom značajnosti α. U slučaju da je $p < \alpha$, odbacujemo nul-hipotezu na nivou značajnosti α i prihvaćamo alternativnu hipotezu $\mathcal{H}_1$. Ako je $p > \alpha$, zaključujemo da nemamo dovoljno argumenata koji bi poduprli odluku o odbacivanju nul-hipoteze.
Za primjenu ovog testa od velike je važnosti ispunjenost pretpostavke o jednakosti varijanci varijabli po tretmanima. Budući da nam stvarne varijance σ_1^2 i σ_2^2 u većini slučajeva nisu poznate, korisno je prije primjene ovog testa testirati hipotezu o jednakosti varijanci. U tu svrhu možemo koristiti tzv. ***F*-test o jednakosti varijanci**.

Nul-hipoteza:

$$\mathcal{H}_0 : \ \sigma_1^2 = \sigma_2^2$$

Test-statistika:

$$V' = \frac{s_{n_1}^2}{s_{n_2}^2} \tag{6.3}$$

Ovdje su s_1^2 i s_2^2 procjene varijanci σ_1^2 i σ_2^2. Ako je nul-hipoteza istinita, test-statistika V' ima F distribuciju s $(n_1 - 1)$ i $(n_2 - 1)$ stupnjeva slobode.

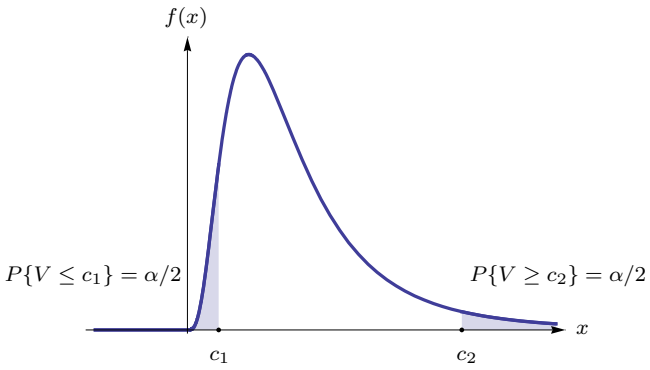
Ako je nul-hipoteza istinita, očekujemo da je na temelju podataka izračunata vrijednost za V' (označit ćemo je s $\hat{v}$) bliska jedinici. Označimo s V slučajnu varijablu koja ima F distribuciju s $(n_1 - 1)$ i $(n_2 - 1)$ stupnjeva slobode. Nul-hipotezu odbacujemo ako za izračunatu vrijednost $\hat{v}$ vrijedi jedna od sljedećih nejednakosti:

$$\hat{v} \leq c_1 \qquad \text{ili} \qquad \hat{v} \geq c_2,$$

gdje su c_1 i c_2 pozitivni realni brojevi takvi da je

$$P(V \leq c_1) = P(V \geq c_2) = \frac{\alpha}{2},$$

gdje je α je nivo značajnosti testa (slika 6.7).



Slika 6.7: $P\{V \leq c_1\} + P\{V \geq c_1\} = \alpha$

Brojeve c_1 i c_2 određujemo kalkulatorom vjerojatnosti pri čemu je ključno za distribuciju odabrati F distribuciju sa stupnjevima slobode (eng. degrees of freedom, df) $(n_1 - 1)$ i $(n_2 - 1)$. Npr. ako $n_1 = n_2 = 11$, tada su oba stupnja slobode F distribucije jednaka 10 pa je za $\alpha = 0.05$ $c_1 = 0.27$ i $c_2 = 3.72$, a za $\alpha = 0.1$ je $c_1 = 0.34$ i $c_2 = 2.98$. Dakle, ako je

$$\hat{v} \leq c_1 \text{ ili } \hat{v} \geq c_2$$

na nivou značajnosti α odbacujemo nul-hipotezu $\mathcal{H}_0$ i prihvaćamo alternativnu hipotezu o postojanju razlike među varijancama σ_1^2 i σ_2^2. Ako je

$$v \in (c_1, c_2),$$

tada nemamo dovoljno argumenata koji bi poduprli odluku o odbacivanju hipoteze o jednakosti varijanci.

Primjer 6.4. *Neka su $s_1 = 3.2$ i $s_2 = 3$ procjene standardih devijacija slučajnih varijabli X_1 i X_2 kojima modeliramo neko obilježje u prvom i drugom tretmanu, redom. Pretpostavimo da su procjene s_1 i s_2 dobivene na temelju uzoraka veličina $n_1 = n_2 = 30$. Da bismo na nivou značajnosti $\alpha = 0.05$ proveli F-test i donijeli odluku koja se tiče jednakosti varijanci σ_1^2 i σ_2^2, računamo vrijednost test statistike V':*

$$\hat{v} = \frac{s_{n_1}^2}{s_{n_2}^2} = \frac{3.2^2}{3^2} \approx 1.14.$$

Pomoću kalkulatora vjerojatnosti slijedi da je za slučajnu varijablu V koja ima F distribuciju s oba stupnja slobode jednaka 29

$$P\{V \leq 0.48\} = P\{V \geq 2.1\} = 0.025$$
$$\{V \leq 0.48\} + P\{V \geq 2.1\} = 0.05,$$

pa je $c_1 = 0.48$, a $c_2 = 2.1$. Budući da je izračunata vrijednost $\hat{v} = 1.14$ sadržana u intervalu $(c_1, c_2) = (0.48, 2.1)$, na nivou značajnosti $\alpha = 0.05$, nemamo dovoljno argumenata koji bi poduprli tvrdnju o odbacivanju nul-hipoteze. Dakle, ne možemo tvrditi da su varijnce različite.

Primjer 6.5. *Neko poduzeće bavi se izdavačkom djelatnošću. Svoje proizvode na prodajna mjesta dostavlja koristeći usluge dvaju transportnih poduzeća. Upravu poduzeća zanima razlikuju li se očekivana vremena trajanja dostave za ta dva poduzeća ili ne. Da bi se donio zaključak koji daje odgovor na pitanje uprave, potrebno je testirati hipotezu o jednakosti očekivanog vremena trajanja dostave proizvoda za ta dva transportna poduzeća. U tu je svrhu analitičar zabilježio trajanje* 30 *dostava koje je obavilo prvo i* 30 *dostava koje je obavilo drugo transportno poduzeće te na temelju tih podataka procijenio očekivanje promatranih slučajnih uzoraka:*

prvo transportno poduzeće: $n_1 = 30$, $\overline{x}_{n_1} = 16$ *sati*, $s_1 = 3.2$ *sata*
drugo transportno poduzeće: $n_2 = 30$, $\overline{x}_{n_2} = 18$ *sati*, $s_2 = 3$ *sata.*

Pretpostavimo da vremena trajanja dostave proizvoda u organizaciji prvog i drugog transportnog poduzeća možemo modelirati normalnim slučajnim varijablama. U primjeru 6.4 proveli smo F-test o jednakosti varijanci za ovaj slučaj i pokazali da, na nivou značajnosti $\alpha = 0.05$, ne odbacujemo nul-hipotezu. Dakle, za testiranje hipoteze o jednakosti očekivanog vremena trajanja dostave za dva promatrana transportna poduzeća, tj. za testiranje hipoteze

$$\mathcal{H}_0 : \mu_1 = \mu_2,$$

možemo koristiti statistički test temeljen na test statistici 6.2 koja u ovom slučaju prima vrijednost

$$\hat{t} = \frac{16 - 18}{s_p \left(\frac{1}{30} + \frac{1}{30}\right)} \approx -2.49,$$

gdje je

$$s_p = \sqrt{\frac{(30-1)3.2^2 + (30-1)3^2}{30 + 30 - 2}} = 3.1.$$

Uočimo da za procjene $\overline{x}_{n_1} = 16$ i $\overline{x}_{n_2} = 18$ nepoznatih očekivanja μ_1 i μ_2 vrijedi nejednakost $\overline{x}_1 < \overline{x}_2$, tj. nejednakost $\overline{x}_1 - \overline{x}_2 < 0$, što odgovara alternativnoj hipotezi

$$\mathcal{H}_1 : \mu_1 - \mu_2 < 0.$$

Pripadna p-vrijednost je

$$p = P\{T < \hat{t}\} = P\{T < -2.49\} \approx 0.0077.$$

Budući da je za nivo značajnosti $\alpha = 0.05$ očito $p < \alpha$, slijedi da na nivou značajnosti $\alpha = 0.05$ odbacujemo nul-hipotezu i prihvaćamo alternativnu hipotezu da je očekivano vrijeme trajanja dostave za prvo transportno poduzeće kraće od očekivanog vremena trajanja dostave za drugo poduzeće.

6.1.2 Usporedba očekivanja — vezani uzorci

Često u praksi imamo potrebu za uspoređivanjem varijabli u vezanim tretmanima. Npr. ako želimo uspoređivati rezultate testa za iste bolesnike prije i nakon liječenja. Prethodni test ovdje nije adekvatan jer smo svjesni da mjerena vrijednost varijable u svakom pojedinom slučaju drugog tretmana može ovisiti o tome kolika je bila vrijednost varijable odgovarajućeg slučaja u prvom tretmanu. Dakle, pretpostavka o nezavisnosti varijabli po tretmanima nije opravdana. U ovakvim primjerima slučajevi se moraju pratiti u paru, a zaključci o postojanju razlika među tretmanima donose se na osnovu razlika varijabli u pojedinim tretmanima kao što je prikazano u tablici 6.1.

ispitanik	tretman 1	tretman 2	razlika
1	x_1	y_1	$d_1 = x_1 - y_1$
2	x_2	y_2	$d_2 = x_1 - y_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	x_n	y_n	$d_n = x_1 - y_n$

Tablica 6.1: Razlike vrijednosti varijabli u promatranim tretmanima.

Dakle, slučajni uzorak koji se ovdje promatra sastoji se od n uređenih parova slučajnih varijabli $(X_1, Y_1), \ldots, (X_n, Y_n)$ pomoću kojih definiramo slučajne varijable razlika $D_i = X_i - Y_i$, $i \in \{1, \ldots, n\}$, gdje su slučajne varijable $X_1, \ldots, X_n$ nezavisne i jednako distribuirane (isto vrijedi za slučajne varijable $Y_1, \ldots, Y_n$). Pretpostavimo da su i slučajne varijable $D_1, \ldots, D_n$ također nezavisne i jednako distribuirane. Očekivanje slučajne varijable razlika $D_i = X_i - Y_i$, $i \in \{1, \ldots, n\}$, može se dobiti kao razlika očekivanja μ_1 i μ_2 slučajnih varijabli X_i i Y_i, tj.

$$\mu_D = \mu_1 - \mu_2.$$

Testiranje hipoteze

$$\mathcal{H}_0 : \ \mu_1 - \mu_2 = 0$$

sada se svodi na testiranje ekvivalentne hipoteze

$$\mathcal{H}_0 : \mu_D = 0$$

koja se odnosi na očekivanje slučajne varijable razlika. Testovi kojima možemo testirati ovako postavljenu hipotezu opisani su u poglavlju Statističko zaključivanje — jedna varijabla.

Uočimo da sada, uz procjene za parametre varijabli svakog pojedinog tretmana, trebamo i procjene za parametre varijable razlika koje ćemo koristiti za testiranje hipoteze. Procjene za očekivanje razlike i varijance razlike su:

$$\overline{d}_n = \overline{x}_n - \overline{y}_n, \qquad s_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (d_i - \overline{d}_n)^2.$$

Primjer 6.6. (igre.sta)
U primjeru 6.3 opisali smo istraživanje provedeno u jednoj školi o tome što djeca misle i osjećaju prema sebi. Provjerimo možemo li na nivou značajnosti $\alpha = 0.05$ prihvatiti hipotezu o postojanju razlike u očekivanoj ocjeni djece prije i nakon tretmana igrama.
Budući da prilikom provođenje testa o razlici očekivanja p-vrijednost iznosi 0.009, na nivou značajnosti $\alpha = 0.05$ odbacujemo nul-hipotezu u korist alternativne hipoteze o povećanju očekivane ocjene djece prije i nakon tretmana igrama.

6.1.3 Usporedba proporcija u velikim uzorcima

Kao što je već objašnjeno do sada, problem procjene proporcije i problem procjene vjerojatnosti pojavljivanja događaja mogu se uklopiti u statistički model s istim tipom distribucije. Ovdje se bavimo utvrđivanjem postojanja razlika među vjerojatnostima pojavljivanja izabranog događaja u dvije nezavisne populacije. Primjerima ćemo pokazati kako se ta procedura može iskoristiti za utvrđivanje postajanja razlika u proporcijama.

Dakle, želimo na temelju učestalosti pojave nekog događaja u uzorcima iz dvije populacije usporediti vjerojatnosti pojavljivanja tog događaja u te dvije populacije. Za modeliranje ćemo iskoristiti Bernoullijevu slučajnu varijablu. Jedan problem tog tipa prikazan je u primjeru 6.7.

Primjer 6.7. *Na temelju tisuću dimenzionalnog reprezentativnog uzorka ($n_1 = 1000$) stanovnika jednog grada utvrđeno je da je proporcija ljudi u uzorku koji redovito vježbaju $\widehat{p}_1 = 25\%$ dok je u nekom drugom gradu na temelju 2000 dimenzionalnog uzorka ($n_2 = 2000$) utvrđeno je da je proporcija redovitih vježbača $\widehat{p}_2 = 28\%$. Evidentno je da je proporcija ljudi koji redovito vježbaju u uzorku iz drugog grada veća od proporcije u uzorku iz prvog grada. Mi želimo utvrditi možemo li na temelju toga zaključiti da je proporcija stanovnika koji redovito vježbaju u drugom gradu veća nego u prvom gradu.*

U tu svrhu iskoristit ćemo uobičajeni postupak modeliranja kod zaključivanja o proporciji. Prilikom uzimanja uzorka (ispitavanja odabranih osoba vježbaju li ili ne) označimo s 1 odgovor "da", a s 0 odgovor "ne". Za modeliranje uzoraka koristimo dvije Bernoullijeve slučajne varijable

$$X_1 = \begin{pmatrix} 0 & 1 \\ 1-p_1 & p_1 \end{pmatrix}, \quad X_2 = \begin{pmatrix} 0 & 1 \\ 1-p_2 & p_2 \end{pmatrix}, \quad p_1, p_2 \in (0,1),$$

gdje je p_1 vjerojatnost pojave promatranog događaja u prvoj populaciji (odgovara proporciji osoba koje redovito vježbaju u prvoj populaciji), a p_2 vjerojatnost pojave istog događaja u drugoj populaciji (odgovara proporciji osoba koje redovito vježbaju u drugoj populaciji). Korištenjem relativne frekvencije kao procjenitelja za vjerojatnost, na temelju uzoraka stanovnika dvaju promatanih gradova procjenjujemo parametre p_1 i p_2 s $\widehat{p}_1 = 25\%$ i $\widehat{p}_2 = 28\%$. Svjesni smo da su procjenitelji slučajne varijable. Njihove realizacije, tj. procjene ne daju točnu vrijednost parametara. Možemo li, na temelju informacija koje imamo, reći da je u drugom gradu veća proporcija ljudi koji redovito vježbaju?

Da bismo odgovorili na pitanje postavljeno u primjeru 6.7, služimo se sljedećim

testom:

Nul-hipoteza:

$$H_0 : \ p_1 = p_2$$

Test-statistika:

$$Z' = \frac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{\widehat{p}(1-\widehat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$$

$$\widehat{p} = \frac{n_1\widehat{p}_1 + n_2\widehat{p}_2}{n_1 + n_2}$$

Ovdje su n_1 i n_2 veličine uzoraka iz dviju promatranih populacija, a $\widehat{p}_1$ i $\widehat{p}_2$ proporcije realizacija događaja od interesa u prvom i drugom uzorku, redom. Za velike uzorke i ako je nul-hipoteza istinita, slučajna varijabla Z' ima približno standardnu normalnu distribuciju.

Ako je nul-hipoteza istinita, očekujemo da će se Z' realizirati blizu nule. Kao i do sada, ne smijemo zanemariti činjenicu da postoji mala vjerojatnost realizacije Z' u intervalu daleko od nule i ako je nul-hipoteza istinita, što je osnova za računanje p-vrijednosti. Označimo sa $\hat{z}$ realizaciju od Z' na uzorku (tj. $\hat{z}$ je izračunata vrijednost za Z' na temelju podataka). Ako sa Z označimo slučajnu varijablu koja ima standardnu normalnu distribuciju, na osnovi $\hat{z}$ možemo odrediti p-vrijednost na sljedeći način:

- $p = P\{Z \geq \hat{z}\}$ ako je alternativna hipoteza oblika $\mathcal{H}_1 : p_1 - p_2 > 0$
- $p = P\{Z \leq \hat{z}\}$ ako je alternativna hipoteza oblika $\mathcal{H}_1 : p_1 - p_2 < 0$.

Tako izračunatu p-vrijednost uspoređujemo s nivoom značajnosti α. U slučaju da je $p < \alpha$, odbacujemo nul-hipotezu i na nivou značajnosti α prihvaćamo alternativnu hipotezu $\mathcal{H}_1$. Ako je $p > \alpha$, zaključujemo da nemamo dovoljno argumenata koji bi poduprli odluku o odbacivanju nul-hipoteze.

Primjer 6.8. *Provedimo navedeni test za problem iz primjera 6.7 uz nivo značajnosti* $\alpha = 0.05$. *Budući da je*

$$\widehat{p}_1 < \widehat{p}_2,$$

alternativna hipoteza je oblika

$$\mathcal{H}_1 : p_1 - p_2 < 0.$$

Za $n_1 = 1000$, $\widehat{p}_1 = 0.25$, $n_2 = 2000$ *i* $\widehat{p}_2 = 0.28$ *je*

$$\widehat{p} = \frac{1000 \cdot 0.25 + 2000 \cdot 0.28}{1000 + 2000} = 0.27$$

pa je vrijednost test statistike Z'

$$\hat{z} = \frac{0.25 - 0.28}{\sqrt{0.27(1-0.27)\left(\frac{1}{1000} + \frac{1}{2000}\right)}} \approx -1.74,$$

a p-vrijednost

$$p = P\{Z \leq -1.74\} \approx 0.041.$$

Budući da je $p < \alpha$, odbacujemo nul-hipotezu i na nivou značajnosti $\alpha = 0.05$ prihvaćamo alternativnu hipotezu koja kaže da je u drugom gradu veća proporcija ljudi koji redovito vježbaju.

Primjer 6.9. *U jednom slučajnom uzorku od 100 stalnih kupaca dane trgovine 43 kupca plaćaju Master karticom ($n_1 = 100$, $\widehat{p}_1 = 0.43$), a u drugom slučanom uzorku koji također broji 100 kupaca njih 58 plaća Visa karticom ($n_2 = 100$, $\widehat{p}_2 = 0.58$). Zanima nas možemo li na razini značajnosti $\alpha = 0.05$ tvrditi da je proporcija kupaca te trgovine koja za plaćanje koristi Visa karticu veća od proporcije kupaca koji koriste Master karticu.*

Budući da je

$$\widehat{p}_1 < \widehat{p}_2,$$

alternativna hipoteza je oblika

$$\mathcal{H}_1 : p_1 - p_2 < 0$$

pa za vrijednost test statistike Z' dobivamo:

$$\widehat{p} = \frac{100(0.43 + 0.58)}{100 + 100} = \frac{101}{200} = 0.505,$$

$$\hat{z} = \frac{0.43 - 0.58}{\sqrt{0.505(1-0.505)\left(\frac{1}{100} + \frac{1}{100}\right)}} \approx -2.12.$$

Pripadna p-vrijednost je

$$p = P\{Z \leq \hat{z}\} = P\{Z \leq -2.12\} = 0.017.$$

Za nivo značajnosti $\alpha = 0.05$ slijedi da je $p < \alpha$ pa zaključujemo da odbacujemo nul-hipotezu i na razini značajnosti $\alpha = 0.05$ prihvaćamo alternativnu hipotezu da je proporcija kupaca te trgovine koji za plaćanje koriste Visa karticu veća od proporcije kupaca koji za plaćanje koriste Master karticu.

6.2 Dvodimenzionalan slučajni vektor

U prethodnom poglavlju uveli smo pojam vezanih uzoraka kod kojih se, za svaki pojedinačni slučaj, bilježi vrijednost jednog obilježja u dva različita tretmana. Uočimo da to rezultira tablicom u kojoj imamo unesene vrijednosti (realizacije) tog obilježja u svakom pojedinom slučaju (tablica 6.2).

broj	tretman 1	tretman 2
1	x_1	y_1
2	x_2	y_2
$\vdots$	$\vdots$	$\vdots$
n	x_n	y_n

Tablica 6.2: Tablica vrijednosti obilježja za n mjerenja u svakom od dva tretmana.

Za ovakve podatke kažemo da su realizacije **slučajnog vektora** (X, Y), gdje je X slučajna varijabla kojom modeliramo realizacije prvog tretmana, a Y slučajna varijabla kojom modeliramo ralizacije drugog tretmana. Slične tablice pojavljuju se ako na istim jedinkama bilježimo realizacije dviju varijabli, bilo da one opisuju istu karakteristiku bilo neku drugu karakteristiku. Npr. ako za skupinu osoba na kojoj vršimo ispitivanje mjerimo istovremeno tjelesnu masu i visinu, bilježimo realizacije dviju različitih karakteristika osobe, ali opet imamo vezane varijable. Naime, sasvim je prirodno da masa osobe nije potpuno neovisna o visini, međutim nije ni jednoznačno određena visinom osobe. U ovakvim slučajevima od interesa je ustanoviti postoje li neke ovisnosti među varijablama koje se prate u paru ili su one neovisne jedna o drugoj. Da bismo to bili u stanju, potrebno je prvo naučiti osnovne pojmove vezane uz slučajni vektor. U ovom poglavlju opisat ćemo slučajni vektor, njegovu distribuciju i osnovne karakteristike u diskretnom slučaju.

Treba također naglasiti da slučajni vektor ne mora uvijek biti uređeni par slučajnih varijabli, tj. ne mora biti dvodimenzionalan. On može biti uređena n-torka slučajnih varijabli (tj. n-dimenzionalan slučajni vektor) kao npr. kada za svaki pojedini slučaj bilježimo realizacije više različitih karakteristika, a ne samo dvije. Međutim, za naše potrebe i osnove statističke analize kojom utvrđujemo postojanje ovisnosti među varijablama bit će dovoljno razmatranje dvodimenzionalnog slučaja.

6.2.1 Tablica distribucije diskretnog slučajnog vektora

Jedna realizacija dvodimenzionalnog slučajnog vektora uvijek je uređeni par realnih brojeva. Ako je takav slučajni vektor ujedno diskretan, onda realizacija može biti samo konačno ili prebrojivo mnogo, kao i kod diskretne slučajne varijable. Radi jednostavnosti promatat ćemo samo slučajne vektore s konačnim skupom svih mogućih vrijednosti. Da bismo zadali distribuciju takvog slučajnog vektora, potrebno je zadati pripadnu vjerojatnost na skupu svih njegovih mogućih realizacija. Postupak zadavanja distribucije slučajnog vektora opisat ćemo prvo na jednom primjeru.

Primjer 6.10. *Tvornica bombona koristi dvije linije za pakiranje bombona u vrećice. Svaka od linija povremeno ne zavari vrećicu na odgovarajući način pa se pakiranje ne može poslati u prodaju. Radi analize uzroka ovih problema analitičar želi saznati distribuciju broja pogrešno zavarenih pakiranja u jednom satu na svakoj liniji posebno, ali i njihovu zajedničku distribuciju. Naime, analitičar želi saznati pojavljuje li se povećan broj loše zavarenih pakiranja istovremeno na obje linije pa možda uzroke treba tražiti u npr. povremenim smetnjama u električnom napajanju i sličnim mogućim zajedničkim uzrocima.*

U tu svrhu analitičar je brojao pogrešno zavarena pakiranja sa svake linije tijekom 400 *sati i dobio podatke koje je bilježio u tablicu 6.3.*

sat	prva linija - broj grešaka	druga linija - broj grešaka
1	0	0
2	1	0
3	2	2
⋮	⋮	⋮
400	3	1

Tablica 6.3: Frekvencije pogrešno zavarenih vrećica na prvoj i drugoj liniji po satima.

Dobivene podatke pregledno (sumarno) možemo prikazati korištenjem **zajedničke tablice frekvencija** *oblika 6.4.*

		druga linija					
		0	1	2	3	4	zbroj
prva linija	0	22	12	13	12	7	66
	1	20	24	14	30	10	98
	2	15	20	30	10	7	82
	3	6	5	10	32	20	73
	4	5	7	13	31	25	81
zbroj		68	68	80	115	69	400

Tablica 6.4: Zajednička tablica frekvencija pogrešno zavarenih vrećica na obje linije.

Označimo li s X slučajnu varijablu kojom opisujemo broj pogrešno zavarenih vrećica po satu s prve linije, a Y s druge linije, vidimo da je skup svih mogućih realizacija pripadnog slučajnog vektora (X, Y) skup $\mathcal{R}(X,Y) = \{(0,0),(0,1),\ldots,(0,4),(1,0),\ldots,(1,4),\ldots,(4,4)\}$ i da se on lako može opisati korištenjem oznaka na gornjoj i lijevoj margini zajedničke tablice frekvencija 6.4. Iz tablice frekvencija 6.4 možemo odrediti empirijsku distribuciju slučajne varijable X (tablica 6.5) i Y (tablica 6.6) koje mogu poslužiti za procjenu stvarne (nepoznate) distribucije slučajnih varijabli X i Y.

vrijednost od X	0	1	2	3	4
relativna frekvencija	0.165	0.245	0.205	0.1825	0.2025

Tablica 6.5: Empirijska distribucija slučajne varijable X.

vrijednost od Y	0	1	2	3	4
relativna frekvencija	0.17	0.17	0.2	0.2875	0.1725

Tablica 6.6: Empirijska distribucija slučajne varijable Y.

Razmislite: kolika je procjena vjerojatnosti da na drugoj liniji budu 4 *loše zavarene vrećice bombona po satu, a koliko na prvoj liniji?*

Također, iz zajedničke tablice frekvencija 6.4 možemo izračunati relativnu frekvenciju pojavljivanja svakog uređenog para iz skupa mogućih realizacija slučajnog vektora (X, Y) *dijeljenjem odgovarajućih frekvencija ukupnim brojem slučajeva u uzorku, tj. s* 400 *(tablica 6.7).*

		Y				
		0	1	2	3	4
	0	0.0550	0.0300	0.0325	0.0300	0.0175
	1	0.0500	0.0600	0.0350	0.0750	0.0250
X	2	0.0375	0.0500	0.0750	0.0250	0.0175
	3	0.0150	0.0125	0.0250	0.0800	0.0500
	4	0.0125	0.0175	0.0325	0.0775	0.0625

Tablica 6.7: Zajednička tablica relativnih frekvencija pogrešno zavarenih vrećica na obje linije.

Ovako dobivena zajednička tablica relativnih frekvencija 6.7 odgovara **empirijskoj tablici distribucije diskretnog slučajnog vektora** (X, Y) *pa se može koristiti ako želimo npr. procijeniti koliko iznosi vjerojatnost da na prvoj liniji ne bude pogrešno zavarenih pakiranja, a istovremeno na drugoj liniji budu* 4 *pogreške, tj. za procjenu vjerojatnosti pojavljivanja odgovarajućih parova* $\{X = x\} \cap \{Y = y\}$. *Uočimo da se empirijske distribucije slučajne varijable* X *i slučajne varijable* Y *mogu dobiti sumiranjem odgovarajućih redaka, odnosno stupaca iz zajedničke tablice relativnih frekvencija 6.7, kao što je prikazano u tablici 6.8.*

		Y					
		0	1	2	3	4	zbroj
	0	0.0550	0.0300	0.0325	0.0300	0.0175	0.165
	1	0.0500	0.0600	0.0350	0.0750	0.0250	0.245
X	2	0.0375	0.0500	0.0750	0.0250	0.0175	0.205
	3	0.0150	0.0125	0.0250	0.0800	0.0500	0.1825
	4	0.0125	0.0175	0.0325	0.0775	0.0625	0.2025
zbroj		0.17	0.17	0.2	0.2875	0.1725	1

Tablica 6.8: Zajednička tablica relativnih frekvencija s marginama.

Općenito, distribucija dvodimenzionalnog slučajnog vektora (X, Y), pri čemu su $\{x_1, \ldots, x_m\}$ vrijednosti koje može primiti slučajna varijabla X (prva komponenta ovog vektora), a $\{y_1, \ldots, y_n\}$ slučajna varijabla Y (druga komponenta ovog vektora), dana je **tablicom distribucije** 6.9.

		Y			
		y_1	y_2	$\ldots$	y_n
X	x_1	$p(x_1, y_1)$	$p(x_1, y_2)$	$\ldots$	$p(x_1, y_n)$
	x_2	$p(x_2, y_1)$	$p(x_2, y_2)$	$\ldots$	$p(x_2, y_n)$
	$\vdots$	$\vdots$	$\vdots$		$\vdots$
	x_m	$p(x_m, y_1)$	$p(x_m, y_2)$	$\ldots$	$p(x_m, y_n)$

Tablica 6.9: Tablica distribucije dvodimenzionalnog diskretnog slučajnog vektora.

Broj $p(x_i, y_j)$ je vjerojatnost da slučajna varijabla X primi vrijednost x_i i slučajna varijabla Y vrijednost y_j, tj. vjerojatnost da se dogode oba događaja $\{X = x_i\}$ i $\{Y = y_j\}$:

$$p(x_i, y_j) = P(\{X = x_i\} \cap \{Y = y_j\}) = P\{X = x_i, Y = y_j\}.$$

Uočimo da se distribucije slučajnih varijabli koje čine ovaj slučajni vektor (tj. posebno distribucija od X i distribucija od Y) mogu također dobiti iz tablice distribucije slučajnog vektora zbrajanjem vjerojatnosti u odgovarajućim redovima, odnosno stupcima. Te distribucije zovemo **marginalne distribucije** slučajnog vektora (X, Y) te ih dodajemo u zajedničku tablicu distribucije kako je prikazano tablicom 6.10.

		Y				
		y_1	y_2	$\ldots$	y_n	zbroj
X	x_1	$p(x_1, y_1)$	$p(x_1, y_2)$	$\ldots$	$p(x_1, y_n)$	$p_X(x_1)$
	x_2	$p(x_2, y_1)$	$p(x_2, y_2)$	$\ldots$	$p(x_2, y_n)$	$p_X(x_2)$
	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
	x_m	$p(x_m, y_1)$	$p(x_m, y_2)$	$\ldots$	$p(x_m, y_n)$	$p_X(x_m)$
	zbroj	$p_Y(y_1)$	$p_Y(y_2)$	$\ldots$	$p_Y(y_n)$	1

Tablica 6.10: Tablica distribucije diskretnog slučajnog vektora s istaknutim marginalnim distribucijama.

Pri tome smo koristili oznake:

$$p_Y(y_1) = P\{Y = y_1\}, \ldots, p_Y(y_n) = P\{Y = y_n\}$$

$$p_X(x_1) = P\{X = x_1\}, \ldots, p_X(x_m) = P\{X = x_m\}.$$

Kao što je već navedeno, **empirijsku distribuciju diskretnog slučajnog vektora** dobijemo tako da elemente zajedničke tablice frekvencija dobivene temeljem nezavisnih mjerenja realizacija slučajnog vektora (X, Y) podijelimo ukupnim brojem mjerenja N.

Primjer 6.11. (djeca.sta)
U bazi podataka djeca.sta *nalazi se dio podataka o nekim ocjenama novorođenčeta, načinu poroda i majci iz istraživanja koje je provedeno u jednoj bolnici (pogledati zadatak 6.11). Varijabla* uzv *sadrži jednu ocjenu ultrazvučnog pregleda mozga novorođenčeta (u skali od* 1 *do* 4*), a varijabla* konvulzije *informaciju o tome je li novorođenče imalo konvulzije (vrijednost D) ili ne (vrijednost N). Označimo s X slučajnu varijablu kojom modeliramo pojavu konvulzija, a Y slučajnu varijablu kojom modeliramo ocjenu ultrazvučnog pregleda. Empirijska distribucija slučajnog vektora* (X, Y) *i empirijske distribucije njegovih margina X i Y prikazane su tablicom na slici 6.8.*

	Summary Frequency Table (djeca.sta) Table: konvulzije(2) x uzv(4)					
	konvulzije	uzv 1	uzv 2	uzv 3	uzv 4	Row Totals
Count	N	182	12	59	36	289
Total Percent		57.41%	3.79%	18.61%	11.36%	91.17%
Count	D	14	0	9	5	28
Total Percent		4.42%	0.00%	2.84%	1.58%	8.83%
Count	All Grps	196	12	68	41	317
Total Percent		61.83%	3.79%	21.45%	12.93%	

Slika 6.8: Empirijska distribucija slučajnog vektora (X, Y) kojim modeliramo pojavu konvulzija i ocjenu ultrazvučnog nalaza novorođenčeta.

Pomoću empirijske distribucije 6.8 možemo procijeniti vjerojatnosti realizacija slučajnog vektora (X, Y)*:*

- *procjena vjerojatnosti da su konvulzije bile prisutne i da je ocjena ultrazvuka* 1 *(tj. P{X=D, Y=1}) iznosi* 0.0442
- *procjena vjerojatnosti da su konvulzije bile prisutne i da je ocjena ultrazvuka* 4 *(tj. P{X=D, Y=4}) iznosi* 0.0158
- *procjena vjerojatnosti da konvulzije nisu bile prisutne i da je ocjena ultrazvuka* 4 *(tj. P{X=N, Y=4}) iznosi* 0.1136
- *procjena vjerojatnosti da su konvulzije bile prisutne (tj. P{X=D}) iznosi* 0.0883
- *procjena vjerojatnosti da je ocjena ultrazvuka* 4 *(tj. P{Y=4}) iznosi* 0.1293.

6.2.2 Uvjetne distribucije. Nezavisnost

Često se javlja potreba za proučavanjem distribucije jedne komponente slučajnog vektora ako je poznata realizacija njegove druge komponente. Takve distribucije nazivamo uvjetnim distribucijama slučajnog vektora.

Teorijske uvjetne distribucije i koncept zavisnosti dviju slučajnih varijabli definirat ćemo na temelju tablice distribucije 6.10 dvodimenzionalnog slučajnog vektora (X, Y) koristeći princip koji je ilustriran na empirijskoj distribuciji slučajnog vektora. Ako je teorijska distribucija dvodimenzionalnog slučajnog vektora (X, Y) dana tablicom distribucije 6.10, tada **uvjetne distribucije slučajne varijable Y uz uvjet da se dogodio** $\{X = x_i\}$ za one $x_i \in \mathcal{R}(X)$ za koje je $P\{X = x_i\} = p_X(x_i) \neq 0$ dobijemo kao niz brojeva

$$p_{\{Y|X=x_i\}}(y_j) = \frac{P\{X = x_i, Y = y_j\}}{P\{X = x_i\}} = \frac{p(x_i, y_j)}{p_X(x_i)}, \quad j = 1, \ldots, n.$$

Analogno, za one $y_j \in \mathcal{R}(Y)$ za koje je $P\{Y = y_j\} = p_Y(y_j) \neq 0$ dobivamo uvjetne distribucije od X uz uvjet da se dogodio dodađaj $\{Y = y_j\}$

$$p_{\{X|Y=y_j\}}(y_j) = \frac{P\{X = x_i, Y = y_j\}}{P\{Y = y_j\}} = \frac{p(x_i, y_j)}{p_Y(y_j)}, \quad i = 1, \ldots, m.$$

Za slučajne varijable X i Y, čija je zajednička distribucija dana tablicom 6.10, kažemo da su nezavisne ako za sve $i = 1, \ldots, m$, $j = 1, \ldots, n$ vrijedi da je

$$p(x_i, y_j) = p_X(x_i) \cdot p_Y(y_j),$$

tj. vjerojatnosti iz distribucije slučajnog vektora mogu se dobiti množenjem odgovarajućih vjerojatnosti iz marginalnih distribucija. U suprotnom kažemo da su slučajne varijable X i Y zavisne.

Koncept nezavisnosti slučajnih varijabli X i Y usko je vezan uz uvjetne distribucije, tj. distribucije uvjetovanih slučajnih varijabli

$$X|Y = y_j, \quad y_j \in \mathcal{R}(Y), \quad j = 1, \ldots, , n$$

$$Y|X = x_i, \quad x_i \in \mathcal{R}(X), \quad i = 1, \ldots, m.$$

Naime, ako su slučajne varijable X i Y nezavisne, tada za svaki $x_i \in \mathcal{R}(X)$ za koji je $p_X(x_i) \neq 0$ i za svaki $y_j \in \mathcal{R}(Y)$ vrijedi da je

$$p_{\{Y|X=x_i\}}(y_j) = \frac{P\{X = x_i, Y = y_j\}}{P\{X = x_i\}} =$$

$$= \frac{p(x_i, y_j)}{p_X(x_i)} = \frac{p_X(x_i) \cdot p_Y(y_j)}{p_X(x_i)} = p_Y(y_j).$$

Dakle, ako su X i Y nezavisne, tada vrijedi:

- za svaki $x_i \in \mathcal{R}(X)$ za koji je $p_X(x_i) \neq 0$ slučajne varijable Y i $Y|X = x_i$ imaju jednake distribucije
- za svaki $y_j \in \mathcal{R}(Y)$ za koji je $p_Y(y_j) \neq 0$ slučajne varijable X i $X|Y = y_j$ imaju jednake distribucije.

Primjer 6.12. (citanje.sta)
Baza podataka citanje.sta *sadrži rezultate istraživanja o čitateljskim navikama stanovnika jednog grada. Varijabla* citanje *sadrži informaciju o tome pročita li ispitanik svaka tri mjeseca barem jednu knjigu (*1 - *pročita,* 0 - *ne pročita), varijabla* spol *sadrži informaciju o spolu ispitanika (Z - žena, M - muškarac), a varijabla* obrazovanje *stupanj obrazovanja svakog ispitanika (NSS - niža stručna sprema, SSS - srednja stručna sprema, VSS - visoka stručna sprema).*
*Neka je (X, Y) slučajni vektor gdje je X slučajna varijabla koja se realizira jedinicom ako stanovnik tog grada svaka tri mjeseca pročita barem jednu knjigu, a inače se realizira nulom, a Y slučajna varijabla kojom modeliramo stručnu spremu stanovnika tog grada (*1 - *NSS,* 2 - *SSS,* 3 - *VSS). Ako želimo analizirati čitateljske navike stanovnika tog grada s obzirom na njihovo obrazovanje, zapravo trebamo proučavati slučajnu varijablu X uvjetovanu na poznatu (danu) vrijednost slučajne varijable Y. Tako dolazimo do tablica frekvencija 6.11, 6.12 i 6.13.*

X	0	1	zbroj
frekvencija pod uvjetom $Y = 1$ (NSS)	48	16	64

Tablica 6.11: Tablica frekvencija slučajne varijable X uvjetovane na $\{Y = 1\}$.

X	0	1	zbroj
frekvencija pod uvjetom $Y = 2$ (SSS)	426	51	477

Tablica 6.12: Tablica frekvencija slučajne varijable X uvjetovane na $\{Y = 2\}$.

X	0	1	zbroj
frekvencija pod uvjetom $Y = 3$ (VSS)	184	19	203

Tablica 6.13: Tablica frekvencija slučajne varijable X uvjetovane na $\{Y = 3\}$.

Frekvencije iz tablica 6.11, 6.12 i 6.13 možemo interpretirati kao frekvencije realizacija novih slučajnih varijabli $X|Y = 1$ (X u uvjetima $Y = 1$), $X|Y = 2$ (X u uvjetima $Y = 2$) i $X|Y = 3$ (X u uvjetima $Y = 3$). Njihove distribucije redom zovemo: **uvjetna distribucija od X uz uvjet da je $Y = 1$**, **uvjetna distribucija od X uz uvjet da je $Y = 2$** *i* **uvjetna distribucija od X uz uvjet da je $Y = 3$**. *Ako se te uvjetne distribucije razlikuju od distribucije slučajne varijable X, možemo to interpretirati kao činjenicu da čitateljske navike stanovnika (varijabla X) ovise o stupnju obrazovanja, tj. to sugerira da su X i Y zavisne slučajne varijable.*
Kao što smo već naučili, u statistici su stvarne distribucije uglavnom nepoznate pa ih treba procijeniti na temelju podataka. Tako je i sa stvarnim uvjetnim distribucijama. U tu svrhu pomoću tablica frekvencija 6.11, 6.12 i 6.13 računamo empirijske distribucije navedenih uvjetovanih slučajnih varijabli (tablice 6.14, 6.15 i 6.16).

X	0	1	zbroj
relativna frekvencija pod uvjetom $Y = 1$ (NSS)	0.75	0.25	1

Tablica 6.14: Empirijska distribucija slučajne varijable $X|Y = 1$.

X	0	1	zbroj
relativna frekvencija pod uvjetom $Y = 2$ (SSS)	0.89	0.11	1

Tablica 6.15: Empirijska distribucija slučajne varijable $X|Y = 2$.

X	0	1	zbroj
relativna frekvencija pod uvjetom $Y = 3$ (VSS)	0.91	0.09	1

Tablica 6.16: Empirijska distribucija slučajne varijable $X|Y = 3$.

Tablica na slici 6.9 sadrži empirijsku distribuciju slučajnog vektora (X, Y) (plavi postoci), njegove marginalne distribucije (ljubičasti postoci), empirijsku distribuciju slučajne varijable X uvjetovanu na poznatu vrijednost slučajne varijable Y (crveni postoci u istom redu tablice) te empirijsku distribuciju slučajne varijable Y uvjetovanu na poznatu vrijednost slučajne varijable X (zeleni postoci u istom stupcu tablice).

	Summary Frequency Table (citanje.sta) Table: obrazovanje(3) x citanje(2)			
	obrazovanje	citanje 0	citanje 1	Row Totals
Count	**NSS**	**48**	**16**	**64**
Column Percent		**7.29%**	**18.60%**	
Row Percent		**75.00%**	**25.00%**	
Total Percent		**6.45%**	**2.15%**	**8.60%**
Count	**SSS**	**426**	**51**	**477**
Column Percent		**64.74%**	**59.30%**	
Row Percent		**89.31%**	**10.69%**	
Total Percent		**57.26%**	**6.85%**	**64.11%**
Count	**VSS**	**184**	**19**	**203**
Column Percent		**27.96%**	**22.09%**	
Row Percent		**90.64%**	**9.36%**	
Total Percent		**24.73%**	**2.55%**	**27.28%**
Count	**All Grps**	**658**	**86**	**744**
Total Percent		**88.44%**	**11.56%**	

Slika 6.9: Tablica distribucije slučajnog vektora (X, Y) iz primjera 6.12, njegove marginalne i uvjetne distribucije.

Ako pretpostavimo da empirijska distribucija slučajnog vektora (X, Y) dobro opisuje njegovu stvarnu distribuciju, možemo procijeniti npr. sljedeće vjerojatnosti:

ako biramo među ispitanicima koji svaka tri mjeseca pročitaju barem jednu knjigu, procjena vjerojatnosti da izaberemo osobu s visokom stručnom spremom, tj. vjerojatnosti $P\{Y = 3|X = 1\}$, iznosi 0.22

ako biramo među ispitanicima s nižom stručnom spremom, procjena vjerojatnosti da izaberemo osobu koja svaka tri mjeseca pročita barem jednu knjigu, tj. vjerojatnosti $P\{X = 1|Y = 1\}$, *iznosi* 0.25.

Analizom tablice 6.9 dolazimo do zaključka da se odgovarajuće empirijske uvjetne i empirijske marginalne distribucije slučajnog vektora (X, Y) *ne podudaraju pa to može sugerirati da stupanj obrazovanja i čitateljske navike ispitanika iz populacije koju promatramo nisu nezavisne varijable. Međutim, nezavisnost slučajnih varijabli definirana je na temelju stvarnih, a ne empirijskih distribucija. Prema tome, zaključak sugeriran empirijskim distribucijama može biti pogrešan. U sljedećem poglavlju opisat ćemo postupak testiranja hipoteze o nezavisnosti dviju slučajnih varijabli i tako riješiti nedoumicu koja je ovdje prisutna.*

Primjer 6.13. (citanje.sta)
Ako želimo analizirati čitateljske navike s obzirom na spol stanovnika tog grada, tada trebamo procijeniti distribuciju slučajne varijable X *uvjetovane na vrijednost slučajne varijable koja se realizira jedinicom ako je osoba ženskog spola (vrijednost* Z *varijable* spol*), a dvojkom ako je osoba muškog spola (vrijednost* M *varijable* spol*). Označimo tu slučajnu varijablu sa* Z. *Empirijske distribucije uvjetovanih slučajnih varijabli* $X|Z = 1$ *i* $X|Z = 2$ *dane su u tablici 6.10.*

	Summary Frequency Table (citanje.STA) Table: spol(2) x citanje(2)			
	spol	citanje 0	citanje 1	Row Totals
Count	Z	313	55	368
Row Percent		85.05%	14.95%	
Count	M	345	31	376
Row Percent		91.76%	8.24%	
Count	All Grps	658	86	744

Slika 6.10: Empirijske distribucije uvjetovanih slučajnih varijabli $X|Z = 1$ i $X|Z = 2$.

Sada možemo procijeniti vjerojatnosti sljedećeg tipa:

ako je slučajno odabrani stanovnik tog grada žena, tada procjena vjerojatnosti da ona svaka tri mjeseca pročita bar jednu knjigu, tj. vjerojatnosti $P\{X = 1|Z = 1\}$, *iznosi* 0.15

ako je slučajno odabrani stanovnik tog grada muškarac, tada procjena vjerojatnosti da on svaka tri mjeseca pročita bar jednu knjigu, tj. vjerojatnosti $P\{X = 1|Z = 2\}$, *iznosi* 0.08.

6.3 Analiza zavisnosti

U prethodnom poglavlju koristili smo podatke dobivene kao realizacije dvodimenzionalnog diskretnog slučajnog vektora te smo uveli pojmove uvjetnih distribucija slučajnog vektora i zavisnosti slučajnih varijabli. Na temelju podataka odredili smo empirijsku distribuciju slučajnog vektora (X, Y), marginalne empirijske distribucije, kao i uvjetne empirijske distribucije koje koristimo za procjenu odgovarajućih

stvarnih distribucija. Međutim, zavisnost slučajnih varijabli definirana je na temelju pravih, a ne empirijskih distribucija. Prirodno je da procjene odstupaju od stvarnih distribucija pa se postavlja pitanje kako temeljem prikupljenih podataka provjeriti jesu li slučajne varijable, koje su margine slučajnog vektora, zavisne ili ne. U ovom poglavlju opisat ćemo statistički test kojim možemo testirati hipotezu o nezavisnosti slučajnih varijabli.

Da bi test bio jasno prezentiran, prikazat ćemo zajedničku tablicu frekvencija slučajnog uzorka dvodimenzionalnog slučajnog vektora (X, Y) tablicom 6.17.

		Y				
		y_1	y_2	$\dots$	y_n	*zbroj*
X	x_1	$n(x_1, y_1)$	$n(x_1, y_2)$	$\dots$	$n(x_1, y_n)$	$n_X(x_1)$
	x_2	$n(x_2, y_1)$	$n(x_2, y_2)$	$\dots$	$n(x_2, y_n)$	$n_X(x_2)$
	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
	x_m	$n(x_m, y_1)$	$n(x_m, y_2)$	$\dots$	$n(x_m, y_n)$	$n_X(x_m)$
	zbroj	$n_Y(y_1)$	$n_Y(y_2)$	$\dots$	$n_Y(y_n)$	N

Tablica 6.17: Zajednička tablica frekvencija slučajnog vektora (X, Y).

Procjenu za teorijsku distribuciju ovog slučajnog vektora dobijemo na temelju zajedničke empirijske distribucije, kako je ilustrirano u prethodnom poglavlju, a prikazat ćemo je (u skladu s teorijskom distribucijom iz tablice 6.9) tablicom 6.18.

		Y				
		y_1	y_2	$\dots$	y_n	*zbroj*
X	x_1	$\widehat{p}(x_1, y_1)$	$\widehat{p}(x_1, y_2)$	$\dots$	$\widehat{p}(x_1, y_n)$	$\widehat{p}_X(x_1)$
	x_2	$\widehat{p}(x_2, y_1)$	$\widehat{p}(x_2, y_2)$	$\dots$	$\widehat{p}(x_2, y_n)$	$\widehat{p}_X(x_2)$
	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
	x_m	$\widehat{p}(x_m, y_1)$	$\widehat{p}(x_m, y_2)$	$\dots$	$\widehat{p}(x_m, y_n)$	$\widehat{p}_X(x_m)$
	suma	$\widehat{p}_Y(y_1)$	$\widehat{p}_Y(y_2)$	$\dots$	$\widehat{p}_Y(y_n)$	1

Tablica 6.18: Zajednička empirijska distribucija slučajnog vektora (X, Y).

Uočimo da je stvarna tablica distribucije (tablica 6.9) slučajnog vektora (X, Y) dana na analogan način kao empirijska, samo su pripadne vjerojatnosti označene s p bez "kapice".

Kod dovoljno velikih veličina uzorka, za testiranje nul-hipoteze da su slučajne varijable X i Y nezavisne, tj. nul-hipoteze

$$\mathcal{H}_0: \quad p(x_i, y_j) = p_X(x_i) \cdot p_Y(y_j), \quad \forall i = 1 \ldots, m,\ j = 1, \ldots, n,$$

možemo koristiti χ^2 test. On se temelji na usporedbi očekivanih frekvencija po poljima tablice u uvjetima istinitosti nul-hipoteze s frekvencijama koje u tom polju stvarno imamo na osnovi podataka. Očekivana frekvencija ij-tog polja tablice u uvjetima istinitosti nul-hipoteze je

$$E_{ij} = N\widehat{p}_X(x_i)\widehat{p}_Y(y_j) = \frac{n_X(x_i)n_Y(y_j)}{N},$$

dok je eksperimentalna (utvrđena) frekvencija

$$n_{ij} = n(x_i, y_j).$$

Ako su X i Y nezavisne slučajne varijable, test-statistika

$$\chi^2 = \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{(n_{ij} - E_{ij})^2}{E_{ij}}$$

ima χ^2 distribuciju s $(n-1)(m-1)$ stupnjeva slobode. Na temelju realizacije test statistike određujemo pripadnu p-vrijednost na uobičajeni način i usporedbom dobivene p-vrijednosti s nivoom značajnosti α donosimo odluku:

ako je $p < \alpha$, odbacujemo nul-hipotezu i na razini značajnosti α prihvaćamo alternativnu hipotezu, tj. kažemo da podaci potvrđuju postojanje zavisnosti između varijabli X i Y na nivou značajnosti α

ako je $p > \alpha$, nemamo dovoljno argumenata koji bi poduprli odluku o odbacivanju nul-hipoteze, tj. kažemo da podaci ne daju potvrdu o postojanju zavisnosti među varijablama X i Y.

Veličina uzorka koja je dovoljna za primjenu ovog testa analizirana je detaljno u statističkoj literaturi i može se odrediti na nekoliko različitih načina. Tako npr. znamo da je uzorak dovoljno velik ako su očekivane frekvencije u svakom polju tablice frekvencija veće od 5.

Valja napomenuti da zavisnost slučajnih varijabli još uvijek ne znači i uzročnu vezu. Naime, može se dogoditi da varijable nisu uzročno povezane, ali imaju neku zajedničku varijablu koja je s objema u uzročnoj vezi (analizirajte u tom kontekstu primjer 6.10.)

Primjer 6.14. (citanje.sta)
Sjetimo se primjera 6.12 u kojemu smo govorili o istraživanju čitalačkih navika stanovnika jednog grada. Analiza tablice na slici 6.9 sugerirala je postojanje zavisnosti između slučajnih varijabli X (čitalačke navike, varijabla citanje*) i Y (stručna sprema, varijabla* obrazovanje*). Ako sa Z označimo slučajnu varijablu kojom modeliramo spol, možemo analizirati i zavisnost slučajnih varijabli X i Z. Tablice na slici 6.11 prikazuju p-vrijednosti provedenih χ^2 testova.*

Statistic	Chi-square	df	p
Pearson Chi-square	12.62149	df=2	p=.00182
M-L Chi-square	10.23795	df=2	p=.00598

(a) citanje i obrazovanje

Statistic	Chi-square	df	p
Pearson Chi-square	8.168828	df=1	p=.00426
M-L Chi-square	8.258259	df=1	p=.00406

(b) citanje i spol

Slika 6.11: p-vrijednosti χ^2 testa za testiranje hipoteze o nezavisnosti slučajnih varijabli iz primjera 6.14.

Kako su u oba slučaja p-vrijednosti manje on zadanog nivoa značajnosti $\alpha = 0.05$, zaključujemo da u oba slučaja odbacujemo nul-hipotezu i na nivou značajnosti 0.05 prihvaćamo alternativnu hipotezu koja kaže da su slučajne varijable X i Y, odnosno X i Z, zavisne. Dakle, na nivou značajnosti $\alpha = 0.05$ možemo tvrditi da je slučajna varijabla kojom modeliramo čitateljske navike zavisna o slučajnim varijablama kojima modeliramo spol i stručnu spremu.

6.4 Jednostavna linearna regresija

Ako imamo parove podataka iz dvije neprekidne slučajne varijable i želimo zaključivati o postojanju zavisnosti između njih, metoda iz prethodnog poglavlja nije prikladna. Naime, da bismo primijenili navedenu metodu, trebali bismo varijable kategorizirati, a postupak kategorizacije nerijetko može značajno utjecati na statističke zaključke s obzirom da se u tom postupku uvijek gubi dio informacija. Prije nego što se upustimo u zaključivanje o zavisnosti između dvije slučajne varijable, promotrit ćemo dva prirodna tipa veza među varijablama.

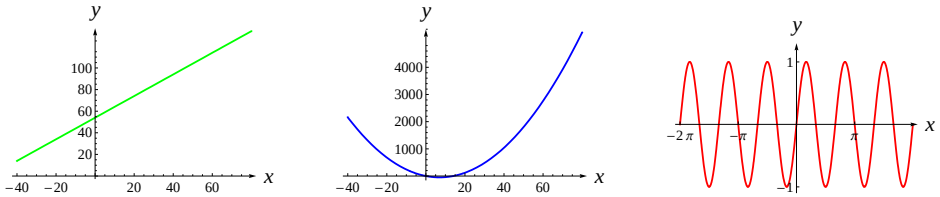
6.4.1 Deterministička veza

Deterministička veza između dvije varijable jest veza zadana pravilom oblika

$$y = f(x)$$

gdje je y zavisna varijabla, x nezavisna varijabla, a $f\colon \mathbb{R} \to \mathbb{R}$ zadana funkcija. Na primjer, pravilima $y = x + 54$, $y = x^2 - 14x$ i $y = \sin(3x)$ zadane su determinističke veze među varijablama x i y jer za svaku dopuštenu vrijednost nezavisne varijable x

možemo izračunati točnu vrijednost zavisne varijable y. Grafovi ovih triju funkcija prikazani su na slici 6.12.



(a) $y = x + 54$ (b) $y = x^2 - 14x$ (c) $y = \sin(3x)$

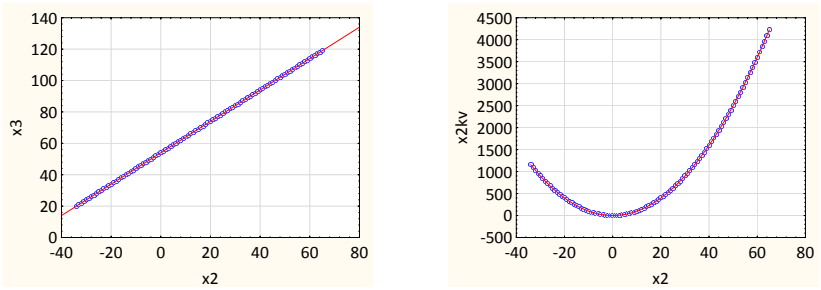
Slika 6.12: Grafovi jedne linearne funkcije, jednog polinoma drugog stupnja i jedne trigonometrijske funkcije.

Primjer 6.15. (regresija.sta)
Baza podataka regresija.sta*, između ostalih varijabli, sadrži simulirane vrijednosti varijable* x2*, varijable* x3 *čije su vrijednosti dobivene dodavanjem broja* 54 *svakoj vrijednosti varijable* x2 *(*x3 = x2 + 54*) i varijable* x2kv *čije su vrijednosti dobivene pomoću pravila* x2kv = (x2)^2 - 14· x2*. Vidimo da se ovdje radi o determinističkim vezama među varijablama:*

veza između varijabli x2 *(nezavisna varijabla) i* x3 *(zavisna varijabla) je linearna*

veza između varijabli x2 *(nezavisna varijabla) i* x2kv *(zavisna varijabla) je polinomijalna drugog stupnja.*

*Parovi(*x2, x2kv*) i (*x2, x3*) podataka iz baze* regresija.sta *prikazani su na slici 6.13.*



(a) x3 = x2 + 54 (b) x2kv = (x2)^2 - 14· x2

Slika 6.13: Parovi podataka (x2, x3) i (x2, x2kv) za sve simulirane vrijednosti varijable x2 iz baze regresija.sta.

6.4.2 Statistički model s aditivnom greškom

U statističkim analizama nije realno očekivati determinističke veze. To ćemo najlakše uočiti ako promatramo **dijagram raspršenosti** podataka (eng. scatter plot)

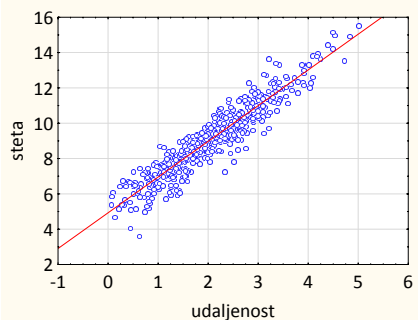
kojim je dan prikaz uređenih parova podataka iz dviju slučajnih varijabli u koordinatnom sustavu.

Primjer 6.16. (pozar.sta)
Baza podataka pozar.sta *sadrzi podatke o* 100 *požara na nekom području:*

varijabla udaljenost *sadrži udaljenost u kilometrima od mjesta požara do najbližeg vatrogasnog centra*

varijabla steta *sadrži štetu uzrokovanu požarom izraženu u tisućama kuna.*

Intuicija nam govori da materijalna šteta uzrokovana požarom ovisi o blizini vatrogasnog centra, tj. da veća udaljenost vatrogasnog centra od mjesta požara sugerira veću štetu. Jezikom matematike to bi značilo da je sugerirana funkcijska veza između udaljenosti mjesta požara od najbližeg vatrogasnog centra i štete nastale požarom rastuća. To možemo provjeriti crtajući dijagram raspršenosti vrijednosti varijabli udaljenost *i* steta *(slika 6.14).*



Slika 6.14: Dijagram raspršenosti vrijednosti varijabli udaljenost i steta.

Vidimo da se parovi vrijednosti varijabli udaljenost *(neovisna varijabla) i* steta *(ovisna varijabla) grupiraju oko pravca koji ima pozitivan koeficijent smjera. To sugerira da bi rastuća linearna veza među ovim varijablama bila dobar odabir za modeliranje zavisnosti među ovim varijablama, ali vidimo da ne možemo odrediti pravac tako da svi podaci leže na njemu. Prirodna je pretpostavka da se među ovim varijablama može uspostaviti funkcijska veza do na neku grešku.*

Regresijska metoda modeliranja koju ćemo opisati u ovom poglavlju pretpostavlja da možemo uspostaviti funkcijsku vezu do na dodanu grešku, tj. da će veza između neovisne varijable x i ovisne slučajne varijable $Y(x)$ biti oblika

$$Y(x) = f(x) + \varepsilon, \tag{6.4}$$

gdje pretpostavljamo da je ε slučajna varijabla koja opisuje grešku u modeliranju. Koristeći se činjenicom da mnogo nezavisnih slučajnih smetnji u pravilu ima normalnu distribuciju, u primjenama se u klasičnom načinu modeliranja prihvaća da je model adekvatan ako je u njemu postignuta normalna distribuiranost grešaka ε,

uz ostale zahtjeve o kojima će biti riječi u ovom poglavlju. Primjer 6.17 uvodi nas u problematiku ovakvog modeliranja.

Primjer 6.17. *Iz medicinskih istraživanja poznato je da krvni tlak čovjeka ima tendenciju porasta s porastom dobi. Htjeli bismo, temeljem prikupljenih podataka, argumentirati tu činjenicu te modelirati vezu između krvnog tlaka i dobi. U tu svrhu označimo s x dob ispitanika, a s $Y(x)$ slučajnu varijablu kojom modeliramo krvni tlak za dob x. Krvni tlak za osobu dobi x moramo modelirati kao slučajnu varijablu s obzirom da je prirodno da osobe iste dobi nemaju i isti krvni tlak. Pretpostavimo da krvni tlak u populaciji za dob x, možemo modelirati kao normalnu slučajnu varijablu s očekivanjem $\mu(x)$ i varijancom σ^2. Na taj način svakoj dobi x pripada odgovarajuća normalna razdioba $\mathcal{N}(\mu(x), \sigma^2)$ krvnog tlaka $Y(x)$. Činjenica da se starenjem povećava krvni tlak trebala bi se odraziti na funkciju $x \mapsto \mu(x)$ koja dobi pridružuje očekivanu vrijednost krvnog tlaka u toj dobi. Ova bi funkcija, prema očekivanjima, trebala biti rastuća.*

Dakle, cilj je na temelju sparenih mjerenja $(x_1, y_1), \ldots, (x_n, y_n)$ dvaju obilježja ustanoviti prirodu ovisnosti slučajnih varijabli $Y_1, \ldots Y_n$ (čije su realizacije realni brojevi $y_1, \ldots y_n$) o neovisnoj varijabli x (čije su izmjerene vrijednosti $x_1, \ldots x_n$). Ako je matematički model oblika

$$Y_i = f(x_i) + \varepsilon_i, \quad i = 1, \ldots, n,$$

gdje je $t \mapsto f(t)$ realna funkcija jedne realne varijable, a $\varepsilon_1, \ldots, \varepsilon_n$ međusobno nezavisne slučajne varijable t.d. je $E\,\varepsilon_i = 0$ i $\mathrm{Var}(\varepsilon_i) = \sigma^2$, onda govorimo o **regresijskom modelu**.
Prvi korak u uspostavljanju ovakvih veza među varijablama Y i x prikaz je podataka u dijagramu raspršenosti iz kojeg se lako vidi grupiraju li se sparena mjerenja oko pravca (linearna zavisnost) ili neke krivulje (neka druga funkcijska zavisnost - polinomijalna ($n \geq 2$), logaritamska, ...).

6.4.3 Regresijski pravac

Pretpostavimo da je graf funkcije $f(x)$ u modelu 6.4 pravac. To znači da $f(x)$ možemo algebarski prikazati formulom $f(x) = \alpha + \beta x$. Slobodni koeficijent α zove se odsječak na y-osi, a koeficijent β uz neovisnu varijablu x zove se koeficijent smjera i važan je iz sljedećeg razloga:

ako je $\beta < 0$ funkcija $f(x) = \alpha + \beta x$ je padajuća

ako je $\beta > 0$ funkcija $f(x) = \alpha + \beta x$ je rastuća.

U kontekstu ovog statističkog modela graf funkcije $f(x) = \alpha + \beta x$ nazivamo **regresijskim pravcem**, a koeficijente α i β **regresijskim parametrima**.

6.4.4 Statistički model

Linearni regresijski model može se zapisati u obliku

$$Y_i = \alpha + \beta x_i + \varepsilon_i, \quad i = 1, \ldots, n.$$

Ovdje su:

$x_1, x_2, \ldots, x_n$ vrijednosti varijable x koje je analitičar **izabrao/izmjerio** u svrhu studije.

$Y_1, Y_2 \ldots, Y_n$ slučajne varijable (njihove izmjerene vrijednosti su $y_1, \ldots, y_n$).

$\varepsilon_1, \ldots, \varepsilon_n$ predstavljaju varijable greške koja je dodana na linearnu vezu $(\alpha + \beta x_i)$. Ovo su **nemjerljive slučajne varijable** za koje pretpostavljamo da su međusobno nezavisne i da sve imaju normalnu distribuciju s očekivanjem 0 i istom varijancom σ^2.

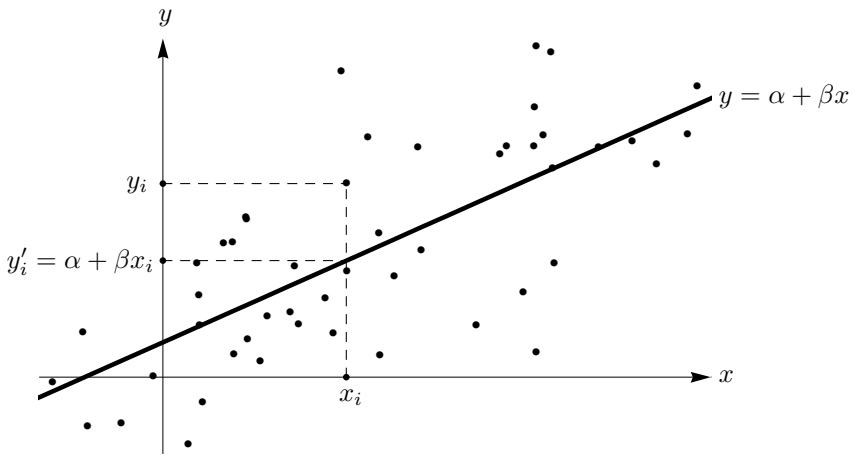
α i β su **nepoznati parametri** linearne veze koje treba odrediti u postupku modeliranja, tj. **procijeniti**. To zapravo znači da trebamo **procijeniti regresijski pravac** $y = \alpha + \beta x$.

6.4.5 Metoda najmanjih kvadrata

Problem procjene nepoznatih parametara α i β možemo identificirati s procjenom nepoznatog regresijskog pravca $y = \alpha + \beta x$. Pretpostavimo da je u dijagramu raspršenosti dodan graf pravca $y = \alpha + \beta x$ (slika 6.15).

Za svaku izmjerenu vrijednost x_i možemo odrediti broj $y_i' = \alpha + \beta x_i$ koji odgovara vrijednosti očekivanja ovisne varijable u x_i. Taj broj zovemo **teorijska vrijednost** ovisne varijable u x_i (eng. predicted value). Izmjerena ili eksperimentalna vrijednost ovisne varijable u x_i (eng. observed value) je y_i. Ona se u pravilu razlikuje od teorijske vrijednosti pa točke (x_i, y_i), $i = 1, \ldots, n$, uglavnom ne leže na regresijskom pravcu.

Da bi model bio dobar, trebale bi razlike među izmjerenim i teorijskim vrijednosti ovisne varijable, tj. razlika između y_i i $(\alpha + \beta x_i)$ biti što manje. U skladu s tom idejom regresijske parametre α i β standardno procjenjujemo **metodom najmanjih kvadrata**.



Slika 6.15: Regresijski pravac $y = \alpha + \beta x$.

Ideja metode najmanjih kvadrata je minimizacija sume kvadrata odstupanja teorijskih od eksperimentalnih vrijednosti, tj. procjene $\widehat{\alpha}$ i $\widehat{\beta}$ regresijskih parametara α i β trebamo odrediti tako da vrijedi:

$$D(\widehat{\alpha},\widehat{\beta}) = \sum(\text{eksperimentalne vrijednosti} \ - \ \text{teorijske vrijednosti})^2 =$$

$$= \sum_{i=1}^{n} \Big(y_i - (\widehat{\alpha} + \widehat{\beta}x_i)\Big)^2 = \min_{(\alpha,\beta)\in\mathbb{R}^2} \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2 = \min_{(\alpha,\beta)\in\mathbb{R}^2} D(\alpha,\beta).$$

Dakle, α i β biramo tako da za vrijednosti $\widehat{\alpha}$ i $\widehat{\beta}$, $D(\widehat{\alpha},\widehat{\beta})$ prima minimalnu vrijednost koju može postići po svim mogućim vrijednostima (α,β). Takve procjene $\widehat{\alpha}$ i $\widehat{\beta}$ nazivamo **procjenama u smislu metode najmanjih kvadrata** (eng. least square estimates) regresijskih parametara α i β. Jasno je da je u tom smislu procjena nepoznatog regresijskog pravca $y = \alpha + \beta x$ upravo pravac $\widehat{y} = \widehat{\alpha} + \widehat{\beta}x$.

Za zapis procjena $\widehat{\alpha}$ i $\widehat{\beta}$ parametara α i β potrebne su sljedeće veličine:

$$\overline{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \overline{y}_n = \frac{1}{n}\sum_{i=1}^{n} y_i,$$

$$s_x^2 = \sum_{i=1}^{n}(x_i - \overline{x}_n)^2, \quad s_y^2 = \sum_{i=1}^{n}(y_i - \overline{y}_n)^2, \quad s_{xy} = \sum_{i=1}^{n}(x_i - \overline{x}_n)(y_i - \overline{y}_n).$$

Korištenjem tih veličina procjene za nepoznate regresijske parametre β i α u smislu najmanjih kvadrata su:

$$\widehat{\beta} = \frac{s_{xy}}{s_x^2}, \quad \widehat{\alpha} = \overline{y}_n - \widehat{\beta}\,\overline{x}_n,$$

tj. regresijski pravac $y = \alpha + \beta x$ procjenjujemo pravcem

$$\widehat{y} = \widehat{\alpha} + \widehat{\beta} x$$

koji ćemo zvati **procjena regresijskog pravca**.

Uočimo da, koristeći formulu procijenjenog regresijskog pravca $\widehat{y} = \widehat{\alpha} + \widehat{\beta}x$, za svaku vrijednost x možemo izračunati pripadnu procjenu teorijske vrijednosti, tj. vrijednost $\widehat{y}$. Te vrijednosti zovemo **predikcije**. To znači da za svaku vrijednost x_i nezavisne varijable možemo izračunati iznos odstupanja procijenjene teorijske vrijednosti $\widehat{y}_i$ od izmjerene vrijednosti y_i ovisne varijable:

$$e_i = y_i - \widehat{y}_i = y_i - (\widehat{\alpha} + \widehat{\beta} x_i).$$

Tako dobivene vrijednosti $e_i, \ldots, e_n$ zovemo **rezidualima** i možemo ih smatrati procjenama grešaka $\varepsilon_1, \ldots, \varepsilon_n$ iz modela $Y_i = \alpha + \beta x_i + \varepsilon_i$. Suma kvadrata svih reziduala upravo je minimalna postignuta vrijednost za D, tj. $D(\widehat{\alpha}, \widehat{\beta})$, i predstavlja jednu mjeru kvalitete modela koju označavamo SSE[1]:

$$SSE = \sum_{i=1}^{n} e_i^2.$$

Primjer 6.18. (pozar.sta)
U primjeru 6.16 analizom dijagrama raspršenosti 6.14 zaključili smo da se parovi vrijednosti varijabli udaljenost *i* steta *grupiraju oko pravca. Metodom najmanjih kvadrata možemo odrediti jednadžbu tog pravca:* $y = 4.9275 + 2.0224x$.

Promotrimo prvi redak baze podataka pozar.sta. *U njemu je zabilježena vrijednost* $x_1 = 1.27$ *varijable* udaljenost *i odgovarajuća vrijednost varijable* steta $y_1 = 7.54$. *Pomoću procjene regresijskog pravca sada lako možemo izračunati predikciju ovisne varijable koja odgovara vrijednosti* $x_1 = 1.27$*:*

$$\widehat{y}_1 = \widehat{\alpha} + \widehat{\beta} x_1 = 2.0224 \cdot 1.27 + 4.9275 = 7.496.$$

Odgovarajući rezidual tada iznosi

$$e_1 = y_1 - (\widehat{\alpha} + \widehat{\beta} x_1) = 7.54 - (2.0224 \cdot 1.27 + 4.9275) = 0.044.$$

Reziduale za sve parove (x_i, y_i) *eksperimenatlnih vrijednosti možemo dobiti u programskom paketu Statistica (slika 6.16).*

[1] SSE je kratica za sum of squares of errors.

	Observed, Predicted, and Residual Values		
Case number	steta Observed	steta Predicted	steta Resids
1	7,54138	7,49736	0,04402
2	9,53428	9,84628	-0,31200
3	10,44098	11,39049	-0,94951
4	9,61824	10,19044	-0,57220
5	6,36490	6,62169	-0,25679
6	11,13802	10,72266	0,41536
7	7,34543	7,42164	-0,07621
8	6,03117	6,75742	-0,72625
9	12,18975	12,33009	-0,14034
10	10,56394	10,82848	-0,26454
11	9,69733	9,58729	0,11004

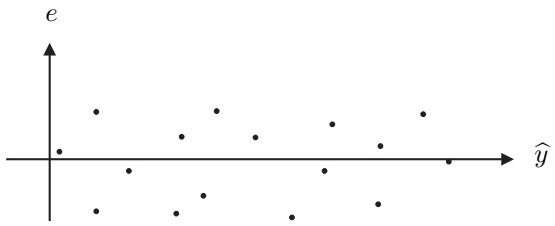
Slika 6.16: Tablica nekoliko prvih reziduala za varijablu steta iz baze podataka pozar.sta.

6.4.6 Statističko zaključivanje

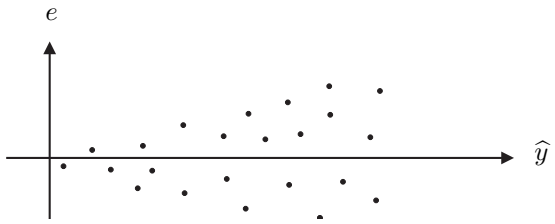
Da bismo mogli koristiti ovako dobiven model potrebno je prvo napraviti analizu prihvatljivosti modela kojom istražujemo jesu li ispunjene osnovne pretpostavke klasičnog regresijskog modela. Sjetimo se, greške modela trebaju biti međusobno nezavisne i jednako distribuirane slučajne varijable s normalnom distribucijom. Dio analize modela koji se provodi u tu svrhu obično se naziva **analiza reziduala**.

Analiza reziduala

Detaljna analiza reziduala složen je postupak koji prelazi okvire ove knjige. Za potrebe osnovne statističke analize ovdje navodimo samo nekoliko vizualnih provjera reziduala na temelju kojih se može naslutiti da postoji sumnja u istinitost pretpostavki modela, što automatski znači da je takav model neprihvatljiv za bilo kakvu daljnu interpretaciju ili korištenje.
Prva pretpostavka koju greške modela $\varepsilon_1, \dots \varepsilon_n$ trebaju ispunjavati jest pretpostavka o jednakosti varijanci. O tome zaključujemo na temelju procjena grešaka modela, tj. na temelju reziduala $e_1, \dots, e_n$. Zbog toga se umjesto o zaključivanju o jednakosti varijanci grešaka često govori o analizi homogenosti reziduala. Grafički prikaz reziduala u ovisnosti o predikcijama, tj. dijagram raspršenosti za točke $(\widehat{y}_i, e_i)$, $i = 1, \dots, n$, može pomoći kod uočavanja nehomogenosti reziduala. Ako u tom dijagramu uočavamo sustavno povećanje ili smanjenje raspršenosti vezano uz vrijednosti $\widehat{y}$, to je znak da varijance nisu homogene. Nekoliko ilustrativnih primjera dano je slikama 6.17 i 6.18.



Slika 6.17: Parovi $(\widehat{y}_i, e_i)$ koji sugeriraju homogenost varijanci reziduala.

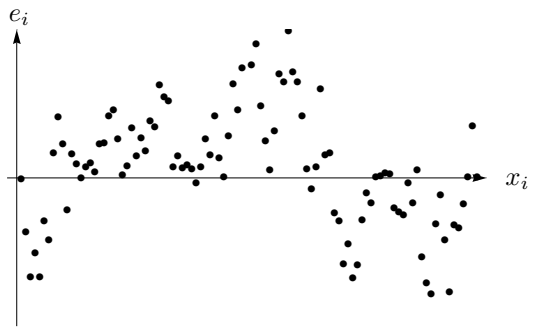


Slika 6.18: Ovakav raspored parova $(\widehat{y}_i, e_i)$ sugerira stalan rast varijance, dakle varijance nisu homogene.

Druga pretpostavka koja se tiče grešaka jest pretpostavka da su slučajne varijable $\varepsilon_1, \ldots \varepsilon_n$ normalno distribuirane s očekivanjem 0 i varijancom σ^2. Normalnost distribucije grešaka možemo provjeriti provođenjem Kolmogorov-Smirnovljeva i Shapiro-Wilkova testa na rezidualima $e_1, \ldots e_n$ te grafički (analizom stupčastog dijagrama reziduala).

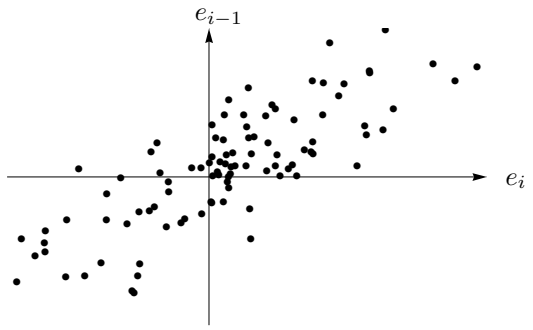
Treća pretpostavka koja se tiče slučajnih varijabli grešaka modela jest pretpostavka o njihovoj nezavisnosti. Zavisnost grešaka može se manifestirati na razne načine. Ovdje navodimo samo dva primjera u kojima je očigledno da treba sumnjati u nezavisnost reziduala, a problem se može uočiti pomoću prikladno izabranog dijagrama raspršenosti.

Prvi slučaj odnosi se na strukturu podataka u dijagramu raspršenosti reziduala u odnosu na vrijednosti neovisne varijable. Slikom 6.19 prikazan je jedan takav slučaj. Uočavamo niz pozitivnih reziduala nakon čega slijedi niz negativnih reziduala, zatim nešto duži niz pozitivnih reziduala, itd.



Slika 6.19: Ovakav raspored parova (x_i, e_i) sugerira međusobnu zavisnost grešaka modela.

Drugi slučaj odnosi se na strukturu podataka u dijagramu raspršenosti parova susjednih reziduala. Pretpostavimo da su podaci numerirani tako da je $x_1 < x_2 < \ldots < x_n$. Slikom 6.20 prikazan je jedan dijagram raspršenosti susjednih reziduala, tj. parova (e_i, e_{i-1}), $i = 2, \ldots, n$. Ovakav dijagram jasno sugerira negativnu vezu između susjednih grešaka modela.



Slika 6.20: Ovakav raspored parova (e_i, e_{i-1}) sugerira međusobnu zavisnost grešaka modela .

Ako nemamo razloga sumnjati u ispravnost pretpostavki modela, možemo ga koristiti za zaključivanje o vezi između neovisne i ovisne varijable. Pri tome su za primjene posebno zanimljivi odgovori na pitanja je li koeficijent smjera pravca različit od nule te koliki je udio varijablinosti ovisne varijable objašnjen modelom, a koliko je dio ostao neobjašnjen.

Zaključivanje o koeficijentu smjera regresijskog pravca

U ovom nas slučaju najviše zanima je li model $Y_i = \alpha + \beta x_i + \varepsilon_i$ bolji od nul-modela $Y_i = \alpha + \varepsilon_i$, tj. modela u kojemu je $\beta = 0$. Potrebno je utvrditi koji od navedena dva modela bolje opisuje promjene u očekivanju slučajnih varijabli Y_i u ovisnosti o vrijednostima x_i. Naime, ako je $\beta = 0$, takav regresijski pravac bio bi paralelan s x-osi pa promjena vrijednosti neovisne varijable ne bi rezultirala promjenom očekivanja ovisne varijable. U svrhu analize možemo koristiti statistički test čije su hipoteze

$$\mathcal{H}_0 : \beta = 0,$$

$$\mathcal{H}_1 : \beta > 0, \quad \text{odnosno} \quad \mathcal{H}_1 : \beta < 0,$$

ovisno o tome je li procjena $\widehat{\beta}$ nepoznatog parametra β pozitivna ili negativna. Ovaj se test temelji na test-statistici čiju vrijednost $\hat{t}$ za eksperimentalne vrijednosti x_i i y_i računamo formulom

$$\hat{t} = \frac{s_x \cdot \widehat{\beta}}{s} \sqrt{n-1},$$

gdje je

$$s_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x}_n)^2}, \quad s = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}}, \tag{6.5}$$

a $\widehat{\beta}$ procjena regresijskog koeficijenta β metodom najmanjih kvadrata. Ako je nul-hipoteza istinita, test-statistika ima Studentovu distribuciju s $(n-2)$ stupnja slobode. Na temelju realizacije $\hat{t}$ test statistike računamo pripadnu p-vrijednost na sljedeći način:

$p = P\{T \geq \hat{t}\}$ ako je alternativna hipoteza oblika $\mathcal{H}_1 : \beta > 0$

$p = P\{T \leq \hat{t}\}$ ako je alternativna hipoteza oblika $\mathcal{H}_1 : \beta < 0$.

Ovdje je T slučajna varijabla koja ima Studentovu distribuciju s $(n-2)$ stupnja slobode. Tako izračunatu p-vrijednost uspoređujemo s nivoom značajnosti α i donosimo odluku kako slijedi:

ako je $p < \alpha$, odbacujemo nul-hipotezu i na razini značajnosti α prihvaćamo alternativnu hipotezu, tj. podaci potvrđuju da se promjene u vrijednosti nezavisne varijable odražavaju na promjene u očekivanju zavisne varijable na nivou značajnosti α

ako je $p > \alpha$, nemamo dovoljno argumenata tvrditi da se promjene u vrijednosti nezavisne varijable odražavaju na promjene u očekivanju zavisne varijable na nivou značajnosti α.

Dio varijablinosti objašnjen modelom

Ovdje se bavimo pitanjem koliki je dio promjena u eksperimentalnim vrijednostima ovisne varijable objašnjen dobivenim modelom. U tu svrhu možemo koristiti broj koji se zove **koeficijent determinacije**.

On se standardno označava s R^2 i definiran je izrazom

$$R^2 = \frac{s_{xy}^2}{s_x^2 s_y^2}, \quad R^2 \in [0, 1].$$

Koeficijent determinacije R^2 daje nam informaciju o tome u kolikoj mjeri je rasipanje eksperimentalnih vrijednosti ovisne varijable objašnjeno linearnom funkcijom $x \mapsto \alpha + \beta x$, a u kolikoj se mjeri radi o tzv. rezidualnom ili neobjašnjenom rasipanju (tu informaciju očitavamo iz broja $(1 - R^2)$).

Velika vrijednost koeficijenta determinacije (slučaj kada je R^2 blizu 1) ukazuje na to da linearan model objašnjava velik dio raspršenosti u eksperimentalnim vrijednostima ovisne varijable, tj. da je samo mali dio ostao neobjašnjen modelom i treba ga pripisati slučajnoj grešci. Modeli kod kojih je R^2 mali nisu informativni za opis varijable Y korištenjem vrijednosti neovisne varijable x jer opisuju samo mali dio varijablnosti u podacima iz Y, dok je veliki dio ostao neobjašnjen modelom.

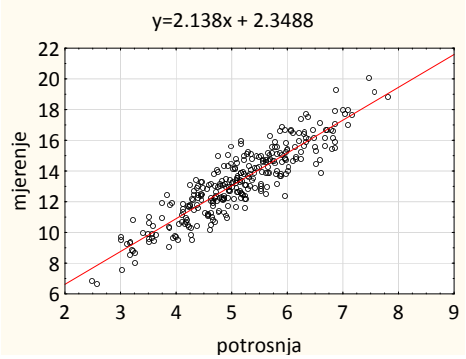
Primjer 6.19. (automobili.sta)
Varijabla potrosnja *baze podataka* automobili.sta *sadrži podatke o potrošnji goriva novog modela automobila pri brzini od* 110 *km/h za* 300 *nezavisnih mjerenja, dok varijabla* mjerenje *sadrži vrijednosti nekog parametra izmjerenog na tehničkom pregledu tog automobila nakon svake od tih vožnji, a za kojeg se pretpostavlja da bi kod tehnički ispravnog automobila trebao biti linearno povezan s prosječnom potrošnjom automobila pri velikim brzinama.*

Stoga ćemo izraditi linearan regresijski model u kojemu je varijabla potrosnja *neovisna varijabla, a varijabla* mjerenje *ovisna varijabla te ispitati njegovu prikladnost za modeliranje veze između spomenutih varijabli. Za početak, promotrimo dijagram raspršenosti vrijednosti varijabli* potrosnja *i* mjerenje *(slika 6.21).*

Sa slike 6.21 vidimo da se parovi vrijednosti varijabli potrosnja *i* mjerenje *grupiraju oko regresijskog pravca* $y = 2.138 \cdot x + 2.3488$. *Cilj je ovog primjera provjeriti je li linearan regresijski model*

$$Y = 2.138 \cdot x + 2.3488 + \varepsilon \tag{6.6}$$

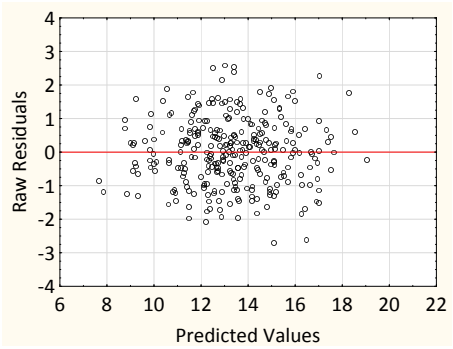
dobar izbor za opisivanje ovisnosti vrijednosti varijable mjerenje *o potrošnji goriva u opisanim uvjetima. U tu svrhu ispitajmo detaljnije svojstva tog modela.*



Slika 6.21: Dijagram raspršenosti vrijednosti varijabli potrosnja i mjerenje.

Analiza reziduala - homogenost varijanci grešaka $\varepsilon_1, \ldots, \varepsilon_n$

O homogenosti varijanci reziduala zaključujemo analizom grafičkog prikaza 6.22 na kojem su prikazani parovi $(\widehat{y}_i, e_i)$ *prediktiranih vrijednosti ovisne varijable i pripadnih reziduala.*

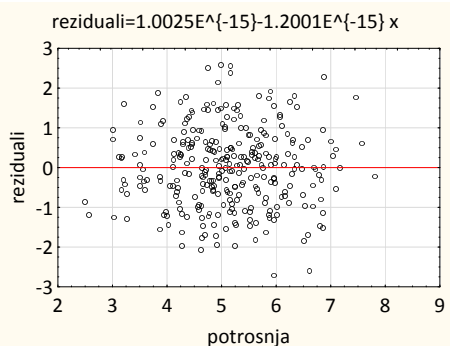


Slika 6.22: Analiza homogenosti varijanci reziduala u modelu 6.6.

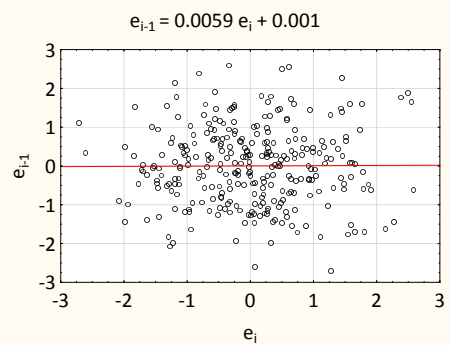
Grafički prikaz 6.22 sugerira homogenost varijanci reziduala.

Analiza reziduala - nezavisnost grešaka $\varepsilon_1, \ldots, \varepsilon_n$

O nezavisnosti grešaka zaključujemo na temelju dijagrama raspršenosti 6.23 reziduala u odnosu na vrijednosti neovisne varijable i dijagrama raspršenja 6.24 susjednih reziduala, tj. parova (e_i, e_{i-1}), $i = 2, \ldots, 300$.



Slika 6.23: Dijagrama raspršenosti reziduala u odnosu na vrijednosti nezavisne varijable u modelu 6.6.

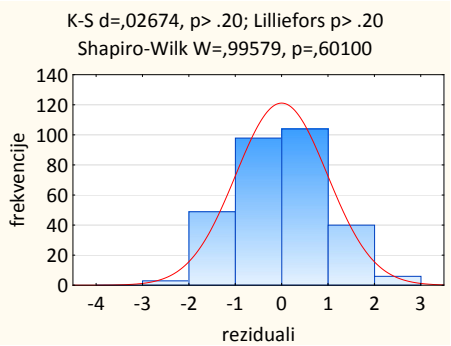


Slika 6.24: Dijagram raspršenosti susjednih reziduala u modelu 6.6.

Dijagrami raspršenosti 6.23 i 6.24 sugeriraju nezavisnost grešaka $\varepsilon_1, \ldots, \varepsilon_n$.

Analiza reziduala - normalnost grešaka $\varepsilon_1, \ldots, \varepsilon_n$

Provođenjem Kolmogorov-Smirnovljeva i Shapiro-Wilksova testa za normalnost slijedi da su pripadne p-vrijednosti (vidi sliku 6.25) u oba slučaja veće od nivoa značajnosti $\alpha = 0.05$. *Dakle, nemamo dovoljno argumenata da bismo odbacili hipotezu o normalnosti grešaka* $\varepsilon_1, \ldots, \varepsilon_n$.



Slika 6.25: Analiza normalnosti grešaka u modelu 6.6.

O varijablinosti objašnjenoj modelom

Iznos koeficijenta determinacije R^2 u programskom paketu Statistica dobivamo u sastavu tablice 6.26.

Dependent Variable	Multiple R	Multiple R2
mjerenje	0,903208	0,815784

Slika 6.26: Koeficijent determinacije linearnog regresijskog modela 6.6.

Iz tablice 6.26 vidimo da je $R^2 \approx 0.816$. To znači da je približno 81.6% rasipanja eksperimentalnih vrijednosti y_i oko procjene regresijskog pravca objašnjeno linearnim regresijskim modelom, a ostatak od 19.4% rasipanja modelom je nebjašnjeno (tzv. rezidualno) rasipanje.

O koeficijentu smjera pravca

Ovom analizom donosimo odluku o tome opisuje li model 6.6 vezu između potrošnje automobila pri velikim brzinama i vrijednosti promatranog parametra bolje od nul-modela, tj. modela u kojem je $\beta = 0$. Budući da je $\widehat{\beta} > 0$, problem se svodi na provođenje statističkog testa čije su hipoteze

$$\mathcal{H}_0 : \beta = 0, \qquad \mathcal{H}_1 : \beta > 0,$$

a pripadna test-statistika T u uvjetima istinitosti nul-hipoteze ima Studentovu T distribuciju s $(n-2)$ stupnja slobode. Vrijednost $\hat{t}$ test statistike T možemo izračunati pomoću formule 6.4.6. Pripadnu p-vrijednost $p = P\{t \geq \hat{t}\}$ tada računamo u kalkulatoru vjerojatnosti i uspoređujemo ga sa zadanim nivoom značajnosti α, npr. $\alpha = 0.05$. Vrijednost $\hat{t}$ i pripadnu p-vrijednost možemo dobiti i u proramskom paketu Statistica (tablica na slici 6.27).

Effect	mjerenje Param.	mjerenje Std.Err	mjerenje t	mjerenje p
Intercept	2,348824	0,306415	7,66549	0,000000
potrosnja	2,137996	0,058854	36,32725	0,000000

Slika 6.27: Vrijednost $\hat{t}$ test statistike i pripadna p-vrijednost t-testa za adekvatnost modela 6.6.

Budući da je $p \approx 0$, pa je manji od zadanog nivoa značajnosti α, slijedi da odbacujemo nul-hipotezu na razini značajnosti α i prihvaćamo alternativnu hipotezu koja kaže da je model 6.6 bolji od nul-modela.

Na temelju provedene analize reziduala, zaključivanja o koeficijentu smjera regresijskog pravca i koeficijenta determinacije zaključujemo da je linearan regresijski model dobar izbor za opisivanje zavisnosti između potrošnje goriva novog modela automobila pri velikim brzinama i vrijednosti promatranog parametra izmjerenog na tehničkom pregledu.

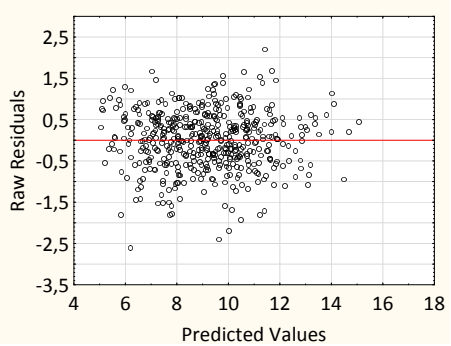
Primjer 6.20. (pozar.sta)
U primjeru 6.18 procijenili smo regresijski pravac između varijabli udaljenost *i* steta*. Cilj je ovog primjera provjeriti je li linearan regresijski model*

$$Y = 2.0224 \cdot x + 4.9275 + \varepsilon \tag{6.7}$$

dobar izbor za opisivanje zavisnosti štete prouzročene požarom o udaljenosti mjesta požara do najbližeg vatrogasnog centra. U tu svrhu napravimo za model 6.7 analizu reziduala.

Analiza reziduala - homogenost varijanci grešaka $\varepsilon_1, \ldots, \varepsilon_n$

O homogenosti varijanci reziduala zaključujemo analizom grafičkog prikaza 6.28 na kojem su prikazani parovi $(\widehat{y}_i, e_i)$ prediktiranih vrijednosti ovisne varijable i pripadnih reziduala.

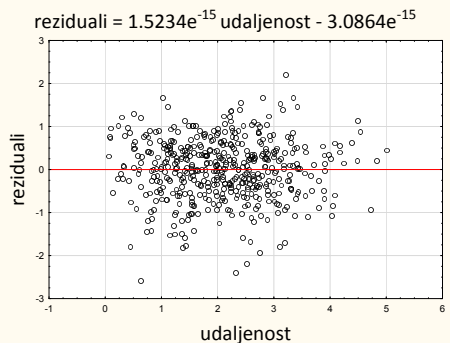


Slika 6.28: Analiza homogenosti varijanci reziduala u modelu 6.7.

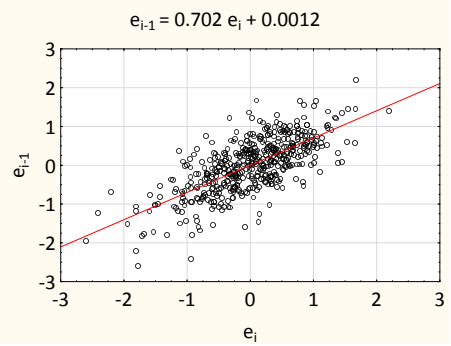
Grafički prikaz 6.28 sugerira homogenost varijanci reziduala.

Analiza reziduala - nezavisnost grešaka $\varepsilon_1, \ldots, \varepsilon_n$

O nezavisnosti grešaka zaključujemo na temelju dijagrama raspršenosti 6.29 reziduala u odnosu na vrijednosti nezavisne varijable i dijagrama raspršenja 6.30 susjednih reziduala, tj. parova (e_i, e_{i-1}), $i = 2, \ldots, 100$.



Slika 6.29: Dijagram raspršenosti reziduala u odnosu na vrijednosti nezavisne varijable u modelu 6.7.



Slika 6.30: Dijagram raspršenosti susjednih reziduala u modelu 6.7.

Dijagram raspršenosti 6.30 ne sugerira nezavisnost grešaka $\varepsilon_1, \ldots, \varepsilon_n$.

Na temelju provedene analize reziduala zaključujemo da linearan regresijski model nije dobar izbor za opisivanje zavisnosti štete prouzročene požarom o udaljenosti mjesta požara do najbližeg vatrogasnog centra.

6.5 Koeficijent korelacije

Koeficijent korelacije jedna je numerička karakteristika dvodimenzionalnog slučajnog vektora koja može poslužiti za analizu zavisnosti među njegovim komponentama.

Neka je (X, Y) dvodimenzionalan slučajni vektor kojemu svaka komponenta ima varijancu. Koeficijent korelacije je broj definiran izrazom:

$$\rho_{XY} = \frac{E(X-\mu)(Y-\nu)}{\sigma_X \sigma_Y},$$

gdje su

$$\mu = EX, \quad \nu = EY, \quad \sigma_X = \sqrt{Var\, X}, \quad \sigma_Y = \sqrt{Var\, Y}.$$

O koeficijentu korelacije valja znati sljedeće činjenice:

- $\rho_{XY} \in [-1, 1]$
- ako su X i Y nezavisne slučajne varijable tada je $\rho_{XY} = 0$
- $Y = aX + b$, gdje je $a > 0$, onda i samo onda ako je $\rho_{XY} = 1$
- $Y = aX + b$, gdje je $a < 0$, onda i samo onda ako je $\rho_{XY} = -1$.

Ako je $\rho_{XY} = 0$, kažemo da su slučajne varijable X i Y nekorelirane.

Navedena svojstva koeficijenta korelacije upućuju na činjenicu da zavisnost između slučajnih varijabli X i Y možemo potvrditi ako pokažemo da je njihov koeficijent korelacije različit od 0. Osim toga, ako je koeficijent korelacije 1 ili -1, onda znamo i tip veze između X i Y, tj. u tim slučajevima ta je veza linearna.

Za procjenu koeficijenta korelacije možemo koristiti nekoliko procjenitelja. Ovdje ćemo spomenuti samo procjenitelja koji se zove Pearsonov korelacijski koeficijent i koristi se kod neprekidnih slučajnih varijabli. Ako su $(x_1, y_1), \ldots, (x_n, y_n)$ parovi nezavisnih realizacija slučajnog vektora (X, Y), onda se iznos Pearsonova korelacijskog koeficijenta računa pomoću izraza

$$r = \frac{\sum_{i=1}^{n} (x_i - \overline{x}_n)(y_i - \overline{y}_n)}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x}_n)^2} \sqrt{\sum_{i=1}^{n} (y_i - \overline{y}_n)^2}}.$$

Ako usporedimo izraz za Pearsonov korelacijski koeficijent r s izrazima navedenim u poglavlju o linearnoj regresiji, možemo uočiti da je r^2 jednak koeficijentu determinacije za koji je rečeno da mjeri jakost linearne veze između varijabli u linearnom regresijskom modelu.

Da bismo korištenjem procjene koeficijenta korelacije potvrdili zavisnost slučajnih varijabli, potrebno je odbaciti statističku hipotezu

$$\mathcal{H}_0 : \rho_{XY} = 0.$$

Ovdje navodimo jedan od testova koji se može koristiti u tu svrhu. On je kreiran pod pretpostavkom normalnosti distribucije slučajnog vektora (X, Y) korištenjem Pearsonova korelacijskog koeficijenta. Za testiranje navedene nul-hipoteze računamo vrijednost test statistike po formuli:

$$\hat{t} = \frac{\sqrt{n-2}\, r}{\sqrt{1-r^2}}.$$

Ako je nul-hipoteza istinita, statistika kojoj smo tako izračunali realizaciju ima Studentovu distribuciju s $(n-1)$ stupnjeva slobode. Označimo li s T slučajnu varijablu koja ima Studentovu distribuciju s $(n-1)$ stupnjeva slobode, pripadnu p-vrijednost određujemo na uobičajeni način:

$p = P\{T \geq t\}$ ako je alternativna hipoteza oblika $\mathcal{H}_1 : \rho_{XY} > 0$

$p = P\{T \leq t\}$ ako je alternativna hipoteza oblika $\mathcal{H}_1 : \rho_{XY} < 0$.

Tako izračunatu p-vrijednost uspoređujemo s nivoom značajnosti α i donosimo odluku:

ako je $p < \alpha$, odbacujemo nul-hipotezu i na razini značajnosti α prihvaćamo alternativnu hipotezu, tj. kažemo da su slučajne varijable X i Y zavisne

ako je $p > \alpha$, nemamo dovoljno argumenata koji bi poduprli odluku o odbacivanju nul-hipoteze, tj. kažemo da nemamo dovoljno argumenata tvrditi da su X i Y zavisne varijable.

Primjer 6.21. (pozar.sta)
Vratimo se primjeru 6.16. Označimo s X slučajnu varijablu kojom modeliramo udaljenost mjesta požara do najbližeg vatrogasnog centra, a Y slučajnu varijablu kojom modeliramo štetu nastalu požarom. Budući da ne znamo stvarne distribucije slučajnih varijabli X i Y, ne možemo izračunati koeficijent korelacije ρ_{XY}. No na temelju podataka sadržanih u bazi pozar.sta *možemo, koristeći*

Pearsonov korelacijski koeficijent, procijeniti koeficijent korelacije slučajnih varijabli X i Y. U tablici na slici 6.31 prikazane su procjene očekivanja i varijanci slučajnih varijabli X i Y, njihov koeficijent korelacije te vrijednost test statistike i pripadna p-vrijednost statističkog testa kojim se testira hipoteza

$$\mathcal{H}_0 : \rho_{XY} = 0.$$

	Correlations (pozar.sta) Marked correlations are significant at p < .05000					
Var. X & Var. Y	Mean	Std.Dv.	r(X,Y)	r2	t	p
udaljenost	2.080108	0.956215				
steta	9.134346	2.058874	0.939286	0.882257	60.47009	0.00

Slika 6.31: Procjena koeficijenta korelacije slučajnih varijabli X i Y iz primjera 6.21.

Procjena koeficijenta korelacije ρ_{XY} je

$$r \approx 0.94 > 0.$$

p-vrijednost testa kojim se testira hipoteza o nekoreliranosti slučajnih varijabli X i Y manja je od 0.01 *pa to znači da odbacujemo hipotezu o nekoreliranosti slučajnih varijabli X i Y i na nivou značajnosti $\alpha = 0.01$ prihvaćamo alternativnu hipotezu koja kaže da su X i Y zavisne varijable. Uočimo da se u tablici nalazi i vrijednost kvadrata procjene koeficijenta korealcije (oznaka r2) - to je upravo koeficijent determinacije R^2.*

6.6 Zadaci

Zadatak 6.1. Možete li u sljedećim zadacima na nivou značajnosti $\alpha = 0.05$ potvrditi da su varijance σ_1^2 i σ_2^2 različite (varijance su procijenjene s $s_{n_1}^2$ i $s_{n_2}^2$)?

a) $s_{n_1} = 1989$, $n_1 = 50$, $s_{n_2} = 1843$, $n_2 = 30$, $\alpha = 0.05$.

b) $s_{n_1} = 250$, $n_1 = 20$, $s_{n_2} = 300$, $n_2 = 16$, $\alpha = 0.05$.

Rješenje. *U oba zadatka nemamo dovoljno argumenata koji bi poduprli odluku o odbacivanju nul-hipoteze da su varijance jednake pa ne možemo potvrditi da su varijance različite.*

Zadatak 6.2. Ekonomisti u nekoj zemlji odlučili su provjeriti jesu li očekivane cijene u eurima uvoznih automobila više u njihovoj zemlji nego u matičnoj zemlji određenog proizvođača. Prikupljen je uzorak od 50 cijena u promatranoj zemlji i 30 cijena u matičnoj zemlji za isto razdoblje. Na temelju tih uzoraka procijenjena očekivanja i standardne devijacije slučajnih varijabli kojima se modelira cijena tog tipa automobila su:

promatrana zemlja	$n_1 = 50$, $\overline{x}_{n_1} = 16545$, $s_{n_1} = 1989$
matična zemlja proizvođača	$n_2 = 30$, $\overline{x}_{n_2} = 17243$, $s_{n_2} = 1843$.

Možemo li na nivou značajnosti $\alpha = 0.05$ potvrditi postojanje razlika u očekivanim cijenama automobila?

Rješenje. $p = 0.0613$ *pa na nivou značajnosti* $\alpha = 0.05$ *nemamo dovoljno argumenata koji bi poduprli odluku o odbacivanju nul-hipoteze (jednakost očekivanja!) pa ne možemo potvrditi postojanje razlika u očekivanim cijenama automobila.*

Zadatak 6.3. Menadžment jednog velikog medicinskog centra želi provjeriti postoji li razlika u očekivanoj godišnjoj neto-plaći između bolničarki i bolničara. Na temelju uzoraka bolničarki i bolničara procijenjena su očekivanja i standardne devijacije slučajnih varijabli kojima se modeliraju njihove plaće:

$$\begin{array}{ll} \text{bolničarke:} & n_1 = 20,\ \overline{x}_{n_1} = 23750,\ s_{n_1} = 250 \\ \text{bolničari:} & n_2 = 16,\ \overline{x}_{n_2} = 23800,\ s_{n_2} = 300. \end{array}$$

Možemo li na nivou značajnosti $\alpha = 0.05$ potvrditi postojanje razlika u očekivanim plaćama bolničarki i bolničara pod uvjetima da su zadovoljene pretpostavke o jednakosti varijanci i o normalnoj distribuiranosti slučajnih varijabli kojima modeliramo njihove plaće?

Rješenje. $p = 0.2944$ *pa na nivou značajnosti* $\alpha = 0.05$ *nemamo dovoljno argumenata koji bi poduprli odluku o odbacivanju nul-hipoteze pa ne možemo potvrditi postojanje razlika u očekivanim plaćama.*

Zadatak 6.4. (student.sta)
Studentska služba jednog sveučilišta želi vidjeti postoji li razlika u očekivanoj dobi među studentima koji studiraju na klasičan način i studenata koji studiraju putem interneta (e-learning). Prikupljeni podaci o dobi za 50 studenata iz svake kategorije nalaze se u bazi student.sta. Možemo li na nivou značajnosti $\alpha = 0.05$ potvrditi postojanje razlika u očekivanoj dobi studenata?

Rješenje. $p = 0.214$ *pa na nivou značajnosti* $\alpha = 0.05$ *nemamo dovoljno argumenata koji bi poduprli odluku o odbacivanju nul-hipoteze.*

Zadatak 6.5. (burza.sta)
U bazi podataka burza.sta zabilježene su cijene nekih dionica na dvije burze smještene u dva različita grada - gradu A i gradu B. U jednom financijskom časopisu pročitali smo da je očekivana cijena dionice viša na burzi u gradu A u odnosu na očekivanu cijenu na burzi u gradu B. Možemo li na nivou značajnosti $\alpha = 0.05$ potvrditi postojanje razlika u očekivanoj cijeni dionice na promatranim burzama?

Rješenje. $p = 0.0067$ *pa na nivou značajnosti* $\alpha = 0.05$ *odbacujemo nul-hipotezu i možemo tvrditi da su očekivane cijene dionice na burzi u gradu A statistički značajno različite od očekivane cijene na burzi u gradu B.*

Zadatak 6.6. (manager.sta)
Jedna grupa istraživača razvila je indeks koji mjeri uspjeh menadžera, pri čemu veći indeks sugerira veću uspješnost menadžera. Neki istraživač želi usporediti taj indeks za dvije grupe menadžera. Jedna grupa menadžera ima mnogo interakcija s ljudima izvan svog radnog okruženja (telefoniranje, razgovori, sastanci i sl.), dok druga grupa ima vrlo rijetke kontakte izvan svog okruženja. U bazi podataka manager.sta nalaze se indeksi za uzorak menadžera iz grupe koja ima mnogo interakcija (varijabla mnogo interakcija) i indeksi za uzorak menadžera iz grupe koja ima malo interakcija (varijabla malo interakcija). Možemo li na nivou značajnosti $\alpha = 0.05$ potvrditi postojanje razlika u očekivanim indeksima uspješnosti menadžera iz te dvije grupe pod uvjetima da su zadovoljene pretpostavke o jednakosti varijanci i o normalnoj distribuiranosti slučajnih varijabli kojima modeliramo indekse?

Rješenje. $p = 0$ *pa na nivou značajnosti* $\alpha = 0.05$ *odbacujemo nul-hipotezu te tvrdimo da su očekivani indeksi uspješnosti za te dvije grupe menadžera statistički značjno različiti.*

Zadatak 6.7. (potrosac.sta)
Marketinški stratezi željeli bi predvidjeti prijem nove vrste paste za zube kod potrošača prema njihovoj dobi. U bazi podataka potrosac.sta raspolažemo podacima o dobi u godinama za 20 potrošača koji su kupili novu pastu za zube (varijabla korisnici) i 20 potrošača koji ju još uvijek nisu kupili (varijabla nisu korisnici). Možemo li na nivou značajnosti $\alpha = 0.01$ potvrditi postojanje razlike u očekivanoj dobi potrošača iz te dvije grupe pod pretpostavkom da normalna distribucija dobro opisuje distribuciju slučajnih varijabli kojima modeliramo njihovu dob te su varijance jednake?

Rješenje. $p = 0.0296$ *pa na nivou značajnosti* $\alpha = 0.01$ *odbacujemo nul-hipotezu i potvrđemo različitost očekivane dobi potrošača koji su kupili i onih koji još uvijek nisu kupili novu vrstu paste za zube.*

Zadatak 6.8. (restorani.sta)
Pretpostavimo da je neki veliki lanac restorana uložio puno novca u reklamu te menadžer želi usporediti očekivanu dnevnu zaradu tog lanca restorana prije i nakon tog ulaganja. U bazi podataka restorani.sta nalaze se informacije o dnevnoj zaradi za 22 restorana prije ulaganja u marketing (varijabla prije) i nakon ulaganja u marketing (varijabla poslije). Možemo li na nivou značajnosti $\alpha = 0.05$ potvrditi postojanje razlike u očekivanoj dnevnoj zaradi tog lanca restorana prije i nakon ulaganja u marketing pod pretpostavkom da razlike dnevnih zarada prije i nakon ulaganja u reklamu možemo modelirati normalnom slučajnom varijablom?

Rješenje. $p = 0.005845$ *pa na nivou značajnosti* $\alpha = 0.05$ *odbacujemo nul-hipotezu i potvrđujemo različitost očekivane dnevne zarade lanca restorana prije i nakon ulaganja u marketing.*

Zadatak 6.9. (vitamini.sta)
Jedan liječnik tvrdi da se uzimanjem specijalnog vitamina može povećati snaga dizača utega. Kako bi se provjerila njegova tvrdnja odabrano je osam dizača utega kojima je izmjerena snaga. Nakon dva tjedna treninga podržanih upotrebom specijalnog vitamina ti isti dizači utega su opet testirani i dobiveni su sljedeći rezultati:

prije tretmana:	210, 230, 182, 205, 262, 253, 219, 216
poslije tretmana:	219, 236, 179, 204, 270, 250, 222, 216.

Možemo li na nivou značajnosti $\alpha = 0.05$ potvrditi postojanje razlike u očekivanoj snazi dizača utega prije i nakon tretmana specijalnim vitaminima pod pretpostavkom da razliku izmjerene snage prije i nakon tretmana možemo modelirati normalnom slučajnom varijablom?

Rješenje. $p = 0.43$ *pa na nivou značajnosti* $\alpha = 0.05$ *nemamo dovoljno argumenata za odbacivanje nul-hipoteze o jednakosti očekivanja.*

Zadatak 6.10. U sklopu studije organizacije rada poduzeća ispituje se efikasnost zaposlenika u proizvodnom procesu. Ispitivanje se provodi mjerenjem produktivnosti rada na uzorku radnika. Radi mogućeg povećanja produktivnosti na radnim mjestima radnika u uzorku izmijenjen je red radnih operacija i prostorni razmještaj sredstava rada. Rezultati mjerenja produktivnosti rada prije i nakon izmjena dani su u sljedećoj tablici:

prije izmjena:	45, 34, 42, 28, 35, 39, 50, 41, 27, 29,
poslije izmjena:	49, 40, 43, 32, 40, 39, 51, 42, 30, 24.

Možemo li na nivou značajnosti $\alpha = 0.05$ potvrditi postojanje razlike u očekivanoj produktivnosti radnika prije i nakon uvedenih izmjena pod pretpostavkom da razliku izmjerene produktivnosti prije i nakon izmjena možemo modelirati normalnom slučajnom varijablom?

Rješenje. $p = 0.077$ *pa na nivou značajnosti* $\alpha = 0.05$ *nemamo dovoljno argumenata za odbacivanje nul-hipoteze o jednakosti očekivanja.*

Zadatak 6.11. (gradjevina.sta)
Varijable placa2008 i placa2009 u bazi podataka gradjevina.sta sadrže prosječne neto-plaće u eurima u 2008. i 2009. godini za 100 građevinskih poduzeća srednje veličine u nekoj zemlji. Možemo li na nivou značajnosti $\alpha = 0.05$ prihvatiti hipotezu o postojanju razlike u očekivanoj prosječnoj plaći u građevinskim poduzećima srednje veličine u toj zemlji u 2008. i 2009. godini pod pretpostavkom da razlike prosječnih plaća u 2008. i 2009. godini možemo modelirati normalnom slučajnom varijablom?

Rješenje. $p = 0.164$ *pa na nivou značajnosti* $\alpha = 0.05$ *nemamo dovoljno argumenata za odbacivanje nul-hipoteze o jednakosti.*

Zadatak 6.12. Istraživač proučava uzorke dvaju tipova automobila koji pripadaju istoj klasi, ali potječu od različitih proizvođača. Na temelju uzorka koji broji 400 automobila prvog proizvođača utvrdio je da se 53 automobila pokvarilo tijekom prve godine korištenja, dok je na temelju uzorka od 500 automobila drugog proizvođača utvrdio da ih se pokvarilo čak 78. Možemo li na nivou značajnosti $\alpha = 0.05$ tvrditi da je vjerojatnije da će se tijekom prve godine korištenja pokvariti automobil drugog proizvođača nego automobil prvog proizvođača?

Rješenje. $p = 0.16$ *pa na nivou značajnosti* $\alpha = 0.05$ *nemamo dovoljno argumenata za odbacivanje nul-hipoteze o jednakosti vjerojatnosti.*

Zadatak 6.13. Raspolažete sljedećim podacima:

	menadžeri	MBA studenti
veličina uzorka	162	109
proporcija muškaraca	0.95	0.689
proporcija ljudi u braku	0.912	0.534

a) Možemo li na nivou značajnosti $\alpha = 0.01$ tvrditi da je proporcija muškaraca među menadžerima veća nego među MBA studentima?

b) Možemo li na nivou značajnosti $\alpha = 0.01$ tvrditi da je proporcija menadžera koji su u braku veća nego proporcija studenata koji su u braku?

Rješenje. *U oba zadatka je* $p < 0.00001$ *pa odbacujemo nul-hipotezu i na nivou značajnosti* $\alpha = 0.01$ *potvrđujemo alternativnu hipotezu.*

Zadatak 6.14. Financijski analitičar ispituje proporciju tekućih računa s negativnim saldom većim od dopuštenog u prosincu u dvije poslovnice jedne banke. Njegova je pretpostavka da je proporcija takvih računa u poslovnici II manja nego u poslovnici I. U uzorku koji broji 562 računa poslovnice I 75 ih je s nedopuštenim prekoračenjem, a u uzorku koji broji 462 računa poslovnice II 44 ih je s nedopuštenim prekoračenjem. Možete li na razini značajnosti $\alpha = 0.05$ potvrditi pretpostavku financijskog analitičara?

Rješenje. $p = 0.029$ *pa odbacujemo nul-hipotezu i na nivou značajnosti* $\alpha = 0.05$ *potvrđujemo hipotezu financijskog analitičara.*

Zadatak 6.15. (gradjevina.sta)
Varijable zaposleni2008 i zaposleni2009 u bazi podataka gradjevina.sta sadrže broj zaposlenika u 2008. i 2009. godini za 100 slučajno izabranih građevinskih poduzeća srednje veličine u nekoj zemlji. Možete li na razini značajnosti $\alpha = 0.05$ potvrditi hipotezu koja kaže da je proporcija takvih poduzeća koja zapošljavaju više od 150 radnika veća u 2009. godini nego u 2008. godini?

Rješenje. $p = 0.4245$ *pa na nivou značajnosti* $\alpha = 0.05$ *nemamo dovoljno argumenata za odbacivanje nul-hipoteze. Dakle, na toj razini značajnosti ne možemo potvrditi hipotezu navedenu u zadatku.*

Zadatak 6.16. Klub ljubitelja rock-glazbe želi provjeriti postoji li razlika u proporcijama članova mlađih od 30 godina i onih starijih od 30 godina koji osim rocka vole i klasičnu glazbu. U svrhu ovog istraživanja ispitano je 56 članova mlađih od 30 i 65 članova starih barem 30 godina. Od ukupnog ispitanog broja klasiku voli slušati 14 članova mlađih od 30 i 15 članova starih barem 30 godina. Možete li na razini značajnosti $\alpha = 0.05$ potvrditi hipotezu koja kaže da se navedene proporcije razlikuju?

Rješenje. $p = 0.7975$ *pa na nivou značajnosti* $\alpha = 0.05$ *nemamo dovoljno argumenata za odbacivanje nul-hipoteze o jednakosti.*

Zadatak 6.17. (djeca.sta)
Varijablom apgar1 dana je jedna ocjena vitalnosti novorođenčeta odmah nakon poroda, dok je varijablom komplikacije dana informacija o tome je li tijekom trudnoće bilo komplikacija ili ne. Označimo s X slučajnu varijablu kojom modeliramo ocjenu vitalnosti apgar, a Y slučajnu varijablu kojom modeliramo prisutnost komplikacija u trudnoći. Odredite empirijsku distribuciju slučajnog vektora (X, Y) i njegove marginalne empirijske distribucije te riješite sljedeće zadatke:

a) procijenite vjerojatnost da je ocjena vitalnosti 1 i da su komplikacije bile prisutne

b) procijenite vjerojatnost da je ocjena vitalnosti 4 i da su komplikacije bile prisutne

c) procijenite vjerojatnost da je ocjena vitalnosti 4 i da komplikacije nisu bile prisutne

d) procijenite vjerojatnost pojave komplikacija u trudnoći u promatranoj populaciji

e) procijenite vjerojatnost pojave ocjene vitalnosti 4 u promatranoj populaciji novorođenčadi.

Zadatak 6.18. (citanje.sta)
Baza podataka citanje.sta, koja sadrži rezultate istraživanja o čitalačkim navikama stanovnika jednog grada, opisana je u primjeru 6.12.

a) Procijenite distribuciju slučajnog vektora (X, Y), gdje X označava slučajnu varijablu koja se realizira jedinicom ako stanovnik tog grada svaka tri mjeseca pročita barem jednu knjigu, a inače se realizira nulom, a Y slučajnu varijablu kojom modeliramo stručnu spremu stanovnika toga grada (za NSS Y se realizira jedinicom, za SSS dvojkom, a za VSS trojkom).

b) Pretpostavite da empirijska distribucija slučajnog vektora (X, Y) odgovara njegovoj stvarnoj distribuciji te procijenite sljedeće vjerojatnosti:

- vjerojatnost da slučajno odabrani ispitanik iz populacije koju promatramo u tom gradu svaka tri mjeseca pročita barem jednu knjigu i ima srednju stručnu spremu, tj. vjerojatnost $P\{X = 1, Y = 2\}$
- vjerojatnost da slučajno odabrani ispitanik iz populacije koju promatramo u tom gradu svaka tri mjeseca pročita barem jednu knjigu, tj. vjerojatnosti $P\{X = 1\}$
- vjerojatnost da slučajno odabrani ispitanik iz populacije koju promatramo u tom gradu ima srednju stručnu spremu, tj. vjerojatnosti $P\{Y = 2\}$.

Rješenje.

a) Empirijska distribucija slučajnog vektora (X, Y) dana je tablicom 6.32.

	Summary Frequency Table (citanje.STA) Table: citanje(2) x obrazovanje(3)				
	citanje	obrazovanje NSS	obrazovanje SSS	obrazovanje VSS	Row Totals
Count	0	48	426	184	658
Total Percent		6.45%	57.26%	24.73%	88.44%
Count	1	16	51	19	86
Total Percent		2.15%	6.85%	2.55%	11.56%
Count	All Grps	64	477	203	744
Total Percent		8.60%	64.11%	27.28%	

Slika 6.32: Empirijska distribucija slučajnog vektora (X, Y) iz primjera **??**.

b) Procjena vjerojatnosti $P\{X = 1, Y = 2\}$ *iznosi* 0.0685*, procjena vjerojatnosti* $P\{X = 1\}$ *iznosi* 0.1156*, procjena vjerojatnosti* $P\{Y = 2\}$ *iznosi* 0.6411*.*

Zadatak 6.19. (planovi.sta)
U bazi podataka planovi.sta nalaze se podaci o dobi (varijabla dob), spolu (varijabla spol: 1 - muškarac, 2 - žena) i planovima za posao nakon diplomiranja (varijabla poslovni plan: 1 - raditi puno radno vrijeme, 2 - raditi pola radnog vremena, 3 - uopće ne raditi) za uzorak od 129 studenata jednog sveučilišta. Zanima nas postoji li razlika u planovima za posao s obzirom na spol ispitanika. Možete li na razini značajnosti $\alpha = 0.1$ potvrditi zavisnost slučajnih varijabli kojima modeliramo spol ispitanika i planove za posao nakon diplomiranja?

Rješenje. *Dobivena p-vrijednost manja je od nivoa značajnosti* $\alpha = 0.1$ *pa zaključujemo da odbacujemo nul-hipotezu i na nivou značajnosti* $\alpha = 0.1$ *možemo reći da podaci potvrđuju postojanje zavisnosti između slučajnih varijabli kojima modeliramo spol ispitanika i planove za posao nakon diplomiranja.*

Zadatak 6.20. U primjerima 6.10 i 6.11 testirajte hipotezu o nezavisnosti.

Zadatak 6.21. U primjeru 6.10 procijenite svih pet uvjetnih distribucija za Y uz uvjet da se dogodi $\{X = i\}$, $i = 0, 1, 2, 3, 4$. Mijenjaju li se te distribucije promjenom događaja na koji uvjetujemo? Možete li to objasniti i povezati s pojmom zavisnosti i nezavisnosti slučajnih varijabli X i Y?

Zadatak 6.22. (krv.sta)
U bazi podataka krv.sta nalaze se podaci o mjerenim vrijednostima nekoliko različitih analiza krvi u definiranoj populaciji bolesnih osoba. Analitičar želi istražiti može li se odrediti veza između izmjerenih vrijednosti ovih analiza. Utvrđivanje veze i jasno uspostavljanje zakona koji ih povezuje smanjilo bi broj potrebnih pretraga krvi. Naime, trebalo bi napraviti samo one koje su međusobno neovisne, dok bi se ostale mogle na osnovi njih prognozirati. Za podatke iz baze prikažite svake dvije varijable u dijagramu raspršenja i kratko ga analizirajte.

Zadatak 6.23. Skicirajte grafove funkcija

$$f(x) = 2x - 1, \quad f(x) = \frac{1}{2}x + 3, \quad f(x) = -2x$$

i komentirajte značenje koeficijenata α i β. Koji koeficijent opisuje iznos povećanja vrijednosti ovisne varijable za jediničnо povećanje vrijednosti neovisne varijable?

Zadatak 6.24. (krv.sta, regresija.sta)

a) Koristeći bazu podataka krv.sta procijenite regresijski pravac između varijabli CD4 i CD8. Odredite vrijednosti reziduala. Ponovite postupak za još nekoliko parova varijabli.

b) Koristeći bazu podataka regresija.sta procijenite regresijski pravac između varijabli x1 i x2. Odredite vrijednosti reziduala i prokomentirajte dobiveni rezultat.

Zadatak 6.25. (statistika.sta)
Mnogi studenti odlučili su ispit iz Statistike položiti putem kolokvija. Pri tome se postignuti bodovi na sva četiri kolokvija zbrajaju i na temelju zbroja bodova donosi se odluka o tome ima li student pravo izaći na usmeni dio ispita. U bazi podataka statistka.sta nalazi se zbroj bodova prva dva kolokvija (varijabla kol-1-2) i ukupan broj bodova nakon svih provedenih kolokvija (varijabla ukupno). Koju ćete od ovih varijabli promatrati kao neovisnu, a koju kao ovisnu varijablu? Odredite procjenu regresijskog pravaca te odgovorite na sljedeća pitanja:

- Što o linearnom regresijskom modelu možete reći na temelju analize reziduala?
- Što o linearnom regresijskom modelu možete reći na temelju koeficijenta smjera procijenjenog regresijskog pravca?
- Koliki je dio promjena u izmjerenim vrijednostima ovisne varijable objašnjen linearnim regresijskim modelom?

Zadatak 6.26. (ptsp.sta)
Baza podataka ptsp.sta sadrži podatke o ispitanicima kojima je dijagnosticiran posttraumatski stresni poremećaj. Na primjer, varijabla ptspb2 sadrži rezultate testova nakon terapije nekim lijekom, a varijabla ptspb odražava stanje prije provedene terapije. Koju ćete od ovih varijabli promatrati kao neovisnu, a koju kao ovisnu varijablu? Odredite procjenu regresijskog pravaca te odgovorite na sljedeća pitanja:

- Što o linearnom regresijskom modelu možete reći na temelju analize reziduala?
- Što o linearnom regresijskom modelu možete reći na temelju koeficijenta smjera procijenjenog regresijskog pravca?
- Koliki je dio promjena u eksperimentalnim vrijednostima ovisne varijable objašnjen linearnim regresijskim modelom?

Analogno napravite za parove varijabli ptspc i ptspc2, te ptspd i ptspd2.

Zadatak 6.27. (gradjevina.sta)
Varijable godisnja placa2009 i troskovi2009 u bazi podataka gradjevina.sta sadrže podatke o prosječnoj godišnjoj plaći zaposlenika i ukupnim troškovima u 2009. godini za 100 građevinskih poduzeća srednje veličine u nekoj zemlji. Ako znamo da se plaće zaposlenika uračunavaju u ukupne troškove poduzeća, koju ćete od ovih varijabli promatrati kao neovisnu, a koju kao ovisnu varijablu? Odredite procjenu regresijskog pravaca te odgovorite na sljedeća pitanja:

- Što o linearnom regresijskom modelu možete reći na temelju analize reziduala?
- Što o linearnom regresijskom modelu možete reći na temelju koeficijenta smjera procijenjenog regresijskog pravca?
- Koliki je dio promjena u eksperimentalnim vrijednostima ovisne varijable objašnjen linearnim regresijskim modelom?

Zadatak 6.28. (gradjevina.sta)
Koristeći bazu podataka gradjevina.sta procijenite koeficijent korelacije za varijable godisnja placa2009 i troskovi2009. Rezultat usporedite s rezultatima regresijske analize za isti par varijabli.

Zadatak 6.29. (krv.sta, regresija.sta)
Koristeći baze podataka krv.sta i regresija.sta procijenite koeficijent korelacije za sve parove varijabli. Rezultat usporedite s rezultatima regresijske analize za iste parove varijabli.

Zadatak 6.30. (regresija1.sta, regresija2.sta, regresija3.sta, regresija4.sta)
Koristeći baze regresija1.sta, regresija2.sta, regresija3.sta i regresija4.sta procijenite regresijski pravac između varijabli tih baza podataka. Što uočavate? Možete li na ovoj razini donijeti grubu ocjenu o primjerenosti korištenja linearnog modela za opisivanje veze među ovim varijablama? Koji bi model bio prikladniji i zašto?

Zadatak 6.31. (gorivo.sta)
U bazi podataka gorivo.sta varijabla udaljenost sadrži podatke o udaljenosti radnog mjesta od mjesta stanovanja za 100 slučajno odabranih zaposlenika jednog poduzeća, a varijabla troskovi iznos u kunama koji ti zaposlenici troše na gorivo da bi se dovezli do posla. Procijenite regresijski pravac između varijabli udaljenost i troskovi te odgovorite na sljedeća pitanja:

- Što o linearnom regresijskom modelu možete reći na temelju analize reziduala?
- Što o linearnom regresijskom modelu možete reći na temelju koeficijenta smjera procijenjenog regresijskog pravca?
- Koliki je dio promjena u eksperimentalnim vrijednostima ovisne varijable objašnjen linearnim regresijskim modelom?

Zadatak 6.32. (glukoza.sta)
Koristeći bazu podataka glukoza.sta, čije su varijable opisane u primjeru 2.2, procijenite regresijski pravac između varijabli dob i koncentracija te odgovorite na sljedeća pitanja:

- Što o linearnom regresijskom modelu možete reći na temelju analize reziduala?
- Što o linearnom regresijskom modelu možete reći na temelju koeficijenta smjera procijenjenog regresijskog pravca?
- Koliki je dio promjena u eksperimentalnim vrijednostima ovisne varijable objašnjen linearnim regresijskim modelom?

Zadatak 6.33. (apartmani.sta)
U bazi podataka apartmani.sta varijabla udaljenost sadrži podatke o udaljenosti apartmana do najbliže plaže za 100 slučajno izabranih apartmana u nekom turističkom mjestu, a varijabla cijena cijenu apartmana po danu izraženu u kunama. Procijenite regresijski pravac između varijabli udaljenost i cijena te odgovorite na sljedeća pitanja:

- Što o linearnom regresijskom modelu možete reći na temelju analize reziduala?
- Što o linearnom regresijskom modelu možete reći na temelju koeficijenta smjera procijenjenog regresijskog pravca?
- Koliki je dio promjena u eksperimentalnim vrijednostima ovisne varijable objašnjen linearnim regresijskim modelom?

Zadatak 6.34. (servis.sta)
U bazi podataka servis.sta varijabla broj km sadrži podatke o prijeđenom broju kilometara za 100 automobila istog tipa prije obavljenog prvog servisa, a varijabla servis kn cijenu servisa nakon tog broja kilometara. Procijenite regresijski pravac između varijabli broj km i servis kn te odgovorite na sljedeća pitanja:

- Što o linearnom regresijskom modelu možete reći na temelju analize reziduala?
- Što o linearnom regresijskom modelu možete reći na temelju koeficijenta smjera procijenjenog regresijskog pravca?
- Koliki je dio promjena u eksperimentalnim vrijednostima ovisne varijable objašnjen linearnim regresijskim modelom?

Poglavlje 7

Zadaci za vježbu

Zadatak 7.1. U razredu koji broji 25 učenika zaključne ocjene iz matematike na kraju školske godine raspodijeljene su na sljedeći način: tri učenika imaju peticu, sedam učenika četvorku, osam učenika trojku, pet učenika dvojku, a dva učenika moraju pristupiti popravnom ispitu (imaju jedinicu). Ocjene učenika sadržane su u varijabli ocjena baze podataka razred.sta. Sljedeće zadatke riješite samostalno te rezultate provjerite korištenjem programskog paketa Statistica.

1. Sastavite tablicu frekvencija i relativnih frekvencija za varijablu ocjena.
2. Koristeći Statisticu grafički prikažite frekvencije i relativne frekvencije (stupčastim i kružnim dijagramima).
3. Izračunajte aritmetičku sredinu, mod, raspon te varijancu i standardnu devijaciju ovog skupa podataka.
4. Izračunajte numeričke karakteristike ovog skupa podataka koje su vam potrebne za kutijasti dijagram na bazi medijana te ga nacrtajte.

Zadatak 7.2. (desno.sta)
Baza podataka desno.sta sadrži dio podataka iz istraživanja kojim se proučava učestalost korištenja desne ruke u skupini dešnjaka, ljevaka i ambidekstera jedne populacije. Varijabla sum sadrži ocjenu učestalosti korištenja desne ruke u deset izabranih radnji i može primiti vrijednosti od 0 do 30. Varijabla objektivno sadrži informaciju o tome je li osoba dešnjak, ljevak ili ambidekster. Sve opisane varijable možemo modelirati diskretnim slučajnim varijablama koje primaju vrijednosti iz prikladno konstruiranih skupova - odredite te skupove, tj. slike tih slučajnih varijabli. Uz pretpostavku o jednakosti stvarnih i empirijskih distribucija tih slučajnih varijabli riješite sljedeće zadatke.

1. Procijenite vjerojatnost da slučajnim izborom osobe iz ove populacije odaberemo dešnjaka.
2. Procijenite vjerojatnost da slučajnim izborom osobe iz ove populacije odaberemo ljevaka.
3. Procijenite vjerojatnost da slučajnim izborom osobe iz ove populacije odaberemo osobu čija je učestalost korištenja desne ruke manja ili jednaka 10.
4. Procijenite vjerojatnost da slučajnim izborom osobe iz ove populacije odaberemo osobu čija je učestalost korištenja desne ruke barem 10.

5. Procijenite vjerojatnost da slučajnim izborom osobe iz ove populacije odaberemo osobu čija učestalost korištenja desne ruke nije 20.
6. Procijenite vjerojatnost da slučajnim izborom osobe iz ove populacije odaberemo osobu čija je učestalost korištenja desne ruke veća od 20.
7. Procijenite vjerojatnost da slučajnim izborom osobe iz ove populacije odaberemo osobu čija je učestalost korištenja desne ruke 30.
8. Uz pretpostavku da stvarna distribucija slučajne varijable kojom modeliramo varijablu sum odgovara empirijskoj distribuciji te varijable, odredite njeno očekivanje, varijancu i standardnu devijaciju.
9. Uz pretpostavku da stvarna distribucija slučajne varijable kojom modeliramo varijablu sum odgovara empirijskoj distribuciji te uz oznaku $\mu = EX$, $\sigma^2 = Var\, X$ odredite sljedeće vjerojatnosti: $P\{|X - \mu| \leq \sigma\}$, $P\{|X - \mu| \leq 2\sigma\}$ i $P\{|X - \mu| \leq 3\sigma\}$.
10. Uz pretpostavku da stvarna distribucija slučajne varijable kojom modeliramo varijablu sum odgovara empirijskoj distribuciji, odredite jedan medijan te slučajne varijable. Također, odredite $P\{|X - m| \leq \sigma\}$, $P\{|X - m| \leq 2\sigma\}$ i $P\{|X - m| \leq 3\sigma\}$, gdje je m medijan koji ste odabrali. Diskutirajte o razlikama u odnosu na prethodno pitanje.

Zadatak 7.3. (tlak.sta)
Baza podataka tlak.sta sadrži podatke o krvnom tlaku utvrđene anketom na reprezentativnom uzorku pacijenata jedne klinike:

varijable spol i dob sadrže informacije o spolu i broju godina za svakog ispitanika

varijable sistolicki-tlak i dijastolicki-tlak sadrže vrijednosti sistoličkog i dijastoličkog tlaka za svakog ispitanika

varijabla tlak klasificira vrijednosti sistoličkog i dijastoličkog tlaka u tri kategorije: N - nizak tlak, O - normalan tlak, P - povišen tlak

varijabla puls sadrži broj otkucaja srca u minuti (puls) za svakog ispitanika

varijabla opce-stanje sadrži subjektivnu ocjenu (u standardnoj skali od 1 do 5) vlastitog zdravstvenog stanja svakog ispitanika.

Na temelju podataka sadržanih u ovoj bazi riješite sljedeće zadatke:

1. Odredite tablice frekvencija i relativnih frekvencija, nacrtajte i proanalizirajte stupčaste dijagrame frekvencija i relativnih frekvencija te kružni dijagram s prikazom relativnih frekvencija za podatke sadržane u varijabli opce-stanje. Kolike su frekvencija i relativna frekvencija ispitanika koji su svoje opće zdravstveno stanje ocijenili barem ocjenom 4?
2. Odredite tablice frekvencija i relativnih frekvencija za podatke sadržane u varijabli opce-stanje posebno za kategoriju ispitanika ženskog spola i kategoriju ispitanika muškog spola te nacrtajte pripadne stupčaste dijagrame frekvencija i relativnih frekvencija. Također nacrtajte stupčaste dijagrame frekvencija i relativnih frekvencija za podatke sadržane u varijabli opce-stanje kategorizirane po vrijednostima varijable tlak (N, O, P). Proanalizirajte dobivene stupčaste dijagrame.
3. Odredite i ukratko protumačite sljedeće numeričke karakteristike podataka sadržanih u varijabli dob: aritmetičku sredinu, medijan, donji i gornji kvartil, mod, raspon i standardnu

devijaciju. Je li mod jedinstven? Koliko iznosi maksimalno odstupanje podataka sadržanih u varijabli dob od njihove aritmetičke sredine? Nacrtajte i detaljno proanalizirajte kutijasti dijagram na bazi medijana za podatke sadržane u varijabli dob. Obrazložite svoj odgovor.

4. Nacrtajte i detaljno proanalizirajte kutijasti dijagram na bazi medijana za podatke sadržane u varijabli dob. Obrazložite svoj odgovor.
5. Crtanjem i analizom kutijastog dijagrama na bazi medijana neosjetljivog na stršeće vrijednosti i kutijastog dijagrama na bazi medijana osjetljivog na stršeće vrijednosti donesite zaključak o tome pojavljuju li se među podacima sadržanima u varijabli puls stršeće vrijednosti ili ne. Ako ste se uvjerili u njihovo postojanje, korištenjem kategoriziranih tablica frekvencija odredite sve prisutne stršeće vrijednosti među podacima u varijabli puls. Kako biste neutralizirali njihov utjecaj na numeričke karakteristike podataka?

Zadatak 7.4. (glukoza.sta)
Baza podataka glukoza.sta opisana je u primjeru 2.2. Poznato je da na nivou značajnosti $\alpha = 0.05$ možemo prihvatiti hipotezu o normalnoj distribuiranosti podataka sadržanih u varijablama `dob` i `glukoza`.

1. Intervalom pouzdanosti 95% procijenite očekivanu koncentraciju glukoze.
2. Postavite potrebne hipoteze i prikladnim testom provjerite je li na nivou značajnosti $\alpha = 0.05$ očekivana koncentracija glukoze statistički značajno veća od 5.5 mMol/L.
3. Intervalom pouzdanosti 95% procijenite proporciju ispitanika kod kojih je koncentracija glukoze u krvi između 4 i 6 mMol/L.
4. Postavite potrebne hipoteze i prikladnim testom provjerite je li na nivou značajnosti $\alpha = 0.05$ proporcija ispitanika kod kojih je koncentracija gluoze veća od 8 mMol/l statistički značajno različita od 0.1.
5. Protumačite sve dobivene rezultate u kontekstu promatranog problema.

Zadatak 7.5. (uvis.sta)
Baza podataka uvis.sta sadrži bodove koje su studenti treće godine preddiplomskog studija matematike prikupili na kolokvijima iz Uvoda u vjerojatnost i statistiku (UVIS):

varijable kol-1 i kol-2 sadrže bodove s redovnog prvog i drugog kolokvija

varijable kol-P1 i kol-P2 sadrže bodove s popravnih kolokvija

varijable konacno-1 i konacno-2 sadrže konačne bodove prikupljene na prvom i drugom kolokviju

varijabla ukupno-1-2 sadrži ukupan broj bodova nakon provedenih redovnih i popravnih kolokvija

varijabla ocjena sadrži prijedlog konačne ocjene iz kolokvija

varijabla stanovanje sadrži informacije o mjestu stanovanja studenata kategorizirane na sljedeći način - Osijek (student stanuje u Osijeku), Drugo mjesto (student stanuje u nekom drugom mjestu).

Na temelju podataka dostupnih u ovoj bazi riješite sljedeće zadatke:

1. Kojeg su tipa varijable kol-1 i ocjena?
2. Odredite empirijsku distribuciju varijable ocjena.

3. Procijenite vjerojatnost da je student kolokvij iz UVIS-a položio ocjenom većom od 2, ali manjom od 5.
4. Nacrtajte stupčasti dijagram frekvencija i relativnih frekvencija za podatke koji su sadržani u varijabli ocjena.
5. Za podatke sadržane u varijabli kol-1 odredite vrijednosti aritmetičke sredine, moda (je li mod ovog niza podataka jedinstven?), donjeg kvartila, medijana i gornjeg kvartila. Ukratko protumačite značenje svake od navedenih numeričkih karakteristika.
6. Skicirajte i proanalizirajte kutijasti dijagram na bazi medijana za podatke sadržane u varijabli kol-2.
7. Provođenjem prikladnih statističkih testova provjerite možemo li na nivou značajnosti $\alpha = 0.05$ tvrditi da je varijabla kol-2 normalno distribuirana.
8. Provođenjem prikladnog statističkog testa provjerite je li na razini značajnosti $\alpha = 0.05$ očekivani broj bodova na prvom popravnom kolokviju (varijabla kol-P1) statistički značajno veći od $\mu_0 = 42.17391$. Koji ste test odabrali i zašto?
9. Provođenjem prikladnog statističkog testa provjerite je li na razini značajnosti $\alpha = 0.05$ proporcija studenata koji su na drugom popravnom kolokviju (varijabla kol-P2) prikupili više od 80 bodova statistički značajno različita od $p_0 = 0.1$. Koji ste test odabrali i zašto?

Zadatak 7.6. (uvis.sta)
Analizirajte bazu podataka uvis.sta opisanu u zadatku 7.5.

1. Analizirajte razlike među rezultatima na redovnim i popravnim kolokvijima za sve studente te posebno za studente koji stanuju u Osijeku i studente koji stanuju u nekom drugom mjestu.
2. Analizirajte veze između rezultata na kolokvijima i prijedloga konačne ocjene iz kolokvija? Što možete zaključiti o utjecaju popravnih kolokvija na konačnu ocjenu?
3. Napravite usporedbu predloženih konačnih ocjena za studente koji žive u Osijeku i studente koji žive u drugim mjestima.

Odaberite prikladne mjere da biste ilustrirali tvrdnje te ih potkrijepite prikladnim statističkim testovima.

Zadatak 7.7. (slobodno-vrijeme.sta)
Baza podataka slobodno-vrijeme.sta sadrži podatke o slobodnom vremenu ispitanika jedne ankete:

varijable Spol i Godine sadrže informacije o spolu, odnosno godinama starosti ispitanika

varijable TV i Kava sadrže podatke koliko sati dnevno ispitanici gledaju televiziju, odnosno koliko šalica kave dnevno popiju

varijabla Hobiji sadrži informacije o tome ima li ispitanik neki hobi ili ne

varijabla Zadovoljan sadrži informacije o tome koliko je ispitanik zadovoljan iskorištenošću svoga slobodnog vremena (1 - nisam zadovoljan, 2 - nije loše, 3 - poprilično sam zadovoljan, 4 - zadovoljan sam, 5 - prezadovoljan sam).

Na temelju podataka dostupnih u ovoj bazi riješite sljedeće zadatke:

1. Kojeg su tipa varijable Hobiji i Zadovoljan?
2. Odredite empirijsku distribuciju varijable Zadovoljan.
3. Procijenite vjerojatnost da je ispitanik poprilično zadovoljan ili zadovoljan iskorištenošću svoga slobodnog vremena.
4. Nacrtajte stupčasti dijagram frekvencija i relativnih frekvencija za podatke koji su sadržani u varijabli Zadovoljan.
5. Za podatke sadržane u varijabli Godine odredite očekivani broj godina ispitanika, najčešći broj godina te maksimalno odstupanje od očekivanog broja godina.
6. Skicirajte i proanalizirajte kutijasti dijagram na bazi aritmetičke sredine za podatke sadržane u varijabli TV.
7. Provođenjem prikladnih statističkih testova provjerite možemo li na nivou značajnosti $\alpha = 0.05$ tvrditi da je varijabla TV normalno distribuirana.
8. Provođenjem prikladnog statističkog testa provjerite je li na razini značajnosti $\alpha = 0.01$ očekivani broj ispijenih kava (varijabla Kava) statistički značajno veći od $\mu_0 = 1$. Koji ste test odabrali i zašto?

Zadatak 7.8. (slobodno-vrijeme.sta)
Analizirajte bazu podataka slobodno-vrijeme.sta koja je opisana u zadatku 7.7.

1. Analizirajte spolnu i starosnu strukturu uzorka u ovom primjeru te varijablu TV za sve kategorije varijable Spol i prikladno kategorizirane vrijednosti varijable Godine.
2. Analizirajte varijablu TV za različite kategorije varijable Hobiji za sve ispitanike zajedno te posebno za ispitanike muškog i posebno za ispitanike ženskog spola. Napravite usporedbe rezultata za muški i ženski spol.
3. Promatrajte dvije dobne skupine ispitanika - ispitanike mlađe od 30 godina i one stare barem 30 godina. Napravite usporedbu zadovoljstva iskorištenošću svog slobodnog vremena među tim dvjema dobnim skupinama. Za navedene dobne skupine napravite usporedbe varijable Zadovoljstvo s obzirom na različite kategorije varijable Hobiji.

Odaberite prikladne mjere da biste ilustrirali tvrdnje te ih potkrijepite prikladnim statističkim testovima.

Zadatak 7.9. (zdravlje.sta)
Baza podataka zdravlje.sta opisana je u zadatku 2.4. Na temelju podataka dostupnih u ovoj bazi riješite sljedeće zadatke:

1. Kojeg su tipa varijable dodatno-zdravstveno i cijena?
2. Odredite empirijsku distribuciju varijable zdravlje.
3. Procijenite vjerojatnost da slučajno odabrani ispitanik svoje zdravstveno stanje smatra barem dobrim.
4. Nacrtajte stupčasti dijagram frekvencija i relativnih frekvencija za podatke sadržane u varijabli spol.

5. Za podatke sadržane u varijabli godine odredite broj godina koji se nalazi na centralnoj poziciji uređenog niza podataka, očekivani broj godina ispitanika te maksimalno odstupanje od očekivanog broja godina.
6. Skicirajte i proanalizirajte kutijasti dijagram na bazi medijana za podatke sadržane u varijabli cijena.
7. Provođenjem prikladnih statističkih testova provjerite možemo li na nivou značajnosti $\alpha = 0.01$ tvrditi da je varijabla cijena normalno distribuirana.
8. Provođenjem prikladnog statističkog testa provjerite je li na razini značajnosti $\alpha = 0.05$ očekivani broj pregleda u tekućoj akademskoj godini (varijabla broj-pregleda) statistički značajno različit od $\mu_0 = 4$. Koji ste test odabrali i zašto?

Zadatak 7.10. (zdravlje.sta)
Analizirajte bazu podataka zdravlje.sta koja je opisana u zadatku 2.4.

1. Analizirajte varijablu zdravlje posebno za kategoriju ispitanika koji imaju dodatno zdravstveno osiguranje te posebno za kategoriju ispitanika koji ga nemaju. Napravite usporedbu rezultata. Isti postupak ponovite posebno za muškarce, a posebno za žene te napravite usporedbu dobivenih rezultata.
2. Na prikladan način kategorizirajte vrijednosti varijable godine te napravite usporedbu očekivane cijene najskupljeg zdravstvenog pregleda među tako napravljenim dobnim skupinama.
3. Procijenite zajedničku distribuciju slučajne varijable koja modelira broj zdravstvenih pregleda i slučajne varijable koja se realizira jedinicom u slučaju da ispitanik ima dodatno zdravstveno osiguranje, a nulom ako ga nema. Procijenite sve marginalne i uvjetne distribucije tog dvodimenzionalnog slučajnog vektora. Obratite pažnju na proporcije ispitanika koje se odnose na najveći broj zdravstvenih pregleda u dobivenim marginalnim i uvjetnim empirijskim distribucijama te napravite usporedbe koje smatrate korisnima i zabilježite svoje zaključke.

Odaberite prikladne mjere da biste ilustrirali tvrdnje te ih potkrijepite prikladnim statističkim testovima.

Zadatak 7.11. (novi-stan.sta)
Baza podataka novi-stan.sta sadrži podatke potrebne banci da odobri kredit klijentu za kupnju novog stana:

varijable Spol i Godine sadrže informacije o spolu, odnosno godinama starosti klijenta

varijabla God-rad-staža sadrži podatke o godinama radnog staža klijenta

varijabla Stručna sprema sadrži informacije o stručnoj spremi klijenta

varijabla Krediti sadrži informacije o broju do sada odobrenih kredita tog klijenta

varijabla Kvadratura sadrži informacije o željenoj kvadraturi stana (50, 75, 100 ili 120 m^2)

varijabla Smještaj sadrži informacije o tome živi li trenutno klijent u Osijeku ili izvan njega

varijabla Broj djece sadrži informacije o broju djece klijenta.

Na temelju podataka dostupnih u ovoj bazi riješite sljedeće zadatke:

1. Kojeg su tipa varijable Smještaj i Kvadratura?
2. Odredite empirijsku distribuciju varijable Stručna sprema.
3. Procijenite vjerojatnost da stranka ima više od dva djeteta.
4. Nacrtajte stupčasti dijagram frekvencija i stupčasti dijagram relativnih frekvencija za podatke koji su sadržani u varijabli Spol.
5. Za podatke sadržane u varijabli Godine odredite vrijednosti aritmetičke sredine, moda (je li jedinstven), varijance i standardne devijacije. Ukratko protumačite značenje svake od navedenih numeričkih karakteristika.
6. Skicirajte kutijasti dijagram na bazi medijana za podatke sadržane u varijabli Godine.
7. Provođenjem prikladnih statističkih testova provjerite možemo li na nivou značajnosti $\alpha = 0.05$ tvrditi da je varijabla Krediti normalno distribuirana.
8. Provođenjem prikladnog statističkog testa provjerite je li na razini značajnosti $\alpha = 0.05$ proporcija klijenata koji imaju dvoje djece (varijabla Broj djece) statistički značajno manja od $p_0 = 0.4$. Koji ste test odabrali i zašto?

Zadatak 7.12. (novi-stan.sta)
Analizirajte bazu podataka novi-stan.sta opisanu u zadatku 7.11.

1. Analizirajte razlike varijable Krediti između muškaraca i žena. Jesu li se u prosjeku više (pri čemu se misli na broj zaduživanja, ne na njihov iznos) kreditno zaduživali muškarci ili žene? Analizirajte distribuciju broja zaduživanja posebno za svaki spol.
2. Analizirajte broj zaduživanja klijenata ženskog spola za različite kategorije stručne spreme. Isti postupak provedite i za klijente muškog spola.
3. Pod uvjetom da klijent živi u Osijeku, analizirajte ovisi li željena kvadratura stana o broju djece klijenta. Isti postupak provedite i pod uvjetom da klijent ne živi u Osijeku.

Odaberite prikladne mjere da biste ilustrirali tvrdnje te ih potkrijepite prikladnim statističkim testovima.

Zadatak 7.13. (kredit.sta)
Baza podataka kredit.sta sadrži podatke o kreditnoj povijesti klijenata jedne američke komercijalne banke. U nastavku je opisano značenje svih varijabli.

varijabla KO predstavlja ocjenu klijenta na sljedeći način: L - loš; D- dobar

varijabla RAC sadrži podatke o stanju računa klijenta; BR - klijent nema otvoren račun u banci; N - klijent nema sredstava na računu; <=$300 - stanje na računu je pozitivno i manje ili jednako od 300; >$300 - klijent ima iznos na računu veći od $300

varijabla T prestavlja trajanje otplate kredita (u mjesecima)

varijabla NK opisuje namjenu kredita: NA - novi automobil; RA - rabljeni automobil; NM - namještaj; TV - televizor; KA - kućanski aparati; P - popravak; O - odmor; PKV - prekvalifikacija; POS - posao; D - drugo

varijabla IK predstavlja iznos kredita

varijabla PS predstavlja trajanje zaposlenosti klijenta na trenutnom radnom mjestu: NZ - nezaposlen; <1 god - manje od 1 godine; 1-5 god - između 1 i 5 godina; 5-8 god - između 5 i 8 godina; > 8 god - više od 8 godina

varijabla BR opisuje bračno stanje klijenta: RAZ - razveden; ZR - zivi rastavljeno; SM - samac; BRU - živi u bračnoj zajednici ili kao udovac/ica

varijabla S predstavlja spol klijenta: M - muško; Z - žensko

varijabla DOB predstavlja starosnu dob klijenta.

Na temelju podataka dostupnih u ovoj bazi riješite sljedeće zadatke:

1. Kojeg su tipa varijable RAC i IK?
2. Odredite empirijsku distribuciju varijable NK.
3. Procijenite vjerojatnost da je klijent ostvario kredit čija otplata traje najviše 20, a najmanje 10 godina.
4. Nacrtajte stupčasti dijagram frekvencija i stupčasti dijagram relativnih frekvencija za podatke koji su sadržani u varijabli PS.
5. Odredite tablicu frekvencija i relativnih frekvencija za podatke sadržane u varijabli RAC posebno za kategoriju ispitanika ženskog spola, a posebno za kategoriju ispitanika muškog spola.
6. Nacrtajte zajednički stupčasti dijagram frekvencija i relativnih frekvencija tipa `Overlaid` svih podataka sadržanih u varijabli RAC kategoriziran prema spolu klijenta.
7. Za podatke sadržane u varijabli DOB odredite vrijednosti aritmetičke sredine, moda (je li jedinstven), varijance i standardne devijacije. Protumačite značenje svake od navedenih numeričkih karakteristika.
8. Skicirajte i protumačite kutijasti dijagram na bazi medijana za podatke sadržane u varijabli T.
9. Je li moguće na osnovi tablice frekvencija i relativnih frekvencija te stupčastog dijagrama numeričke varijable IK dobiti dovoljno informacija o iznosima kredita klijenata promatrane banke. Obrazložite svoj odgovor.
10. Iskoristite izmjerene vrijednosti iste varijable iz baze podataka kredit-score.sta. Mijenjajte broj intervala na koji dijelite skup vrijednosti. Proučavajte što se događa i pribilježite svoj zaključak.
11. Kategorizaciju izmjerenih vrijednosti varijable IK napravite na način koji vam izravno daje procjenu vjerojatnosti da je klijent ostvario kredit u iznosu od najviše $10000, ali ne manje od $5000.

Zadatak 7.14. (djelatnici.sta)
Baza podataka djelatnici.sta opisana je u zadatku 2.4. Na temelju opisanih podataka riješite sljedeće zadatke:

1. Kojeg su tipa varijable Obrazovanje i Visina?
2. Odredite empirijsku distribuciju varijable Obrazovanje.

3. Procijenite vjerojatnost da djelatnik radi na odjelu za transport ili isporuku.
4. Nacrtajte stupčasti dijagram frekvencija i stupčasti dijagram relativnih frekvencija za podatke koji su sadržani u varijabli Obrazovanje.
5. Odredite tablicu frekvencija i relativnih frekvencija za podatke sadržane u varijabli Obrazovanje posebno za kategoriju djelatnika ženskog spola, a posebno za kategoriju djelatnika muškog spola.
6. Nacrtajte zajednički stupčasti dijagram frekvencija i relativnih frekvencija svih podataka sadržanih u varijabli Odjel kategoriziran prema varijabli Obrazovanje.
7. Za podatke sadržane u varijabli Rukovodstvo odredite vrijednosti raspona, donjeg i gornjeg kvartila te medijana. Protumačite značenje svake od navedenih numeričkih karakteristika.
8. Skicirajte i protumačite kutijasti dijagram na bazi medijana za podatke sadržane u varijabli Dob.
9. Iskoristite izmjerene vrijednosti varijable Placa_prije. Kategorizirajte varijablu na 5 jednakih podintervala (napišite tablicu relativnih frekvencija i skicirajte stupčasti dijagram relativnih frekvencija). Mijenjajte broj intervala na koji dijelite skup vrijednosti. Proučavajte što se događa i pribilježite svoj zaključak.
10. Kategorizaciju izmjerenih vrijednosti varijable Visina napravite na način koji vam izravno daje procjenu vjerojatnosti da je visina djelatnika u intervalu $[165, 180\rangle$. Koliko iznosi procjena vjerojatnosti? Napišite tablicu relativnih frekvencija kategorizirane varijable.
11. Kojim tipovima slučajnih varijabli modeliramo varijable ove baze podataka?
12. Intervalom pouzdanosti 95% procijenite očekivanje slučajne varijable kojom je modelirana dob djelatnika tvornice A.
13. Intervalom puzdanosti 95% procijenite vjerojatnost da je djelatnik tvornice A viši od 170 cm.
14. Možete li na razini značajnosti $\alpha = 0.05$ tvrditi da je očekivana visina djelatnika tvornice A manja od 170 cm?
15. Možete li na razini značajnosti $\alpha = 0.05$ tvrditi da je vjerojatnost da je djelatnik tvornice A stariji od 30 godina manja od 0.5?
16. Možete li na razini značajnosti $\alpha = 0.05$ tvrditi da slučajna varijabla kojom je modelirana dob djelatnika promatrane tvornice nije normalno distribuirana?
17. Možete li na razini značajnosti $\alpha = 0.05$ tvrditi da se distribucija slučajne varijable kojom je modelirano radno mjesto (varijabla Odjel) djelatnika tvornice A razlikuje od distribucije zadane tablicom teorijskih frekvencija

Obrazovanje	TR	P	IS
Frekvencija	20	40	40

18. Ispitajte može li se zavisnost između visine mjesečne neto-plaće prije i nakon reorganizacije sustava poslovanja tvornice A opisati jednostavnim linearnim regresijskim modelom:
    - Koju varijablu promatrate kao ovisnu, a koju kao neovisnu (prediktornu) varijablu? Procijenite koeficijente pripadnog regresijskog pravca i proanalizirajte dobiveni rezultat.
    - Kako se računaju reziduali? Možete li na razini značajnosti $\alpha = 0.05$ tvrditi da reziduali nisu normalno distribuirani?

- Što o linearnom regresijskom modelu možete reći na temelju koeficijenta smjera procjenjenog regresijskog pravca?
- Koliki je dio promjena u eksperimentalnim vrijednostima ovisne varijable objašnjen linearnim regresijskim modelom?

Zadatak 7.15. (rakovi.sta)
Baza podataka rakovi.sta sadrži podatke o jednom biološkom istraživanju u kojem su bilježene reprezentativne karakteristike ženki bodljaša *Carpilius convexus* koje uključuju broj satelita (tj. broj mužjaka bodljaša prihvaćenih za gnijezdo koje grade ženke), stanje bodlje, boju, težinu, itd. U nastavku je opisano značenje svih varijabli.

varijabla I pokazuje ima li bodljaš satelite ili ne: 1 - bodljaš ima više od 0 satelita; 0- bodljaš nema satelita

varijabla B označava boju jedinke iz uzorka: SS - srednje svijetla; S - svijetla; ST - srednje tamna; T - tamna

varijabla KR prestavlja stanje bodlji: 2D - obje bodlje u dobrom stanju; 1D - jedna bodlja je u dobrom stanju dok je druga u lošem; 0D - obje bodlje su u lošem stanju

varijabla D predstavlja širinu karapakse ženke bodljaša u centimetrima

varijabla NS predstavlja broj satelita kod jedinke iz uzorka

varijabla M predstavlja masu jedinke iz uzorka (u kg).

Na temelju podataka dostupnih u ovoj bazi riješite sljedeće zadatke:

1. Kojeg su tipa varijable B i D?
2. Odredite empirijsku distribuciju varijable KR.
3. Procijenite vjerojatnost da je broj satelita kod ženke bodljaša *Carpilius convexus* veći od 5.
4. Nacrtajte stupčasti dijagram frekvencija i stupčasti dijagram relativnih frekvencija za podatke koji su sadržani u varijabli B.
5. Odredite tablicu frekvencija i relativnih frekvencija za podatke sadržane u varijabli KR posebno za kategoriju jedinki koje imaju satelite, a posebno za kategoriju jedinki koje nemaju satelite.
6. Nacrtajte zajednički stupčasti dijagram frekvencija i relativnih frekvencija tipa `Separate` svih podataka sadržanih u varijabli B kategoriziranih prema tome imaju li odgovarajuće jedinke iz uzorka satelite ili ne.
7. Za podatke sadržane u varijabli M odredite vrijednosti aritmetičke sredine, moda (je li jedinstven?), varijance i standardne devijacije. Protumačite značenje svake od navedenih numeričkih karakteristika.
8. Skicirajte i protumačite kutijasti dijagram na bazi medijana za podatke sadržane u varijabli NS.
9. Kategorizirajte varijablu D tako da procijenite vjerojatnost da je širina karapakse veća ili jednaka od 26 a manja od 28.

Zadatak 7.16. Prema jednoj anketi provedenoj u RH, da bi posjetitelj ZOO-vrta bio zadovoljan mnogobrojnošću vrsta, u ZOO-vrtu trebalo bi biti 15% divljih mačaka, 20% ptica, 10% majmuna, 15% glodavaca, 20% morskih životinja te 20% ostalih životinja. Podaci o broju životinja u jednom novootvorenom ZOO-vrtu dani su u sljedećoj tablici:

divlje mačke	ptice	majmuni	glodavci	morske ž.	ostale ž.
24	36	22	32	60	26

.

Razlikuje li se ova distribucija na nivou značajnosti $\alpha = 0.05$ statistički značajno od distribucije predviđene anketom? Koji ste test koristili?

Zadatak 7.17. Prema podacima iz 2007. godine tjedna prodaja cipela u jednoj osječkoj trgovini cipela bila je oblika: 10% prodano je ponedjeljkom, 13% utorkom, 15% srijedom, 17% četvrtkom, 20% petkom te 25% subotom. Prošli tjedan zabilježene su sljedeće frekvencije:

pon	uto	sri	čet	pet	sub
16	20	40	26	52	46

.

Vlasnika trgovine zanima odstupaju li na nivou značajnosti $\alpha = 0.05$ prošlotjedni podaci statistički značajno od prošlogodišnjeg tjednog standarda.

Zadatak 7.18. Po istraživanjima Državne uduge ljubitelja sladoleda, da bi sladokusci slastičarnicu ocijenili ocjenom izvrstan mora im biti ponuđeno 30% voćnih vrsta sladoleda, 40% mliječnih vrsta sladoleda, 20% miješanih vrsta sladoleda i 10% *light* vrsta sladoleda (bez obzira jesu li po sastavu voćni, mliječni ili mješoviti). Frekvencije spomenutih kategorija sladoleda u poznatoj osječkoj slastičarnici Petar Pan dane su u sljedećoj tablici:

kategorija sladoleda	voćni	mliječni	mješoviti	*light*
frekvencija	12	10	5	3

.

Razlikuje li se ova distribucija na nivou značajnosti $\alpha = 0.1$ od distribucije dobivene istraživanjem? Koji ste test koristili?

Zadatak 7.19. Po istraživanjima Nacionalne organizacije knjižničara dobro opremljenom smatramo knjižnicu u kojoj 40% knjižničnog fonda čini beletristika, 35% klasici, 20% stručna literatura i 5% rijetke i vrijedne knjige (bez obzira jesu li klasici ili stručne knjige). Frekvencije spomenutih kategorija knjiga u promatranoj knjižnici dane su u sljedećoj tablici:

kategorija knjiga	beletristika	klasici	stručne knjige	rijetke i vrijedne knjige
frekvencija	430	330	200	40

.

Razlikuje li se ova distribucija na nivou značajnosti $\alpha = 0.05$ statistički značajno od distribucije dobivene istraživanjem? Koji ste test koristili?

Zadatak 7.20. Vlasnika poznate slastičarnice koja prodaje najbolje krempite u gradu zanima postoji li dio dana u kojemu se kod građana budi veća želja za konzumacijom tog kolača. Počevši od 10:00 sati odabrao je 5 vremenskih intervala duljine 2 sata i bilježio broj ljudi koji su kupili krempitu. Na razini značajnosti $\alpha = 0.05$ provjerite konzumiraju li građani krempite više u nekom od ponuđenih vremenskih intervala ili ih konzumiraju jednoliko tijekom cijelog mjerenog perioda.

Vremenski interval	10 - 12	12 - 14	14 - 16	16 - 18	18 - 20
Broj kupaca	16	24	30	20	10

.

Zadatak 7.21. Voditelj pjevačkog zbora nastoji poštovati zahtjev o jednakoj zastupljenosti prvog, drugog i trećeg glasa u svom zboru. Trenutačno zbor broji 90 pjevača, čije su frekvencije po glasovima dane u sljedećoj tablici:

Glas	Prvi	Drugi	Treći
Broj pjevača	33	35	23

.

Razlikuje li se ova distribucija na nivou značajnosti $\alpha = 0.05$ od zahtijevane distribucije? Koji ste test koristili?

Zadatak 7.22. Jednog liječnika hitne medicine zanima postoji li dio dana u kojemu ljudi frekventnije traže hitne medicinske intervencije. U svrhu svog istraživanja dan je podijelio na 4 jednaka vremenska intervala (svaki u trajanju od 6 sati) i prikupio sljedeće podatke:

Vremenski interval	0:00 - 6:00	6:00 - 12:00	12:00 - 18:00	18:00 - 24:00
Broj intervencija	20	27	31	22

Na razini značajnosti $\alpha = 0.01$ provjerite jesu li hitne liječničke intervencije češće u određeno doba dana ili su jedoliko distribuirane tijekom cijelog dana.

Zadatak 7.23. (sport.sta)
U bazi podataka sport.sta nalaze se rezultati istraživanja o bavljenju sportom (varijabla sport: 0 - osoba se u slobodno vrijeme ne bavi sportom; 1 - osoba se u slobodno vrijeme bavi sportom) s obzirom na spol ispitanika (varijabla Spol: Z - osoba je ženskog spola; M - osoba je muškog spola). Riješite sljedeće zadatke:

1. Odredite zajedničku tablicu frekvencija varijabli sport i spol te procijenite zajedničku tablicu distribucije ovih varijabli.
2. Pomoću zajedničke tablice distribucije varijabli sport i spol procijenite empirijske distribucije varijabli sport i spol.
3. Procijenite uvjetnu distribuciju varijable sport posebno za svaku vrijednost varijable spol.
4. Možemo li na nivou značajnosti $\alpha = 0.05$ govoriti o nezavisnosti varijabli sport i spol? Koji ste test koristili?

Zadatak 7.24. (kupovina.sta)
Baza kupovina.sta sadrži podatke o broju bodova koje je kupac skupio tijekom dosadašnje kupovine u nekom trgovačkom centru (varijabla broj-bodova) i iznosu popusta u kunama koje mu isti trgovački centar poklanja u sljedećoj kupovini (varijabla popust-kn) za 100 promatranih kupaca. Koju ćete od ovih varijabli promatrati kao neovisnu, a koju kao ovisnu varijablu? Odredite procjenu regresijskog pravaca te odgovorite na sljedeća pitanja:

1. Što o linearnom regresijskom modelu možete reći na temelju analize reziduala?
2. Što o linearnom regresijskom modelu možete reći na temelju koeficijenta smjera procijenjenog regresijskog pravca?
3. Koliki je dio promjena u eksperimentalnim vrijednostima ovisne varijable objašnjen linearnim regresijskim modelom?

Bibliografija

[1] Bain, L.E, Engelhardt, M. *Introduction to Probability and Mathematical statistics*, Duxbury, 2009.

[2] Bhattacharyya, G. K., Johnson, R. A. *Statistical Concepts and Methods*, Wiley, New York, 1977.

[3] Daniel, W.W., Terrell, J.C. *Business Statistics*, Houghton Mifflin Company, Boston, 1989.

[4] Elezović, N. *Diskretna vjerojatnost*, Element, Zagreb, 2007.

[5] Elezović, N. *Slučajne varijable*, Element, Zagreb, 2007.

[6] Elezović, N. *Statistika i procesi*, Element, Zagreb, 2007.

[7] Freund, J. E. *Mathematical Statistics*, Prentice Hall, 1992.

[8] Ilijašević, M., Pauše, Ž. *Riješeni primjeri i zadaci iz vjerojatnosti i statistike*, "Zagreb", Samobor, 1990.

[9] Iversen, G. R. *Statistics, the conceptual Approach*, Springer, Berlin, 1997.

[10] Ivanović, B. *Teorijska statistika* Jugoslavenski institut za ekonomska istraživanja, Beograd, 1966.

[11] Jazbec, A. *Osnove statistike*, Šumarski fakultet, Zagreb, 2008.

[12] Jukić, D., Scitovski, R. *Matematika I* Elektrotehnički fakultet, Odjel za matematiku, Prehrambeno-tehnološki fakultet, Osijek, 2000.

[13] Jamnik, R. *Matematična statistika*, Državna založba Slovenije, Ljubljana, 1980.

[14] Javor, P. *Uvod u matematičku analizu*, Školska knjiga, Zagreb, 1988.

[15] Lehmann, E.L. *Testing Statistical Hypotheses*, J. Wiley, 1959.

[16] Lehman, E. L., Casella, G. *Theory of Point Estimation*, Springer, 1998.

[17] Lipschutz, S., Schiller, J. *Introduction to Probability and Statistics*, Schaum's Outline Series, McGraw-Hill, New York – Toronto, 1998.

[18] McClave, J. T., Benson, P. G., Sincich, T. *Statistics for Bussiness and Economics*, Prentice Hall, London, 2001.

[19] McPherson, G. *Applying and Interpreting Statistics*, Springer, Berlin, 2001.

[20] Mittelhammer, R.C. *Mathematical Statistics for Economics and Bussines*, Springer, New York, 1996.

[21] Pauše, Ž. *Uvod u matematičku statistiku*, Školska knjiga, Zagreb, 1993.

[22] Pauše, Ž. *Vjerojatnost, informacija, stohastički procesi*, Školska knjiga, Zagreb, 1988.

[23] Pavlić, I. *Statistička teorija i primjena*, Tehnička knjiga, Zagreb, 1985.

[24] Pogany, T. *Teorija vjerojatnosti, zbirka riješenih ispitnih zadataka*, Odjel za pomorstvo Sveučilišta u Rijeci, Rijeka, 1999.

[25] Rawlings, J. O., Pantula, S. G., Dicky, D. A. *Applied Regression Analysis*, Springer, Berlin, 1998.

[26] Sarapa, N. *Teorija vjerojatnosti*, Školska knjiga, Zagreb, 1988.

[27] Sarapa, N. *Vjerojatnost i statistika I. dio: osnove vjerojatnosti - kombinatorika*, Školska knjiga, Zagreb, 1995.

[28] Sarapa, N. *Vjerojatnost i statistika II. dio: osnove statistike - slučajne varijable*, Školska knjiga, Zagreb, 1996.

[29] Seber G.A.F, Lee A.J. *Linear Regression Analysis*, Wiley, Hoboken-New Jersey, 2003.

[30] Serdar, V., Šošić, I. *Uvod u statistiku*, Školska knjiga, Zagreb, 1986.

[31] Triola, M.F. *Elementary Statistics*, The Benjamin/Cummings Publishing company, Inc. 1989.

[32] Vranić, V. *Vjerojatnost i statistika*, Tehnička knjiga, Zagreb, 1971.

[33] Vranjković, P. *Zbirka zadataka iz vjerojatnosti i statistike*, Školska knjiga, Zagreb, 1990.

Indeks