Vjerojatnost i statistika

Građevinski fakultet, Sveučilište J.J. Strossmayera u Osijeku

Statistički praktikum 1: Prikupljanje i organizacija podataka. Deskriptivna statistika.

14. prosinca 2015.

Primjer

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 - SVE jedinke koje su predmet istraživanja
- uzorak - dio populacije na kojemu je osigurano kvalitetno provođenje istraživanja
- repezentativan uzorak - dio populacije u kojem su zastupljene tipične osobine cijele populacije

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 zaista bilo praktično, ali takav uzorak nije reprezentativan za zaključivanje o cijeloj populaciji.

- način izbora (reprezentativnog) uzorka: slučajan uzorak
- slučajan uzorak - svaka jedinka ima jednaku vjerojatnost ulaska u uzorak

- 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).
- 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).

- **veličine (obilježja) promatrane na jedinkama** obuhvaćenim nekim istraživanjem nazivamo **varijablama** - modeliramo ih korištenjem **slučajnih varijabli**
- **vrijednosti varijable izmjerene na jedinkama iz uzorka** (tj. vrijednosti zabilježene u stupac baze podataka) - **nezavisne realizacije slučajne varijable** kojom modeliramo promatranu veličinu (obilježje)
- slučajna varijabla - u potpunosti zadana svojom **distribucijom**
- poznavanje distribucije omogućuje izračunavanje **vjerojatnosti** vezanih uz realizacije slučajne varijable i njezinih **numeričkih karakteristika** (očekivanje, varijanca, standardna devijacija...)
- **nepoznata distribucija** slučajne varijable - problem

Primjer

Raspolažemo podacima o realizaciji slučajne varijable X koja opisuje potrošnju goriva novog modela automobila pri brzini od 110 km/h na autocesti u 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?

Odgovor na ova pitanja dolazi kasnije...!!!

Kvalitativne varijable

- njihove vrijednosti nisu, po svojim svojstvima korištenim u istraživanju, realni brojevi, već ih svrstavamo u kategorije
- kategorije mogu biti definirane u skladu s potrebama statističkog istraživanja

Primjer

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).*

Dakle, "spol osobe" je jedna kvalitativna varijabla, a pripadne kategorije su "muški" i "ženski" spol.

Numeričke varijable

- vrijednosti numeričkih slučajnih varijabli su realni brojevi
- kategorije kvalitativnih varijabli mogu se izražavati brojevima, ali to ih ne čini numeričkim varijablama (npr. kategoriju "ženski spol" možemo označiti oznakom "1", a kategoriju "muški spol" oznakom "2", što može biti korisno prilikom unošenja podataka u bazu.
- razlikujemo **diskretne** i **neprekidne**

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

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°C.*

Primjer

Sljedeće numeričke varijable su neprekidne:

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

Radi prikaza podataka i nekih statističkih analiza vrijednosti numeričke varijable također se mogu svrstati u kategorije.

Primjer

(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. Kako 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, kategorizacija 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 na sljedećem slajdu.*

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

Tablica: Primjer kategorizacije diskretne numeričke varijable automobili.

Ordinalne varijable

- prema karakteru su kvalitativne, ali među kategorijama može se uspostaviti prirodan poredak
- tipičan primjer takve varijable je "stručna sprema osobe"

Primjer

Baza podataka matematika.sta sadrži podatke prikupljene anketiranjem studenata nakon održanih predavanja, vježbi, kolokvija te usmenog ispita iz jednog matematičkog kolegija.

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

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: Kategorizacija studenata prema broju izostanaka s predavanja/vježbi.

Zadatak

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.*

Zadatak

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).*

Zadatak

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.*

Zadatak

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).*

Metode opisivanja kvalitativnih podataka

- kvalitativne varijable primaju vrijednosti koje su razvrstane u kategorije

Primjer

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 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

Pitanja...

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 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?

Frekvencija

Stoga se postavlja pitanje kako izmjeriti zastupljenost pojedine kategorije u uzorku?

- **frekvencija** kategorije - osnovna mjera kojom opisujemo zastupljenost jedne kategorije u uzorku
- neka varijabla X ima k kategorija (npr. $k = 4$)
- oznake kategorija su nam $x_1, x_2, \ldots, x_k$
- frekvencija kategorije x_i - broj izmjerenih vrijednosti varijable koje pripadaju kategoriji x_i, $i = 1, \ldots, k$
- oznaka: f_i, $i = 1, \ldots, k$

Relativna frekvencija

- frekvencija pojedine kategorije ovisi o broju izvršenih mjeranja, tj. dimenziji uzorka
- koristimo **relativnu frekvenciju**
- Relativna frekvencija kategorije x_i je broj izmjerenih vrijednosti varijable koje pripadaju kategoriji x_i podijeljen s ukupnim brojem izmjerenih vrijednosti za ispitivanu varijablu, $i = 1, \ldots, k$.

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

- n - dimenzija uzorka, f_i frekvencija kategorije x_i, $i = 1, \ldots, k$
- udio kategorije u uzorku, izražava se kao postotak
- frekvencije i relativne frekvencije pojedinih kategorija prikazujemo **tablično** i **grafički**

Tablični prikaz frekvencija i relativnih frekvencija -Primjer

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

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

Kategorizirane tablice frekvencija i relativnih frekvencija

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

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

Pitanja i odgovori

Sada lako možemo odgovoriti na unaprijed postavljena pitanja:

- Koliko ispitanika ženskog spola ima u promatranom uzorku? - 6
- Koliki je udio ispitanika s krvnom grupom 0 u promatranom uzorku? - 20%
- Koliko ispitanika ženskog spola iz promatranog uzorka ima krvnu grupu A? - 2
- Koliki udio od ispitanika muškog spola iz promatranog uzorka ima krvnu grupu B ili AB? - 50%

Primjer

koristimo programski paket Statistica - baza: **krvne-grupe.sta**

- Tablične prikaze frekvencija i relativnih frekvencija varijabli krvna grupa i spol

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

- kategorizirane tablice frekvencija i relativnih frekvencija varijable spol kategorizirane prema krvnoj grupi ispitanika

  1. način: Statistics → Basic Statistics→ Freq. Tables → Variables (odabrati varijablu **spol**) → Select Cases → 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.

  2. način: Statistics → Basic Statistics→ Freq. Tables → Variables (odabrati varijablu **spol**) → By Group... → pod Grouping Variable(s) odabrati varijablu **krvna_grupa** → OK → Summary.

Zadatak

koristimo programski paket Statistica - baza: **hormon.sta**

Značenje varijabli:

- 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 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),
- varijabla dijagnoza sadrži dijagnozu ispitanika.

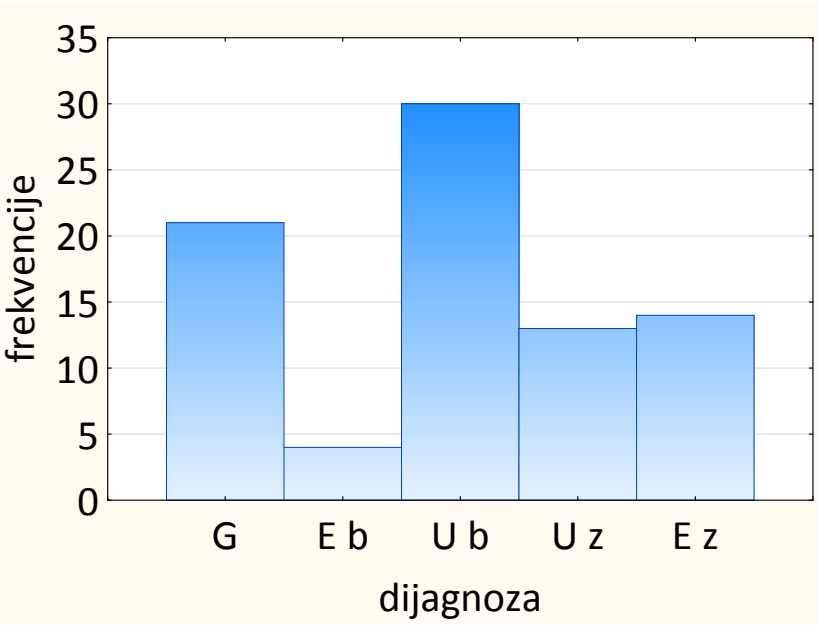
Zadatak:

- odredite tablice frekvencija i relativnih frekvencija svih kategorija za varijable koje smatrate kvalitativnima
- odredite kategorizirane tablice frekvencija i relativnih frekvencija varijable dijagnoza kategorizirane prema tome da li je ispitanik pušač ili nepušač (prema varijabli pusenje)

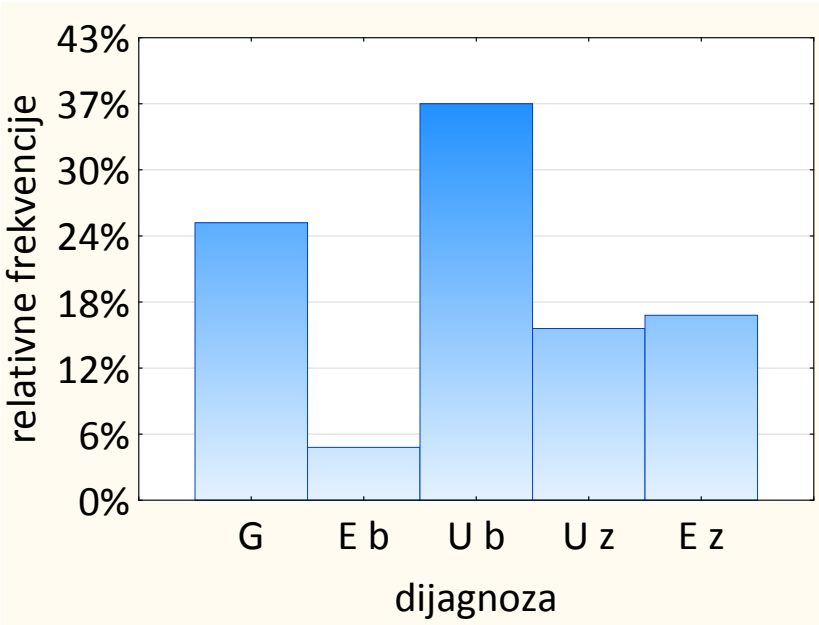
Grafički prikazi frekvencija i relativnih frekvencija

Primjer

- stupčasti dijagrami (histogrami) frekvencija i relativnih frekvencija
- kružni dijagrami (strukturirani krugovi) frekvencia i relativnih frekvencija
- baza: hormon.sta, varijabla: dijagnoza



(a) frekvencije



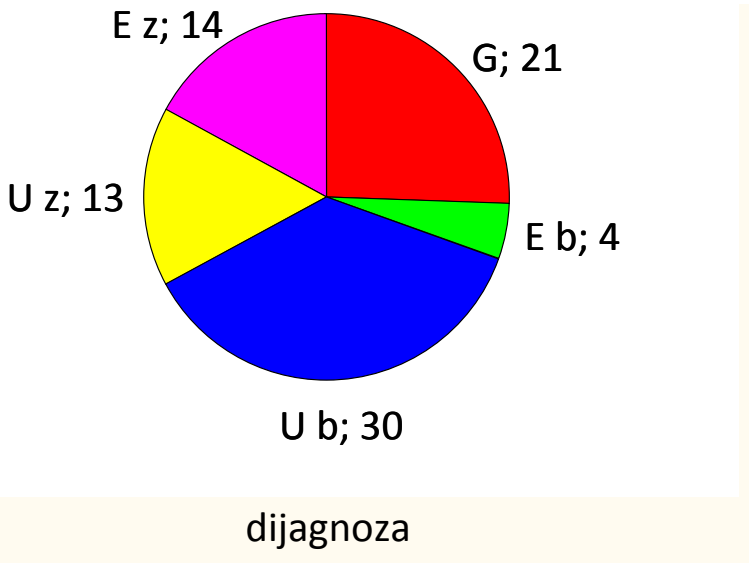
(b) relativne frekvencije

Slika: Histogrami frekvencija i relativnih frekvencija svih kategorija varijable dijagnoza.

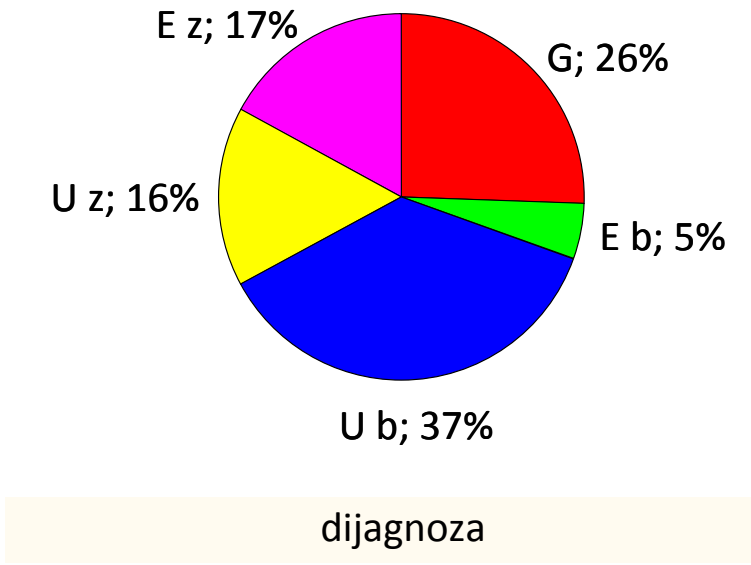
Grafički prikazi frekvencija i relativnih frekvencija

Primjer

- baza: hormon.sta, varijabla: dijagnoza



(a) frekvencije



(b) relativne frekvencije

Slika: Strukturirani krugovi frekvencija i rel. frekvencija svih kategorija varijable dijagnoza.

Grafički prikazi frekvencija i relativnih frekvencija

Kako to napraviti u Statistici?

- histogram frekvencija:

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

- histogram frekvencija i **relativnih** frekvencija može se dobiti i ovako:

  Graphs → Histograms → Choose variables → Advanced → Pod "Fit type" odabrati "Off"→ Pod "Y axis" uključiti "N" za frekvencije, a "% and N" za relativne frekvencije i frekvencije → OK.

- strukturirani krugovi frekvencija i relativnih frekvencija

  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 → Pod "Type" odabrati "2D" → OK.

Primjer

koristimo programski paket Statistica - baza: **djelatnici.sta**

- kvalitativnu varijabla **obrazovanje**:
  - SSS - srednja stručna sprema,
  - VŠSS - viša stručna sprema,
  - VSS - visoka stručna sprema.
- kvalitativnu varijabla **spol** označava spol ispitanika

Zadatak:

- tablica frekvencija i relativnih frekvencija varijable **obrazovanje**
- histogram frekvencija i relativnih frekvencija varijable **obrazovanje**
- strukturirani krug frekvencija i relativnih frekvencija varijable **obrazovanje**
- prethodna tri sa kategorizacijom prema varijabli **spol**

NAPOMENA: kategorizaciju u svim slučajevima dobivamo koristeći opciju By Group

Zadatak

koristimo programski paket Statistica - baza: **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 ultrazvucnog pregleda mozga novorođenčeta (u skali od 1 do 4).

Zadatak

koristimo programski paket Statistica - baza: **djeca.sta**

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?

c) Koliki je udio majki starijih od 35 godina?

Rješenje

a) napraviti u Statistici

b) iz tablica (relativnih) frekvencija varijable **spol** možemo vidjeti da je uzorkom obuhvaćeno 338 novorođenčadi - 160 djevojčica i 178 dječaka. Dakle, u uzorku ima više dječaka.

c) Statistics $\rightarrow$ Basic Statistics/Tables $\rightarrow$ Freq. Tables $\rightarrow$ Variables (izabrati varijablu **majka_dob**) $\rightarrow$ Select Cases $\rightarrow$ označiti Enable Selection Conditions $\rightarrow$ pod Include Cases odabrati opciju "Specific, selected by expression" (u polje za unos teksta upisati majka_dob>35 $\rightarrow$ OK. Majki starijih od 35 godina ima $29/338 \approx 8.58\%$.

Zadatak

koristimo programski paket Statistica - baza: **TV-program.sta**

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

- 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.

napravite sljedeće tablične i grafičke prikaze:

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

b) napravite tablice i nacrtajte 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) nacrtajte kružne dijagrame frekvencija i relativnih frekvencija za podatke sadržane u varijablama spol i P3,

Zadatak

koristimo programski paket Statistica - baza: **TV-program.sta**

d) nacrtajte kružne dijagrame frekvencija i relativnih frekvencija tipa **Separate** - za odvojene histograme kategorija određenih varijabli i **Overlaid** - za paralelne histograme kategorija određenih varijabli, ali ovdje za podatke sadržane u varijabli P3 posebno za kategoriju ispitanika ženskog spola, a posebno za kategoriju ispitanika muškog spola.

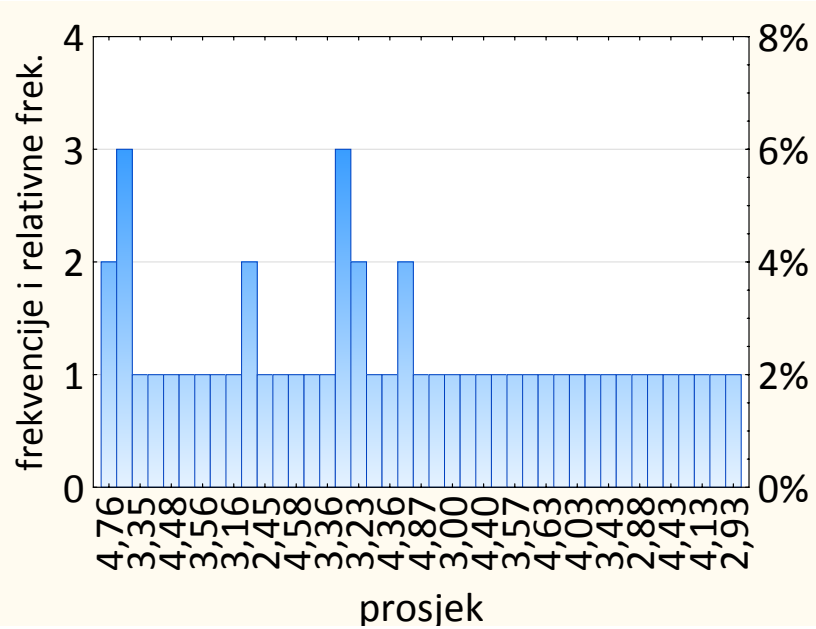
Rješenje:

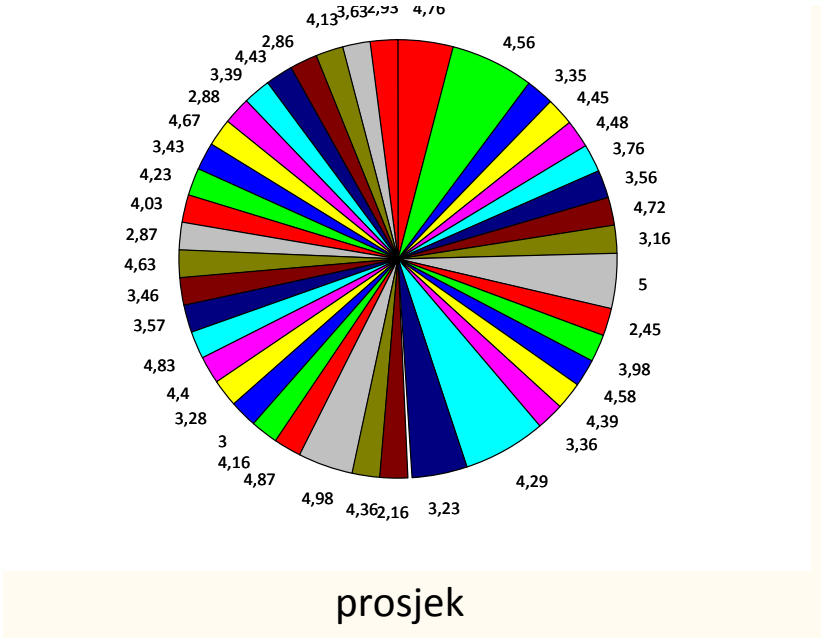
d) Graphs $\rightarrow$ Categorized Graphs $\rightarrow$ Histograms $\rightarrow$ Variables (Variable - P3, X-Category - spol) $\rightarrow$ Layout (Separate - za odvojene histograme kategorija varijable P3 kategoriziranih s obzirom na vrijednosti varijable spol; Overlaid - za prikaz frekvencija kategorija varijable P3 kategoriziranih s obzirom na vrijednosti varijable spol na istom histogramu)

Primjer

kategorizacija numeričkih varijabli koje nisu diskretne

- **Ako numerička varijabla nije diskretna, za prikazivanje skupa izmjerenih vrijednosti obično nam neće puno pomoći frekvencije, histogrami i strukturirani krugovi napravljeni na osnovu svake pojedine izmjerene vrijednosti.**
- histogram frekvencija i relativnih frekvencija varijable **prosjek** iz baze podataka matematika.sta (u programu "Statistica" dodatno odabrati "unique values")





Postupak razvrstavanja numeričkih podataka u kategorije

Razvrstavanje vrijednosti neprekidne slučajne varijable u kategorije moguće je provesti na nekoliko načina, npr.

- skup svih podataka podijeliti na disjunktne intervale, ne nužno jednake duljine
- dakle, nema točno definiranog pravila po kojemu bi trebalo definirati duljine intervala niti nihov broj
- intervala ne smije biti niti previše niti premalo da bi cijeli postupak imao smisla i služio svrsi
- kriterij treba biti temeljen na razumijevanju problema koji proučavamo

Za prikaz frekvencija ili relativnih frekvencija tako kategoriziranih podataka možemo koristiti histogram koji 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.

Zadatak

koristimo programski paket Statistica - baza: **hormon.sta**

- 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 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),
- varijabla dijagnoza sadrži dijagnozu ispitanika.

Zadaci:

a) Odredite tablicu frekvencija i histogram za kontinuiranu numeričku varijablu gastrS iz baze podataka hormon.sta tako da za kategorije uzmete sve međusobno različite izmjerene vrijednosti.

b) Iskoristite izmjerene vrijednosti varijable gastrS te ju razvrstajte na 10 disjunktnih intervala počevši od najmanje vrijednosti do najveće

c) Iskoristite izmjerene vrijednosti varijable gastrS te ju razvrstajte na 15 disjunktnih intervala duljine 10 počevši od 0

Zadatak

koristimo programski paket Statistica - baza: **hormon.sta**

d) Procijenite vjerojatnost da je koncentracija enzima gastrS u krvi ispitanika manja od 45.

Rješenje:

a) Graphs → Histograms → Choose variables → Advanced → Pod "Y axis" uključiti "% and N" → Pod "Intervals" uključiti "unique values"→OK.

b) Graphs → Histograms → Choose variables → Advanced → Pod "Y axis" uključiti "% and N", pod "Intervals" u polje Categories upisati 10 → OK.

c) Graphs → Histograms → Choose variables → Advanced → Pod "Y axis" uključiti "% and N", pod "Intervals" otići na Boundaries, zatim Specify Boundaries, i redom upisati u polje Minimum: 0, Interval: 10 i Maximum: 150 → OK.

d) Statistics → Basic Statistics/Tables → Frequency Tables → Variables, izabrati gastrS → Advanced → u polje "Step Size" upisati 15 (ili bilo koji broj kojemu je 45 višekratnik), "starting at": 0, isključiti: "at minimum" → Summary.

procjenjenu vjerojatnosti pročitati iz "Cumulative Percent": 0.402439

Mjere centralne tendencije podataka

Aritmetička sredina podataka

- **aritmetička sredina** (eng. mean) niza izmjerenih vrijednosti (podataka) $x_1, x_2, \ldots, x_n$ varijable X definirana je izrazom

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

- npr. neka su $1.2, 2.1, 3.2, 4.3, 5.4, 6.5, 7.6, 8.7, 9.8$ izmjerene vrijednosti jedne varijable

- obzirom da ih ima ukupno devet, aritmetička sredina ovog skupa podataka je

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

Mjere centralne tendencije podataka

Medijan podataka

- **medijan** ima značenje izmjerene vrijednosti koja se nalazi na sredini niza podataka kada je on uređen po veličini - barem pola podataka je manje ili jednako medijanu, a istovremeno je barem pola podataka veće ili jednako od medijana
- način njegovog određivanja ovisi o tome imamo li **neparan** ili **paran** broj izmjerenih vrijednosti varijable (podataka)

Mjere centralne tendencije podataka

Medijan podataka - neparan broj podataka

- ukoliko imamo **neparan broj** izmjerenih vrijednosti, onda postoji podatak koja je na srednjoj poziciji u uređenom skupu izmjerenih vrijednosti, pa njega definiramo kao medijan
- npr. neka su $1, 2, 5, 6, 5, 1, 2, 7, 2, 2, 3, 3$ izmjerene vrijednosti jedne varijable
- prvo ove vrijednosti poredamo po veličini: $1, 1, 2, 2, 2, \mathbf{2}, 3, 5, 5, 6, 7$
- obzirom da ih ima ukupno jedanaest, medijan je vrijednost koja je na šestoj poziciji u tako dobivenom nizu, tj. broj 2

Mjere centralne tendencije podataka

Medijan podataka - neparan broj podataka

- ukoliko imamo **paran broj** izmjerenih vrijednosti, 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)
- npr. neka su $1, 2, 5, 6, 5, 1, 2, 7, 2, 2, 3, 3$ izmjerene vrijednosti jedne varijable
- prvo ove vrijednosti poredamo po veličini: $1, 1, 2, 2, 2, \mathbf{2}, \mathbf{3}, 3, 5, 5, 6, 7$
- obzirom da ih ima dvanaest, "sredinu" čine šesti i sedmi podatak, tj. brojevi 2 i 3 - medijan ovog skupa podataka je sredina ta dva broja, tj. medijan je $(2 + 3)/2 = 2.5$

Mjere raspršenosti podataka

Postotna vrijednost, donji i gornji kvartil

- **postotna vrijednost** za neki izabrani broj $p \in \langle 0, 100 \rangle$, označimo je x'_p, definira se poštujući zahtjev da je barem $p\%$ izmjerenih vrijednosti varijable manje ili jednako x'_p, dok je barem $(100 - p)\%$ vrijednosti veće ili jednako x'_p
- **dvadesetpet postotna vrijednost** zove se **donji kvartil**
- **sedamdesetpet postotna vrijednost** zove se **gornji kvartil**
- kao i kod računanja medijana, ako se na traženoj poziciji za računanje postotne vrijednosti nalaze dva podatka u uređenom skupu izmjerenih vrijednosti, postotnu vrijednost određujemo kao njihovu aritmetičku sredinu
- Prvo je potrebno podatke poredati u rastućem poretku i odrediti "poziciju" j koja je ključna za određivanje zadanog postotka 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 $j + 1$.

Mjere raspršenosti podataka

Postotna vrijednost, donji i gornji kvartil

- npr. neka su $1, 2, 5, 6, 6, 1, 3, 7, 3, 3, 3, 3$ izmjerene vrijednosti jedne varijable
- 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 - gornji kvartil je 5.5

Mjere raspršenosti podataka

Najmanja i najveća vrijednost, raspon podataka

- ako su $x_1, x_2, \ldots, x_n$ izmjerene vrijednosti varijable X, označimo najmanju od njih **(minimum)** $x_{\min}$, a najveću od njih **(maksimum)** $x_{\max}$
- **raspon** (eng. range) podataka - razlika najveće i najmanje vrijednosti u skupu izmjerenih vrijednosti varijable (tj. razlika maksimalne i minimalne izmjerene vrijednosti varijable)
- npr. neka su izmjerene vrijednosti jedne varijable $1, 2, 5, 6, 5, 1, 2, 7, 2, 2, 3, 3$ - 1 je najmanja izmjerena vrijednost, a 7 najveća, pa je raspon ovog skupa izmjerenih vrijednosti $7 - 1 = 6$

Mjere raspršenosti podataka

Maksimalno odstupanje od "prosjeka"

- **maksimalno odstupanje izmjerenih vrijednosti varijable od "prosjeka"**, tj. aritmetičke sredine tih izmjerenih vrijednosti - veći od brojeva $(\bar{x}_n - x_{\min})$ i $(x_{\max} - \bar{x}_n)$, tj. broj

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

- npr. neka su $1, 2, 5, 6, 5, 1, 2, 7, 2, 2, 3, 3$ izmjerene vrijednosti neke varijable X:

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

- maksimalno odstupanje izmjerenih vrijednosti ove varijable od prosjeka:

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

Mjere raspršenosti podataka

Varijanca i standardna devijacija podataka

- **varijanca** i **standardna devijacija** karakteriziraju raspršenost podataka oko aritmetičke sredine
- varijanca niza izmjerenih vrijednosti $x_1, x_2, \ldots, x_n$ varijable X definirana je izrazom

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

a standardna devijacija je kvadratni korijen iz varijance, tj.

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

Mjere raspršenosti podataka

Varijanca i standardna devijacija podataka

- npr. neka su izmjerene vrijednosti jedne varijable

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

- iz primjera znamo da je aritmetička sredina ovog skupa podataka približno jednaka 5.42, pa su varijanca i standardna devijacija ovog skupa podataka

$$s_n^2 \approx \frac{1}{9}\sum_{i=1}^{9}(x_i - 5.42)^2 \approx 7.87, \quad s_n \approx \sqrt{\frac{1}{9}\sum_{i=1}^{9}(x_i - 5.42)^2} \approx 2.81$$

Mjere raspršenosti podataka

Mod podataka

- **mod** podataka je vrijednost iz niza izmjerenih vrijednosti varijable X kojoj pripada najveća frekvencija, tj. izmjerena je najviše puta
- mod ne mora biti jedinstven
- npr. neka su $1, 2, 5, 6, 5, 1, 2, 7, 2, 2, 3, 3$ izmjerene vrijednosti jedne varijable - vrijednost 2 je izmjerena najviše puta (četiri puta) pa je 2 mod ovog skupa podataka
- npr. neka su $1, 2, 3, 6, 5, 3, 1, 2, 7, 2, 2, 3, 3$ izmjerene vrijednosti jedne varijable - najviše puta izmjerene dvije vrijednosi, tj. 2 i 3 su obje izmjerene točno četiri puta pa mod ovog skupa podataka nije jedinstven nego su mod i 2 i 3

Grafička metoda opisivanja numeričkih podataka

Kutijasti dijagram

- korištenjem numeričkih karakteristika numeričkih varijabli skup mjerenih vrijednosti može se prikazati grafički pomoću **kutijastog dijagrama** (eng. box plot, boxplot ili box-and-whisker plot)
- kutijastim 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** skupa podataka, ako postoje

Detekcija stršećih vrijednosti

stršeća vrijednost - podatak koji je značajno veći ili manji u odnosu na druge izmjerene vrijednosti jedne varijable i čije je pojavljivanje najčešće vezano uz jedan od sljedećih razloga:

- podatak je ili netočno izmjeren ili krivo unesen u bazu podataka
- podatak dolazi iz druge populacije (ne iz populacije koju promatramo u kontekstu problema kojeg proučavamo)
- podatak je točno izmjeren i unesen u bazu, ali predstavlja rijetku pojavu u populaciji

Primjer

koristimo programski paket Statistica - baza: **trgovacki-centri.sta**

- promatrajući deset trgovačkih centara, zabilježio je cijene proizvoda kod kojega su razlike bile najizraženije.

  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.

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

  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.

Zadatak

koristimo programski paket Statistica - baza: **djelatnici.sta**

- interpretirajte numeričke karakteristike skupa izmjerenih vrijednosti varijable **Placa_prije** u bazi podataka **djelatnici.sta**

Zadatak

koristimo programski paket Statistica - baza: **djelatnici.sta**

1) Kojeg su tipa varijable dane baze?
2) Izradite tablice frekvencija i relativnih frekvencija za podatke sadržane u varijabli **Odjel** te nacrtajte pripadne histograme.
3) Procijenite vjerojatnost da je visina djelatnika veća od 150.
4) Izradite tablice frekvencija i relativnih frekvencija za podatke sadržane u varijabli **Obrazovanje** kategorizirane prema varijabli **Spol** te nacrtajte pripadne strukturirane krugove.
5) Koliki je udio ispitanika ženskoga spola kojima je **Placa_prije** veća od 20000?
6) Kolika je najniža a kolika najviša dob ispitanika?
7) Odredite tablicu frekvencija i relativnih frekvencija, te odgovarajući histogram za varijablu **Placa_poslije** tako da za kategorije uzmete sve međusobno različite izmjerene vrijednosti.
8) Iskoristite izmjerene vrijednosti varijable **Dob** te ju razvrstajte na 9 disjunktnih intervala duljine 13 počevši od 0, a zatim na 8 disjunktnih intervala počevši od najmanje vrijednosti do najveće.
9) Skicirajte kutijasti dijagram na bazi medijana za podatke sadržane u varijabli **Rukovodstvo**.

Literatura

 Benšić, M. i Šuvak, N., *Primijenjena statistika*, Odjel za matematiku, Sveučilište J.J Strossmayera, Osijek, 2012.