Sadržaj

Poglavlje 1

Uvod

Korištenje riječi **statistika** u svakodnevnom životu najčešće je povezano s brojčanim vrijednostima kojima pokušavamo opisati bitne karakteristike nekog skupa podataka. Na službenim web stranicama Državnog zavod za statistiku Republike Hrvatske možemo pročitati (http://www.dzs.hr/ dana 6.6.2009):

Prosječna mjesečna isplaćena neto plaća po zaposlenome u pravnim osobama Republike Hrvatske za srpanj 2009. iznosila je 5 308 kuna.

Minimalna plaća za razdoblje od 1. lipnja 2009. do 31. svibnja 2010. u Republici Hrvatskoj iznosi 2 814.00 kuna.

Stopa registrirane nezaposlenosti za kolovoz 2009. iznosila je 14.2%.

Udio aktivnog stanovništva u radno sposobnom (stopa aktivnosti) iznosi 48%, istovremeno 43.7% radno sposobnih osoba je zaposleno (stopa zaposlenosti), a 8.9% radne snage je nezaposleno (stopa nezaposlenosti).

Temelj statistike, kao znanstvene discipline, kao i svih istraživanja koja se koriste statističkim metodama zaista 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 na jednikama 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 istraživanjem

U ovom primjeru navedena je samo jedna varijabla koja se analizira na jedinkama populacije, tj. uspjeh iz Statistike. Međutim, često nas zanima nekoliko varijabli i/ili veze među njima. Npr. želimo li ispitati ovisi li uspjeh iz Statistike u prethodnom primjeru o spolu, potrebno je u istraživanju populacije za svaku jedinku zabilježiti i vrijednost varijable spol (M ili Ž); želimo li ispitati ovisi li uspjeh iz Statistike o pripadnosti pojedinoj grupi vježbi, potrebno je za svaku jedinku zabilježiti koju grupu vježbi je 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 u tablici 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 nije problem istražiti cijelu populaciju 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 upravo 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. 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 potrebno je detaljno proučiti, 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 temalju prikupljenih podataka na uzorku.

U skladu s ovim razmatranjima, u ovom kolegiju ćemo se baviti nekim **metodama prikupljanja podataka i kreiranja uzorka**, **metodama deskriptivne statistike** i **metodama statističkog zaključivanja**. 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. Taj skup jedinki naziva se **populacija** i kažemo da **populaciju čine sve jedinke koje su predmet istraživanja.**

Primjer 2.1. Istražujemo prehrambene navike i 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, kojeg 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 zastupljne sve tipične karakteristike populacije bitne za istraživanje.

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

Jedan od načina izbora jedinki iz populacije u uzorak je formiranje takozvanog **slučajnog uzorka** poštujući zahtjev da svaka jedinka populacije ima jednaku vjerojatnost (šansu) da uđe u uzorak.

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 su sljedeći:

- Podaci iz javnih izvora (knjige, časopisi, novine, Internet).
- Podaci iz dizajniranog eksperimenta. (Istraživač raspoređuje eksperimentalne jedinke u skupine nad kojima vrši eksperimente te bilježi podatke za varijable koje ga zanimaju.)

  Primjer 2.3. Jedno medicinsko istraživanje proučava snagu nekog lijeka u prevenciji moždanog udara. Skupinu ljudi s kojima će se vršiti istraživanje istraživač dijeli na dvije skupine: tretiranu i kontrolnu. Ljudima u tretiranoj skupini daje se lijek, dok se ljudima u kontrolnoj skupini daje nadomjestak koji izgleda isto kao lijek ali zapravo nije ništa što može imati bilo kakav utjecaj na organizam.

- Podaci iz ankete. (Istraživač sastavlja anketni upitnik, izabire skupinu ljudi koju anketira i na osnovu njihovih odgovora prikuplja podatke.)
- Podaci prikupljeni promatranjem. (Istraživač promatra eksperimentalne jedinke u njihovom prirodnom okruženju i bilježi podatke za varijable od interesa.)

Zadatak 2.1. (stanovnistvo.xls; 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(.xls). Uočimo da dobivena 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 vašeg omiljenog kafića,

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

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 u vašem gradu.

2.3 Tipovi varijabli

U statističkim istraživanjima razlikujemo dva osnovna tipa varijabli koje se međusobno razlikuju po karakteristikama vrijednosti koje varijabla može poprimiti.

2.3.1 Kvalitativne varijable

Karakteristika kvalitativnih varijabli je da njihove vrijednosti nisu, po svojim svojstvima korišenim u istraživanju, realni brojevi. Tipičan primjer takve varijable je spol osobe. Vrijednosti kvalitativne varijable uobičajeno svrstavamo u kategorije. Kategorije kvalitativnih varijabli mogu biti mogu biti definirane u skladu s potrebama statističkog istraživanja.

Primjer 2.4. Sljedeće varijable su 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**. Tipični primjeri numeričkih varijabli su težina i visina osobe. Međutim, treba naglasiti da se i kategorije kvalitativnih varijabli mogu izražavati brojevima što ih ne čini numeričkim varijablama. Npr. spol osobe je jedna kvalitativna varijabla. Kategoriju "ženski spol" možemo označiti npr. "1", a kategoriju "muški spol" "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 varijable su numeričkog tipa:

- postotak prolaznosti na pojedinim ispitima u toku 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 **kontinuirane** varijable.

Diskretne numeričke varijable mogu poprimiti samo konačno ili prebrojivo mnogo vrijednosti

Primjer 2.6. Sljedeće numeričke varijable su diskretne:

- broj bodova na državnoj maturi iz matematike,
- broj ulovljenih komaraca u klopku,
- broj dana u godini s temperaturom zraka većom od 35^oC.

Skup mogućih vrijednosti kontinuiranih numeričkih varijabli je cijeli skup realnih brojeva ili neki interval.

Primjer 2.7. Sljedeće numeričke varijable su kontinuirane:

- postotak prolaznosti na pojedinim ispitima u toku jedne akademske godine,
- temperatura mora,
- vodostaj neke rijeke.

U svrhu prikaza podataka i nekih statističkih analiza, vrijednosti numeričke varijable se tako.er mogu svrstati u **kategorije**. Za razliku od kategorija kvalitativnih varijabli, među kategorijama numeričke varijable uvijek se može prepoznati prirodan poredak.

Primjer 2.8. (auto-centar.sta)
Svrha ovog primjera je prikazati mogućnost kategorizacije diskretne 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 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 diskretna numerička 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

Primjer 2.9. (glukoza.sta)
Svrha ovog primjera je prikazati mogućnost kategorizacije kontinuirane numeričke varijable. Taj se postupak najčešće rješava stvaranjem nove kvalitativne varijable čije su vrijednosti svrstane u nekoliko kategorija. Baza podataka (glukoza.sta) sastoji se od sljedećih varijabli:

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

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

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

Tablica 2.2: Primjer kategorizacije kontinuirane numeričke varijable koncentracija

kategorija - kvalitativna varijabla koja podatke iz varijable **koncentracija glukoze** svrstava u dvije kategorije (svaka kategorija je jedan interval pozitivnih realnih brojeva) na način prikazan u tablici 2.2.

Primjer 2.10. (kolegij.sta)
U ovom primjeru prikazana je kategorizacija jedne diskretne i jedne kontinuirane numeričke varijable. 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 kategorija je jedan konačan skup) na način prikazan u tablici 2.3.

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 2.3: Primjer kategorizacije diskretne numeričke varijable godina-upisa

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

prosjek - kontinuirana numerička varijabla koja sadrži prosječne ocjene iz četiri spomenuta kolegija za svakog od sto promatranih studenata.

uspjeh - diskretna numerička varijabla koja vrijednosti varijeble **prosjek** svrstava u četiri kategorije prema kriteriju prikazanom u tablici 2.4.

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

Tablica 2.4: Primjer kategorizacije kontinuirane numeričke varijable prosjek

Primjer 2.11. (student.sta, student-grupe.sta)
Svrha ovog primjera je 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 - diskretna 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 - diskretna 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 - diskretna 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 2.5.

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

Tablica 2.5: 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 su organizirani 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 (pogledajte tablicu 2.5).

Primjer 2.12. (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 - kontinuirana numerička varijabla koja sadrži podatke o prosječnoj ocjeni studiranja za 49 anketiranih studenata,

polozeno - kvalitativna varijabla koja studente svrstava u dvije kategorije s obzirom na to jesu li položili ispit iz promatranog kolegija prema kriteriju prikazanom u tablici 2.6.

položen/nepoložen ispit	kategorija
položen ispit	1
nepoložen ispit	0

Tablica 2.6: Kategorizacija studenata prema položenosti ispita

predavanja, vjezbe - dvije kvalitativne varijable koje prisutnost studenata na predavanjima/vježbama (p/v) svrstavaju u tri kategorije na način prikazan u tablici 2.7.

tezina-kolegija, materijali - dvije diskretne numeričke varijable koja sadrže subjektivne ocjene (u standardnoj skali od 1 do 5) promatranih studenata za težinu kolegija i dostatnost dostupnih materijala za pripremanje ispita iz promatranog kolegija.

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

Tablica 2.7: Kategorizacija studenata prema broju izostanaka s predavanja/vježbi

Zadatak 2.2. *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 TV 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 njegovom 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 stanovanje 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).*

Ponovimo

- Populaciju čine sve jedinke koje su predmet istraživanja.
- Uzorak je podskup jedinki iz populacije.
- Slučajan uzorak iz populacije formira se tako da svaka jedinka populacije ima jednaku vjerojatnost (šansu) da uđe u uzorak.
- Razlikujemo dva tipa varijabli - kvalitativne i numeričke varijable.
- Vrijednosti numeričkih varijabli su elementi skupa realnih brojeva.
- Vrijednosti kvalitativnih varijabli, po svom karakteru, nisu realni brojevi. Svrstavamo ih u kategorije.
- Među numeričkim varijablama razlikujemo diskretne i kontinuirane varijable.
- Diskretne numeričke varijable mogu poprimiti samo konačno ili prebrojivo mnogo vrijednosti.
- Skup mogućih vrijednosti kontinuiranih numeričkih varijabli je cijeli skup realnih brojeva ili neki interval.

Poglavlje 3

Deskriptivna statistika

3.1 Metode opisivanja kvalitativnih podataka

Kao što je već naglašeno, 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 dvije kategorije (ženskom spolu (Ž) ili muškom spolu (M)), a prema tipu svoje krvne grupe jednoj od četiri kategorije (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 su 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 od 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 je **frekvencija** kategorije.
Neka varijabla, koju ćemo označiti X, ima k kategorija (recimo $k = 5$ znači da varijabla ima 5 kategorija - npr. klasična skala ocjena). Označimo pojedine kategorije kao $x_1, x_2, \ldots, x_k$, 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

$$f_i.$$

Frekvencija pojedine kategorije ovisi o broju izvršenih mjeranja, tj. dimenziji 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 s ukupnim brojem izmjerenih vrijednosti za ispitivanu varijablu, $i = 1, \ldots, k$. Ako je n dimenzija 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 dimenzije 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 za kategorije 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 za kategorije varijable krvna-grupa.

Primjer 3.3. Od velike važnosti u mnogim istraživanjima su i kategorizirane tablice frekvencija i relativnih frekvencija. Frekvencije i relativne frekvencije za izmjerene vrijednosti varijable i 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 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. U ovom primjeru naučit ćemo kako bazu podataka te tablice frekvencija i relativnih frekvencija napraviti u programskom paketu Statistica. Rezultat postupka u Statistici prikazan je za varijable krvna-grupa i spol iz primjera 3.1. Tablične prikaze frekvencija i relativnih frekvencija u programskom paketu Statistica možemo dobiti provodeći sljedeći postupak (kojeg provodimo slijedeći navedeni niz opcija u izborniku software-a):

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

Rezultat provedbe prethodnog postupka su tablice prikazane slikom 3.1.

	Frequency table: krvna grupa (KrvnaGrupa_Spol.sta)			
Category	Count	Cumulative Count	Percent	Cumulative Percent
A	3	3	30,00000	30,0000
B	3	6	30,00000	60,0000
AB	2	8	20,00000	80,0000
O	2	10	20,00000	100,0000
Missing	0	10	0,00000	100,0000

(a) krvna grupa

	Frequency table: spol (KrvnaGrupa_Spol.sta)			
Category	Count	Cumulative Count	Percent	Cumulative Percent
Ž	6	6	60,00000	60,0000
M	4	10	40,00000	100,0000
Missing	0	10	0,00000	100,0000

(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="O" za krvnu grupu O) → OK.

Rezultat provedbe prethodnog postupka su tablice prikazane slikom 3.2.

	Frequency table: spol (KrvnaGrupa_Spol.sta) Include condition: krvna_grupa="A"			
Category	Count	Cumulative Count	Percent	Cumulative Percent
Ž	2	2	66,66667	66,6667
M	1	3	33,33333	100,0000
Missing	0	3	0,00000	100,0000

(a) kategorija: krvna grupa A

	Frequency table: spol (KrvnaGrupa_Spol.sta) Include condition: krvna_grupa="B"			
Category	Count	Cumulative Count	Percent	Cumulative Percent
Ž	2	2	66,66667	66,6667
M	1	3	33,33333	100,0000
Missing	0	3	0,00000	100,0000

(b) kategorija: krvna grupa B

	Frequency table: spol (KrvnaGrupa_Spol.sta) Include condition: krvna_grupa="AB"			
Category	Count	Cumulative Count	Percent	Cumulative Percent
Ž	1	1	50,00000	50,0000
M	1	2	50,00000	100,0000
Missing	0	2	0,00000	100,0000

(c) kategorija: krvna grupa AB

	Frequency table: spol (KrvnaGrupa_Spol.sta) Include condition: krvna_grupa="O"			
Category	Count	Cumulative Count	Percent	Cumulative Percent
Ž	1	1	50,00000	50,0000
M	1	2	50,00000	100,0000
Missing	0	2	0,00000	100,0000

(d) kategorija: krvna grupa 0

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

Zadatak 3.1. (graf.sta, 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* gastr-S, somat-S *i* somat-Z *sadrže izmjerene koncentracije određenih enzima u krvi 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), a varijabla* dijagnoza *sadrži dijagnozu 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, a varijabla* stupanj *stupnjevanje rezultata provedenih testova s obzirom na dobnu skupinu kojoj ispitanih 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* skupina *iz baze podataka* (nalaz.sta) *prikazane su slikom 3.3.*

	Frequency table: dijagnoza (hormon.STA)			
Category	Count	Cumulative Count	Percent	Cumulative Percent
G	21	21	25,60976	25,6098
E b	4	25	4,87805	30,4878
U b	30	55	36,58537	67,0732
U z	13	68	15,85366	82,9268
E z	14	82	17,07317	100,0000
Missing	0	82	0,00000	100,0000

(a) varijabla dijagnoza (hormon.sta)

	Frequency table: skupina (Nalaz.sta)			
Category	Count	Cumulative Count	Percent	Cumulative Percent
g1: g1	10	10	9,80392	9,8039
g2: g2	5	15	4,90196	14,7059
g3: g3	15	30	14,70588	29,4118
g4: g4	11	41	10,78431	40,1961
g5: g5	11	52	10,78431	50,9804
g6: g6	9	61	8,82353	59,8039
g7: g7	9	70	8,82353	68,6275
g8: g8	11	81	10,78431	79,4118
g9	21	102	20,58824	100,0000
Missing	0	102	0,00000	100,0000

(b) varijabla skupina (nalaz.sta)

Slika 3.3: Frekvencije i relativne frekvencije svih kategorija varijabli dijagnoza i skupina

3.1.2 Grafički prikazi frekvencija i relativnih frekvencija

Frekvencije i relativne frekvencije kategorija kvalitativnih varijabli grafički prikazujemo pomoću **histograma frekvencija** i **histograma relativnih frekvencija**. U istu svrhu može se koristiti i **strukturirani krug** frekvencija i relativnih frekvencija (strukturirani krug se često naziva **kružni dijagram**, a popularni naziv za isti grafički prikaz je "pita").

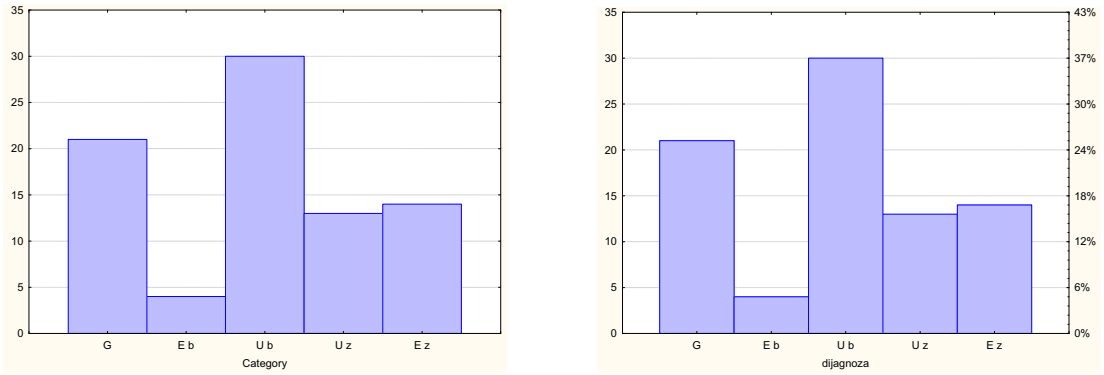
Primjer 3.5. (hormon.sta)
Grafičke prikaze frekvencija i relativnih frekvencija kvalitativnih varijabli prikazat ćemo na primjeru varijable dijagnoza iz baze podataka hormon.sta (koja je opisana u zadatku 3.1). Histogram frekvencija u programskom paketu Statistica možemo dobiti provodeći sljedeći postupak:

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

Histogram 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.

Histogrami frekvencija i relativnih frekvencija kategorija varijable dijagnoza prikazani su slikom 3.4.
Drugi način grafičkog prikazivanja mjera zastupljenosti pojedinih kategorija neke kvalitativne varijable u uzorku su strukturirani krugovi frekvencija i relativnih frekvencija koje u programskom paketu Statistica možemo dobiti provodeći sljedeći postupak:

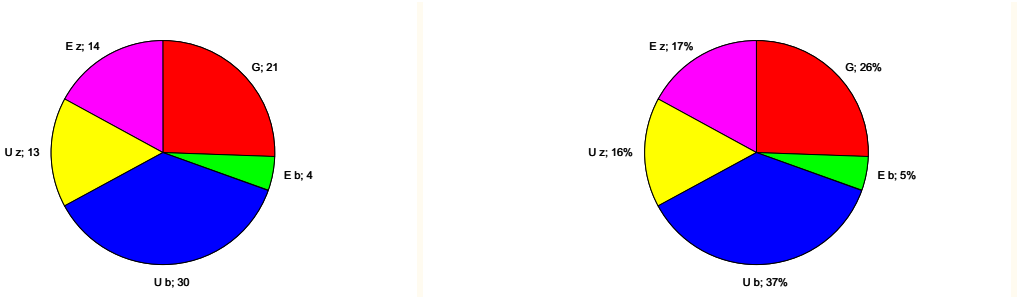


(a) frekvencije (b) relativne frekvencije

Slika 3.4: Histogrami svih kategorija varijable dijagnoza.

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.

Strukturirani krugovi frekvencija i relativnih frekvencija kategorija varijable dijagnoza prikazani su slikom 3.5.



(a) frekvencije (b) relativne frekvencije

Slika 3.5: Strukturirani krugovi svih kategorija varijable dijagnoza.

Zadatak 3.2. (nalaz.sta)
U bazi podataka nalaz.sta *(opisanoj u zadatku 3.1) odredite frekvencije i relativne frekvencije svih kategorija za varijable koje smatrate kvalitativnima.*

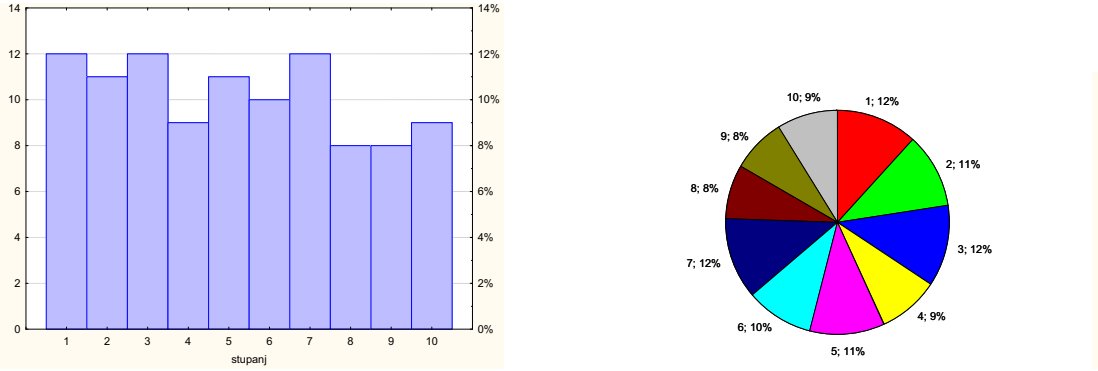
a) Rezultate prikažite grafički koristeći programski paket Statistica.

b) Za koliko ispitanika je 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 a) odredite pripadne relativne frekvencije.

Rješenje.

a) Grafički prikazi frekvencija i relativnih frekvencija kategorija kvalitativne varijable stupanj *prikazani su slikom 3.6.*



Slika 3.6: 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 3.3. (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*), a varijabla* konvulzije *informaciju o konvulzijama kod novorođenčeta (N - konvulzija nije bilo, D - konvulzije su bile prisutne). Odredite frekvencije i relativne frekvencije svih kategorija za varijable koje smatrate kvalitativnima.*

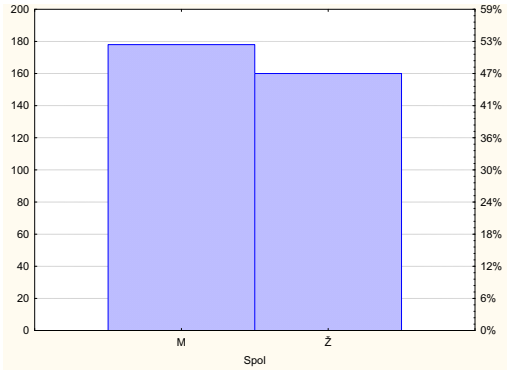
a) Rezultate prikažite tablično i grafički koristeći programski paket Statistica*.*

b) Broji li ovaj uzorak više djevojčica ili dječaka?

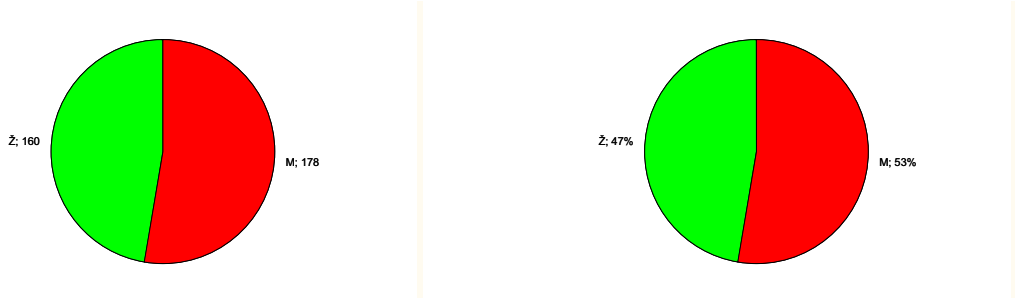
Rješenje.

a) Tablični i grafički prikazi frekvencija i relativnih frekvencija svih kategorija varijable spol *prikazani su na slikama 3.7 i 3.8.*

	Frequency table: Spol (bebe.sta)			
Category	Count	Cumulative Count	Percent	Cumulative Percent
M	178	178	52,19941	52,1994
Ž	160	338	46,92082	99,1202
Missing	3	341	0,87977	100,0000



Slika 3.7: Tablica i histogram frek. i rel. frek. svih kategorija varijable spol.



(a) frekvencije (b) relativne frekvencije

Slika 3.8: Strukturirani krugovi za varijablu spol.

b) Uzorkom je obuhvaćeno 341 *novorođenče, od čega za njih troje nije zabilježen spol. U uzorku od* 338 *novorođenčadi za koje znamo informaciju o spolu ima* 160 *djevojčica i* 178 *dječaka. Pripadne relativne frekvencije su* $160/341 \approx 46.92\%$ *za djevojčice i* $178/341 \approx 53.08\%$ *za dječake. Dakle, u uzorku ima više dječaka.*

Zadatak 3.4. (navike.sta)

U bazi podataka navike.sta *(opisanoj u zadatku 2.2) odredite frekvencije i relativne frekvencije svih kategorija za varijable koje smatrate kvalitativnima.*

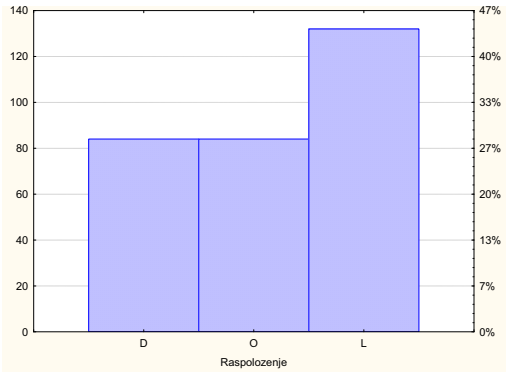
a) Rezultate prikažite tablično i grafički koristeći programski paket Statistica.

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

Rješenje.

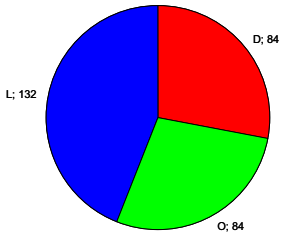
a) Tablični i grafički prikazi frekvencija i relativnih frekvencija svih kategorija varijable raspolozenje *prikazani su na slikama 3.9 i 3.10.*

	Frequency table: Raspolozenje (navike.sta)			
Category	Count	Cumulative Count	Percent	Cumulative Percent
D	84	84	28,00000	28,0000
O	84	168	28,00000	56,0000
L	132	300	44,00000	100,0000
Missing	0	300	0,00000	100,0000

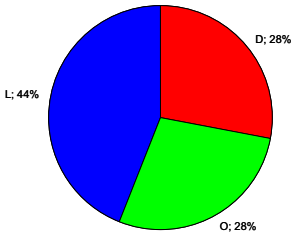


Slika 3.9: Tablica i histogram frek. i rel. frek. svih kategorija varijable raspolozenje.

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.*



(a) frekvencije



(b) relativne frekvencije

Slika 3.10: Strukturirani krugovi za varijablu raspolozenje.

3.2 Metode opisivanja numeričkih podataka

Numerički podaci mogu dolaziti iz diskretne varijable, tj. varijabla se, po svom tipu, može realizirati samo s konačno ili prebrojivo mnogo vrijednosti, ili iz kontinuirane varijable.

Primjer 3.6. (hormon.sta, komarci.sta, matematika.sta)

U bazi podataka hormon.sta (opisanoj u zadatku 3.1) sve numeričke varijable (gastr-S, somat-S, Somat-Z) su kontinuirane.

U bazi podataka komarci.sta (opisanoj u zadatku 3.1) varijable brojM i brojZ su diskretne numeričke varijable, a varijable temperatura i rel-vlaznost kontinuirane numeričke varijable.

U bazi podataka matematika.sta (opisanoj u zadatku 2.12) varijable tezina-kolegija i materijali su diskretne numerička varijable, a varijabla prosjek je kontinuirana numerička varijabla.

Ako su numeričke varijable diskretne, za opis izmjerenih vrijednosti tih varijabli možemo koristiti iste metode kao pri opisivanju kvalitativnih podataka, tj. frekvencije i relativne frekvencije, te ih grafički prkazati histogramima i strukturiranim krugovima.

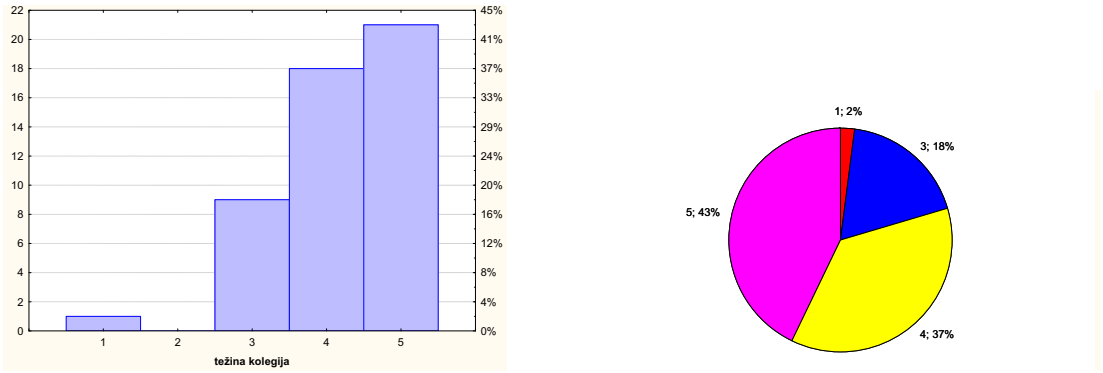
Primjer 3.7. (matematika.sta)
Tablični i grafički prikazi (histogram i strukturirani krug) frekvencija i relativnih frekvencija svih vrijednosti diskretne numeričke varijable tezina-kolegija prikazani su slikama 3.11 i 3.12.

	Frequency table: **težina kolegija** (anketa.sta)			
Category	Count	Cumulative Count	Percent	Cumulative Percent
1	1	1	2,04082	2,0408
3	9	10	18,36735	20,4082
4	18	28	36,73469	57,1429
5	21	49	42,85714	100,0000
Missing	0	49	0,00000	100,0000

Slika 3.11: Tablica frekvencija i relativnih frekvencija za varijablu tezina-kolegija.

Iz prikazanih opisa varijable tezina-kolegija možemo dobiti informacije sljedećeg tipa:

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



Slika 3.12: Grafički prikazi frekvencija i relativnih frekvencija za varijablu tezina-kolegija.

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.

Primjer 3.8. (zdravlje.sta)

Često ima smisla promatrati frekvencije i relativne frekvencije numeričkih varijabli odvojeno po pojedinim kategorijama neke kvalitativne varijable. Na primjer, korisno je analizirati određene zdravstvene karakteristike posebno za osobe ženskog, a posebno za osobe muškog spola. Stoga ćemo u ovom primjeru prikazati takve kategorizirane tablične i grafičke prikaze frekvencija i relativnih frekvencija diskretne numeričke varijable zdravlje po kategorjama kvalitativne varijable spol i iz baze podataka zdravlje.sta koja je opisana u zadatku 2.2.

Prvo ćemo tablično i grafički prikazati frekvencije i relativne frekvencije za podatke sadržane u varijablama zdravlje i spol (slike 3.13, 3.14 i 3.15).

	Frequency table: spol (zdravlje.sta)			
Category	Count	Cumulative Count	Percent	Cumulative Percent
Z: žena	11	11	22,00000	22,0000
M: muškarac	39	50	78,00000	100,0000
Missing	0	50	0,00000	100,0000

(a) varijabla spol

	Frequency table: zdravlje (zdravlje.sta)			
Category	Count	Cumulative Count	Percent	Cumulative Percent
1	4	4	8,00000	8,0000
2	8	12	16,00000	24,0000
3	18	30	36,00000	60,0000
4	12	42	24,00000	84,0000
5	8	50	16,00000	100,0000
Missing	0	50	0,00000	100,0000

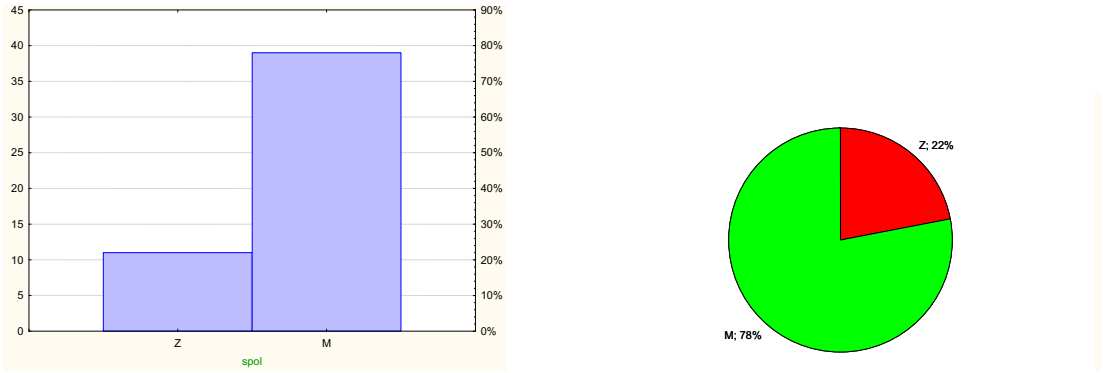
(b) varijabla zdravlje

Slika 3.13: Tablice frekvencija i relativnih frekvencija.

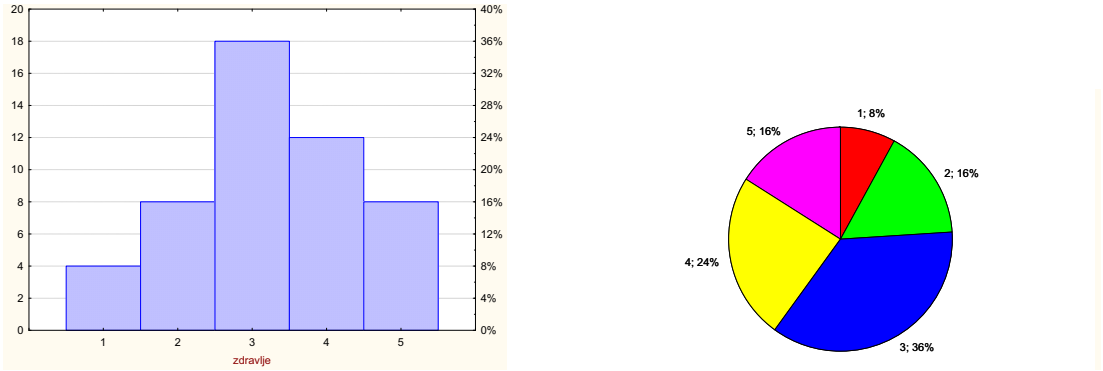
Tablični i grafički prikazi podataka sadržanih u varijabli zdravlje posebno za kategoriju ispitanika ženskog spola, a posebno za kategoriju ispitanika muškog spola prikazani su slikama 3.16, 3.17 i 3.18. Strukturirane krugove relativnih frekvencija sa slike 3.18 u programskom paketu Statistica možemo 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).

Radi usporedbe rezultata po spolu korisno je histograme frekvencija i relativnih frekvencija podataka sadržanih u varijabli zdravlje kategoriziranih prema spolu ispitanika prikazati na jednoj slici, tj. grafu (slika 3.19).



Slika 3.14: Grafički prikazi frek. i rel. frek. svih kategorija varijable spol.



Slika 3.15: Grafički prikazi frek. i rel. frek. podataka iz varijable zdravlje.

Category	Frequency table: zdravlje (zdravlje.sta) Include condition: spol="Z"			
	Count	Cumulative Count	Percent	Cumulative Percent
1	1	1	9,09091	9,0909
2	2	3	18,18182	27,2727
3	5	8	45,45455	72,7273
4	2	10	18,18182	90,9091
5	1	11	9,09091	100,0000
Missing	0	11	0,00000	100,0000

(a) žene (spol=Z)

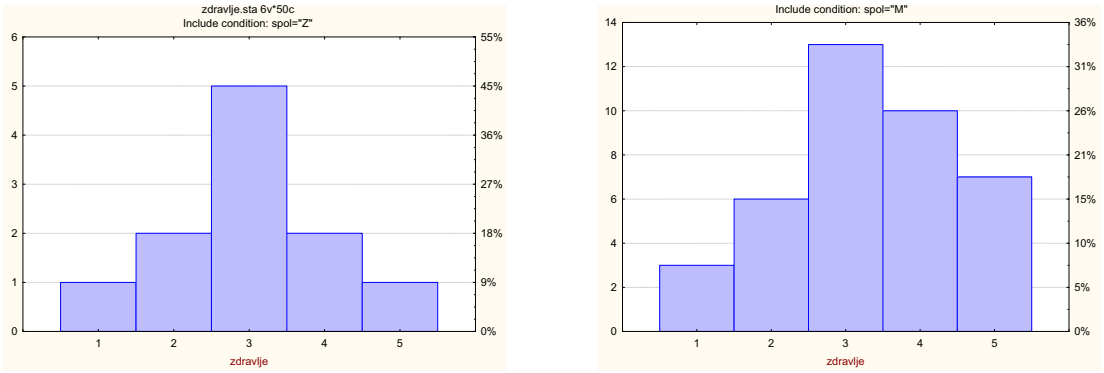
Category	Frequency table: zdravlje (zdravlje.sta) Include condition: spol="M"			
	Count	Cumulative Count	Percent	Cumulative Percent
1	3	3	7,69231	7,6923
2	6	9	15,38462	23,0769
3	13	22	33,33333	56,4103
4	10	32	25,64103	82,0513
5	7	39	17,94872	100,0000
Missing	0	39	0,00000	100,0000

(b) muškarci (spol=M)

Slika 3.16: Tablice frek. i rel. frek. za podatke iz varijable zdravlje kategorizirane prema spolu ispitanika.

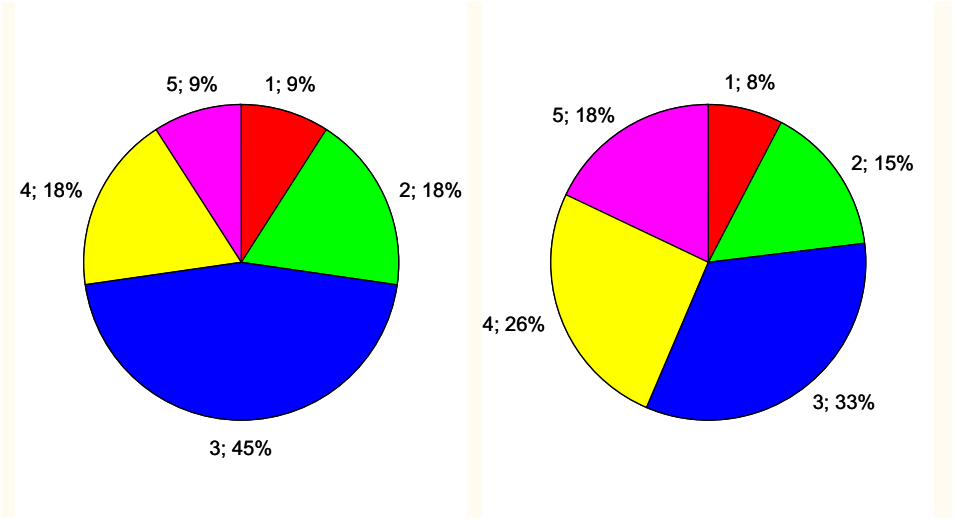
Objedinjene histogramske prikaze frekvencija i relativnih frekvencija neke numeričke varijable čije su vrijednosti kategorizirane po nekom kriteriju možemo dobiti u programskom paketu Statistica provodeći sljedeći postupak:

Graphs → Categorized Graphs → Histograms → Variables (Variable - zdravlje, X-Category - spol) → Layout (Separate - za odvojene histograme 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 histogramu)



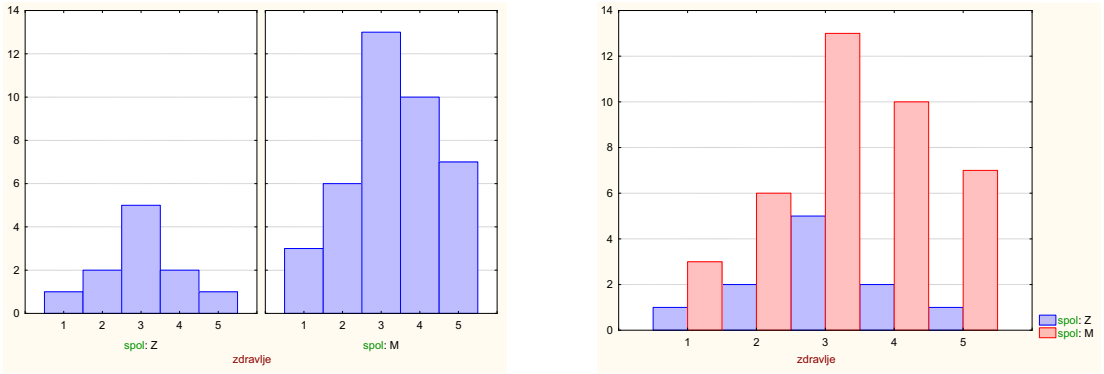
(a) žene (spol=Z) (b) muškarci (spol=M)

Slika 3.17: Histogrami frek. i rel. frek. za podatke iz varijable zdravlje kategorizirane prema spolu ispitanika.



žene (spol=Z) muškarci (spol=M)

Slika 3.18: Strukturirani krugovi rel. frek. za podatke iz varijable zdravlje kategorizirane prema spolu ispitanika.



Slika 3.19: Histogrami frek. i rel. frek. za podatke iz varijable zdravlje kategorizirane prema spolu ispitanika.

Zadatak 3.5. (TV-program.sta)

Za kvalitativne i diskretne numeričke varijable iz baze podataka TV-program.sta *napravite sljedeće*

tablične i grafičke prikaze:

a) *tablice i histograme frekvencija i relativnih frekvencija za podatke sadržane u varijablama* spol *i* P1*,*

b) *tablice i histograme frekvencija i relativnih frekvencija za podatke sadržane u varijabli* P1 *posebno za kategoriju ispitanika ženskog spola, a posebno za kategoriju ispitanika muškog spola,*

c) *zajednički histogram frekvencija i relativnih frekvencija svih podataka sadržanih u varijabli* P1 *kategoriziran prema spolu ispitanika,*

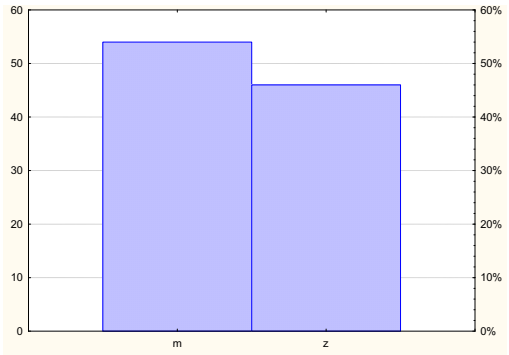
d) *kružne dijagrame frekvencija i relativnih frekvencija za podatke sadržane u varijablama* spol *i* P3*,*

e) *kružne dijagrame 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 za sve kategorije varijabli* spol *i* P1 *prikazani su slikama 3.20 i 3.21, redom.*

	Frequency table: spol (TV_program.sta)			
Category	Count	Cumulative Count	Percent	Cumulative Percent
m	54	54	54,00000	54,0000
z	46	100	46,00000	100,0000
Missing	0	100	0,00000	100,0000

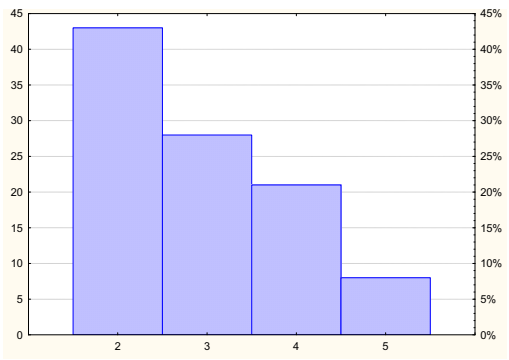


Slika 3.20: Tablica i histogram frek. i rel. frek. za sve kategorije varijable spol.

b) *Tablični i grafički prikazi frekvencija i relativnih frekvencija za sve kategorije varijable* P1 *kategorizirane prema spolu ispitanika prikazani su slikama 3.22 i 3.23.*

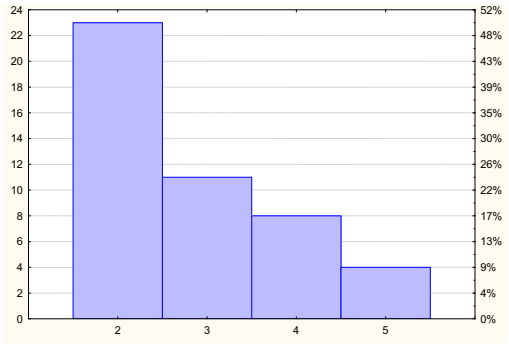
c) *Zajednički prikazi histograma frekvencija i relativnih frekvencija svih podataka sadržanih u varijabli* P1 *kategoriziranih prema spolu ispitanika prikazani su slikom 3.24.*

	Frequency table: HRT1 (TV_program.sta)			
Category	Count	Cumulative Count	Percent	Cumulative Percent
2	43	43	43,00000	43,0000
3	28	71	28,00000	71,0000
4	21	92	21,00000	92,0000
5	8	100	8,00000	100,0000
Missing	0	100	0,00000	100,0000



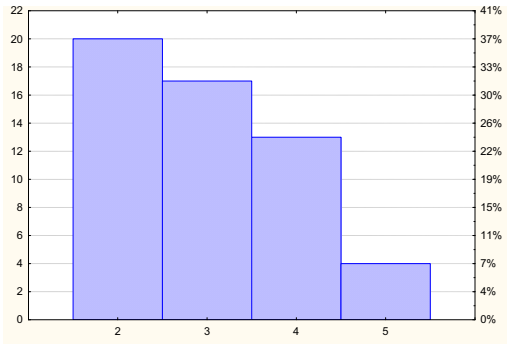
Slika 3.21: Tablica i histogram frek. i rel. frek. za sve vrijednosti varijable P1.

Category	Frequency table: HRT1 (TV_program.sta) Include condition: spol="z" Count	Cumulative Count	Percent	Cumulative Percent
2	23	23	50,00000	50,0000
3	11	34	23,91304	73,9130
4	8	42	17,39130	91,3043
5	4	46	8,69565	100,0000
Missing	0	46	0,00000	100,0000

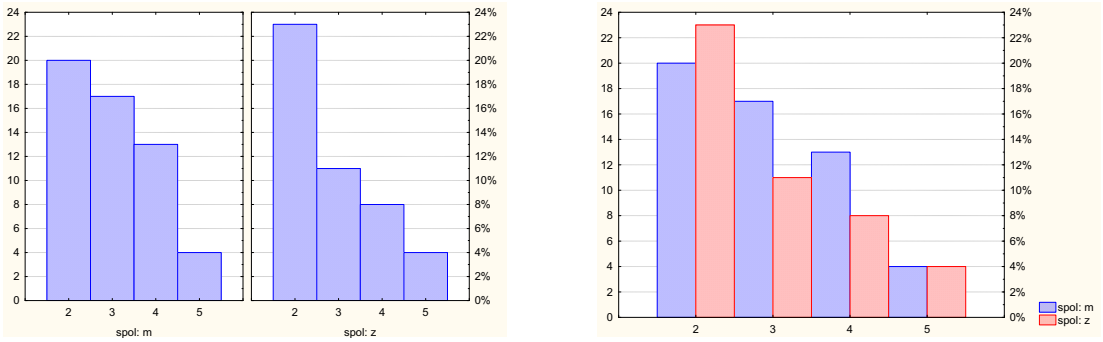


Slika 3.22: Tablica i histogram frek. i rel. frek. za sve vrijednosti varijable P1 za ženski spol.

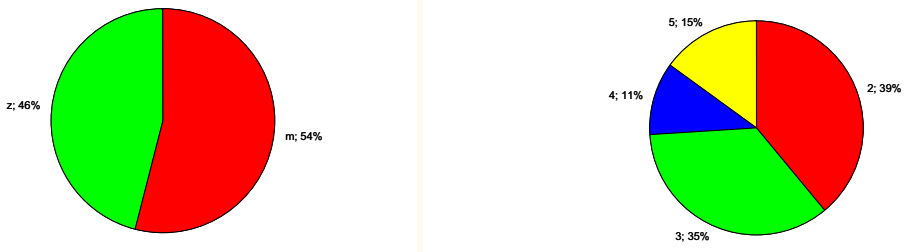
Category	Frequency table: HRT1 (TV_program.sta) Include condition: spol="m" Count	Cumulative Count	Percent	Cumulative Percent
2	20	20	37,03704	37,0370
3	17	37	31,48148	68,5185
4	13	50	24,07407	92,5926
5	4	54	7,40741	100,0000
Missing	0	54	0,00000	100,0000



Slika 3.23: Tablica i histogram frek. i rel. frek. za sve vrijednosti varijable P1 za muški spol.

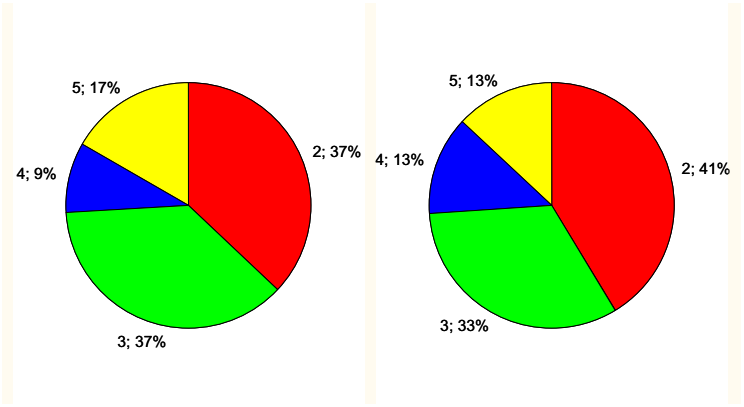


Slika 3.24: Histogrami frek. i rel. frek. za sve vrijednosti varijable HRT1kategorizirane prema spolu ispitanika.



(a) varijablaspol (b) varijablaP3

Slika 3.25: Strukturirani krugovi frek. i rel. frek. podataka iz varijabli spol i P3.



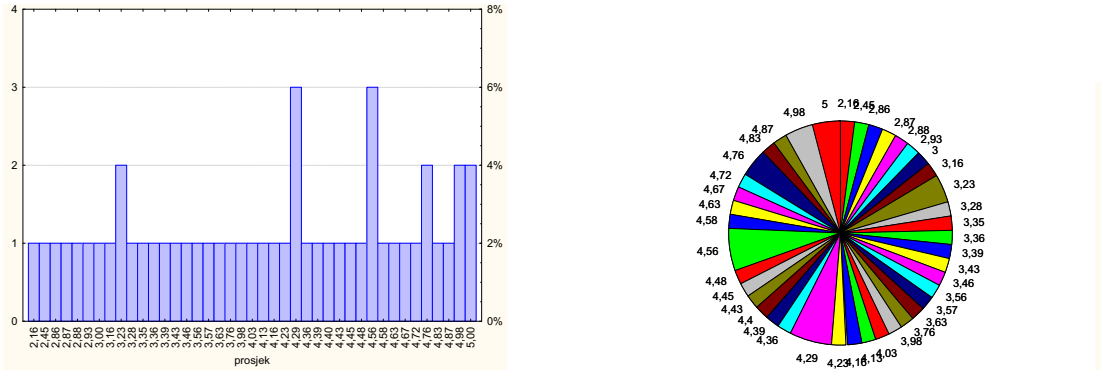
Slika 3.26: Strukturirani krugovi rel. frek. za podatke iz varijable P3 kategorizirane prema spolu ispitanika.

d) Strukturirani krugovi frekvencija i relativnih frekvencija podataka iz varijabli spol *i* P3 *prikazani su slikom 3.25.*

e) Strukturirani krugovi relativnih frekvencija za podatke iz varijable P3 *kategorizirane prema spolu ispitanika prikazani su slikom 3.26.*

Ako numerička varijabla nije diskretna, za prikazivanje skupa izmjerenih vrijednosti neće nam puno pomoći frekvencije, histogrami i strukturirani krugovi napravljeni na osnovu svake pojedine izmjerene vrijednosti.
Razlog tome ilustrirat ćemo sljedećim primjerom.

Primjer 3.9. (matematika.sta) Histogram i strukturirani krug za izmjerene vrijednosti kontinuirane numeričke varijable prosjek iz baze podataka matematika.sta (koja je opisana u primjeru 2.12) prikazani su slikom 3.27. Pri opisivanju ove varijable pretpostavili smo da su sve različite izmjerene vrijednosti varijable prosjek zasebne kategorije. Zbog velikog broja različitih izmjerenih vrijednosti broj kategorija je prevelik pa takvi grafički prikazi najčešće ne daje željene informacije.



Slika 3.27: Grafički prikazi frek. i rel. frek svih izmjerenih vrijednosti varijable prosjek.

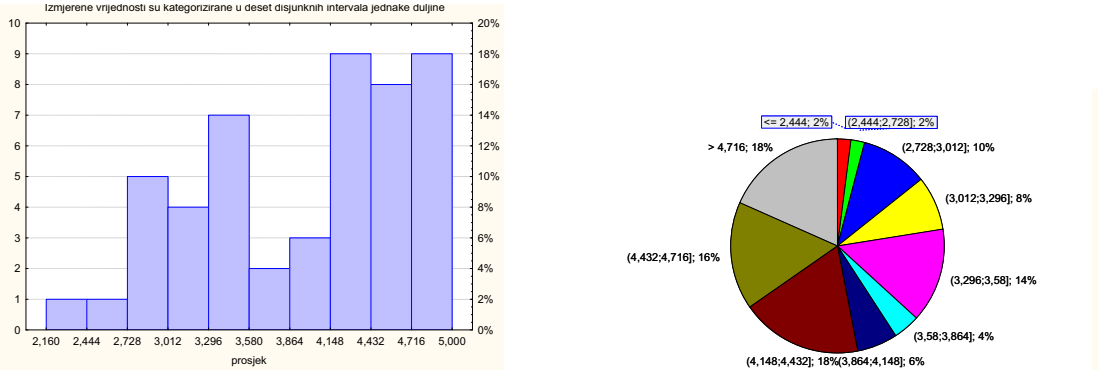
U svrhu dobivanja preglednih i korisnih histograma i strukturiranih krugova za podatke iz kontinuiranih numeričkih varijabli izmjerene vrijednosti ćemo **kategorizirati**, tj. razvrstati izmjerene vrijednosti u određene kategorije. Ako veliki skup podataka

kategoriziramo (podijelimo) u nekoliko disjunktnih intervala po kriteriju za koji smatramo da će nam dati željene rezultate, tablični i grafički prikazi frekvencija i relativnih frekvencija postaju pregledniji i informativniji.

3.2.1 Postupak razvrstavanja numeričkih podataka u kategorije

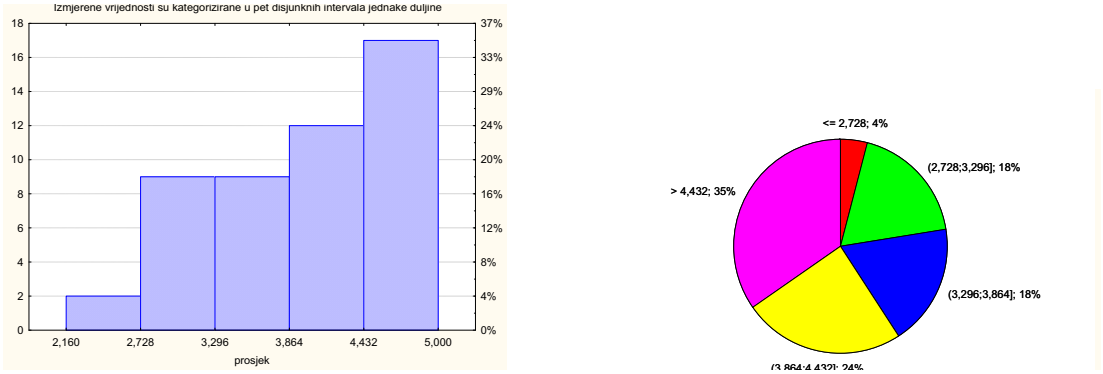
Razvrstavanje izmjerenih vrijednosti kontinuirane numeričke varijable u kategorije moguće je napraviti na mnogo načina. Npr. moguće je skup svih mjerenih vrijednosti (ili nešto veći skup koji sadrži skup svih mjerenih vrijednosti ali kojega je jednostavnije podijeliti na jednake dijelove) podijeliti na disjunktne intervale jednake duljine. Međutim, nije nužno da su intervali uvijek jednake duljine. Nema točno definiranog pravila po kojemu bi trebalo definirati duljine intervala niti nihov broj, ali je jasno da ih ne smije biti niti previše niti premalo da bi cijeli postupak imao smisla i bio koristan za prikazivanje skupa mjerenih vrijednosti. Kriterij za kategorizaciju vrijednosti kontinuirane numeričke varijable treba biti temeljen na razumijevanju problema koji proučavamo, tj. podatke ćemo kategorizirati na način koji nam omogućava efikasno dobivanje potrebnih informacija.

Primjer 3.10. (matematika.sta) Primjerom 3.9 smo pokazali da je teško analizirati varijablu prosjek ako za kategorije uzmemo sve različite izmjerene vrijednosti te varijable. Stoga ćemo provesti kategorizaciju izmjerenih vrijednosti. Dva načina kategorizacije, tj. podjele izmjerenih vrijednosti u disjunkte intervale, rezultiraju histogramima i strukturiranim krugovima prikazanim na slikama 3.28 i 3.30.

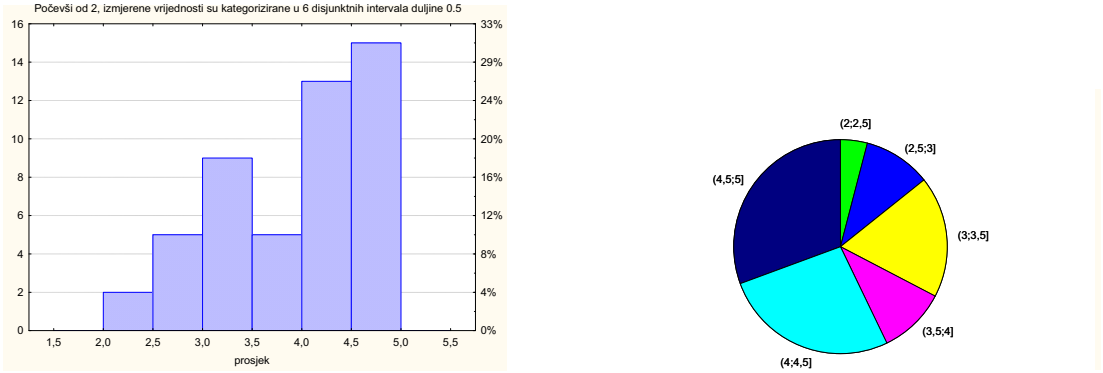


Slika 3.28: Histogram i strukturirani krug za izmjerene vrijednosti varijable prosjek razvrstane u 10 disjunktnih intervala.

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 izmjerene vrijednosti varijable prosjek možemo kategorizirati u šest disjunktnih intervala duljine 0.5, počevši od izmjerene vrijednosti 2.0. Tada iz grafičkih prikaza sa slike 3.30 očitavamo da takvih studenata ima 33, a relativna frekvencija je $33/49 \approx 67.35\%$.



Slika 3.29: Histogram i strukturirani krug za izmjerene vrijednosti varijable prosjek razvrstane u 5 disjunktnih intervala.



Slika 3.30: Histogram i strukturirani krug za izmjerene vrijednosti varijable prosjek razvrstane u 6 disjunktnih intervala počevši od 2.0.

Zadatak 3.6. (hormon.sta)

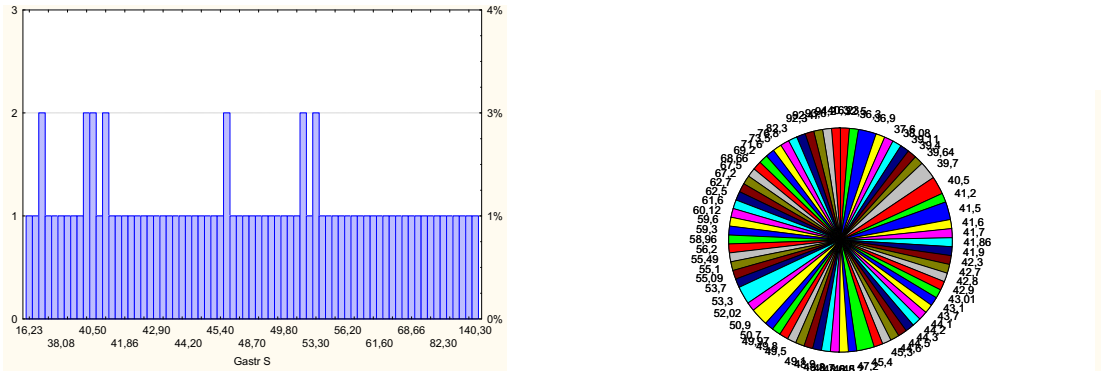
a) Odredite tablicu frekvencija i histogram za kontinuiranu numeričku varijablu gastr-S *iz baze podataka* hormon.sta *(koja je opisana u zadatku 3.1) tako da za kategorije uzmete sve međusobno različite izmjerene vrijednosti.*

b) Iskoristite izmjerene vrijednosti varijable gastr-S *te mijenjajte broj intervala na koji dijelite skup vrijednosti. Proučavajte što se događa i pribilježite vaš zaključak.*

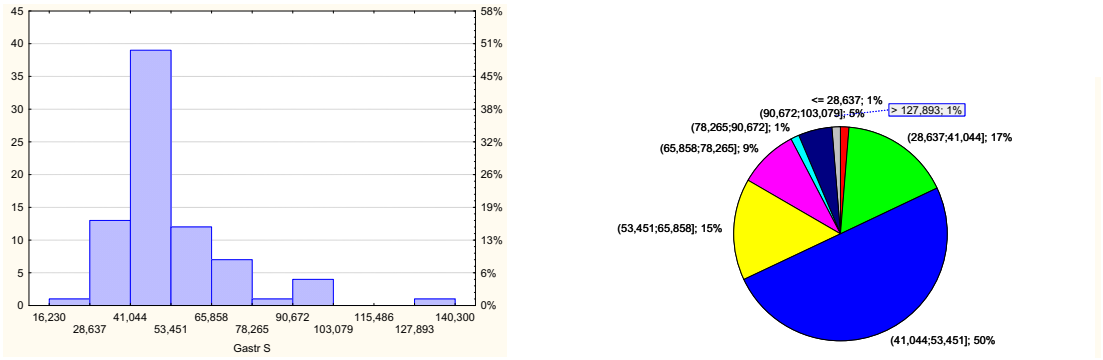
Rješenje.

a) Histogram frekvencija i relativnih frekvencija te strukturirani krug izmjerenih vrijednosti varijable gastr-S u kojima su kao kategorije uzete sve različite izmjerene vrijednosti prikazani su slikom 3.31.

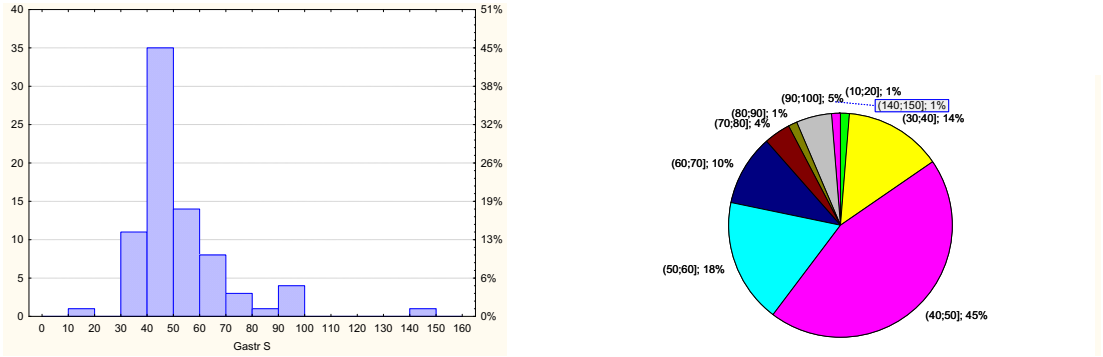
b) Kategorizacija izmjerenih vrijednosti varijable gastr-S na nekoliko disjunktnih intervala daje preglednije grafičke prikaze iz kojih je lakše nalizirati izmjerene vrijednosti i donijeti neke zaključke. Primjeri kategorizacije izmjerenih vrijednosti na 10 i 15 disjunktnih intervala prikazani su slikama 3.32 i 3.33, redom.



Slika 3.31: Histogram i struktirurani krug svih izmjerenih vrijednosti varijable gastr-S.



Slika 3.32: Hist. i str. krug izmjerenih vrij. varijable gastr-S razvrstanih u 10 disjunktnih intervala.



Slika 3.33: Hist. i str. krug izmjerenih vrij. varijable gastr-S razvrstanih u 15 disjunktnih intervala.

3.2.2 Mjere centralne tendencije i raspršenosti podataka

Karakteristika numeričkih varijabli je da među njihovim vrijednostima postoji prirodan uređaj. Na osnovu te činjenice možemo definirati numeričke karakteristike tih varijabli koje imaju logičnu interpretaciju i mogu se iskoristiti u cilju prikazivanja skupa mjerenih vrijednosti. U ovom poglavlju navodimo osnovne numeričke karakteristike te primjerima ilustriramo njihovu interpretaciju u praktičnim problemima.

Aritmetička sredina

Aritmetička sredina niza izmjerenih vrijednosti $x_1, x_2, \ldots, x_n$ varijable X definirana

je izrazom

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

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:

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

Obzirom da ih ima ukupno 9, aritmetička sredina (eng. mean) 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 ima značenje izmjerene vrijednosti koja se nalazi na sredini niza podataka kada je on uređen po veličini, tj. baram pola podataka je manje ili jednako medijanu, a istovremeno je barem pola podataka veće ili jednako od medijana. Način njegovog izračuna ovisi o tome imamo li **neparan** ili **paran** broj izmjerenih vrijednosti varijable. Ukoliko imamo **neparan broj** izmjerenih vrijednosti, onda postoji vrijednost koja je na srednjoj poziciji u uređenom skupu, 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:

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

Obzirom da ih ima ukupno 11, medijan je vrijednost koja je na šestoj poziciji u tako dobivenom nizu, tj. broj 2.

Ukoliko imamo **paran broj** izmjerenih vrijednosti varijable, onda ne postoji podatak koji je na srednjoj poziciji jer srednju poziciju "zauzimaju" dva podatka. Medijan se tada definira kao polovina između ta dva podatka (tj. aritmetička sredina 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:

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

Obzirom da ima 12 podataka, "sredinu" čine šesti i sedmi podatak, tj. vrijednosti 2 i 3. Medijan ovog skupa podataka je sredina ta dva broja, tj. medijan je $(2 + 3)/2 = 2.5$.

Postotna vrijednost, donji i gornji kvartil

Medijan odgovara pedeset postotnoj vrijednosti obzirom da je barem 50% podataka manje ili jednako od medijanu i barem 50% podataka veće ili jednako od medijana. Postotna vrijednost za neki izabrani broj $p \in \langle 0, 100\rangle$, označimo je x'_p, definirana je na temalju zahtjeva da je barem $p\%$ izmjerenih vrijednosti manje ili jednako x'_p, dok je barem $(100 - p)\%$ vrijednosti veće ili jednako x'_p. Dvadesetpet postotna vrijednost zove se donji kvartil, a sedamdesetpet postotna vrijednost zove se gornji kvartil. Analogno kao i kod računanja medijana, ako se na traženoj poziciji za računaje postotne vrijednosti nalaze dva podatka u uređenom skupu izmjerenih vrijednosti, postotnu vrijednost određujemo kao njihovu sredinu. Donji i gornji kvartil su mjere koje spadaju u grupu mjera raspršenosti podataka.

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%). 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 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) $x_{\min}$, a najveću $x_{\max}$.

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 $(\bar{x} - x_{\min})$ i $(x_{\max} - \bar{x})$, tj. broj

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

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

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

Maksimalno odstupanje izmjerenih vijednosti ove varijable od njihovog prosjeka 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:

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

a standardna devijacija je kvadratni korijen varijance, tj.

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

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, zaokružena na dvije decimale, 5.42. Varijanca ovog skupa podataka, zaokružena na dvije decimalne, je

$$s^2 = \frac{1}{9}\sum_{i=1}^{9}(x_i - 5.42)^2 = 8.86,$$

a standardna devijacija

$$s = \sqrt{\frac{1}{9}\sum_{i=1}^{9}(x_i - 5.42)^2} = 2.98.$$

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 vrijednosti 2 i 3 (točno četiri puta svaka). Dakle, mod ovog skupa podataka nije jedinstven. U programskom paketu Statistica za mod ovog skupa izmjerenih vrijednosti bi pisalo mod = multiple, te bismo u tom slučaju sve vrijednosti moda saznali analizom pripadne tablice frekvencija.

Korištenjem numeričkih karakteristika numeričkih varijabli skup mjerenih vrijednosti može se prikazati grafički pomoću **kutijastog dijagrama** (eng. box plot, boxplot ili box-and-whisker plot).

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

trgovački centar	1	2	3	4	5	6	7	8	9	10
cijena proizvoda	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) su tablice prikazane slikom 3.34.

	Descriptive Statistics (cijene_proitvoda.sta)						
Variable	Valid N	Mean	Mode	Frequency of Mode	Range	Variance	Std.Dev.
cijena proizvoda	10	49,66900	Multiple	1	20,00000	34,73377	5,893536

	Descriptive Statistics (cijene_proitvoda.sta)						
Variable	Valid N	Median	Minimum	Maximum	Lower Quartile	Upper Quartile	Range
cijena proizvoda	10	49,58000	39,99000	59,99000	45,52000	52,02000	20,00000

Slika 3.34: Deskriptivna statistika za cijene iz tablice 3.6.

Uočimo da mod nije jedinstven - naime sve su izmjerene vrijednosti međusobno različite, tj. svaka je vrijednost izmjerena točno jednom.
Za analiziranje raspršenosti cijena iz tablice 3.6 korisno je skicirati kutijasti dijagram na bazi medijana (slika 3.35) koji prikazuje odnos numeričkih karakteristika iz donje tablice sa slike 3.34 i kojeg u programskom paketu Statistica možemo napraviti provodeći sljedeći postupak:

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

Primjer 3.21. (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 profesora):

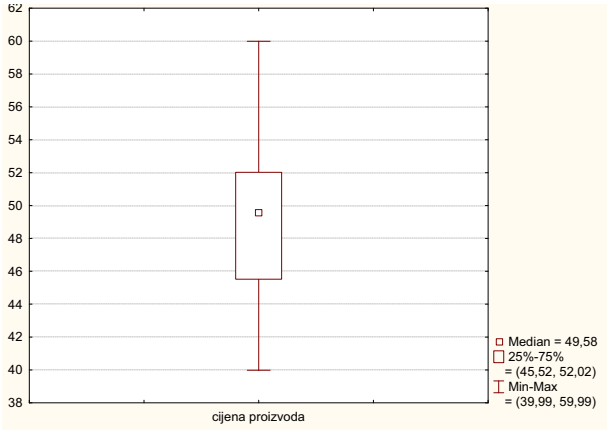
varijabla znanje sadrži ocjene znanja studenta o temi nastavnog sata,

varijabla literatura sadrži ocjene primjerenosti korištene literature za pripremu nastavnog sata,

varijabla predavac sadrži ocjene predavačevog stava i nastupa pred razredom,

varijabla atmosfera sadrži ocjene radne atmosfere na nastavnom satu,

varijabla govor sadrži ocjene studentovog izražavanja tijekom nastavnog sata,



Slika 3.35: Kutijasti dijagram an bazi medijana za cijene iz tablice 3.6.

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.

Sve varijable sadržane u ovoj bazi podataka su diskretne numeričke varijable. S ciljem zaključivanja o uspješnosti budućih profesora iz ove skupine u nastavi analizirajmo npr. varijablu ukupno. Numeričke karakteristike ove varijable prikazane su tablicom 3.36.

	Descriptive Statistics (ocjena.sta)						
Variable	Valid N	Mean	Mode	Frequency of Mode	Range	Variance	Std.Dev.
ukupno	62	8,112903	9,000000	19	6,000000	2,265732	1,505235

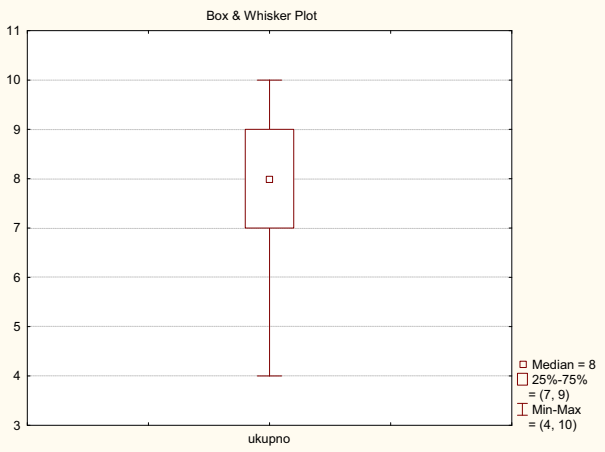
	Descriptive Statistics (ocjena.sta)					
Variable	Valid N	Median	Minimum	Maximum	Lower Quartile	Upper Quartile
ukupno	62	8,000000	4,000000	10,00000	7,000000	9,000000

Slika 3.36: Deskriptivna statistika za varijablu ukupno.

Uočimo da je najčešće (19) predavanje ocijenjeno visokom ocjenom 9 (ocjena 9 je mod ovog skupa izmjerenih vrijednosti) te da je prosječna ocjena predavanja 8.11. Analizu raspršenosti ocjena napravit ćemo pomoću kutijastog dijagrama (slika 3.37).
Analiza kutijastog dijagrama sugerira sljedeće zaključke: nitko od studenata nije dobio ocjenu nižu od četiri. Manje od 25% studenata je dobilo ocjenu 4, 5 ili 6. Zanimljivo je uočiti da je barem 75% predavanja ocijenjeno ocijenom 7 i više dok je barem 50% studenata postiglo ocjenu veću ili jendku 8.

Primjer 3.22. (matematika.sta)

Baza podataka matematika.sta (opisana u primjeru 2.12) sadrži rezultate ankete o kvaliteti izvođenja nekog matematičkog kolegija. Ukoliko nas zanima prilagođenost težine sadržaja kolegija predznanju studenata, analizirat ćemo varijablu tezina-kolegija. Mjere deskriptivne statistike za ovu varijablu prikazane su u slikom 3.38.



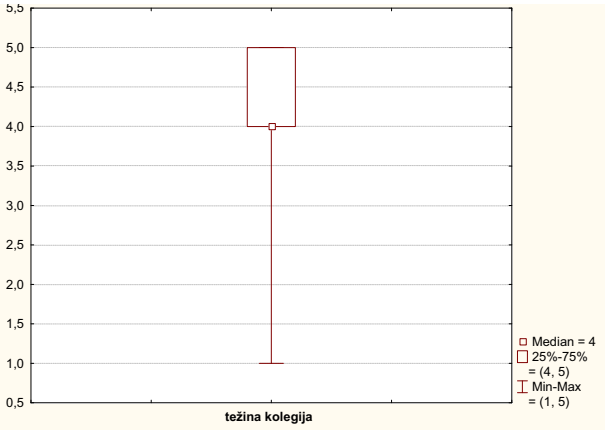
Slika 3.37: Kutijasti dijagram za varijablu ukupno.

Variable	Descriptive Statistics (anketa.sta)						
	Valid N	Mean	Mode	Frequency of Mode	Range	Variance	Std.Dev.
težina kolegija	49	4,183673	5,000000	21	4,000000	0,778061	0,882078

Variable	Descriptive Statistics (anketa.sta)					
	Valid N	Median	Minimum	Maximum	Lower Quartile	Upper Quartile
težina kolegija	49	4,000000	1,000000	5,000000	4,000000	5,000000

Slika 3.38: Deskriptivna statistika za varijablu tezina-kolegija.

Uočimo da je čak 21 ispitanik prilagođenost težine kolegija predznanju studenata ocijenio ocjenom 5 (ocjena 5 je mod ovog skupa izmjerenih vrijednosti) te da je prosječna ocjena 4.18. U svrhu analize raspršenosti ocjena koristimo kutijasti dijagram prikazan na slici 3.39.



Slika 3.39: Kutijasti dijagram za varijablu tezina-kolegija.

Analizom kutijastog dijagrama donosimo sljedećei zaključak: manje od 25% ispitanika je prilagođenost težine kolegija ocijenilo ocjenama 1, 2 ili 3 dok je barem 75% ispitanika prilagođenost težine kolegija ocijenilo ocjenama 4 ili 5.

3.2.3 Detekcija stršećih vrijednosti

Podatak koji je značajno veći ili manji u odnosu na druge izmjerene vrijednosti jedne varijable nazivamo stršeći podatak (engl. outlier). Pojavljivanje stršećih podataka najčešće je vezano uz jedan od sljedećih razloga:

- podatak je netočno izmjeren ili krivo unesen u bazu podataka,
- podatak dolazi iz druge populacije (ne iz populacije koju promatramo u kontekstu problema kojeg 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 outlierom jer se radi o vrlo visokoj koncentraciji glukoze koja se rijetko pojavljuje.

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

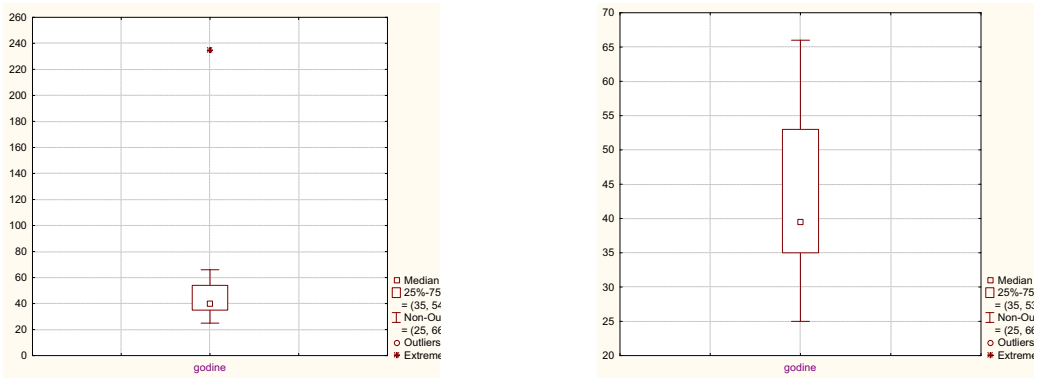
Graphs → 2D Graphs → BoxPlots → Variables → Advanced → pod Whisker odabrati "Non-outlier range" → pod Outliers odabrati "Outl. & Extremes" → OK.

Primjer 3.23. (zdravlje.sta) Baza podataka zdravlje.sta sadrži neke zdravstvene podatke za 51 ispitanika. Kratkom analizom mjera deskriptivne statistike za varijablu godine možemo uočiti da je maksimum skupa izmjerenih vrijednosti 235, što u ovom primjeru znači da naš najstariji ispitanik ima 235 godina (slika 3.40).

	Descriptive Statistics (zdravlje.sta)								
Variable	Valid N	Mean	Median	Mode	Frequency of Mode	Minimum	Maximum	Lower Quartile	Upper Quartile
godine	51	46.60784	40.00000	39.00000	7	25.00000	235.0000	35.00000	54.00000

Slika 3.40: Deskriptivna statistika za varijablu godine.

Taj podatak je stršeća vrijednost ovog skupa izmjerenih vrijednosti. 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 3.41 prikazan je kutijasti dijagram za varijablu godine sa stršećom vrijednišću te kutijasti dijagram koji dobivamo ako iz skupa izmjerenih vrijednosti uklonimo stršeći podatak.
Uklanjanjem stršeće vrijednosti mijenjaju se i vrijednosti mjera deskriptivne statistike. Iz tablica sa slike 3.42 vidimo da su se uklanjanjem stršećeg podatka iz skupa izmjerenih vrijednosti aritmetička sredina (mean) i gornji kvartil smanjili, dok su mod medijan i donji kvartil ostali isti. Općenito, uklanjanjem stršećih podataka mod će najčešće ostati nepromijenjen.



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

Slika 3.41: Kutijasti dijagrami za varijablu godine.

Variable	Descriptive Statistics (zdravlje.sta) Include condition: godine<230								
	Valid N	Mean	Median	Mode	Frequency of Mode	Minimum	Maximum	Lower Quartile	Upper Quartile
godine	50	42.84000	39.50000	39.00000	7	25.00000	66.00000	35.00000	53.00000

Slika 3.42: Deskriptivna statistika za varijablu godine nakon uklanjanja stršeće vrjednosti.

Zadatak 3.7. (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 zaista stršeća 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 zanemarivanja identificirane stršeće vrijednosti?*

Rješenje.

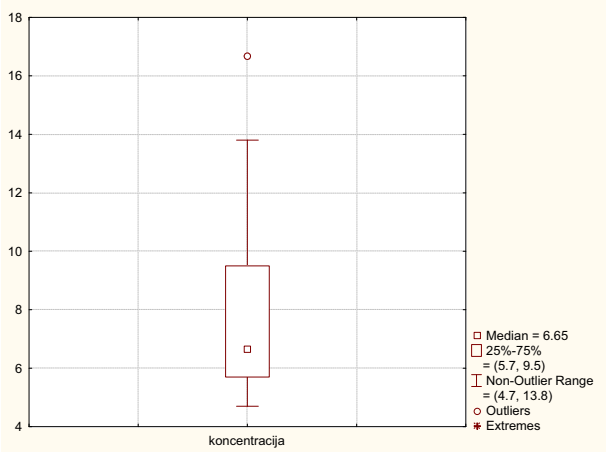
a) Deskriptivna statistika i kutijasti dijagram s označenim stršećim vrijednostima skupa izmjerenih vrijednosti varijable koncentracija *prikazani su na slici 3.43.*

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

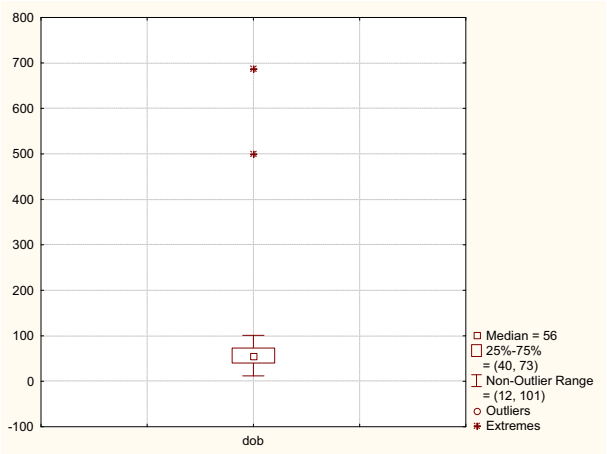
b) Deskriptivna statistika i kutijasti dijagram s označenim stršećim vrijednostima skupa izmjerenih vrijednosti varijable dob *prikazani su na slici 3.44.*

Statistica je kao stršeće vrijednosti među izmjerenim vrijednostima varijable dob *detektirala podatke* 500 *i* 688. *Uklanjanjem tih stršećih podataka dolazi do smanjenja aritmetičke sredine (mean) i medijana izmjerenih vrijednosti.*

Variable	Descriptive Statistics (glukoza.sta)							
	Valid N	Mean	Median	Mode	Frequency of Mode	Minimum	Maximum	Lower Quartile
koncentracija	102	7.697059	6.650000	5.500000	14	4.700000	16.70000	5.700000



Slika 3.43: Deskriptivna statistika za varijablu koncentracija.



Variable	Descriptive Statistics (glukoza.sta)								
	Valid N	Mean	Median	Mode	Frequency of Mode	Minimum	Maximum	Lower Quartile	Upper Quartile
dob	102	66.72549	56.00000	Multiple	4	12.00000	688.0000	40.00000	73.00000

(a) uključene stršeće vrijednosti

Variable	Descriptive Statistics (glukoza.sta) Include condition: dob<110								
	Valid N	Mean	Median	Mode	Frequency of Mode	Minimum	Maximum	Lower Quartile	Upper Quartile
dob	100	56.18000	55.50000	Multiple	4	12.00000	101.0000	40.00000	71.50000

(b) uklonjene stršeće vrijednosti

Slika 3.44: Kutijasti dijagram i deskriptivna statistika za varijablu dob.

3.3 Zadaci za vježbu

Zadatak 3.8. (tlak.sta) Baza podataka tlak.sta sadrži podatke o krvnom tlaku za ispitanike jedne ankete:

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 odgovorite na sljedeća pitanja:

a) Odredite tablice frekvencija i relativnih frekvencija, nacrtajte i proanalizirajte histograme 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?

b) 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 histograme frekvencija i relativnih frekvencija. Također nacrtajte histograme frekvencija i relativnih frekvencija za podatke sadržane u varijabli opce-stanje kategorizirane po vrijednostima varijable tlak (N, O, P). Proanalizirajte dobivene histograme?

c) 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.

d) Nacrtajte i detaljno proanalizirajte kutijasti dijagram na bazi medijana za podatke sadržane u varijabli dob. Obrazložite svoj odgovor.

e) 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 3.9. (komarci.sta)
Otvorite bazu podataka komarci.sta i odredite tablicu frekvencija i histogram 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 vaš zaključak.

Zadatak 3.10. (razred.sta) U razredu koji broji 25 učenika zaključne ocjene iz matematike na kraju školske godine raspodjenjene su na sljedeći način: tri učenika ima 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:

a) Sastavite tablicu frekvencija i relativnih frekvencija za varijablu ocjena.

b) Koristeći Statisticu grafički prikažite frekvencije i relativne frekvencije (histogramima i strukturiranim krugovima).

c) Izračunajte aritmetičku sredinu, mod, raspon te varijancu i standardnu devijaciju ovog skupa podataka.

d) Izračunajte numeričke karakteristike ovog skupa podataka koje su vam potrebne za kutijasti dijagram na bazi medijana te ga nacrtajte.

3.4 Prvi projektni zadatak

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 kontinuiranu 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 vašu bazu podataka.

Ponovimo

- Mjere zastupljenosti kategorije kvalitativne varijable u uzorku su frekvencija i relativna frekvencija kategorije.
- Frekvencija kategorije je broj izmjerenih vrijednosti varijable koje pripadaju toj kategoriji.
- Relativna frekvencija kategorije je broj izmjerenih vrijednosti varijable koje pripadaju toj kategoriji podijeljen s ukupnim brojem izmjerenih vrijednosti za ispitivanu varijablu.
- Frekvencije i relativne frekvencije pojedinih kategorija prikazujemo tablično (tablicom frekvencija i relativnih frekvencija) i grafički (histogramom ili strukturiranim krugom).
- Aritmetička sredina niza izmjerenih vrijednosti $x_1, x_2, \ldots, x_n$ varijable X definirana je izrazom
$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$
- Medijan niza podataka koji se sastoji od neparnog broja izmjerenih vrijednosti je onaj podatak koji se nalazi na srednjoj poziciji niza kada je on uređen po veličini. Ako se niz sastoji od parnog broja podataka onda se medijan definira kao aritmetička sredina dvaju podataka koji zauzimaju centralnu poziju u tom nizu kada je on uređen po veličini.
- Postotna vrijednost x'_p za neki izabrani broj $p \in \langle 0, 100 \rangle$ 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. Dvadesetpet postotna vrijednost zove se donji kvartil, a sedamdesetpet postotna vrijednost zove se gornji kvartil.
- Raspon niza izmjerenih vrijednosti je razlika najvećeg i najmanjeg podatka u nizu.
- Varijanca niza izmjerenih vrijednosti $x_1, x_2, \ldots, x_n$ definirana je izrazom
$$s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2,$$
a standardna devijacija je kvadratni korijen varijance.
- Mod je vrijednost iz niza izmjerenih vrijednosti kojoj pripada najveća frekvencija.
- Podatak koji je značajno veći ili manji u odnosu na druge izmjerene vrijednosti jedne varijable nazivamo stršeći podatak (engl. outlier).