Vjerojatnost i statistika

dr. sc. Martin Lazar

Sveučilište u Dubrovniku
Preddiplomski studij primijenjenog/poslovnog računarstva

2011/2012

Organizacija kolegija

Predavanja: dr. sc. Martin Lazar, sri. 9-12
konzultacije: sri. u 12h, B28
Vježbe: mr. sc. Ivona Milić-Beran, čet 10-12
konzultacije: u B27

OBAVEZE STUDENATA TOKOM NASTAVE

Pohađanje predavanja i vježbi, izrada domaćih zadaća, polaganje tri kolokvija.

Elementi provjere znanja i ocjenivanja su:

- 3 kolokvija (90)
- zadaće (10)

U zagradama su navedeni maksimalni brojevi bodova koje studenti mogu ostvariti u pojedinom vidu provjere znanja.

Dodatnih 10 bodova može se ostvariti za redovito($\geqslant 90\%$), odnosno aktivno pohađanje vježbi i predavanja.

Organizacija kolegija

Kolokviji – sastoje se iz praktičnog (zadatci), i (u manjoj mjeri) teorijskog dijela gradiva.

UVJETI ZA POTPIS:

Prisustvo na 60% predavanja i vježbi, minimalno 5 bodova ostvarenih na zadaćama, te barem 15 (od mogućih 90) bodova skupljenih na kolokvijima.

UVJETI ZA PROLAZNU OCJENU:

Za dobivanje prolazne ocjene student treba ostvariti barem 50 - postotni uspjeh u svakom elementu provjere znanja.
Pri tom se potrebni minimumi (45, odnosno 5 bodova) računaju kumulativno.

Važno: studenti koji nisu položili ispit tokom semestra, ali imaju uvjete za potpis, mogu to ostvariti prijavom i pristupanjem ispitu (pismenom+usmenom) na jednom od ispitnih rokova.

Literatura

- I. Šošić: *Primijenjena statistika*, Školska knjiga, Zagreb, 2004.
- I. Šošić, V. Serdar: *Uvod u statistiku*, Školska knjiga, Zagreb, 1997.
- Ž. Pauše: *Uvod u matematičku statistiku*, Školska knjiga, Zagreb, 1993.
- N. Elezović: *Teorija vjerojatnosti. Zbirka zadataka*, Element, Zagreb, 1995.
- R. Galić: *Vjerojatnost i statistika*, Sveučilište u Osijeku, Osijek, 1999.

Sadržaj

- Statistika - osnovni pojmovi
- Opisna statistika (grafičke i numeričke metode prikaza podataka)
- Vjerojatnost - osnovni pojmovi
- Vjerojatnosne razdiobe
- Statističko zaključivanje - uzorci i testovi
- Regresijska analiza

1. Uvod u statistiku

Osnovni pojmovi

Statistika – znanost o podacima.
Uključuje prikupljanje, klasifikaciju, prikaz, obradu i interpretaciju podataka.

(Statistička) jedinica - objekt (stvar ili osoba) kojem se ispituje (mjeri) neko svojstvo. Sve jedinice zajedno tvore statistički skup.

Mjereno svojstvo se javlja u dva ili više oblika (modaliteta).
Npr. za spol imamo dva modaliteta: muški i ženski.

Obilježje ili varijabla - svojstvo koje se dobije opažanjem ili mjerenjem na statističkom skupu. Ono svojim oblikom varira od jedinice do jedinice statističkog skupa.

Rezultat svakog mjerenja nam predstavlja jedan **podatak**.

Primjer 1.1. Na Sveučilištu se željelo ispitati da li studenti podržavaju pokretanje plesnog tečaja. Od 650 ispitanih studenata, 435 je podržalo takvu ideju.

1. Što je statistička jedinica u ovom primjeru?
   a) Odgovor svakog studenta.
   b) Studenti koji podržavaju prijedlog.
√ c) Pojedini student.

2. Što je varijabla?
   a) Broj studenata koji podržavaju opciju.
   b) Broj studenata koji ne podržavaju opciju.
√ c) Odgovor (mišljenje) pojedinog studenta.

Rezimirajmo:

- statistički skup: Ω
- statistička jedinica: ω
- varijabla: $X : \Omega \to K$ (skup mogućih oblika pridruženih obilježju X),
- Mjerenjem (opažanjem) varijable X dobivamo podatke o nekom svojstvu na promatranom statističkom skupu.

Vrste podataka i varijabli

Ovisno o mjernoj skali obilježja (varijable) dijelimo na:

- kvantitativne – mjerene na numeričkoj skali, dopuštene brojčane operacije (visina, težina, primanja, broj bodova na kolokviju, ...).
- kvalitativne (kategorijalne) – vrijednosti su razredi (kategorije). Nisu dopuštene brojčane operacije (boja kose, spol, ...). U pojedinim slučajevima dopuštene su operacije uspoređivanja: $<, >, =$ (npr. ocjena na ispitu).

Analogno govorimo i o **kvantitativnim i kvalitativnim podacima** (dobivenim mjerenjima odgovarajućih varijabli).

Primjer 1.2.
Odredite kojeg tipa su sljedeći podaci:

1. Visina svakog studenta u razredu.
2. Duljina vremenskog razdoblja kojeg svatko od 30 promatranih pacijenata mora provesti u bolnici.
3. Politička stranka pojedinog zastupnika u Hrvatskom saboru.
4. Veličina stana (u m^2).

Vrste podataka

Primjer 1.3. Razmotrimo opet prvi primjer.
1. Bila su ponuđena tri odgovora, označena s a), b) i c), te je zabilježen odgovor svakog studenta. Kakav je to tip varijable?
2. Zamislimo da smo u istom primjeru ponuđene moguće odgovore označili s 1), 2) i 3), te ponovno bilježili odgovor svakog studenta. Je li u tom slučaju riječ o kvantitativnoj ili kvalitativnoj varijabli?

Iako su u drugom slučaju podaci brojčani, oni nemaju numeričko značenje (npr. nema smisla računati njihov prosjek). Oni su samo oznake razreda u koje svrstavamo odgovore, te stoga predstavljaju kvalitativne podatke.

Osnovni elementi statističke analize

Populacija ili **osnovni skup** – skup svih podataka koji opisuju neki fenomen koji nas zanima.

Primjer: Zanima nas vrijednost prosječnog primanja u Hrvatskoj. Što je populacija u ovom slučaju?
A što statiskički skup, odnosno statistička jedinica?

Uzorak - podskup podataka sakupljenih iz populacije.

Primjer 1.4. Voditelj prodaje tvrtke X želi ispitati starosnu dob kupaca koji kupuju njihove proizvode. U tu svrhu nasumično izaberu 742 kupca i utvrde da je prosjek njihovih godina 42.

1. Što je interesna populacija u ovom slučaju?
   a) Prosjek godina svih kupaca koji kupuju proizvode tvrtke X.
   b) Prosjek godina 742 ispitanih kupaca.
   c) Starosna dob svih kupaca koji kupuju proizvode tvrtke X.

2. Što je uzorak?
   a) Prosjek godina svih kupaca proizvoda tvrtke X.
   b) Starosna dob 742 ispitanih kupaca.
   c) 742 ispitana kupca.

3. Što je statistička jedinica?
   a) Proizvod tvrtke X.
   b) Starosna dob svakog kupca.
   c) Pojedini kupac proizvoda tvrtke X.
   d) Prosjek godina svih kupaca koji kupuju proizvode tvrtke X.

4. Što je varijabla?
   a) Proizvod tvrtke X.
   b) Starosna dob kupca proizvoda tvrtke X.
   c) Starosna dob 742 ispitanih kupaca.
   d) Prosjek godina svih kupaca koji kupuju proizvode tvrtke X.

Grane statistike

- Dizajn eksperimenta (experimental design) – bavi se planiranjem eksperimenta i prikupljanjem podataka.
- Opisna statistika (descriptive statistics) – grana statistike koja se bavi obradom i prikazom podataka.
- Statističko zaključivanje (inferential statistics) – na osnovu uzorka donosi zaključke o cijeloj populaciji.

Potonja se zaniva na nepotpunim podacima, te u sebi sadrži određenu komponentu nesigurnosti.

Međutim, ukoliko se koriste odgovarajuće metode (teroija vjerojatnosti), možemo također dobiti i vjerodostojnost takvih zaključaka.

Primjer 1.5. Vratimo se na prijašnji primjer.

1. Što od sljedećeg bi bio primjer statističkog zaključivanja?
   a) 742 kupca su nasumce izabrana.
   b) Prosjek godina 742 ispitanih kupaca je 42.
   c) Na osnovu uzorka, možemo s 90-postotnom sigurnošću zaključiti da je prosjek kupaca proizvoda tvrtke X između 40 i 44 godine.

Prikupljaje podataka

Reprezentativni uzorak

– posjeduje svojstva slična onima cijele populacije.

Kako izabrati takav uzorak?

Slučajni uzorak (random sample)

– od n statističkih jedinica je onaj uzorak u kojem svaka jedinica ima jednaku mogućnost da bude izabrana za uzorak.

Primjeri:

- Loto 6/45
- generator slučajnih brojeva
- izvlačenje iz šešira.

2. Grafičke metode prikaza podataka

Grafički prikaz kvalitativnih podataka

Kvalitativni (kategorijalni) podatci – nemaju numeričko značenje, mogu se samo razvrstavati u razrede ili kategorije.

Primjer 2.1. Uzet je uzorak od 176 zaposlenika jednog poduzeća, te je bilježena vrsta njihovog obrazovanja. Mogući odgovori su bili: srednja škola (S), bakalaureat(B), magisterij(M), doktorat(D), te ostalo (O).

Što su kategorije u ovom primjeru? S, B; M, D, O.

Što nas zanima?
Broj podataka koji pripada svakoj kategoriji.

Frekvencija ili učestalost

pojedinog razreda je broj podataka koji upadaju u tu kategoriju.

Što nas još zanima?
Postotak od ukupnog broja opažanja koji pripada pojedinoj kategoriji.

Relativna frekvencija ili učestalost

pojedinog razreda je postotak od ukupnog broja podataka koji pripada toj kategoriji. On je jednak omjeru frekvencije i ukupnog broja podataka.

$$\text{Relativna frekvencija} = \frac{\text{frekvencija}}{\text{ukupan broj podataka}}$$

Sažmimo prikupljene podatke u tzv. tablicu frekvencija.

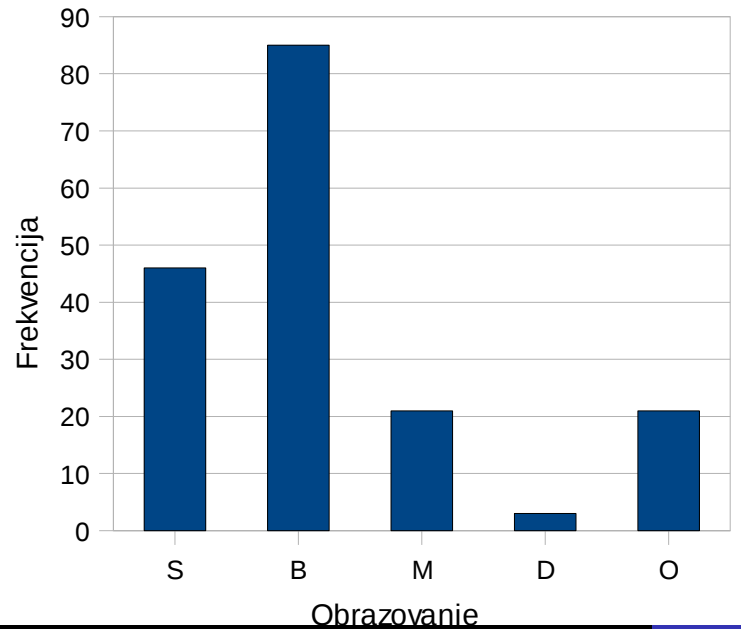
Razredi	**Frekvencija**	**Relativna frekvencija**
Srednja škola	46	26
Bakalaureat	85	48
Magisterij	21	12
Doktorat	3	2
Ostalo	21	12

Dvije najčešće korištene grafičke metode za prikaz kvalitativnih podataka su:
– **stupčasti dijagrami** (frekvencija ili relativnih frekvencija)
– **strukturni dijagrami** (prikazuju relativnu frekvenciju svakog razreda).

Konstrukcija stupčastog dijagrama

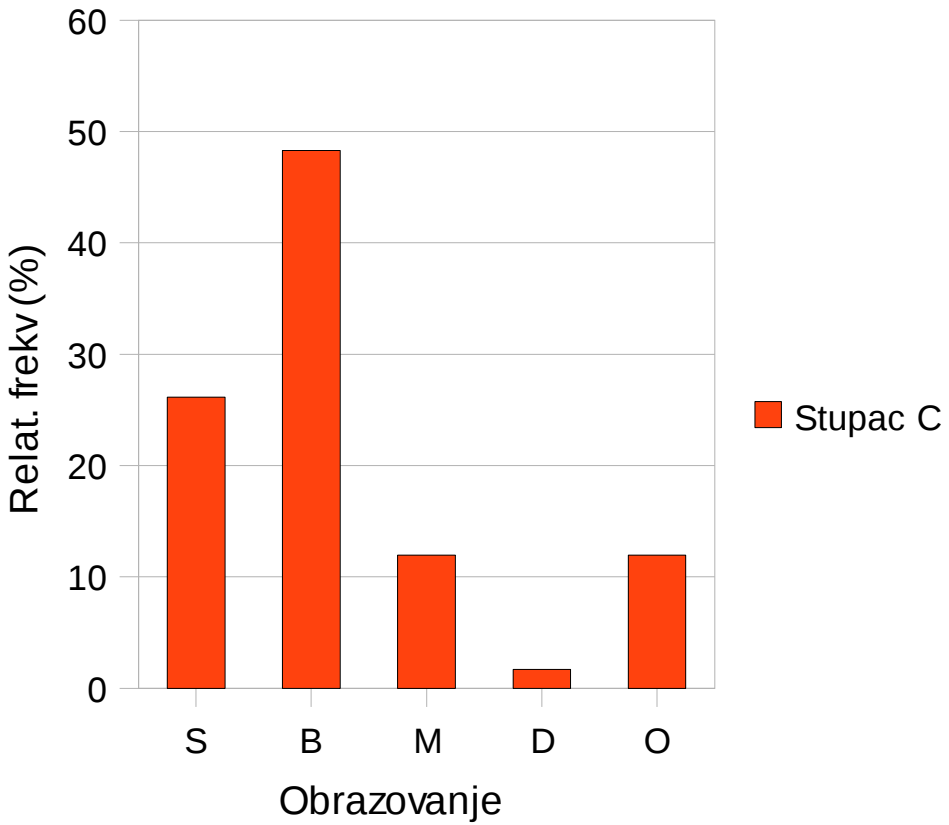
Nacrtajte horizontalnu i vertikalnu os na papiru. Vertikalna os predstavlja (relativnu) frekvenciju pojedinog razreda. Razrede označimo ispod horizontalne osi.
Nacrtajmo stupce jednakih širina iznad svake kategorije. Visina stupca mora biti razmjerna fekvenciji, odnosno relativnoj frekvenciji pojedinog razreda.

Dijagram frekvencija:



Konstrukcija stupčastog dijagrama

Stupčasti dijagram relativnih frekvencija:

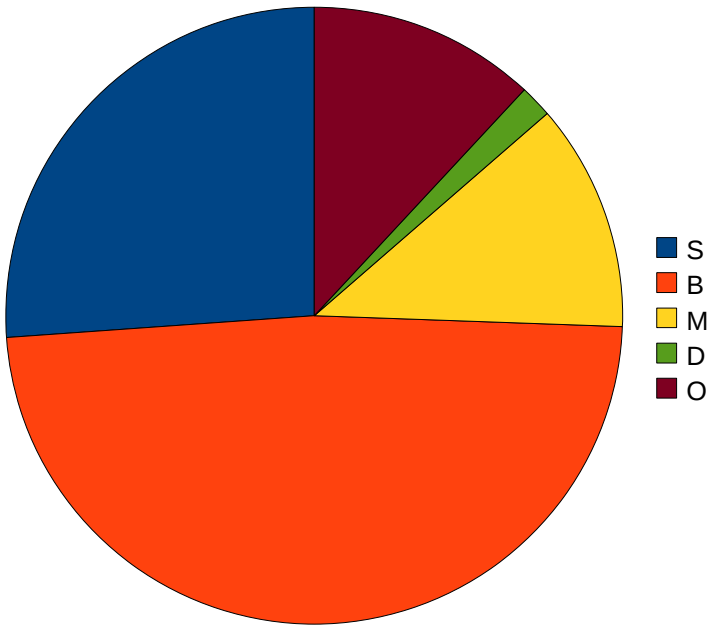


Konstrukcija stupčastog dijagrama - sažetak

- Prikažite podatke u tablici frekvencija. Ona mora sadržavati frekvenciju i relativnu frekvenciju za svaki razred.
- Nacrtajte horizontalnu i vertikalnu os. Ispod horizontalne osi označite razrede. Vertikalna os predstavlja (relativnu) frekvenciju pojedinog razreda.
- Nacrtajte stupce za svaki razred. Visina stupca mora biti razmjerna fekvenciji, odnosno relativnoj frekvenciji pojedinog razreda.

Konstrukcija strukturnog dijagrama

Pizza prikaz (*eng. pie chart*) sastoji se od kruga razdijeljenog na komade, od kojih svaki predstavlja pojedini razred. Veličina (kut) pojedinog komada razmjerna je relativnoj frekvenciji pripadnog razreda.



Grafički prikaz kvantitativnih podataka

Dvije najpopularnije vrste grafičkog prikaza kvantitativnih podataka su

- **S-L prikaz** (stablo i list, *eng. stem and leaf*)
- **histogram**.

Oni prikazuju:
– bilo broj podataka koji padaju u svaki razred (frekvencija razreda)
– bilo postotak od ukupnog broja podataka koji upada u pojedini razred (relativna frekvencija razreda).

S-L prikaz

Za mali broj ($\leqslant 30$) podatka, prikaz se jednostavno radi ručno.

Primjer 2.2. Napravite S-L prikaz za sljedeće podatke:

Cijene (u 1000 kn)				
660	595	1060	500	630
899	1295	749	820	843
710	950	720	575	760
1090	770	682	1016	650
425	367	1480	945	1120

Rješenje:
Nađite najmanju i najveću vrijednost u skupu podataka: 367, 1480 (36 700kn, 1 480 000kn)
Formirajmo razrede (intervale). Npr. prvi razred mogu biti sve cijene između 300 i 400. Tu bi spadale cijene: **3**67, **3**24, **3**56.
Sve one imaju zajedničku početnu znamenku – 3 (na mjestu stotica).

S-L prikaz

Zajedničku znamenku uzimamo za predstavnika razreda - **stablo**.
Preostale dvije znamenke (na mjestu desetica i jedinica) nam tvore **list**.
Tako za svaki podatak – znamenke na mjestu desetica i jedinica predstavljaju list, a preostale predstavljaju stablo.
Npr. za najmnaji i najveći podatak:

Stablo	List
3	67
14	80

S-L prikaz se sastoji od dva stupca:
– u prvom poredamo sva moguća stabla po veličini (od najmanjeg prema najvećem)
– u drugom stavljamo list za svaki podatak u redak uz odgovarajuće stablo.
Tome zatim možemo pridodati (relativne) frekvencije pridružene svakom pojedinom stablu.

S-L prikaz

Stablo	List	Frekvencija	Relat. frekv.(%)
3	67	1	4
4	25	1	4
5	00,75,95	3	12
6	30,50,60,82	4	16
7	10,20,49,60,70	5	20
8	20,43,99	3	12
9	45,50	2	8
10	16,60,90	3	12
11	20	3	12
12	95	1	4
13		0	0
14	80	1	4
	Ukupno:	25	100

S-L prikaz

Prednosti S-L prikaza:

- sačuvani su originalni podaci
- podaci su poredani po veličini
- razredi se jednostavno određuju.

Nedostaci:

- Nedostatak fleksibilnosti u izboru stabla.

Primjer 2.3. Razmotrimo podatke iz zadnjeg primjera. Neka samo zadnja znamenka (na mjestu jedinica) predstavlja list.

Kako bi izgledao prikaz?

Stablo	List
36	7
⋮	⋮
148	0

Koliki je broj razreda? 148-35=113 (za 25 podataka!)
Prikaz ne daje jasnu sliku (informaciju) o podacima!

S-L prikaz

Treća mogućnost je da zadje tri znamenke predstavljaju listove:

Stablo	List
0	$\cdots$
1	$\cdots$

Koliki je sada broj razreda? 2!
Podaci su previše zbijeni. Ni ovaj izbor ne daje jasan prikaz podataka.

S-L prikaz - sažetak

- Nađite najmanju i najveću vrijednost.
- Odredite kako ćete definirati stabla, odnosno listove.
- Poredajte stabla u stupac, polazeći od najmanjeg.
- Za svaki podatak, stavite list u redak pripadnog stabla. Poredajte listove po veličini.

S-L prikaz

Zadatak 2.1. Zadani su sljedeći podatci:

5.9	11.2	1.6	7.4	8.6	1.2	2.1
4	7.35	8.4	8.9	6.7	4.5	6.3
7.6	9.7	3.51	1.1	4.3	3.3	8.4
1.6	8.2	6.5	1.1	5	9.4	6.4

Koristeći znamenku jedinica kao stablo konstruirajte S-L prikaz s pripadnim (relativnim) frekvencijama.

Histogram

Histogrami frekvencija, odnosno relativnih frekvencija:

- podesniji za veće skupove podataka, te mjerenja s većim brojem znamenaka
- slični stupčastim dijagramima.

Primjer 2.4. U tablici su dana 20 mjerenja slučajne varijable. Konstruirajte histogram frekvencija za taj skup podataka.

26	21	32	28	17
22	26	25	30	27
34	32	36	38	39
12	39	31	23	19

Histogram

Odredimo najmanju i najveću vrijednost: 12, 39.
Odredimo interval nad kojim ćemo crtati histogram:
– početna točka intervala $\leqslant$ najmanjeg podatka
– završna točka intervala $\geqslant$ najvećeg podatka.

$$[10,\ 40]$$

Odredimo podintervale (razrede).
Birat ćemo podintervale jednakih širina:

$$\text{širina razreda} = \frac{\text{širina cijelog intervala}}{\text{broj intervala}}$$

Širina intervala = 40-10=30.
Broj intervala – uzmimo ih 6.

Širina razreda$=\frac{30}{6} = 5$.

Histogram

Prvi razred: 10–15
Drugi razred: 15–20
Treći razred: 20–25
itd.
Sto ako se neki podatak nalazi na granici dvaju razreda (npr. 25)?
U koji razred ga svrstati: u treći, u četvrti, u oba...?

Svaki podatak se mora razvrstati točno u jedan razred!

Na koji način razvrstati ovakve granične podatke odlučujemo sami, ali moramo biti dosljedni s pravilom kojeg smo izabrali.
Mi ćemo se držati pravila da podatke na granici dvaju razreda svrstavamo u gornji razred.
Drugim riječima, precizniji zapis naših razreda bi bio:
Prvi razred: $[10, 15\rangle$
Drugi razred: $[15, 20\rangle$
itd.
Šesti (zadnji) razred: $[35, 40]$
Svi razredi su poluotvoreni inetrvali, osim zadnjeg koji je zatvoren.

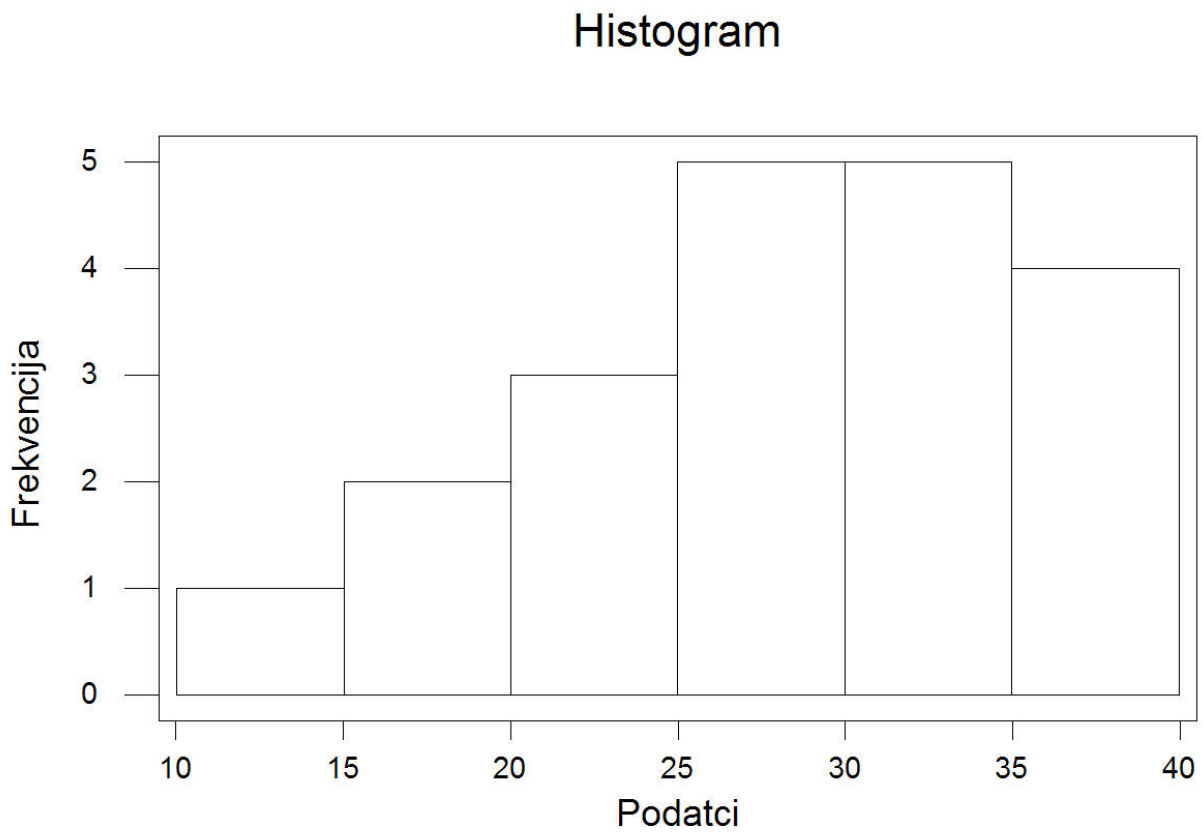
Histogram

Napravimo **tablicu frekvencija**:

Razred	Interval	Brojač	Frekvencija	Relat. frekv.(%)
1	10-15	/	1	5
2	15-20	//	2	10
3	20-25	///	3	15
4	25-30	////	5	25
5	30-35	////	5	25
6	35-40	////	4	20
		Ukupno:	20	100

Sad možemo nacrtatai histogram.

Histogram



Interpretacija:
Površina stupca iznad nekog intervala razmjerna je postotku podataka u tom intervalu.
Većina vrijednosti (70%) se nalazi u posljednja 3 razreda.

Histogram

Na koji način ćemo određivati broj razreda pri konstrukciji histograma?

Iskustveno pravilo za određivanje broja razreda u histogramu.

Broj podataka	Broj razreda
Manji od 25	5 ili 6
25 – 50	7–14
Veći od 50	15–20

Osim toga, koristi se i sljedeće pravilo.

Sturgesovo pravilo za određivanje broja razreda u histogramu.

$$k \approx 1 + 3.3 \log n$$

n – broj podataka

k – broj razreda

Razlike između stupčastog dijagrama i histograma:

- kod prvog razredi nisu povezani
- kod potonjeg razredi su sljedujući intervali na realnoj osi (gornja granica jednog intervala predstavlja ujedno donju granicu sljedećeg).

Interpretacija je slična

- visina stupca je proporcionalna (relativnoj) frekvenciji razreda
- površina pojedinog stupca razmjerna je postotku podataka u pripadnom razredu.

Zadatak 2.2. Konstruirajte histogram relativnih frekvencija za donje podatke. Izaberite odgovarajući broj razreda.

59	53	16	74	86	12	21
40	73	84	89	67	45	63
76	97	35	11	43	33	84
16	82	65	11	50	94	64