Primjena računara u biologiji

Vladimir Filipović

vladaf@matf.bg.ac.rs

Kvalitativni i kvantitativni podaci

Deskriptivna statistika

Deskriptivna statistika

Deskriptivna statistika sadrži metode i procedure za prezentovanje i sumiranje podataka.

Svrha deskriptivne statistike je da pomoću nekoliko brojeva opiše značenje podataka koji stoje iza njih. Podaci se dobijaju na osnovu opservacija na skupu različitih slučajeva koji mogu biti ljudi, životinje, gradovi, škole, različiti događaji ili neka kombinacija svega navedenog.

Deskriptivna statistika je obično prvi korak u analizi podataka, a služi za opisivanje prikupljenih podataka.

Deskriptivna statistika obično prethodi statističkom zaključivanju i predviđanju, ali može biti i krajnji cilj statističke analize. Izvođenjem zaključaka se bavi drugo područje statistike koje se zove statistika zaključivanja.

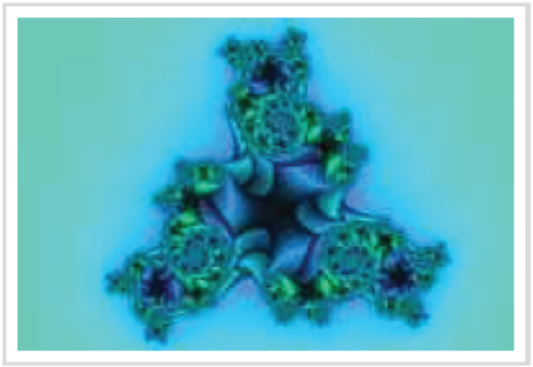
Deskriptivna statistika(2)

Najčešće korišćene procedure u deskriptivnoj statistici su grafičko i tabelarno prikazivanje podataka i izračunavanje mera centralne tendencije i varijabiliteta.

Većina autora svrstava mere korelacije i asocijacije varijabli u deskriptivnu statistiku, jer opisuju vezu između dve ili više varijable.

Kvalitativni podaci

Kvalitativni podaci



Uzorak podataka je ***kvalitativan*** (još se označava i terminom ***kategorijski***) ako njegove vrednosti pripadaju kolekciji poznatih klasa koje se ne preklapaju.

Kvalitativni podaci imaju opisni karakter i ne mogu se predstaviti brojevima.

Primer. Kvantitativni uzorci podataka su ocene kvaliteta robe (A, B, C, D), rejting obveznica (AAA, AAB, ...), veličina odeće, boja materijala utrošenog pri proizvodnji odeće, itd.

Kvalitativni podaci mogu biti:

- nominalni i
- ordinabilni.

Kod **nominalnh** podataka ne postoji mogućnost uređenja (npr. prebivalište, država/region iz koje potiču lajkovi za neku veb stranu itd.)

Kod **ordinabilnih** podataka postoji opšti kritrijum po kome se mogu urediti (npr. stručna sprema, proizvodni lanac, horoskopski znaci itd.).

Kvalitativni podaci (2)

U razmatranjima koja slede, koristi se okvir sa podacima painters, koji sadrži informacije o slikarima do polovine XVIII veka. Ovaj okvir sa podacima se nalazi u sastavu biblioteke MASS – da bi se on mogao koristiti, potrebno je prvo učitati biblioteku MASS.

Učitavanje biblioteke u sistem R se realizuje pomoću funkcije **library**.

Primer. Sledećom naredbom se učitava biblioteka MASS:

```
> library(MASS)      # load the MASS package
```

Po uspešnom učitavanju biblioteke MASS, okvir sa podacima painters je na raspolaganju korisniku, pa se isti može prikazati:

```
> painters
              Composition Drawing Colour Expression School
Da Udine               10       8     16          3      A
Da Vinci               15      16      4         14      A
Del Piombo              8      13     16          7      A
Del Sarto              12      16      9          8      A
Fr. Penni               0      15      8          0      A
Guilio Romano          15      16      4         14      A
                   ...............
```

Kvalitativni podaci (3)

Primer. Poslednja kolona, nazvana School, okvira sa podacima painters sadrži informacije o slikarskoj školi kojoj je pripadao dati klasični slikar. Škole su označene slovima A, B, C itd. i predstavljaju kvalitativne podatke.
Isecanjem po koloni School se mogu jasnije videti informacije o školi u okviru sa podacima painters:

```
> painters$School
 [1] A A A A A A A A A A B B B B B B C C C C C C D D D D
[27] D D D D D D E E E E E E E F F F F G G G G G G G H H
[53] H H
Levels: A B C D E F G H
```

Dodatne informacije o okviru sa podacima painters mogu se naći u dokumentaciji sistema R:

```
> help(painters)
```

Raspodela frekfencija kod kvalitativnih podataka

Raspodela frekfencija za promenljivu koja sadrži podatke je sumarni pregled ***broja javljanja*** podataka u kolekciji kategorija koje se međusobno ne preklapaju.

Primer. U okviru sa podacima painters, raspodela frekefencija za promenljivu School je sumarni pregled informacija koliko slikara pripada kojoj slikarskoj školi. Škole su označene slovima A, B, C itd.
Problem. Odrediti raspodelu frekefencija za promenljivu School u okviru sa podacima painters.

Rešenje. Za određivanje raspodele frekefencija biće primenjena funkcija **table**:

```
> library(MASS)                 # load the MASS package
> school = painters$School      # the painter schools
> school.freq = table(school)   # apply the table function
```

Dakle, raspodela frekfencija za promenljivu School u okviru sa podacima painters je:

```
> school.freq
school
 A  B  C  D  E  F  G  H
10  6  6 10  7  4  7  4
```

Raspodela frekfencija kod kvalitativnih podataka (2)

Problem. Odrediti raspodelu frekefencija za promenljivu School u okviru sa podacima painters. Dobijenu raspodelu frekfencija prikazati kao kolonu, a ne kao vrstu.

Rešenje. Za određivanje raspodele frekefencija biće primenjena funkcija **table**, a za prikaz u obliku kolone funkcija **cbind**.

```
> library(MASS)                 # load the MASS package
> school = painters$School      # the painter schools
> cbind(school.freq)
  school.freq
A          10
B           6
C           6
D          10
E           7
F           4
G           7
H           4
```

Raspodela frekfencija kod kvalitativnih podataka (3)

Zadatak 1. Odrediti raspodelu frekfencija za promenljivu koja opisuje ocenu za kompoziciju (composition score) u okviru sa podacima painters.

Zadatak 2. Odrediti raspodelu frekfencija za promenljivu koja opisuje ocenu za kompoziciju (composition score) u okviru sa podacima painters. Dobijenu raspodelu frekfencija prikazati kao kolonu, a ne kao vrstu.

Zadatak 3. Odredi (programirajući, ne „ručnim“ prebrojavanjem) koja škola u okviru sa podacima painters ima najviše učenika.

Relativna raspodela frekfencija kod kvalitativnih podataka

Relativna raspodela frekfencija za promenljivu koja sadrži podatke je sumarni pregled ***učešća*** podataka u kolekciji kategorija koje se međusobno ne preklapaju.

Veza između frekfencije i relativne frekfencije je data sledećom formulom:

$$Relativna\_frekfencija = \frac{Frekfencija}{Veličina\_uzorka}$$

Primer. U okviru sa podacima painters, relativna raspodela frekefencija za promenljivu School je sumarni pregled učešća broja slikara koji pripadaju datoj slikarskoj školi.
Problem. Odrediti relativnu raspodelu frekefencija za promenljivu School u okviru sa podacima painters.

Rešenje. Radi određivanja relativne raspodele frekefencija za promenljivu School, biće prvo određena raspodela frekfencija za promenljivu School:

```
> library(MASS)                 # load the MASS package
> school = painters$School      # the painter schools
> school.freq = table(school)   # apply the table function
```

Relativna raspodela frekfencija kod kvalitativnih podataka (2)

Rešenje(nastavak). da bi se odredila relativna frekfencija svake od ocena, broj ocena treba podeliti sa veličinom uzorka. Veličina uzorka se može odrediti primenom funkcije nrow na okvir sa podacima.

```
> school.relfreq = school.freq / nrow(painters)
```

Na taj način, promenljiva school.relfreq sadrži relativnu frekfenciju ocena, tj. sledeće vrednosti:

```
> school.relfreq
school
       A        B        C        D        E        F
0.185185 0.111111 0.111111 0.185185 0.129630 0.074074
       G        H
0.129630 0.074074
```

Relativna raspodela frekfencija kod kvalitativnih podataka (3)

Problem. Odrediti relativnu raspodelu frekefencija za promenljivu School u okviru sa podacima painters, tako da se relativne frekfence zaokruže na dve decimale.

Dobijene realtivne frekfencije prikazati u formi vrste i u formi kolone

Rešenje. Relativne frekfencije se računaju na isti način kao u prethodnom

```
> library(MASS)                  # load the MASS package
> school = painters$School       # the painter schools
> school.freq = table(school)    # apply the table function
> school.relfreq = school.freq / nrow(painters)
```

Korišćenjem funkcije **options** sa imenovnim argumentom **digits** se podešava broj cifara za prikaz. Ova funkcija, pored postavljanja novog podešavanja, kao rezulata vraće ranije postavljena podešavanja, što se može iskoristiti da se stara

```
> old = options(digits=1)
> school.relfreq
school
   A    B    C    D    E    F    G    H
0.19 0.11 0.11 0.19 0.13 0.07 0.13 0.07
> options(old)
```

Relativna raspodela frekfencija kod kvalitativnih podataka (4)

Rešenje (nastavak). Prikaz dobijenih rezultata u obliku kolone se postiže korišćenjem funkcije cbind:

```
> old = options(digits=1)
> cbind(school.relfreq)
  school.relfreq
A           0.19
B           0.11
C           0.11
D           0.19
E           0.13
F           0.07
G           0.13
> options(old)    # restore the old option
```

Zadatak 1. Odrediti relativnu raspodelu frekfencija za promenljivu koja opisuje ocenu za kompoziciju u okviru sa podacima painters.

Stubičasti dijagram

Stubičasti dijagram (bar graph) za kvalitativni uzorak podataka se sastoji od paralelnih vertikalnih stubova kojima se grafički prikazuje raspodela frekfenci.

Ova vrsta dijagrama se dominantno koriste kod ordinalnih podataka.

Problem. Oformiti stubičasti dijagram za promenljivu School okvira sa podacima painters.

Rešenje. Prvo treba odrediti raspodelu frekfencija promenljive School:

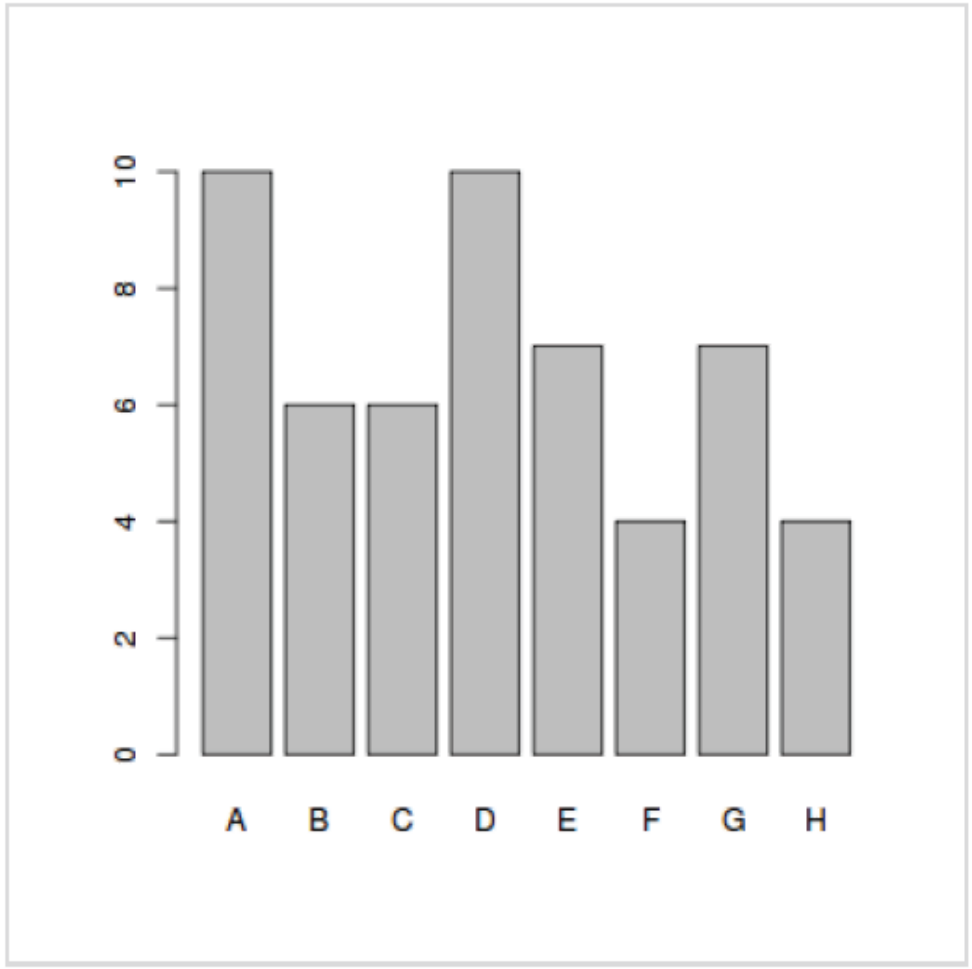
```
> library(MASS)                 # load the MASS package
> school = painters$School      # the painter schools
> school.freq = table(school)   # apply the table function
```

Dijagram se dobija primenom **barplot** funkcije na promenljivu koja sadrži raspodelu frekfencija:

```
> barplot(school.freq)          # apply the barplot function
```

Stubičasti dijagram (2)

Rešenje (nastavak). Dijagram dobijen primenom funkcije barplot ima sledeći izgled:



Stubičasti dijagram (3)

Problem. Oformiti obojeni stubičasti dijagram za promenljivu School okvira sa podacima painters.

Rešenje. Prvo, isto kao u prethodnom primeru, treba odrediti raspodelu frekfencija promenljive School:

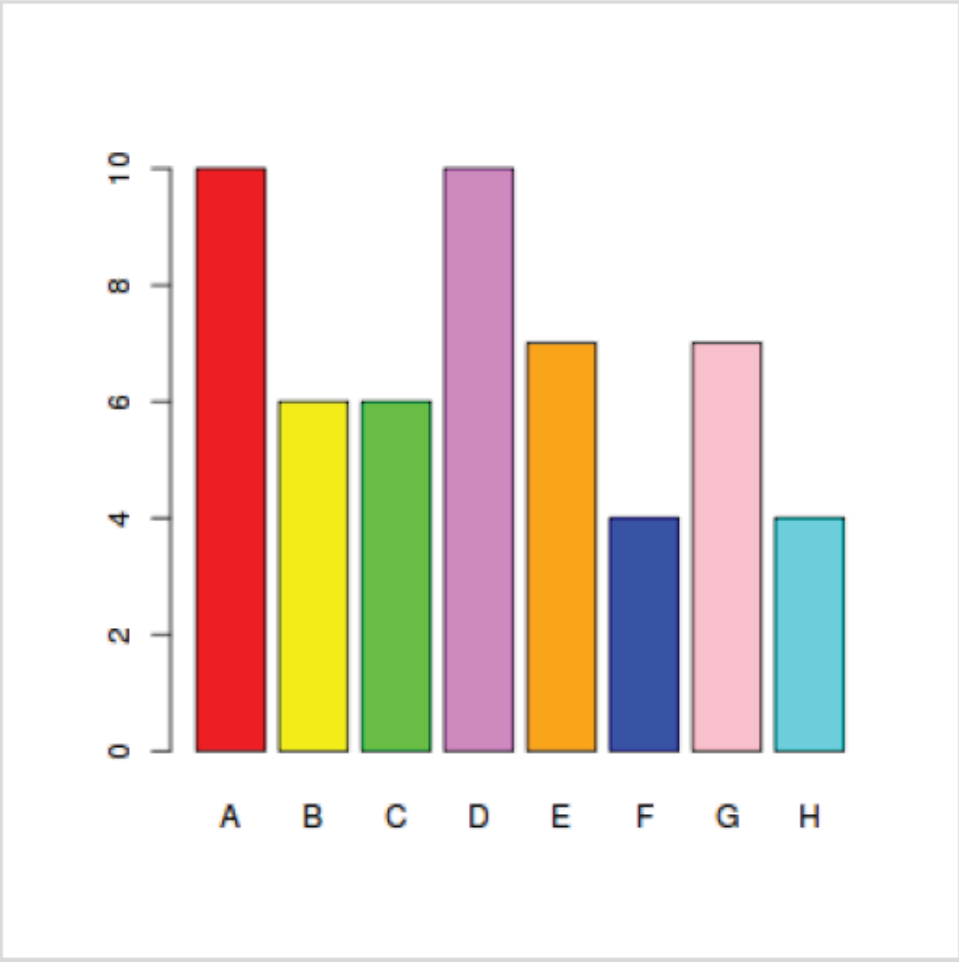
```
> library(MASS)                 # load the MASS package
> school = painters$School      # the painter schools
> school.freq = table(school)   # apply the table function
```

Da bi se obojili stubovi stubičastog dijagrama, potrebno je formirati vektor boja i tako oformljeni vektor proslediti kao imenovani argument **col** prilikom poziva funkcije barplot. U ovom primeru promenljiva colors ima vrednost vektora boja koji se koriste za bojenje stubića:

```
> colors = c("red", "yellow", "green", "violet",
+   "orange", "blue", "pink", "cyan")
> barplot(school.freq,          # apply the barplot function
+   col=colors)                 # set the color palette
```

Stubičasti dijagram (4)

Rešenje (nastavak). Dijagram dobijen primenom funkcije barplot ima sledeći izgled:



Stubičasti dijagram (5)

Zadatak 1. Oformiti stubičasti dijagram za za promenljivu koja opisuje ocenu za kompoziciju u okviru sa podacima painters.

Kružni dijagram

Kružni dijagram ili ***pita (pie chart)*** se sastoji od kružnih isečaka koji graafički prikazuju raspodelu frekfencija.

Družni dijagram se dominantno koristi kod nominalnih podataka.

Problem. Oformiti kružni dijagram za promenljivu School okvira sa podacima painters.

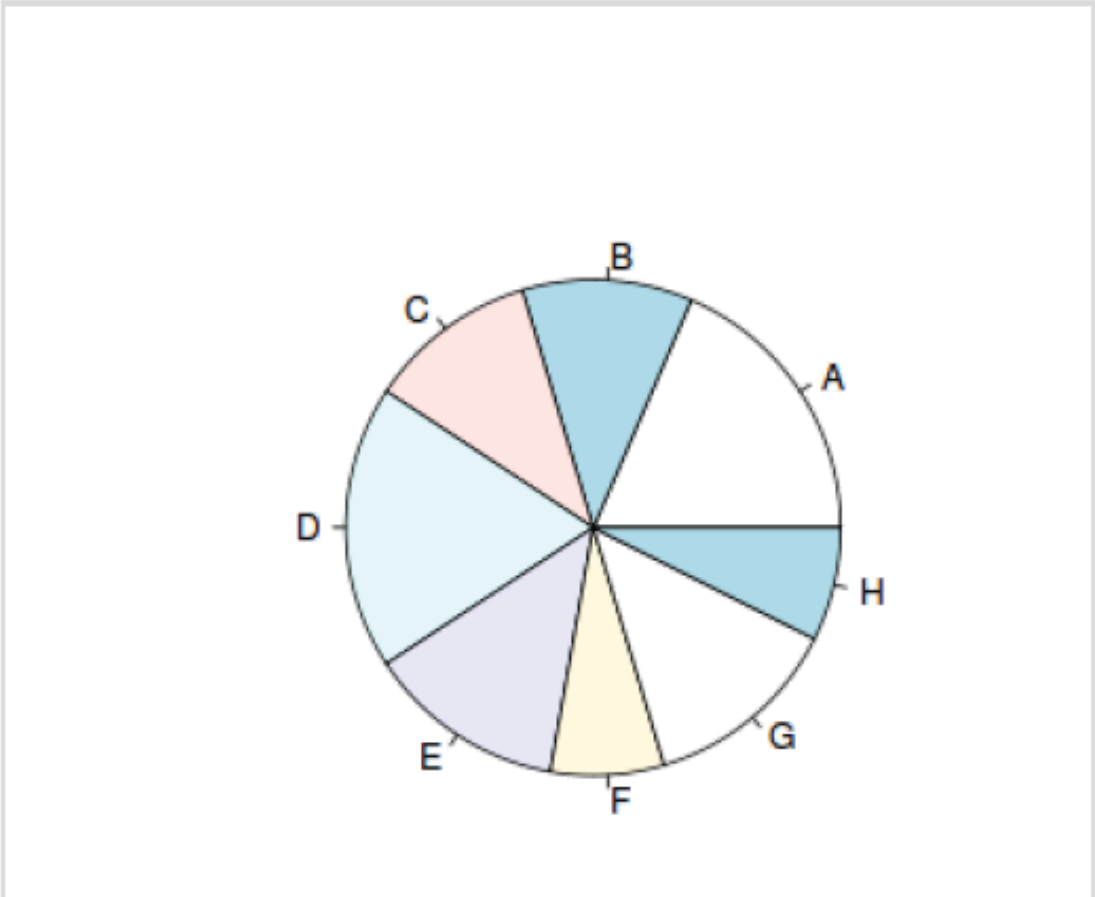
```
> library(MASS)                 # load the MASS package
> school = painters$School      # the painter schools
> school.freq = table(school)   # apply the table function
```

Dijagram se dobija primenom **pie** funkcije na promenljivu koja sadrži raspodelu frekfencija:

```
> pie(school.freq)              # apply the pie function
```

Kružni dijagram (2)

Rešenje (nastavak). Kružni dijagram dobijen primenom funkcije pie ima sledeći izgled:



Kružni dijagram (3)

Problem. Oformiti obojeni kružni dijagram za promenljivu School okvira sa podacima painters.

Rešenje. Prvo, isto kao u prethodnom primeru, treba odrediti raspodelu frekfencija promenljive School:

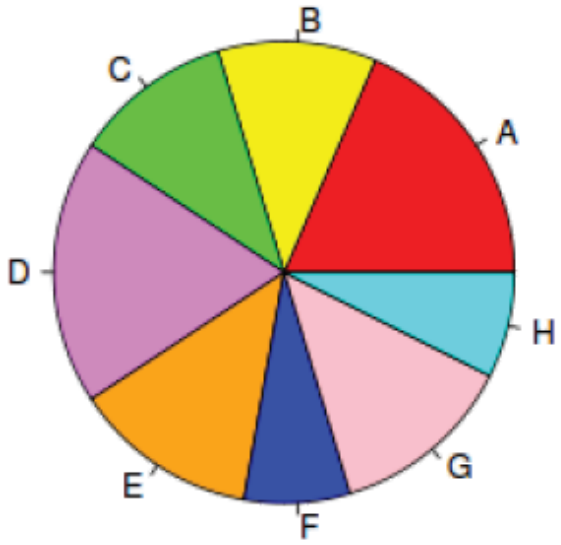
```
> library(MASS)                 # load the MASS package
> school = painters$School      # the painter schools
> school.freq = table(school)   # apply the table function
```

Da bi se obojili isečci kružnog dijagrama, potrebno je formirati vektor boja i tako oformljeni vektor proslediti kao imenovani argument **col** prilikom poziva funkcije pie:

```
> colors = c("red", "yellow", "green", "violet",
+   "orange", "blue", "pink", "cyan")
> pie(school.freq,              # apply the pie function
+   col=colors)                 # set the color palette
```

Kružni dijagram (4)

Rešenje (nastavak). Kružni dijagram dobijen primenom funkcije pie nad neimenovanim argumentom school.freq i imenovanim argumentom (col je ime argumenta) colors koji ima sledeći izgled:



Kružni dijagram (5)

Zadatak 1. Oformiti kružni dijagram za za promenljivu koja opisuje ocenu za kompoziciju u okviru sa podacima painters.

Statistike za kategoriju

Ponekad je potrebno odrediti sredinu, ali ne za sve podatke, već samo za podatke iz date kategorije.

Jedan način da se to postigne je da se prvo izdvoje podaci koji se imaju jednu vrednost date kategorije, a da se potom na te podatke primeni **mean** funkcija.

Primer. Okvira sa podacima painters sadrži informacije o slikarima koji pripadaju različitim slikarskim školama. Svaka od škola u okviru sa podacima može biti opisana različitim statistikama, kao što su prosečne ocene za kompoziciju, za korišćenje boja i za izražajnost.

Prepostavimo da treba utvrditi koja škola ima najvišu prosečnu ocenu za kompoziciju. Jedan način da se dobije ta informacija je da se odvojeno izračuna prosečna ocena za kompoziciju za pripadnike svake od slikarskih škola (A, B, C, D i E), pa da se potom te ocene međusobno uporede.

Statistike za kategoriju (2)

Problem. Odrediti prosečnu ocenu za kompoziciju za pripadnike slikarske škole C.

Rešenje. U prvom koraku se kreira logički vektor c_school, za isecanje onih vrsta okvira sa podacima u kojima su informacije o slikarima iz škole C:

```
> library(MASS)                 # load the MASS package
> school = painters$School      # the painter schools
> c_school = school == "C"      # the logical index vector
```

U drugom koraku se određuju podaci o slikarima iz škole C i postavlja promenljiva c_painters da referiše na njih:

```
> c_painters = painters[c_school, ]  # child data set
```

U trećem koraku se primenom funkcije mean određuje prosek ocene za kompoziciju kod slikara iz škole C:

```
> mean(c_painters$Composition)
[1] 13.167
```

Dakle, dobijeni prosek ocena je 13.167.

Statistike za kategoriju (3)

Ponekad je potrebno odrediti sredinu, ali ne za sve podatke, već samo za podatke iz date kategorije.

Drugi način da se odredi sredina za podatke date katogorije, istovremeno za sve vrednosti koje može da uzme ta kategorija, je primenom funkcije **tapply**.

Primer. Odrediti istovremeno prosečne ocene za kompoziciju pripadnika slikarskih škola, za sve postojeće slikarske škole A, B, C. D i E.

Rešenje. Primenom funkcije tapply se jednom naredbom dolazi do traženog rezultata:

```
> tapply(painters$Composition, painters$School, mean)
     A      B      C      D      E      F      G      H
10.400 12.167 13.167  9.100 13.571  7.250 13.857 14.000
```

Zadatak 1. Odrediti programiranjem (bez „ručnog“ poređenja) školu čiji pripadnici u proseku imaju najvišu ocenu za kompoziciju.

Zadatak 2. Odrediti procenat slikara čija je ocena za korišćenje boja veća ili jednaka od 14.

Kvantitativni podaci

Kvantitativni podaci

Kvantitativni podaci se sastoje od numeričkih vrednosti.

Primer. Kvanititativni podaci su: količina padavina, prihod, temperatura itd.

Kvantitativni podaci mogu biti:

- diskretni i
- neprekidni.

Diskretni podaci predstavljaju one veličine koje uzimaju vrednost iz prebrojivog domena (npr. broj dece u porodici, broj kišnih/sunčanih dana u određenom periodu posmatranja, broj lajkova za datu veb stranu itd.).

Neprekidne su one veličine koje imaju neprekidnu prirodu i koje su merljive (npr. težina, visina, dužina trajanja proizvodnog ciklusa itd.).

Kvantitativni podaci (2)

U razmatranjima koja slede, koristi se ugrađeni okvir sa podacima faithful, koji sadrži informacije o posmatranjima gejzira Old Faithfull u Nacionalnom parku Jelouston, USA.

S obzirom da je faithful ugrađen okvir sa podacima, to korišćenje ovog okvira sa podacima ne zahteva da prethodno bude učitana bilo kakva biblioteka.

Primer. Sledećom naredbom se prikazuje prikaz dela okvira sa podacima, korišćenjem head funkcije:

```
> head(faithful)
  eruptions waiting
1     3.600      79
2     1.800      54
3     3.333      74
4     2.283      62
5     4.533      85
6     2.883      55
```

Okvir sa podacima sadrži dve observacije promenljivih u okviru sa podacima. Prva među njima, nazvana eruptions, sadrži trajanje erupcije gejzira. Druga se zove waiting i ona sadrži trajanje čekanja između dve erupcije.

Raspodela frekfencija kod kvantitativnih podataka

Raspodela frekfencija za promenljivu koja sadrži podatke je sumarni pregled ***broja javljanja*** podataka u kolekciji kategorija koje se međusobno ne preklapaju.

Primer. U okviru sa podacima faithful, raspodela frekefencija za promenljivu eruptions je sumarni pregled trajanja erupcija u skladu sa usvojenom klasifikacijom trajanja.
Problem. Odrediti raspodelu frekefencija za promenljivu eruptions u okviru sa podacima faithful.

Rešenje. Prvo se promenljivoj durations dodeli vektor vrednosti čija se raspodela frekfencija očekuje.

Potom treba odrediti klasifikaciju trajanja erupcija. Da bi se to uspešno uradilo, potrebno je, odrediti u kom opsegu vrednosti se nalaze opservirane vrednosti promenljive eruptions. Opseg se određuje funkcijom **range**.

```
> duration = faithful$eruptions
> range(duration)
[1] 1.6 5.1
```

U ovom slučaju je opseg vrednosti za trajanje erupcija interval [1.6, 5.1]. Sada se može preći na definisanje intervala koji se ne preklapaju kako bi se prebroijlo koliko ima trajanja erupcija u svakom od ovih intervala.

Raspodela frekfencija kod kvantitativnih podataka (2)

Rešenje (nastavak). Potredno je da intervali koji se ne preklapaju „pokriju“ ceo opseg. Poželjno je da intervali budu iste širine, kao i da krajevi tih intervala budu „okrugli“ brojevi.

U konkretnom slučaju je usvojeno da intervali budu širine 0.5, a da krajevi intervala budu smešteni u sekvenci na koju referiše promenljiva breaks. Kreiranje sekvence sa datim granicama je realizovano korišćenjem funkcije **seq**:

```
> breaks = seq(1.5, 5.5, by=0.5)    # half-integer sequence
> breaks
[1] 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0 5.5
```

Po određivanju granica intervala, pozivom funkcije **cut**, se kreiraju intervali sa granicama breaks koji bivaju pridruženi podacima duration. S obzirom da treba kreirati poluotvorene intervale (sadrže levi kraj, ali ne sadrže desni) to je imenovani argument **right** postavljen na **FALSE**. Dobijeni rezulat se dodeljuje promenljivoj duration.cut:

```
> duration.cut = cut(duration, breaks, right=FALSE)
```

Korišššćenjem funkcije **table**, određuje se frekfencija erupcija u svakom od intervala i dodeljuje promenljivoj duration.freq:

```
> duration.freq = table(duration.cut)
```

Raspodela frekfencija kod kvantitativnih podataka (3)

Rešenje (nastavak). Dobijena raspodela frekfencija je:

```
> duration.freq
duration.cut
[1.5,2) [2,2.5) [2.5,3) [3,3.5) [3.5,4) [4,4.5) [4.5,5)
     51      41       5       7      30      73      61
[5,5.5)
      4
```

Problem. Odrediti raspodelu frekefencija za promenljivu eruptions u okviru sa podacima faithful i prikazati je po kolonama.

Rešenje. Određivanje raspodele frekfencija prikazane po vrstama je isto kao u prethodnom primeru i promenljiva duration.freq referiše na tu raspodelu.

Prikaz po kolonama se realizuje pomoću funkcije cbind:

```
> cbind(duration.freq)
        duration.freq
[1.5,2)            51
[2,2.5)            41
[2.5,3)             5
[3,3.5)             7
[3.5,4)            30
[4,4.5)            73
[4.5,5)            61
[5,5.5)             4
```

Raspodela frekfencija kod kvantitativnih podataka (4)

Napomena. Dokumentacija sistema R upućuje da se bolje performanse postižu ako se za kreiranje raspodele frekfencija kod kvantitativnih podataka umesto prethodno opisanog pristupa koristi **hist** funkcija.

Zadatak 1. Odrediti raspodelu frekefencija za promenljivu waiting u okviru sa podacima faithful.
Zadatak 2. Odrediti programskim putem interval za trajanje koji sadrži najviše erupcija.

Histogram

Histogram se sastoji od paralelnih vertikalnih stubova kojima se grafički prikazuje raspodela frekfenci za kvantitativni uzorak podataka. Površina svakog od stubova je proporcionalna frekfenciji podataka iz klase koja odgovara datom stubu.

Ova vrsta dijagrama se dominantno koriste kod neprekidnih podataka.

Problem. Odrediti histogram za trajanje erupcije (tj. promenljivu eruptions) u okviru sa podacima faithful.

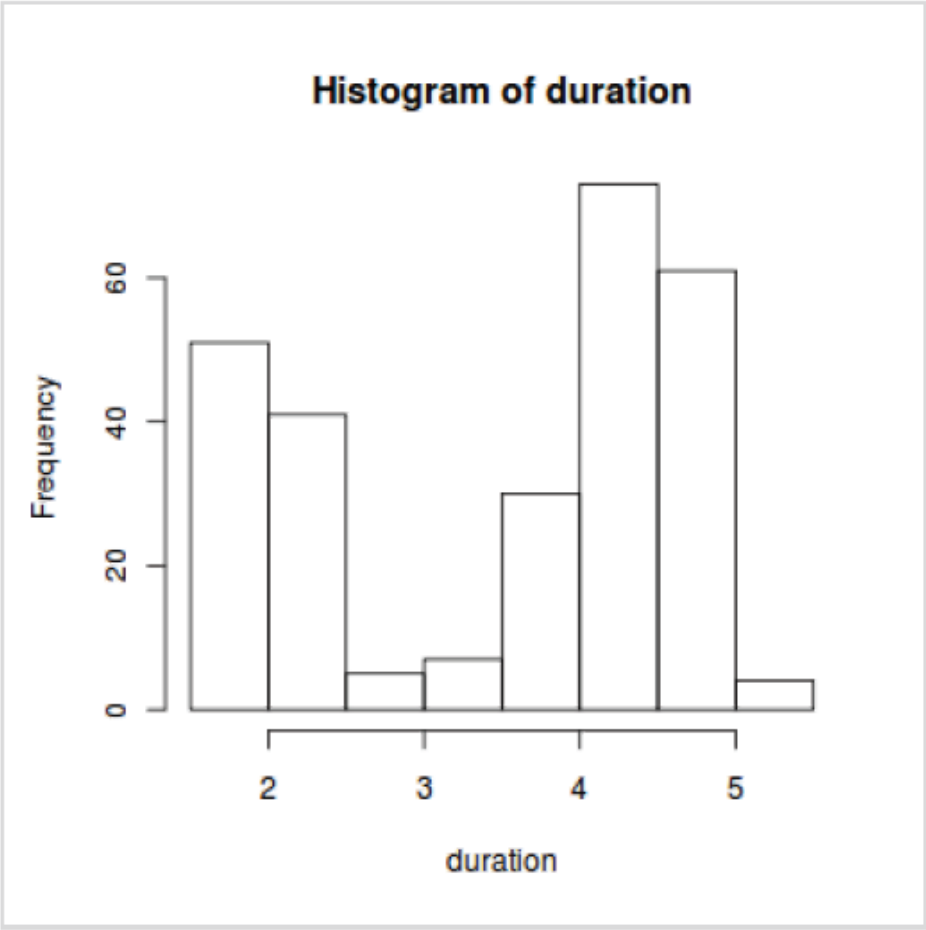
Rešenje. Za određivanje histograma, primenjuje se **hist** funkcija:

```
> duration = faithful$eruptions
> hist(duration,    # apply the hist function
+   right=FALSE)    # intervals closed on the left
```

Promenljiva duration sadrži trajanja erupcija, a intervali koji se koriste za formirnje histograma su poluotvoreni (sadrže levu, ali ne sadrže desnu krajnju tačku), što je postignuto podešavanjem vrednosti za imenovani argument right.

Histogram (2)

Rešenje (nastavak). Rezultat primene hist funkcije je sledeći dijagram:



Histogram (3)

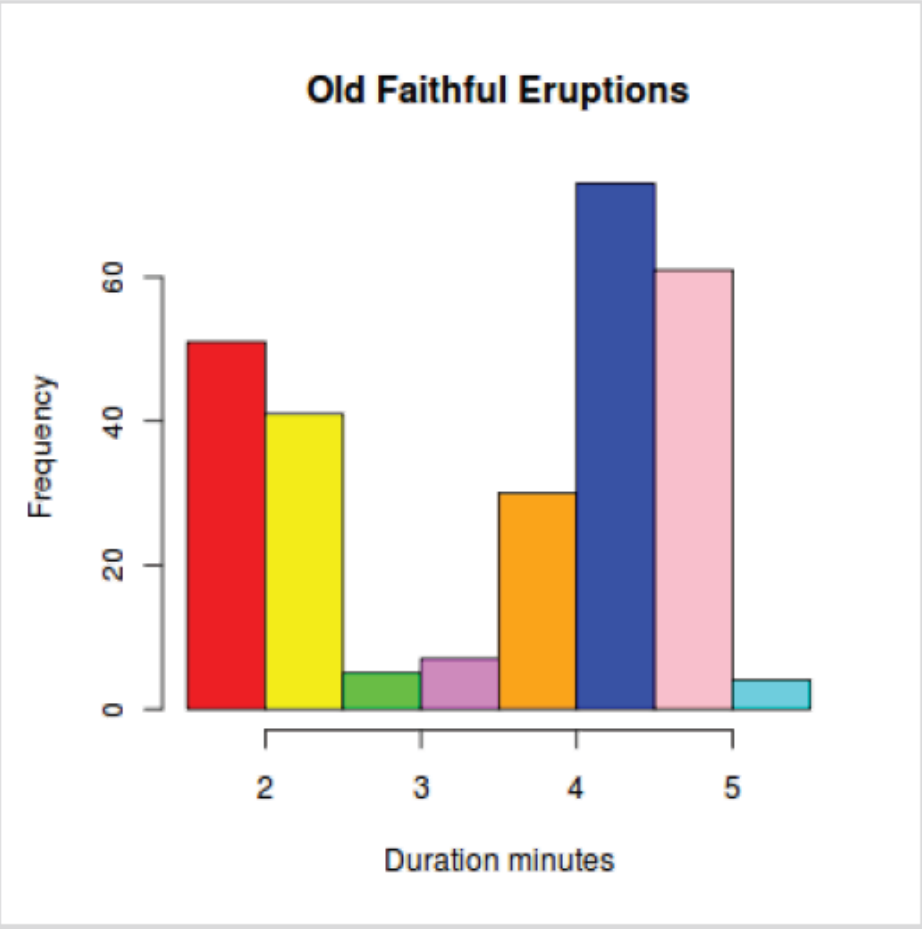
Problem. Odrediti histogram za trajanje erupcija (tj. promenljivu eruptions) u okviru sa podacima faithful. Histogram treba da bude obojen, treba da sadrži naslov i legendu uz apscisu (Ox osu).

Rešenje. Za određivanje histograma, primenjuje se hist funkcija.
Podešavanjem imenovanog argumenta col se definiše koja paleta boja se koristi za bojenje stubova histograma.
Podešavanjem vrednosti za imenovani argument main postiže se postavljanje naslova histograma, a legenda uz Ox osu se postavlja korišćenjem imenovanog argumenta xlab.

```
> colors = c("red", "yellow", "green", "violet", "orange",
+   "blue", "pink", "cyan")
> hist(duration,    # apply the hist function
+   right=FALSE,    # intervals closed on the left
+   col=colors,     # set the color palette
+   main="Old Faithful Eruptions", # the main title
+   xlab="Duration minutes")       # x-axis label
```

Histogram (4)

Rešenje (nastavak). Rezultat primene hist funkcije sa ovim argumentima je sledeći dijagram:



Histogram (5)

Zadatak 1. Odrediti histogram za vreme čakanja u okviru sa podacima faithful.

Relativna raspodela frekfencija kod kvantitativnih podataka

Relativna raspodela frekfencija za promenljivu koja sadrži podatke je sumarni pregled ***učešća*** podataka u kolekciji kategorija koje se međusobno ne preklapaju.

Veza između frekfencije i relativne frekfencije je data sledećom formulom:

$$Relativna\_frekfencija = \frac{Frekfencija}{Veličina\_uzorka}$$

Primer. U okviru sa podacima faithful, relativna raspodela frekefencija za promenljivu duration je sumarni pregled frekfencija trajanja prema u skladu sa prihvaćenom klasifikacijom trajanja erupcija.
Problem. Odrediti relativnu raspodelu frekefencija za promenljivu duration u okviru sa podacima faithful.

Rešenje. Radi određivanja relativne raspodele frekefencija za promenljivu duration, biće prvo određena raspodela frekfencija za promenljivu duration:

```
> duration = faithful$eruptions
> breaks = seq(1.5, 5.5, by=0.5)
> duration.cut = cut(duration, breaks, right=FALSE)
> duration.freq = table(duration.cut)
```

Relativna raspodela frekfencija kod kvantitativnih podataka (2)

Rešenje(nastavak). da bi se odredila relativna frekfencija trajanja erupcije, broj erupcija čije je trajanje u datom intervalu treba podeliti sa veličinom uzorka. Veličina uzorka se može odrediti primenom funkcije nrow na okvir sa podacima.

```
> duration.relfreq = duration.freq / nrow(faithful)
```

Na taj način, promenljiva duration.relfreq sadrži relativnu frekfenciju ocena, tj. sledeće vrednosti:

```
> duration.relfreq
duration.cut
 [1.5,2)  [2,2.5)  [2.5,3)  [3,3.5)  [3.5,4)  [4,4.5)
0.187500 0.150735 0.018382 0.025735 0.110294 0.268382
 [4.5,5)  [5,5.5)
0.224265 0.014706
```

Relativna raspodela frekfencija kod kvantitativnih podataka (3)

Problem. Odrediti relativnu raspodelu frekefencija za promenljivu duration u okviru sa podacima faithful, tako da se relativne frekfence zaokruže na dve decimale.

Dobijene realtivne frekfencije prikazati u formi vrste i u formi kolone.

Rešenje. Relativne frekfencije se računaju na isti način kao u prethodnom primeru:

```
> duration = faithful$eruptions
> breaks = seq(1.5, 5.5, by=0.5)
> duration.cut = cut(duration, breaks, right=FALSE)
> duration.freq = table(duration.cut)
> duration.relfreq = duration.freq / nrow(faithful)
```

Korišćenjem funkcije options sa imenovnim argumentom digits se podešava broj cifara za prikaz:

```
> old = options(digits=1)
> duration.relfreq
duration.cut
[1.5,2) [2,2.5) [2.5,3) [3,3.5) [3.5,4) [4,4.5) [4.5,5)
   0.19    0.15    0.02    0.03    0.11    0.27    0.22
[5,5.5)
   0.01
> options(old)    # restore the old option
```

Relativna raspodela frekfencija kod kvantitativnih podataka (4)

Rešenje (nastavak). Prikaz dobijenih rezultata u obliku kolone se postiže korišćenjem funkcije cbind:

```
> old = options(digits=1)
> cbind(duration.freq, duration.relfreq)
        duration.freq duration.relfreq
[1.5,2)            51             0.19
[2,2.5)            41             0.15
[2.5,3)             5             0.02
[3,3.5)             7             0.03
[3.5,4)            30             0.11
[4,4.5)            73             0.27
[4.5,5)            61             0.22
[5,5.5)             4             0.01
> options(old)    # restore the old option
```

Zadatak 1. Odrediti relativnu raspodelu frekfencija za promenljivu koja opisuje čekanje između dve erupcije u okviru sa podacima faithfull.

Raspodela kumulativnih frekfencija

Raspodela kumulativnih frekfencija se određuje na osnovu broja (tj. frekfencije) observiranih kvantitativnih podataka čije je obeležje manje ili jednako od datog nivoa.

Primer. U okviru sa podacima faithful, raspodela kumulativnih frekfencija za promenljivu eruptions predstavlja ukupan broj erupcija čije je trajanje kraće ili jednako elementima skupa izabranih nivoa.
Problem. Odrediti raspodelu kumulativnih frekfencija za trajanje erupcija (tj. promenljivu eruptions) u okviru sa podacima faithful.

Rešenje. Prvo se odredi raspodela frekfencija za trajanje erupcija, na isti način kao u prethodnim primerima:

```
> duration = faithful$eruptions
> breaks = seq(1.5, 5.5, by=0.5)
> duration.cut = cut(duration, breaks, right=FALSE)
> duration.freq = table(duration.cut)
```

Potom se primeni **cumsum** funkcija nad raspodelom frekfencija i tako se dobiju kumulativne frekfencije:

```
> duration.cumfreq = cumsum(duration.freq)
```

Raspodela kumulativnih frekfencija(2)

Rešenje (nastavak). Na dobijene kumulativne frekfencije za trajanje erupcije referiše promenljiva duration.cumferq. Dakle, dobijena je raspodela:

```
> duration.cumfreq
[1.5,2) [2,2.5) [2.5,3) [3,3.5) [3.5,4) [4,4.5) [4.5,5)
     51      92      97     104     134     207     268
[5,5.5)
    272
```

Problem. Odrediti raspodelu kumulativnih frekfencija za trajanja (tj. promenljivu eruptions) u okviru sa podacima faithful i prikazati je kao kolonu.

Rešenje. Isto kao u prethodnom primeru, računa se raspodela kumulativnih frekfencija za trajanje erupcije. Kolonski prikaz se dobija pomoću cbind:

```
> cbind(duration.cumfreq)
        duration.cumfreq
[1.5,2)               51
[2,2.5)               92
[2.5,3)               97
[3,3.5)              104
[3.5,4)              134
[4,4.5)              207
[4.5,5)              268
[5,5.5)              272
```

Raspodela kumulativnih frekfencija(3)

Zadatak 1. Odrediti raspodelu kumulativnih frekfencija za period čekanja izmedju erupcija u okviru sa podacima faithful.

Dijagram kumulativnih frekfencija

Dijagram kumulativnih frekfencija za kvantitativnu promenljivu je kriva koja prikazuje raspodelu kumulativnih frekfencija te kvantitativne promenljive.

Primer. Za sve tačke dijagrama kumulativnih frekfencija za promenljivu eruptions okviru sa podacima faithful važi: y koordinata je jednaka ***ukupnom*** broju erupcija čije je trajanje ***kraće ili jednako*** od vrednosti x koordinate.

Problem. Odrediti dijagram kumulativnih frekfencija za trajanje erupcije u okviru sa podacima faithful.

Rešenje. Prvo se odredi raspodela frekfencija za trajanje erupcija, na isti način kao u prethodnim primerima:

```
> duration = faithful$eruptions
> breaks = seq(1.5, 5.5, by=0.5)
> duration.cut = cut(duration, breaks, right=FALSE)
> duration.freq = table(duration.cut)
```

Potom se primeni cumsum funkcija nad raspodelom frekfencija i tako se dobije vektor sa kumulativnom raspodelom frekfencija, a potom se tom vektoru doda nula kao polazni element i rezultujući vektor dodeli promenljivoj cumfreq0:

```
> cumfreq0 = c(0, cumsum(duration.freq))
```

Dijagram kumulativnih frekfencija (2)

Rešenje (nastavak). Potom se pozivom funkcije **plot** izvrši iscrtavanje tačaka dijagrama:

```
> plot(breaks, cumfreq0,             # plot the data
+   main="Old Faithful Eruptions",  # main title
+   xlab="Duration minutes",        # x-axis label
+   ylab="Cumulative eruptions")   # y-axis label
```

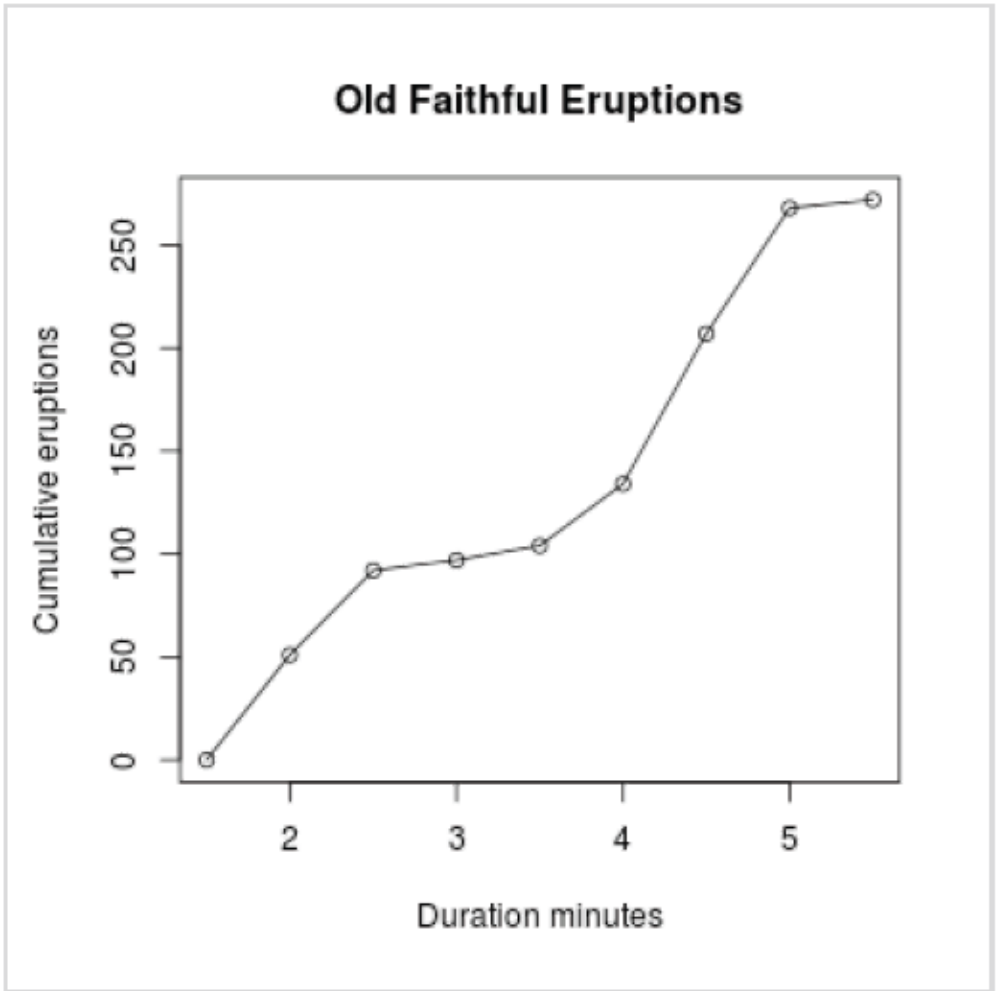
Uočava se da je prvi argument funkcije plot sekvenca sa podeonim tačkama na Ox osi, a da je drugi argument kumulativna raspodela proširena nulom na početku. Imenovani argument main referiše na naslov dijagrama, xlab na legendu uz Ox osu, a ylab na legendu uz Oy osu.
Na kraju, potrebno je još povući duži između istaknutih tačaka dijagrama, što se postiže funkcijom **lines**:

```
> lines(breaks, cumfreq0)           # join the points
```

Argumenti funkcije lines su vektor x koordinata i vektor y koordinata tačaka koje se povezuju.

Dijagram kumulativnih frekfencija (3)

Rešenje (nastavak). Dobijeni dijagram kumulativnih frekfencija ima sledeći oblik:



Dijagram kumulativnih frekfencija (4)

Zadatak 1. Odrediti dijagram kumulativnih frekfencija za period čekanja izmedju erupcija u okviru sa podacima faithful.

Raspodela kumulativnih relativnih frekfencija

Raspodela kumulativnih relativnih frekfencija za kvantitativnu promenljivu je odnos suma frekfencija ispod datog nivoa.

Odnos između kumulativne frekfencije i kumulativne relativne frekfencije dat je sledećom formulom:

$$Kumulativna\ relativna\ frekfencija = \frac{Kumulativna\ frekfencija}{Veličina\ uzorka}$$

Primer. U okviru sa podacima faithful, raspodela kumulativnih relativnih frekfencija za trajanja erupcija je odnos broja erupcija čije je trajanje kraće ili jednako datom nivou i ukupnog broja erupcija.

Problem. Odrediti raspodelu kumulativnih relativnih frekfencija za trajanja erupcije (kolona eruptions) u okviru sa podacima faithful.

Rešenje. Prvo se odredi raspodela frekfencija za trajanja erupcija, na isti način kao u prethodnim primerima:

```
> duration = faithful$eruptions
> breaks = seq(1.5, 5.5, by=0.5)
> duration.cut = cut(duration, breaks, right=FALSE)
> duration.freq = table(duration.cut)
```

Raspodela kumulativnih relativnih frekfencija (2)

Rešenje (nastavak). Onda se korišćenjem funkcije cumsum odredi raspodela kumulativnih frekfencija za trajanja erupcija:

```
> duration.cumfreq = cumsum(duration.freq)
```

Potom se, korišćenjem funkcije nrow odredi veličina uzorka, pa se sa veličinom uzorka podele sve kumulativne frekfencije i tako se dobiju kumulativne relativne frekfencije:

```
> duration.cumrelfreq = duration.cumfreq / nrow(faithful)
```

Raspodela dobijenih kumulativnih relativnih frekfencija u ovom slučaju je:

```
> duration.cumrelfreq
[1.5,2) [2,2.5) [2.5,3) [3,3.5) [3.5,4) [4,4.5) [4.5,5)
0.18750 0.33824 0.35662 0.38235 0.49265 0.76103 0.98529
[5,5.5)
1.00000
```

Raspodela kumulativnih relativnih frekfencija (3)

Problem. Odrediti raspodelu kumulativnih relativnih frekfencija za trajanja erupcija u okviru sa podacima faithful i prikazati je u dve decimale.

Rešenje. Isto kao u prethodnom primeru, računa raspodelu kumulativnih relativnih frekfencija za trajanja erupcije – promenljiva duration.cumrelfreq:

```
> duration = faithful$eruptions
> breaks = seq(1.5, 5.5, by=0.5)
> duration.cut = cut(duration, breaks, right=FALSE)
> duration.freq = table(duration.cut)
> duration.cumfreq = cumsum(duration.freq)
> duration.cumrelfreq = duration.cumfreq / nrow(faithful)
```

Prikaz u manje decimala se postiže pozivom funkcije options gde je podešen imenovani argument digits:

```
> old = options(digits=2)
> duration.cumrelfreq
[1.5,2) [2,2.5) [2.5,3) [3,3.5) [3.5,4) [4,4.5) [4.5,5)
   0.19    0.34    0.36    0.38    0.49    0.76    0.99
[5,5.5)
   1.00
> options(old)    # restore the old option
```

Raspodela kumulativnih relativnih frekfencija (4)

Problem. Odrediti raspodelu kumulativnih frekfencija i raspodelu kumulativnih relativnih frekfencija za trajanja (tj. promenljivu eruptions) u okviru sa podacima faithful i prikazati ih u dve decimale i po kolonama.

Rešenje. Na isti načina kao u prethodnom primeru, računaju se raspodela kumulativnih frekfencija i raspodela kumulativnih relativnih frekfencija za trajanja erupcije – to su promenljive duration.cumrelfreq i duration.cumrelfreq.
Broj decimala se podešava pomoću options, a prikaz po kolonama se postiže pozivom funkcije cbind:

```
> old = options(digits=2)
> cbind(duration.cumfreq, duration.cumrelfreq)
        duration.cumfreq duration.cumrelfreq
[1.5,2)               51                0.19
[2,2.5)               92                0.34
[2.5,3)               97                0.36
[3,3.5)              104                0.38
[3.5,4)              134                0.49
[4,4.5)              207                0.76
[4.5,5)              268                0.99
[5,5.5)              272                1.00
> options(old)
```

Raspodela kumulativnih relativnih frekfencija (5)

Zadatak 1. Odrediti raspodelu kumulativnih relativnih frekfencija za period čekanja izmedju erupcija u okviru sa podacima faithful.

Dijagram kumulativnih relativnih frekfencija

Dijagram kumulativnih relativnih frekfencija za kvantitativnu promenljivu je kriva koja prikazuje kumulativnu raspodelu relativnih frekfencija promenljive.

Primer. Za sve tačke dijagramu kumulativnih relativnih frekfencija za promenljivu eruptions okviru sa podacima faithful važi: y koordinata je jednaka ***udelu*** broja erupcija čije je trajanje ***kraće ili jednako*** od vrednosti x koordinate u ukupnom broju posmatranja.

Problem. Odrediti dijagram kumulativnih relativnih frekfencija za trajanja erupcije u okviru sa podacima faithful.

```
> duration = faithful$eruptions
> breaks = seq(1.5, 5.5, by=0.5)
> duration.cut = cut(duration, breaks, right=FALSE)
> duration.freq = table(duration.cut)
> duration.cumfreq = cumsum(duration.freq)
> duration.cumrelfreq = duration.cumfreq / nrow(faithful)
```

Zatim se kumulativne relativne frekfencije prošire nulom na početku:

```
> cumrelfreq0 = c(0, duration.cumrelfreq)
```

Dijagram kumulativnih relativnih frekfencija (2)

Rešenje (nastavak). Potom se pozivom funkcije **plot** izvrši iscrtavanje tačaka:

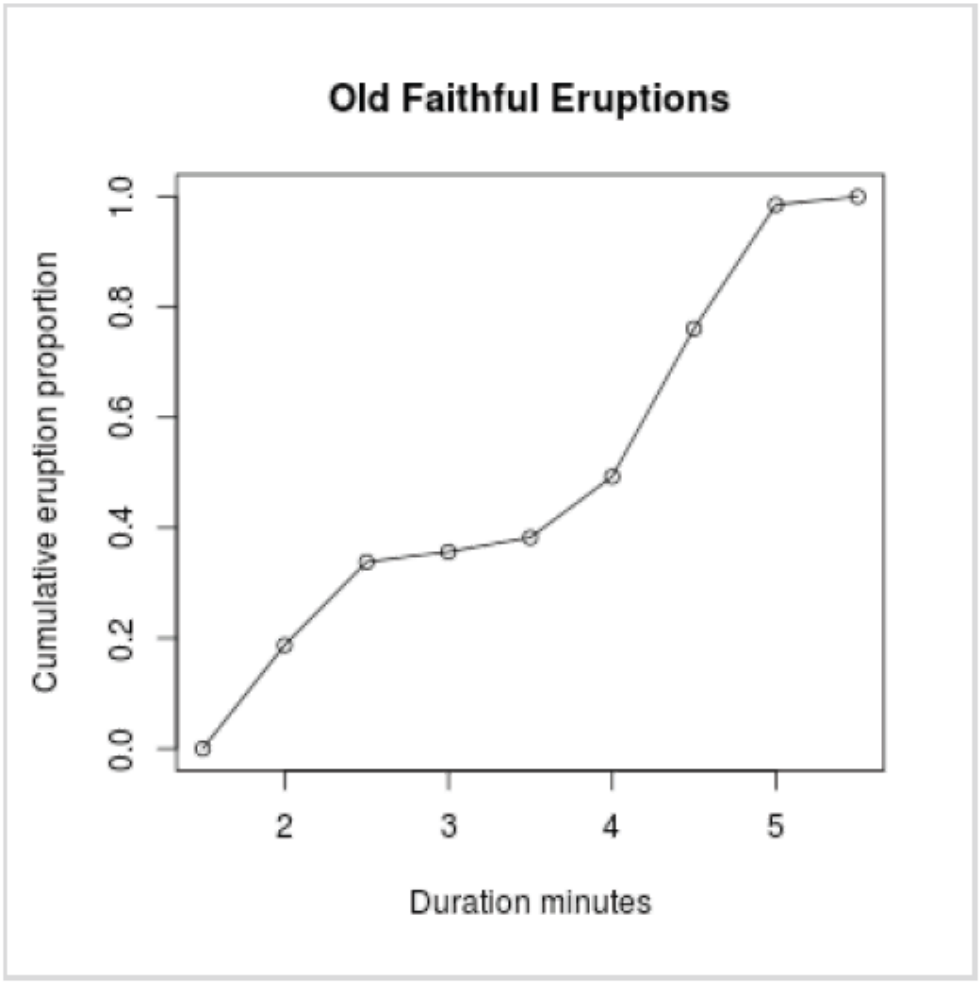
```
> plot(breaks, cumrelfreq0,
+   main="Old Faithful Eruptions",  # main title
+   xlab="Duration minutes",
+   ylab="Cumulative eruption proportion")
```

Prvi argument funkcije plot je sekvenca sa podeonim tačkama na Ox osi, a drugi argument su kumulativne relativne frekfencije proširene nulom. Imenovani argument main referiše na naslov dijagrama, xlab na legendu uz Ox osu, a ylab na legendu uz Oy osu.
Na kraju se povlače duži između tačaka dijagrama prikazanih pomoću plot, što se postiže funkcijom **lines**:

```
> lines(breaks, cumrelfreq0)        # join the points
```

Dijagram kumulativnih relativnih frekfencija (3)

Rešenje (nastavak). Dijagram kumulativnih relativnih frekfencija je:



Dijagram kumulativnih relativnih frekfencija (4)

Zadatak 1. Odrediti dijagram kumulativnih relativnih frekfencija za period čekanja izmedju erupcija u okviru sa podacima faithful.

XY dijagram

XY dijagram, ili ***dijagram disperzije (scatter plot)*** uparuje vrednosti dve kvantitativne promenljive u okviru sa podacima i prikazuje ih kao tačke u dvodimenzionalnoj ravni.

Primer. U okviru sa podacima faithful se uparivanjem vrednosti promenljivih eruptions i waiting za sve observacije dobijaju uređeni parovi (x,y). Ti parovi predstavljaju koordinate tačkaka na XY dijagramu.

Problem. Odrediti i kao kolone prikazati uparene vrednosti trajanja erupcije (promenljiva eruptions) i čekanja između erupcija (promenljiva waiting) u okviru sa podacima faithful.

Rešenje. Vrednosti za trajanje erupcije i za čekanje se dobijaju isecanjem okvira za podatke po odgovarajućim kolonama. Funkcija cbind povezuje ove dve grupe podataka, a funkcija head omogućuje prikaz jednog dela podataka:

```
> duration = faithful$eruptions      # the eruption durations
> waiting = faithful$waiting         # the waiting interval
> head(cbind(duration, waiting))
     duration waiting
[1,]    3.600      79
[2,]    1.800      54
[3,]    3.333      74
```

XY dijagram (2)

Problem. Odrediti XY dijagram za trajanja erupcije i čekanja između erupcija u okviru sa podacima faithful. Da li dijagram ukazuje na neku vezu među promenljivima?

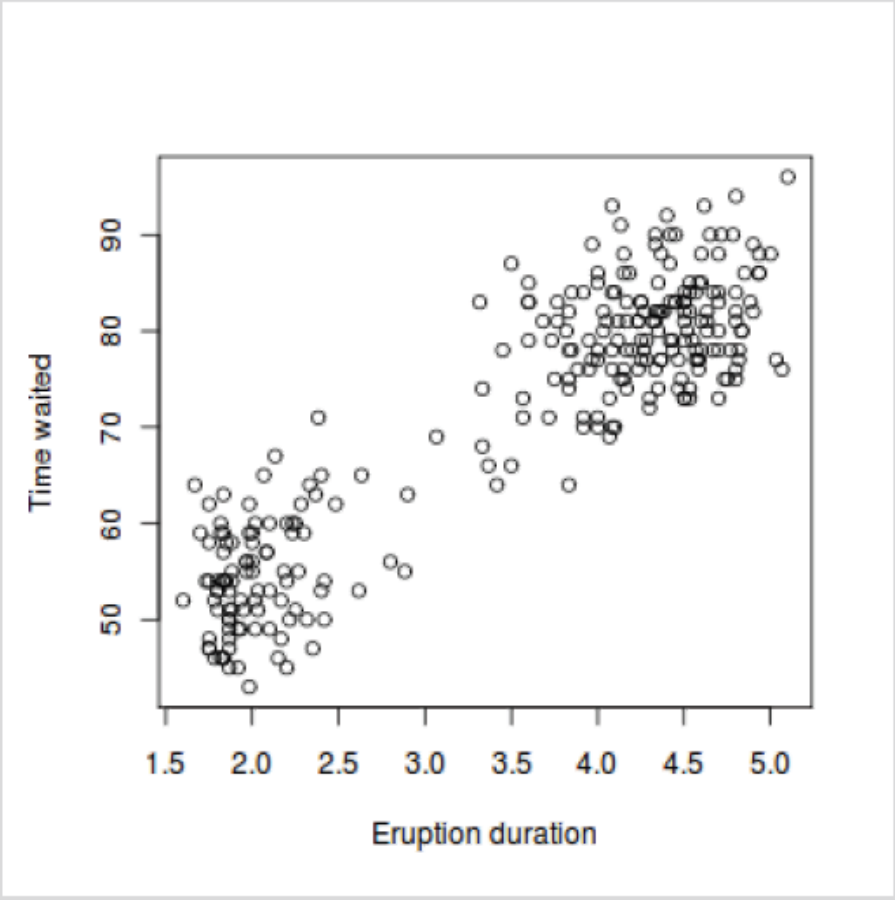
Rešenje. Vrednosti za trajanje erupcije i za čekanje se dobijaju isecanjem okvira za podatke po odgovarajućim kolonama.
Funkcija plot iscrtava tačke koje predstavlaju uparene vrednosti trajanja i čekanja (trajanje se prikazuje na Ox osi, a čekanje na Oy osi):

```
> duration = faithful$eruptions      # the eruption durations
> waiting = faithful$waiting         # the waiting interval
> plot(duration, waiting,            # plot the variables
+   xlab="Eruption duration",        # x-axis label
+   ylab="Time waited")              # y-axis label
```

XY dijagram (3)

Rešenje (nastavak). XY dijagram za trajanje i čekanje u okviru sa podacima faithful ima sledeći oblik:

Dijagram ukazuje na pozitivnu linearnu vezu među ovim promenljivima.



XY dijagram (4)

Problem. Odrediti XY dijagram za trajanja erupcije i čekanja između erupcija u okviru sa podacima faithful, odrediti linearni regresioni model za ove dve promenljive i na dijagramu prikazati liniju trenda.

Rešenje. Vrednosti za trajanje erupcije i za čekanje se dobijaju isecanjem okvira za podatke po odgovarajućim kolonama.
Funkcija plot iscrtava XY dijagram:

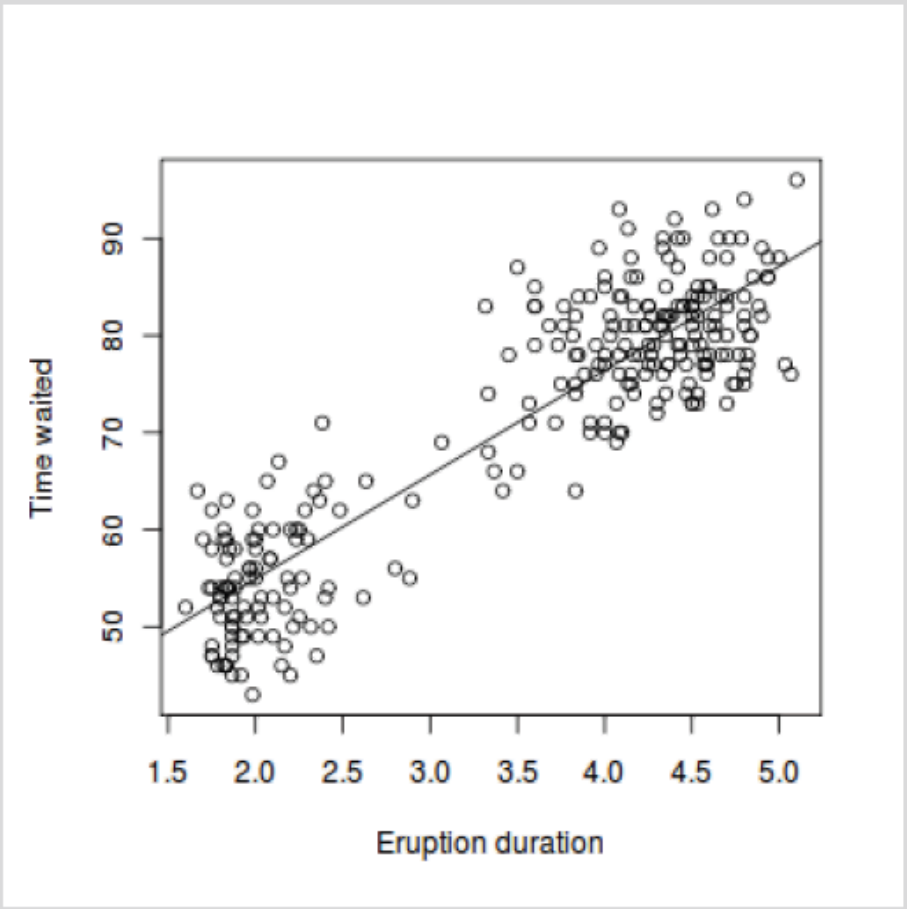
```
> duration = faithful$eruptions      # the eruption durations
> waiting = faithful$waiting         # the waiting interval
> plot(duration, waiting,            # plot the variables
+   xlab="Eruption duration",        # x-axis label
+   ylab="Time waited")              # y-axis label
```

Linearni regresioni model između dve kvantitativne promenjive se kreira pomoću funkcije **lm**, pri čemu se za označavanje relacije koristi simbol **~**.
Iscrtavanje linije na osnovu linearnog modela se postiže **abline** funkcijom. Dakle:

```
> abline(lm(waiting ~ duration))
```

XY dijagram (5)

Rešenje (nastavak). XY dijagram za trajanje i čekanje i linija trenda dobijena na osnovu linearnog modela su:



Korišćeni izvori

Deo materijala ove prezentacije je preuzet sa sajta

http://www.r-tutor.com/

Deo materijala je preuzet sa sajta

http://www.e-statistika.rs