Vjerojatnost i statistika

Prikupljanje i organizacija podataka. Deskriptivna statistika

---

12. prosinca 2023.

Prikupljanje i orgranizacija podataka

Populacija i uzorak

- **populacija** - sve jedinke koje su predmet nekog istraživanja
- **uzorak** - dio populacije na kojemu je osigurano kvalitetno provođenje istraživanja
- **reprezentativan uzorak** - dio populacije na kojemu su zastupljene tipične osobine cijele populacije

Populacija i uzorak - primjer

Pretpostavimo da istražujemo prehrambene navike stanovnika Slavonije i stanovnika Dalmacije.

- **populacija** - svi stanovnici Slavonije i Dalmacije
- **uzorak** - dio stanovnika iz tih područja
- **reprezentativan uzorak** - dio stanovnika koji odražava cijelu populaciju prema nekim značajkama (dob, spol, visina, težina itd.)

Istraživanje ne možemo provesti, primjerice, samo na uzorku djece koja pohađaju srednju školu jer takav uzorak nije reprezentativan. Međutim, ako nas zanimaju samo prehrambene navike studenata iz tih područja, onda populaciju čine samo studenti iz Slavonije i Dalmacije.

Uzorak se kreira na način da svaka jedinka ima jednaku vjerojatnost biti odabrana.

Kako pronaći podatke?

- podaci iz javnih izvora
- 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)

Terminologija

- **obilježja** promatrana na jedinkama obuhvaćenim nekim istraživanjem nazivamo **varijablama**, a modeliramo ih pomoću **slučajnih varijabli**
- **vrijednosti varijable** izmjerene na jedinkama iz uzorka smatramo nezavisnim realizacijama slučajne varijable kojom modeliramo promatranu veličinu
- slučajna varijabla u potpunosti je određena svojom **distribucijom**
- poznavanje distribucije omogućuje izračunavanje vjerojatnosti vezanih uz realizacije slučajne varijable i njezinih numeričkih karakteristika (očekivanje, varijanca, standardna devijacija...)
- distribucija slučajne varijable najčešće je nepoznata

Primjer

Raspolažemo podacima o realizaciji slučajne varijable X koja opisuje potrošnju goriva nekog modela automobila pri brzini od 130 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/100km?
- Kolika je očekivana potrošnja goriva u ovim uvjetima?
- Kolika je standardna devijacija slučajne varijable koja opisuje potrošnju goriva u ovim uvjetima?

Tipovi varijabli

Kvalitativne varijable

- njihove vrijednosti po svojim svojstvima nisu realni brojevi, već ih svrstavamo u kategorije
- kategorije se mogu izražavati riječima, kraticama, brojevima, oznakama itd.
- razlikujemo
  - **nominalne** - nema poretka među kategorijama
  - **ordinalne** - postoji prirodan poredak među kategorijama

Primjeri kvalitativnih varijabli

ordinalne:

- stručna sprema (SSS, VŠS, VSS)
- opisna ocjena (ništa, malo, srednje, puno)
- ocjena iz nekog kolegija (1-5)

nominalne:

- boja očiju (plava, smeđa, zelena)
- krvna grupa (A, B, AB, 0)
- spol (M, Ž)

Primjer - matematika.sta

Baza podataka `matematika.sta` sadrži podatke prikupljene anketiranjem studenata nakon održanih predavanja, vježbi, kolokvija te usmenog ispita iz jednog matematičkog kolegija. Sadrži dvije varijable (*predavanja*, *vjezbe*) koje prisutnost studenata na predavanjima/vježbama (p/v) svrstavaju u tri kategorije na sljedeći način

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

Numeričke varijable

- vrijednosti numeričkih slučajnih varijabli su realni brojevi
- razlikujemo
  - **diskretne** - mogu poprimiti konačno ili prebrojivo mnogo vrijednosti
  - **neprekidne** - skup vrijednosti je najčešće interval

Primjeri numeričkih varijabli

diskretne:

- broj bodova na državnoj maturi iz matematike
- broj ulovljenih komaraca u klopku
- broj dana u godini s temperaturom zraka većom od 35°C

neprekidne:

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

Primjer - auto-centar.sta

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
- *kategorija* - kvalitativna varijabla koja podatke iz varijable *automobili* svrstava u pet kategorija

Kako broj prodanih automobila u jednom danu može znatno varirati, zaključujemo da varijabla *automobili* može poprimiti velik broj različitih vrijednosti. Zato je u nekim situacijama korisno kategorizirati ovu varijable prema točno određenom kriteriju.

Primjer - auto-centar.sta

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

Zadatak

Na sličan način analizirajte i odredite tipove varijabli u sljedećim bazama podataka:

a) baza podataka `komarci.sta` sadrži dio rezultata proučavanja komaraca u jednom močvarnom području (210 mjerenja na istoj lokaciji):
   - *brojM* i *brojZ* - broj muških i ženskih jedinki komaraca
   - *mjesec* - mjesečeva mijena (M - mlađak, U - uštap)
   - *doba dana* - doba dana u kojem je mjerenje obavljeno (P - predvečerje, N - noć, S - svitanje)
   - *svjetlost* - tip osvjetljenja pri mjerenju
   - *temperatura*
   - *rel vlaznost*

Zadatak

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

- *dnevne novine* - broj prelistanih dnevnih novina
- *tv vijesti* - broj pogledanih televizijskih vijesti
- *kava* - broj ispijenih kava
- *troskovi* - informacija o troškovima hrane za promatrani dan
- *vrijeme* - vremenske prilike u mjestu stanovanja (O - oblačno, S - sunčano)
- *raspolozenje* - subjektivna ocjena vlastitog raspoloženja (L - loše, D - dobro, O - odlično)

Zadatak

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

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

- *spol*
- *P1*, *P2*, *P3* i *P4* - subjektivne ocjene kvalitete programa televizijskih kanala P1, P2, P3 i P4
- *prosjek* - prosječna ocjena kvalitete programa navedenih televizijskih kanala

Zadatak

e) u bazi podataka `zdravlje.sta` nalaze se neki zdravstveni podaci anketiranih ispitanika:
   - *godine* i *spol*
   - *zdravlje* - subjektivna ocjena vlastitog zdravstvenog stanja
   - *broj pregleda* - ukupan broju zdravstvenih pregleda ispitanika u jednoj godini
   - *dodatno zdravstveno* - podatak o dodatnom zdravstvenom osiguranju (1 - ispitanik je dodatno osiguran, 0 - ispitanik nije dodatno osiguran)
   - *cijena* - cijena najskupljeg zdravstvenog pregleda svakog ispitanika

Deskriptivna statistika

Metode opisivanja kvalitativnih podataka

Sljedeća 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

Metode opisivanja kvalitativnih podataka

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

Definicija 1. Neka kvalitativna varijabla X ima k kategorija $x_1, x_2, \ldots, x_k$. **Frekvencija** kategorije x_i se definira kao broj izmjerenih vrijednosti varijable X koje pripadaju kategoriji x_i i označava s n_i, $i = 1, \ldots, k$.

- osnovna mjera kojom opisujemo zastupljenost jedne kategorije u uzorku
- ovisi o broju izvršenih mjerenja, tj. o dimenziji uzorka

Relativna frekvencija

Definicija 2. Relativna frekvencija kategorije x_i se definira kao frekvencija kategorije x_i podijeljena s ukupnim brojem izmjerenih vrijednosti, a označava se $f_i = \frac{n_i}{n}$, $i = 1, \ldots k$.

- udio kategorije u uzorku, izražava se kao postotak
- frekvencije i relativne frekvencije pojedinih kategorija prikazujemo **tablično** i **grafički**

Tablični prikaz frekvencija i relativnih frekvencija

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

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	$2/10 = 0.2 = 20\%$

Tablični prikaz 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

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	$1/4 = 0.25 = 25\%$

Pitanja i odgovori

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

Primjer - krvne-grupe.sta

Tablične prikaze frekvencija i relativnih frekvencija varijabli *spol* i *krvna grupa* dobivamo pomoću

Statistics → Basic Statistics → Frequency Tables → Variables → Summary

Primjer - krvne-grupe.sta

Kategorizirane tablice frekvencija i relativnih frekvencija varijable *spol* kategorizirane prema *krvnoj grupi* ispitanika dobivamo na sljedeće načine

Statistics → Basic Statistics → Frequency Tables → Variables (odabrati **spol**) → By Group... → pod Grouping Variable(s) odabrati **krvna_grupa** → OK → Summary

Statistics → Basic Statistics→ Frequency Tables → Variables (odabrati **spol**) → Select Cases → označiti Enable Selection Conditions → pod Include Cases odabrati "Specific, selected by" → u polje za unos teksta upisati npr. krvna_grupa="A" → OK → Summary

Zadatak - hormon.sta

Baza podataka `hormon.sta` sadrži sljedeće varijable:

- *spol*
- *gastrS*, *somatS* i *somatZ* - izmjerene koncentracije određenih enzima u krvi ispitanika
- *pusenje*, *alkohol* i *kava* - informacija o tome konzumira li ispitanik cigarete, alkohol i kavu (0 - ne, 1 - da)
- *CLOtest* - rezultat testa na zarazu bakterijom helicobacter pilory (0 - negativan test, 1 - pozitivan test)
- *dijagnoza* - dijagnoza ispitanika

Zadatak - hormon.sta

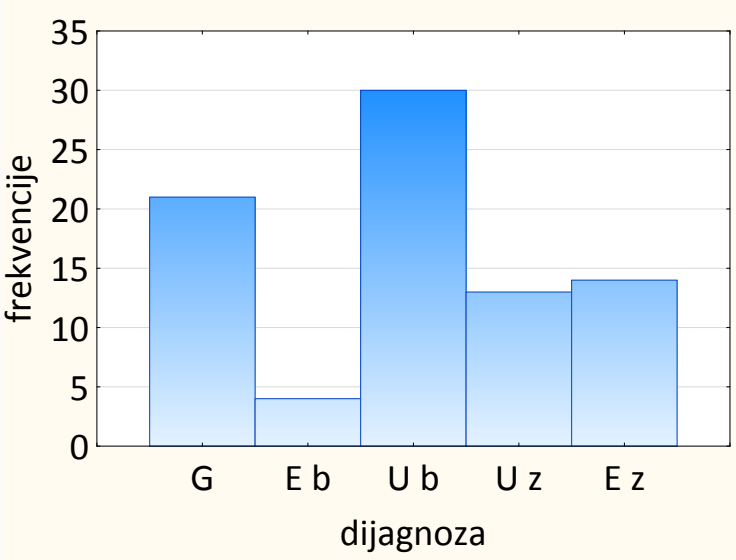
a) odredite tablice frekvencija i relativnih frekvencija svih kategorija za varijable koje smatrate kvalitativnima

b) odredite kategorizirane tablice frekvencija i relativnih frekvencija varijable dijagnoza kategorizirane prema tome da li je ispitanik pušač ili nepušač

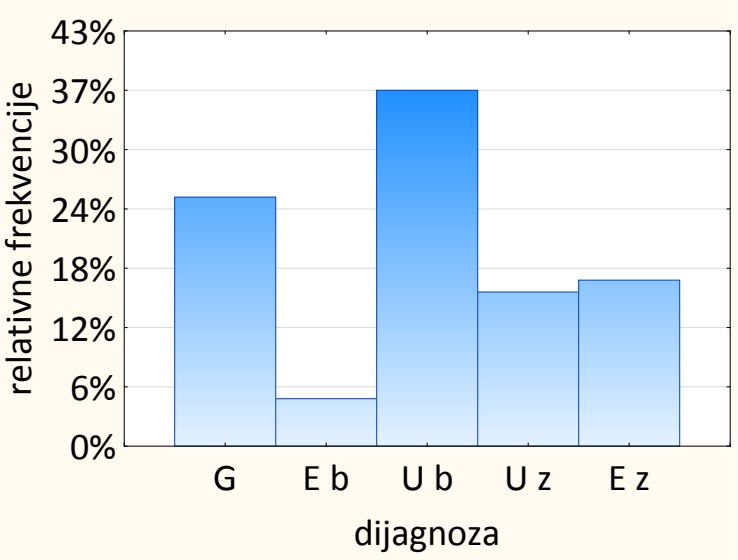
Grafički prikazi

Grafički prikazi frekvencija i relativnih frekvencija

- histogrami (stupčasti dijagrami) frekvencija i relativnih frekvencija



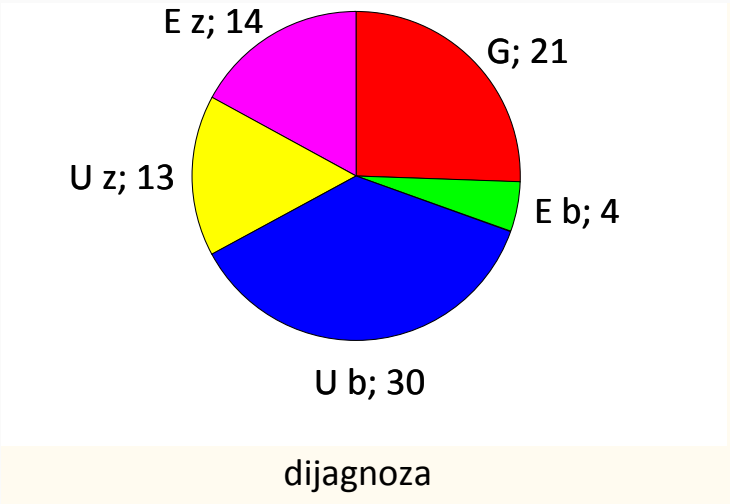
(a) frekvencije



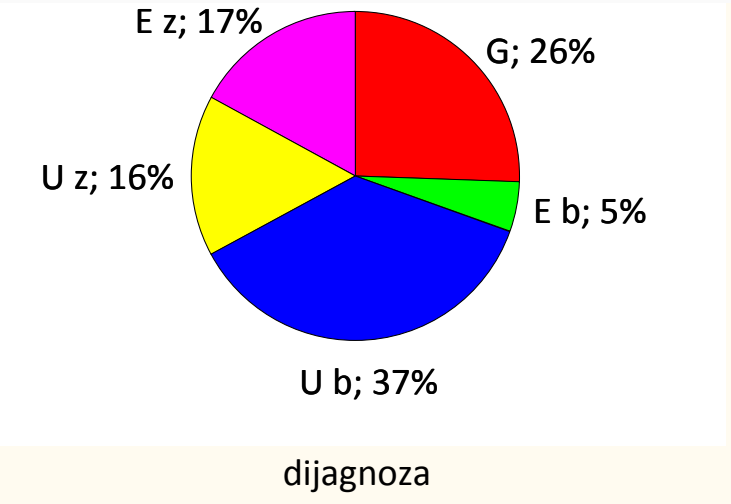
(b) relativne frekvencije

Grafički prikazi frekvencija i relativnih frekvencija

- kružni dijagrami (strukturirani krugovi) frekvencija i relativnih frekvencija



(c) frekvencije



(d) relativne frekvencije

Grafički prikazi frekvencija i relativnih frekvencija

Histogram frekvencija/relativnih frekvencija dobivamo na sljedeće načine:

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

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

Grafički prikazi frekvencija i relativnih frekvencija

Kružni dijagrami frekvencija/relativnih frekvencija dobivaju se pomoću

Graphs $\rightarrow$ 2D Graphs $\rightarrow$ Pie Charts... $\rightarrow$ Choose variables $\rightarrow$ Advanced $\rightarrow$ Pie Legend - odabrati "Text and Value" za dijagram frekvencija, a "Text and Percent" za dijagram relativnih frekvencija $\rightarrow$ "Type" odabrati "2D" $\rightarrow$ "Shape" odabrati "Circle" $\rightarrow$ OK

Primjer - djelatnici.sta

Napravite

a) tablicu frekvencija i tablicu relativnih frekvencija varijable *obrazovanje*,

b) histogram frekvencija i relativnih frekvencija varijable *obrazovanje*,

c) strukturirani krug frekvencija i relativnih frekvencija varijable *obrazovanje*,

d) prethodno troje s kategorizacijom prema varijabli *spol*.

Zadatak - 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. Varijable su

- *spol*
- *nacin_poroda*
- *RM*, *apgar1* i *apgar5* - obilježja novorođenčeta
- *majka_dob*
- *majka_bolest* - informacija o bolesti majke tijekom trudnoće (N/D)
- *komplikacije* - stupanj komplikacija za vrijeme trudnoće (0-7)
- *konvulzije* - informacija o konvulzijama kod novorođenčeta (N/D)
- *uzv* - ocjena ultrazvučnog pregleda mozga novorođenčeta (1-4)

Zadatak - djeca.sta

a) Odredite frekvencije i relativne frekvencije svih kategorija za varijable koje smatrate kvalitativnima. Rezultate prikažite tablično i grafički.

c) Ima li ovaj uzorak više djevojčica ili dječaka?

d) Koliki je udio majki starijih od 35 godina?

Rješenje

a) napraviti u Statistici

b) iz relativnih frekvencija varijable *spol* možemo vidjeti da je uzorkom obuhvaćeno 160 djevojčica i 178 dječaka

c) Statistics $\rightarrow$ Basic Statistics $\rightarrow$ Frequency Tables $\rightarrow$ Variables (odabrati *majka_dob*) $\rightarrow$ Select Cases $\rightarrow$ označiti Enable Selection Conditions $\rightarrow$ "Specific, selected by expression" (u polje za unos teksta upisati majka_dob>35 $\rightarrow$ OK $\rightarrow$ Summary
Majki starijih od 35 godina ima $29/328 \approx 8.84\%$.

Zadatak - TV-program.sta

Baza podataka `TV-program.sta` sadrži sljedeće varijable

- *spol*
- *P1*, *P2*, *P3* i *P4* - subjektivne ocjene kvalitete programa televizijskih kanala P1, P2, P3 i P4
- *prosjek* - prosječna ocjenu kvalitete programa navedenih televizijskih kanala

Zadatak - TV-program.sta

Napravite:

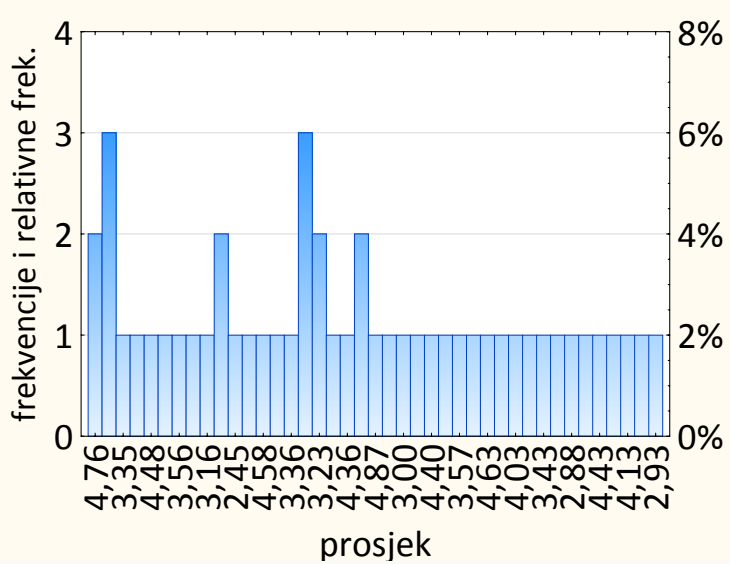
a) tablice i histograme frekvencija i relativnih frekvencija za podatke sadržane u varijablama *spol* i *P1*,

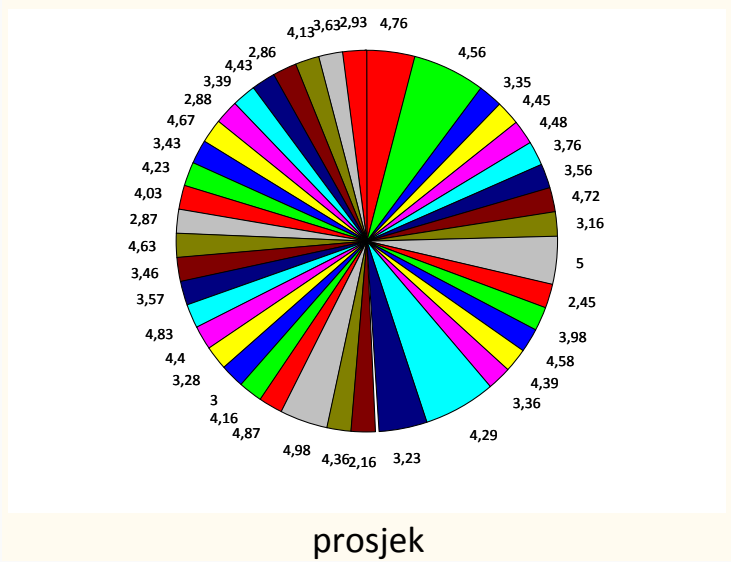
b) tablice i histograme frekvencija i relativnih frekvencija za podatke sadržane u varijabli *P1* posebno za ispitanike ženskog spola i posebno za ispitanike muškog spola,

c) kružne dijagrame frekvencija i relativnih frekvencija za podatke sadržane u varijablama *spol* i *P3*.

Postupak razvrstavanja numeričkih podataka u kategorije

Ako numerička varijabla nije diskretna, za prikazivanje skupa izmjerenih vrijednosti obično nam neće pomoći tablice i dijagrami napravljeni na osnovu pojedine izmjerene vrijednosti.





Postupak razvrstavanja numeričkih podataka u kategorije

Savjeti za razvrstavanje vrijednosti neprekidne slučajne varijable u kategorije:

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

Zadatak - hormon.sta

- *spol*
- *gastrS*, *somatS* i *somatZ* - izmjerene koncentracije određenih enzima u krvi ispitanika
- *pusenje*, *alkohol* i *kava* - informacija o tome konzumira li ispitanik cigarete, alkohol i kavu (0 - ne, 1 - da)
- *CLOtest* - rezultat testa na zarazu bakterijom helicobacter pilory (0 - negativan test, 1 - pozitivan test)
- *dijagnoza* - dijagnoza ispitanika

Zadatak - hormon.sta

a) Odredite tablicu frekvencija i histogram za neprekidnu numeričku varijablu *gastrS* 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.

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

Rješenje - hormon.sta

a) Graphs → Histogram → Advanced → Y axis označiti "% and N" → Fit type označiti "Off" → kod Intervals označiti "unique values" → OK

b) Graphs → Histogram → Advanced → Y axis označiti "% and N" → Fit type označiti "Off" → kod Intervals, u polje Categories upisati 10 → OK

c) Graphs → Histogram → Advanced → Y axis označiti "% and N" → Fit type označiti "Off" → kod Intervals označiti "Boundaries" → Specify Boundaries i redom upisati Minimum: 0, Interval: 10, Maximum: 150 → OK

Rješenje - hormon.sta

d) Statistics → Basic Statistics → Frequency Tables → Variables (odabrati *gastrS*) → Advanced → u polje "Step Size" upisati 15 (ili bilo koji broj kojemu je 45 višekratnik), starting at: 0, isključiti "at minimum" → Summary
Procijenjenu vjerojatnosti pročitati iz "Cumulative Percent": 0.402439