Co-funded by the
Erasmus+ Programme
of the European Union

Dr Đurađ Milošević, Dr Milica Stojković Piperac, Dr Dušanka Cvijanović

NUMERIČKA EKOLOGIJA

Univerzitet u Nišu, Prirodno-matematički fakultet

Univerzitet u Novom Sadu, Prirodno matematički fakultet

Erasmus + Project No ECOBIAS_609967-EPP-1-2019-1-RS-EPPKA2-CBHE-JP
Development of master curricula in ecological monitoring and aquatic bioassessment for Western Balkans HEIs

Sadržaj

Predgovor

Skripta pod nazivom "Numerička ekolgoija" je namenjena studentima master akademskih studija i polaznicima kurseva celoživotnog učenja iz oblasti ekološkog monitoringa voda na univerzitetima u Bosni i Hercegovini (Univerzitet u Sarajevu i Univerzitet u Tuzli) i Crnoj Gori (Univerzitet Donja Gorica), kao i svima onima koji žele da unaprede i prošire svoja znanja iz oblasti analize podataka u ekologiji. Skripta pod nazivom "Numerička ekolgoija'' je nastala kao rezultat ERAZMUS+ projekta „Razvoj master kurikuluma iz ekološkog monitoringa i bioindikacija kopnenih voda na visokoškolskim ustanovama u regionu Zapadnog Balkana'' (Development of master curricula in ecological monitoring and aquatic bioassessment for Western Balkans HEIs -ECOBIAS).

Skripta za Numeričku ekologiju sadrži 8 poglavlja u okviru kojih su prikazane statističke i numeričke tehnike koje se koriste u ekološkim studijama.. Informacije koje pruža skripta, zajedno sa predloženim konceptom praktičnog rada u R programskom jeziku, olakšava razumevanje osnovnih principa u analizi podataka u ekologiji, objašnjavajući kako izabrati i primeniti odgovarajuće metode pri realizaciji postavljenih istraživačkih ciljeva.

Odabir literature kao i njen obim su adekvatno prilagođeni programu visokoškolskih ustanova.

Želimo da zahvalimo uvaženom recenzentu, Prof dr Đorđu Obradoviću, sa Univerziteta SIngidunum, je svojim korisnim savetima doprineo da finalna prezentacija ove skripte bude sistematična i egzaktna, a da istovremeno bude razumljiva za čitaoce.

Autori

1.R programski jezik.

R je programski jezik za statističku analizu i vizuelizaciju podataka, koji zbog svoje pristupačnosti i posebnog fokusa na analizu podataka je jako popularan u ekološkim istaživanjima i praksi. Izuzetna popularnost ovog besplatnog programskog jezika je dovela i do produkcije velike količine onlajn dostupne literature i biblioteka (paketa) koje su pogodne za obradu podataka i statistčku analizu u okviru ekoloških studija.

R programski jezik se može koristiti kroz R konzolu ali se zbog velikog broja funkcija koje nudi, koristi preko integrisanog razvojnog abmijneta (eng. *Integrated Development Environment, EDI*) *-R Studio*. R i R studio se mogu instlairati na operativne sisteme *Windows MAC OSX* i *Linux platforme*.

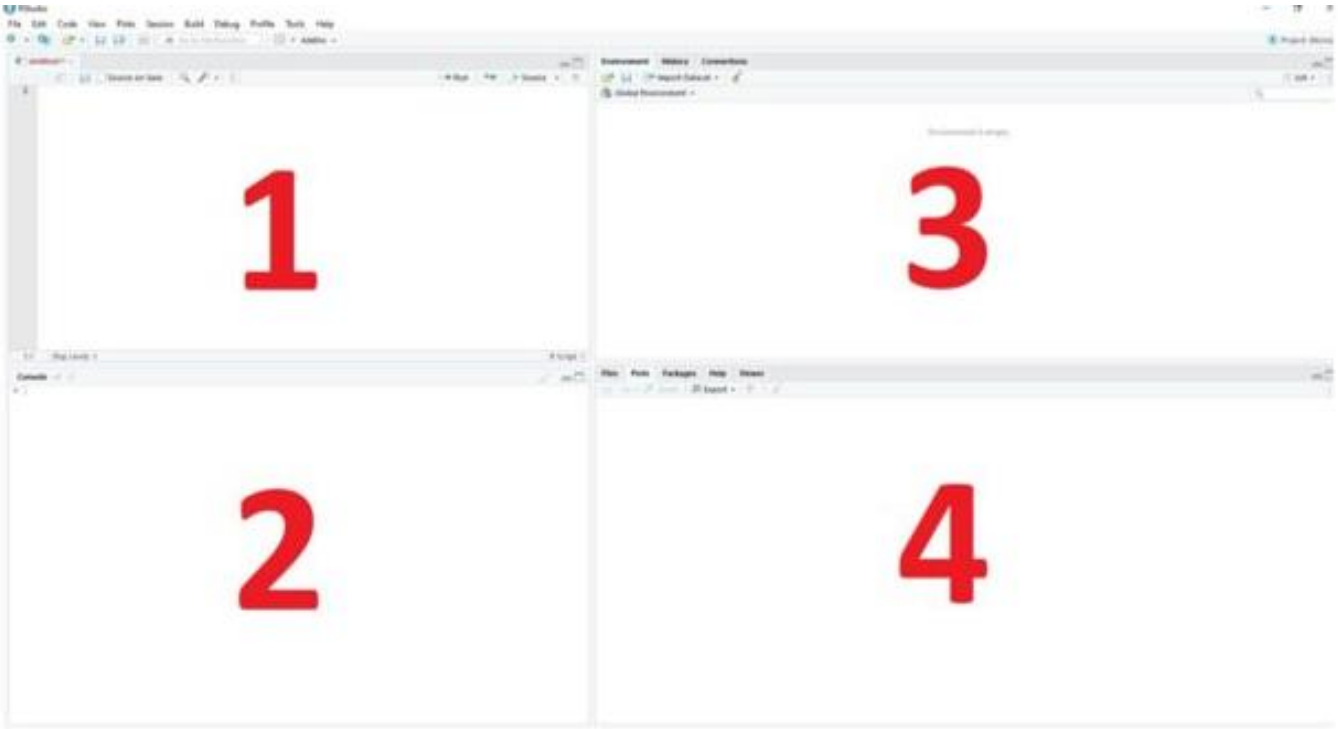
R konzola je besplatna i može se preuzeti i instalirati na sledećoj web stranici:

https://cran.r-project.org/

Ukoliko se R konzola instalira na računar koji koristi *Windows* kao operativni sistem, kliknuti na *Windows* opciju i izabrati 32-bit ili 64-bit verziju R-a. R studio je takođe besplatan i može se preuzeti sa sledeće web strane:

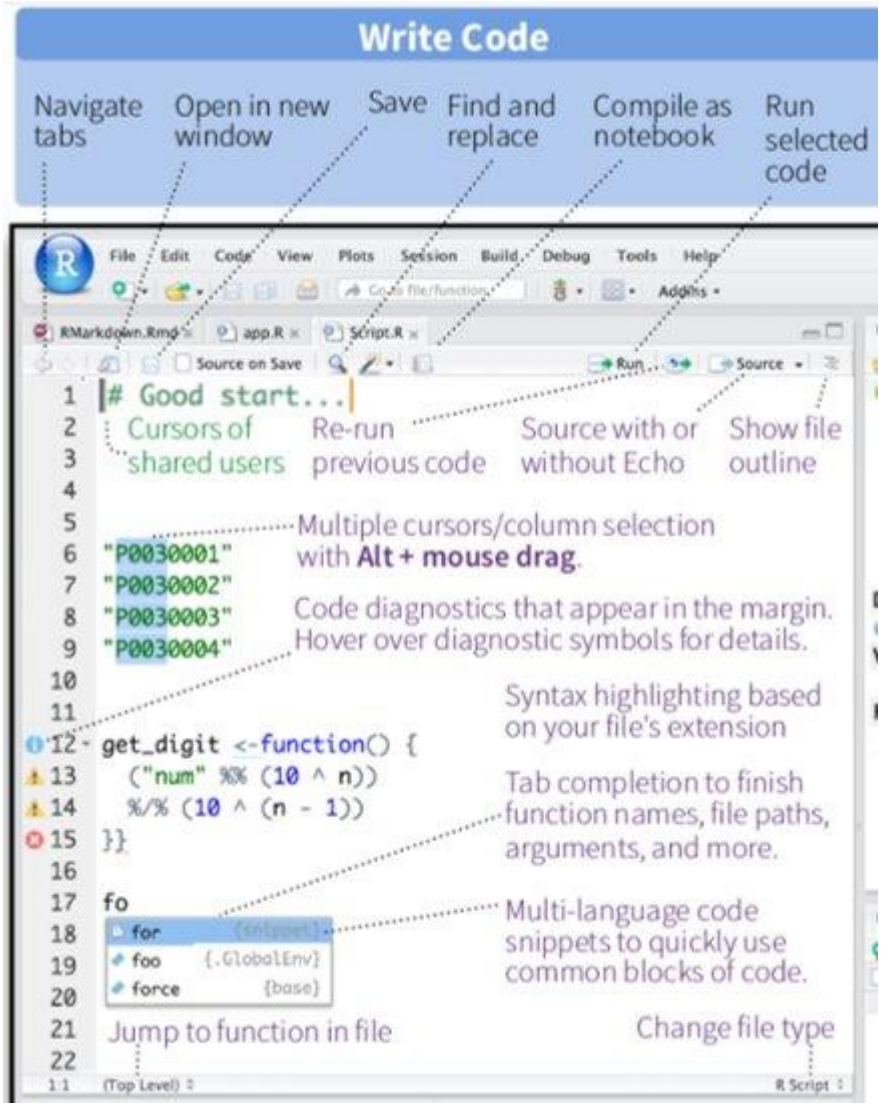
https://rstudio.com/products/rstudio/#Desktop

R studio interfejs je sačinjen od četiri prozora: 1) editor R skripte/koda, 2) R konzola, 3) *Environment/History/Connection/Git* i 4) *Files/Plots/Packages/Help* (Slika 1.1).



Slika 1.1 Interfejs R studija: 1) editor R skripte/koda, 2) R konzola, 3) *Environment/History/Connection/Git* i 4) *Files/Plots/Packages/Help*

U editoru R studija (1) je moguće pisati kod ali i otvarati i menjati postojeće kodove koji su snimljeni u obliku R skripte (eng. *R script*). U skladu sa funkcijom prozora, dostupni su razni alati koji mogućavaju jednostavnije pisanje i korigovanje kodova (slika 1.2)

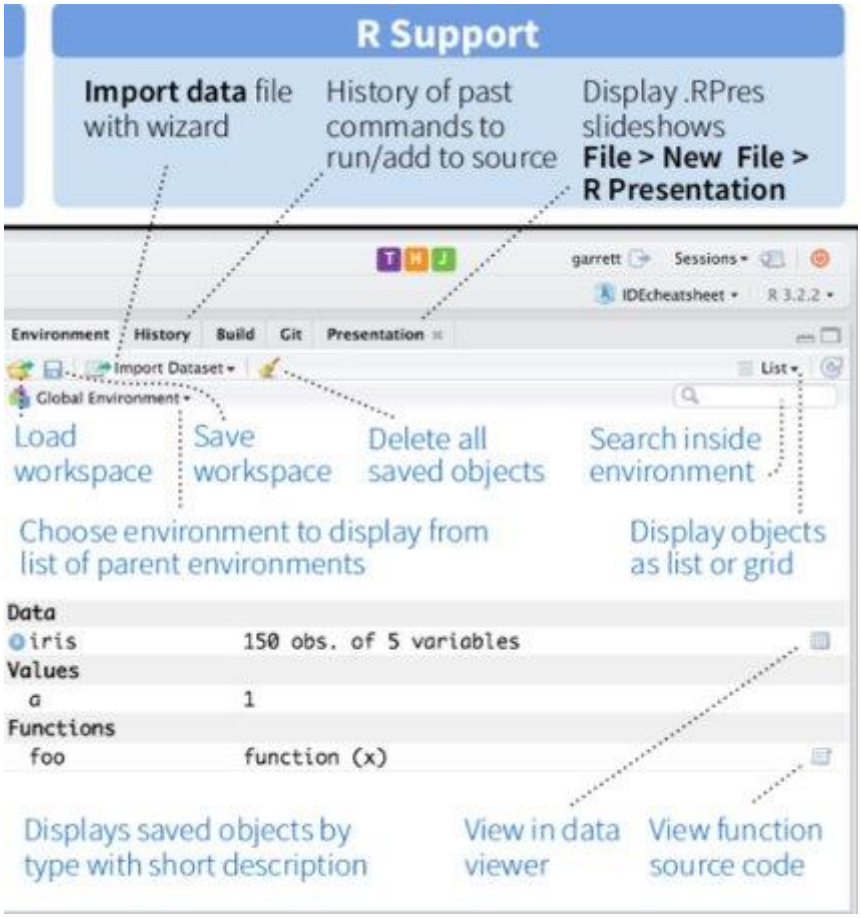


Slike 1.2- Editor R skripte/koda u R studiju sa sivm dostupnim alatima

R konzola (2) je mesto gde se izvršavaju komande i prikazuju rezultati. (slika 1.3). U ovom prozoru su prikazani svi rezultati nakon izvršenog koda. U R konzoli je moguće direktno napisati komandu ali je ne i sačuvati s obzirom da se nakon zatvaranja sesije svi podaci iz R konzole gube.

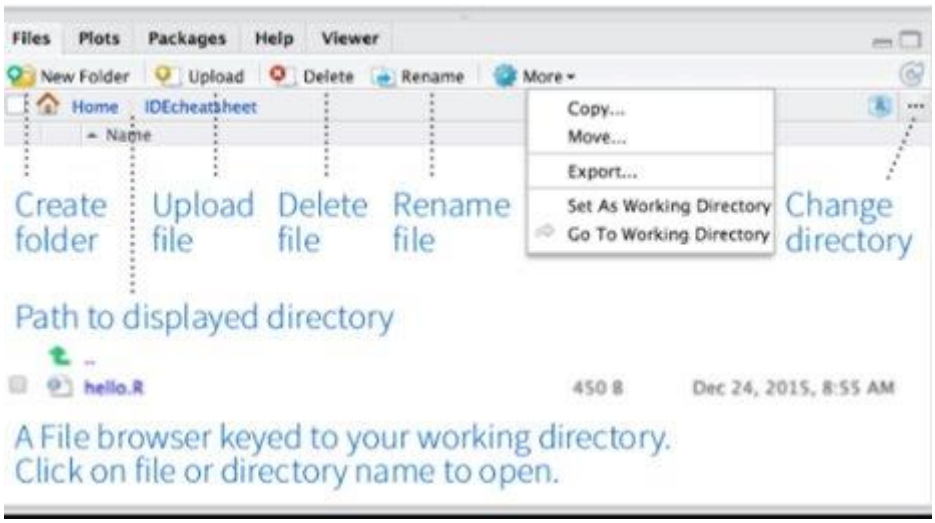
Slika 1.3 R konzola u R studiju sa sivm dostupnim alatima

U prozoru *Environment* je prikazan spisak svih aktivnih R objekata koji su formirani tokom R sesije, ukljucujuci podatke, pakete i funkcije (Slika 1.4). U prozoru *History* se nalaze sačuvane sve prethodne komande dok prozor *Connections* predstavlja konektor za potencijalne konekcije (ODBC I Spark).



Slika 1.4 *Environment/History/Connection/Git* u R stdiju sa svim dostupnim alatima

U četvrtom prozoru je spisak svih fajlova (*Files*) koji se nalaze u okvir radnog direktorijuma, kao I sve vizuelizacije (*Plots*) koje su kreirane tokom R sesije I gde je moguće sačuvati iste u određenoj rezoluciji, dimenzijama I formatu (JPEG, TIF, PDF; Slika 1.5) U prozoru *Packages* su prikazane sve eksterne R biblioteke (paketi) koje su instalirane u lokalnom sistemu. Pored svake biblioteke se nalazi kvadratić čijim čekiranjem je moguće učitati željenu biblioteku dok klikom na *Packages/Install packages* je moguće instalirati nove eksterne biblioteke.



Slika 1.5 *Files/Plots/Packages/Help* u R studiju sa svim dostupnim alatima

Dvostrukim klikom na ikonu R studija koja se nalazi na radnoj povrsini (eng. *Desktop*) aktivira se nova R sesije. U R konzoli je neophodno pre početka programiranja definisati radni direktorijum u kome se nalaze svi fajlovi koji se koriste u R sesiji. R radni direktorijum se određuje funkcijom *setwd()* dok se informacije o trenutnom random direktorijumu i njegovom sadržaju, ukoliko postoji, dobijaju funkcijom *getwd()* I *is()*:

```
setwd("c:/Numericka_ekologija")
> getwd()
[1] "c:/Numericka_ekologija"
> list.files()
[1]  "Baza_Moravica.csv"                    "Koviljski_Dunavac_ribe.csv"
"Tabela.csv"
```

1.1 Osnove pisanja koda u R-u.

Komunikacija sa R-om koja se realizuje u R konzoli ili editoru R skripte/koda, i sprovodi preko varijabli, operatera i funkcija. Komande koda se unose u R konzolu iza znaka > koji se zove odzivnik (eng. *Prompt*). Najednostavnija uporeba R-a je u vidu kalkulatora:

```
 > 2+200
[1] 202
> 2+300*4
[1] 1202
> (2+300)*4
[1] 1208
```

Druga linija prethodnog koda sadrži dve funkcije u kojima R prepoznaje hierarhiju (množenje je starija operacija od sabiranja). Kako bi se to promenilo, potrebno je koristiti zagrade (na primer u trećoj liniji koda). Izvršavanje svake linije koda u editoru R skripte/koda se može sprovesti prečicom *Ctrl+L*.

Kreiranje objekta u R-u je moguće znakom <- (ili =). Na taj način se određena količina podataka (informacija) čuva u radnoj memoriji R-a. Kako bi bio defnisan, objekat (varijabla) je potrebno da sadrži ime (identifikator) i vrednost:

```
> g = 2
> g
[1] 2
```

Ukoliko bi prethodna linija koda bila opisana rečima, kod bi mogao da se pročita kao: ime objekta "g" dobija vrednost 2. Sadržinu objekta se može pregledati jednostavnim kucanjem njegovog imena kao što je to prikazano u drugoj liniji prethodnog koda. Znak za kreiranje objekta se u R studiju može jenostavnije pisati prečicom *Alt*+– (zajedno se kuca *Alt* i znak minus). S obzirom da R ne prepoznaje razmake i nove redove u kodu, korišćenje razmaka je preporučljivo jer povećava preglednost napisanog koda.

Ime objekta mora početi slovom, a pored toga može sadržati i brojeve, donju crtu (_) i tačku (.). Međutim, naziv objekta ne sme sadržati zapetu jer R koristi taj interpunkcijski znak za razdvajanje argumenata u okviru funkcija. Preporučuje se da se prilikom pisanja imena objekta koristi notacija *snake_case*, koja podrazumeva upotrebu donje crte (_) umesto razmaka.

```
Broj_studenata_numericka_ekologija <- 6
```

```
> Broj_studenata_numericka_ekologija
[1] 6
```

S obizirom da R razlikuje mala slova od velikih, lako može doći do greške prilikom pozivanja objekta:

```
> broj_studenata_numericka_ekologija
Error: object 'broj_studenata_numericka_ekologija' not found
```

Ukoliko je naziv objekta predugačak, kako bi se izbegla greška prilikom kucanja, preporučuje se korišćenje alata *Tab*. R studio u tom slučaju nudi padajuću listu sa mogućim završecima svih objekata, aktivnih u R sesiji.

Prethodnim kodom objekat "Broj_studenata_numericka_ekologija" je sačuvan u radnoj memoriji sa određenom vrednošću koja mu je dodeljena. Ukoliko se istom objektu dodeli nova vrednost, R će je memorisati preko prethodne i zadržati u radnoj memoriji samo poslednju dodeljenu vrednost:

```
>Broj_studenata_numericka_ekologija                    =6;
Broj_studenata_numericka_ekologija=10
> Broj_studenata_numericka_ekologija
[1] 10
```

Interpunkcijski znak tačka i zarez (;) se koristi u kodu za razdvajanje komandi koje su napisane u istom redu.

Tokom pisanja koda, ako je izraz nepotpun (obično nedostaju upareni znakovi () ili "") R će prikazati znak plus (+), znak za nastavak koda. Ukoliko zbog nekog razloga je potrebno prekinuti izraz koji nije u potpunosti izvršen, to se može uraditi pritiskom na dugme *ESC*.

```
> x="Doobrodosli_na_UDG
+ "
> x
[1] "Doobrodosli_na_UDG"
```

1.2 R operatori

Operatori u R-u izvode različite matematičke i logičke operacije. Prema operacijama koje izvode, operatri se mogu klasifikovati u 4 kategorije: aritmetički, operatori poređenja, logički i operator dodele.

Aritmetički operatori izvode matematiče operacije (Tabela 1.1) gde je rezultat izvršavanja broj. Redosled izvršavanja operacija je isti kao u matematici (videti poglavlje 1.1).

Tabela 1.1 Aritmetički operatori u R-u

Operator	**Opis**	**Primer**
+	Plus	2+3=5
–	Minus	8-2=6
*	Puta	2*3=6
/	Podeljeno	10/5=2
^	Eksponent	2^5=32
%%	Ostatak od deljenja	7%%3=1
%/%	Rezultat deljenja bez decimalnog ostatka	7%/%3=2

Operatori poređenja se koriste za poređenje dve vrednosti (Tabela 1.2). Rezultat primene ove grupe operatora može biti „tačno“ (eng *TRUE*) ili „netačno“ (*FALSE*). Na primer, ukoliko se u R-u definišu dva objekta *x* i *y* sa konkretnim vrednostima, operatorima poređenja je moguće porediti zadate vrednosti:

```
> x <- 5
> y <- 16
> x<y [1] TRUE
> x>y [1] FALSE
> x<=5 [1] TRUE
> y>=20 [1] FALSE
> y == 16 [1] TRUE
> x != 5 [1] FALSE
```

Tabela 1.2 Relacioni operatori u R-u

Operator	**Opis**

<	**Manje od**
>	**Veće od**
<=	**Manje ili jednako**
>=	**Veće ili jednako**
==	**Jednako**
!=	**različito**

Često se prilikom programiranja u R-u, pogotovu kod početnika, pravi greška prilikom ispitivanja jednakosti, kada se za tu svrhu upotrebljava znak = umesto ==. Zbog toga se preporučuje da se umesto = prilikom definisanja objekata u R-u koristi znak <-.

Konačno, logički operatori se koriste za izvođenje složenijih izraza kada se proverava više uslova, odnosno istovremeno sprovodi više poređenja (Tabela 1.3).

```
n1 <- seq(-5, 5, by=2)
n1
[1] -5 -3 -1 1 3 5
(n1 > 0) & (n1 > 3)
[1] FALSE FALSE FALSE FALSE FALSE TRUE
```

* Objašnjenje funkcije *seq* se nalazi u odeljki 1.3. R funkcije.

Tabela 1.3. Logički operateri u R-u

Operator	**Opis**
!	ne
&	i
|	ili

1.3 R funkcije

Funkcija predstavlja deo koda koji izvršava određeni zadatak. U programiranju se često koriste funkije jer sadrže set instrukcija koje se ponavljaju u kodu ili obavljaju kompleksan zadatak, čineći potprogram koji se poziva po potrebi. Funkcija može da prihvata argumente i parametre i kao izlazni podatak vraća konkretne vrednosti. Argumenti funkcije se mogu podeliti u dva velika skupa: podaci na kojima se obavlja izračunavanje i argumenti koji upravljalju detaljima tog izračunavanja. U najvećem broju slučajeva funkcija ima svoje ime i argumente koje čine telo funkcije i koji su napisani u zagradi ():

```
Ime_funkcije (argument1=vrednost1, argument2=vrednost2, …)
```

Na primer, funkcija *t.test()* kojom se u R-u sprovodi statistčko testiranje hipoteze sadzi argumente *x* i *y* koji predstavljaju ulazne podatke za testiranje dok se detalji statistčkog testa definišu argumentima: *alternative*-defnisanje alternativne hipoteze (*HA*), *mu*-definisanje teorijske srednje vrednosti populacije (μ_0), *paired*-defnisanje tipa t-testa, *var.equal*-definisanje pretpostavke o homogenosti varijansi i *conf.level* -određivanje intervala poverenja. U telu funkcije se prvo definišu argumenti ulaznih podataka, a nakon toga argumenti koji definišu detalje izračunavanja. Ukoliko se neki argument ne napiše u kodu funkcija će primeniti podrazumevanu (*default*) vrednost datog argumenta. Na primer, u okviru *t.test* () funkcije, za argument *alternative* se podrazumeva dvostrana alternativna hipoteza ("*two sided*").

Prilikom pisanja koda koriste se uglaste zagrade {}, u okviru kojih se definiše zadatak funkcije, tačnije na koji način će funkcija da manipuliše ulaznim podacima.

function.name <- function(arguments)

```
{
computations on the arguments
some other code
}
```

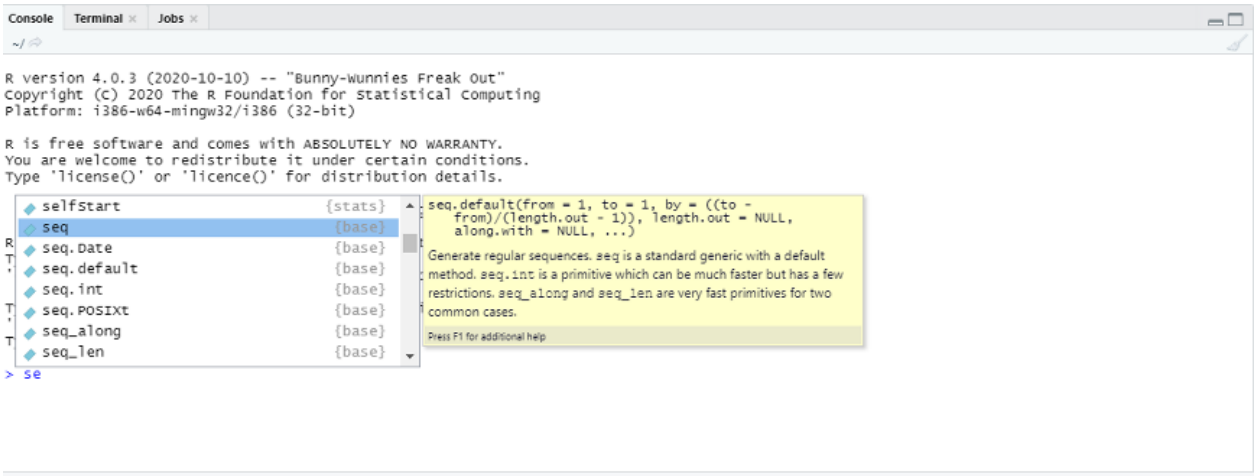
Na primer ukoliko želimo da definišemo funkciju koja će da izračunava kvadratnu vrednost ulaznih podataka, to se može realizovati sledećim kodom:

```
myFirstFun<- function(n)
{
  # Compute the square of integer `n`
  n*n
}
# Assign `10` to `k`


k <- 10
# Call `myFirstFun` with that value


m <- myFirstFun(k)


# Call `m`
m
100
```

S druge strane ukoli je potrebno prikazati telo funckije, to je moguće realizovati funkcijom body()

```
> body(myFirstFun)
{
    n * n
}
```

R ima veliki broj ugrađenih funkcija koje se pozivaju imenom funkcije. Na primer generisanje nizova brojeva se može sprovesti funkcijom *seq()*. Kako bi se što jednostavnije pozvale funckcije, R nudi opciju *tab* kojom se formiram padajući prozor sa funkcijama koje počinju na zadato slovo, a nakon odabira funkcije, taster F1 generiše plutajući prozor u komu su prikazani svi detalji, relevantni za datu funkciju (Slika 1.6).



Slika 1.6 Alati koji pojednostavljuju primenu funkcija prilikom programiranja: Tab-Padajući prozor i F1- plutajući prozor

1.4 R Biblioteke (paketi)

R biblioteke ili paketi (eng. R packages) predstavljaju kolekcije funkcija i setove podataka koji su razvijeni od strane istraživačke zajednice sa ciljem realizacije određene grupe zadataka. Na primer, biblioteka *Vegan* je napravljena sa ciljem deskriptivne analize potadataka o ekološkim zajednicama i ekolozi ga naješće koriste. *Vegan* sarži funkcije za analizu biološkog diverziteta i strukture zajednica na osnovu indeksa sličnosti. Do sada je razvjeno oko 10.000 paketa koji su skladišteni u repozitorijumima čiji je pristup besplatan. Najpoznatiji repozitorijum je *CRAN*. Postoji nekoliko ugrađenih funkcija koje omogućavaju jenostavno korišćenje R biblioteka u okviru R studija. Osnovne informacije o paketu je moguće dobiti pomoću funkcije *packageDescription()* ili *help (package="ime_paketa"):*

```
> packageDescription("vegan")
Package: vegan
Title: Community Ecology Package
Version: 2.5-6
```

```
Author: Jari Oksanen, F. Guillaume Blanchet, Michael Friendly,
Roeland Kindt, Pierre Legendre, Dan McGlinn, Peter R.
            Minchin, R. B. O'Hara, Gavin L. Simpson, Peter
Solymos, M. Henry H. Stevens, Eduard Szoecs, Helene Wagner

Maintainer: Jari Oksanen <jhoksane@gmail.com>

Depends: permute (>= 0.9-0), lattice, R (>= 3.4.0)

Suggests: parallel, tcltk, knitr

Imports: MASS, cluster, mgcv

VignetteBuilder: utils, knitr

Description: Ordination methods, diversity analysis and other
functions for community and vegetation ecologists.

License: GPL-2

BugReports: https://github.com/vegandevs/vegan/issues

URL:                                  https://cran.r-project.org,
https://github.com/vegandevs/vegan

NeedsCompilation: yes

Packaged: 2019-08-31 07:21:37 UTC; jarioksa

Repository: CRAN

Date/Publication: 2019-09-01 14:30:02 UTC

Built:  R  4.0.3;  x86_64-w64-mingw32;  2020-11-19  19:25:44  UTC;
windows

--             File:             C:/Program             Files/R/R-
4.0.3/library/vegan/Meta/package.rds
```

Instaliranje, uklanjanje i učitavanje određenog paketa se može relaizovati funkcijama *install.packages(), remove.packages(),* odnosno *libray().*

S obzirom da svaka R biblioteka ima veliki broj fajlova (funkcija i setova podataka), sadržina paketa se može istražiti funkcijom *is():*

```
> is("vegan")
[1] "character"          "vector"            "data.frameRowLabels"
"SuperClassMethod"    "index"                "atomicVector"
```

Nakon toga je moguće istražiti svaku funkciju u okviru biblioteke ponaosob pomoću funkije *help (funkcija, package="ime_paketa"):*

```
> help (cca, package=vegan)
```

2. Tipovi podataka u ekološkim studijama

Sprovođenje ekološke studije podrazumeva prikupljanje velikog broja podataka koji najčešće opisuju strukturu i dinamiku sistema različitih ekoloških nivoa (populacija, zajednica i ekosistema). Ekološke studije skroro uvek uključiju i korišćenje metode testiranja hipoteze. Ovaj pristup predstavlja prikaz eksplicitnih hipoteza o ekološkim procesima koje se prihvataju ili odbaucju na osnovu prikupljenih podataka. Kako bi se uspešno primenila ova metoda, neophodno je da svaka studija sadrži i odgovarajući statistički dizajn kojim će se na ispravan način testirati postavljene hipoteze. To podrazumeva statističku analizu skupova kvantitativnih podataka kako bi se odredilo da li postoji značajna razlika između njih koja je veća od razlike izazvane faktorom slučajnosti.

S obzirom da su svi ekološki sistemi veoma kompleksni, podaci koji ih opisuju takođe preslikavaju tu kompleksnost, što direktno utiče na varijabilnost podataka. Kako bi uspešno opisali kompleksnot analiziranih procesa i sistema, neophodno je da se na adekvatan način opiše varijabilnost podataka. Prilikom prikupljanja ekoloških podataka, svaka karakteristika koja se prati predstavlja **promenljivu**, a vrednost te promenljive je **podatak**. Na primer, promenljiva može predstavljati broj jedinki u populaciji ili broj vrsta u zajednici, kao i koncentracija nitrata u vodi.

Kako bi se opisali obrasci variranja promenljivih, koristeći adekvatne deskriptivne statističke metode, neophodno je definisati karakteristike sakupljenih podataka. Na osnovu tipa podataka, promenljive se mogu predstaviti na 4 skale:

1) **Promenljive na racionalnoj skali** imaju kvantitativnu prirodu gde se svakom merenju dodeljuje numerička vrednost. Na primer, dužina karakpaksa šumske kornjače (*Testudo hermani*) ili masa potočne mrene (*Leuciscus cephalus*) u populaciji varira, gde između jedinice mere postoje intervali konstantne veličine. Razlika između dve jedinke kornjače duzine 20mm i 21mm i dužine 81mm i 82mm ili dve potočne mrene težine 50 i 60 i 100 i 110 su iste veličine i iznose 1mm, donsono 10g . U oba slučaja merne skale imaju fiksiranu nulu koja se tumači kao odsustvo posmatranog svojstva i omogućava poređenje vrednosti promenljive pomoću količnika. Na priamer, karapaks od 60mm je duplo duži od karapaksa dužine 30mm ili duplo teže potočne mrene od mrena mase 50g su teške 100g.

2) **Promenljive na intervalnoj skali** poseduju intervale konstantne veličine između jedinica ali ne i nulu koja ima prirodno značenje (predstavlja odsustvo posmatranog svojstva). Na primer, temperatura nekog područja, merena u Celzijusima ili Farenhajtima može da prikaže razliku u temperaturi ali poređenje vrednosti u smislu povišene temperature od određenog broja puta nije ispravno tumačenje rezultata jer vrdnost 0 °C ne predstavlja odsustvo temperature. U intervalnoj skali nula se može fiksirati arbitrarno. Na primer, vreme kao cirkularna intervalna skala ima arbitrarnu nultu tačku u ponoć (00.00). Ta činenjica onemogućava smisleno poređenje vremenskih odnosa tokom dana.
3) **Promenljive na ordinalnoj skali** prikazuju relativne razlike između merenja a ne kvantitativne (numeričke), definišući rangirane kategorije sa određenom relacijom poretka. Za razliku od racionalne i intervalne skale, ordinalni podaci nemaju konstantne intervale između uzatopnih vrednosti. Tačnije, razlike između susednih kategorija nemaju iste razlike u redu veličine. Na primer, izlovljene tri jedinke potočne mrene od 30g, 200g i 500g se mogu grupisati u tri rangirane kategorije, lake, srednje i teške., respektivno. U slučaju šumske kornjače, umesto da se dužina karapaksa prikazuje na mernoj skali u cm, moguće je grupisati jedinke populacije u 4 grupe, juevnilne (do 40mm) male (od 40 do 80 mm), srednje (od 81 do 170 mm) i velike (preko 171 mm).
4) **Promenljive na nominalnoj skali.** grupišu podatke u kategorije između kojih ne postoji relacija poretka (nerangirane kategorije). U ekološkim studijama, kvalitativvne promenljive koje se najčešće prate su pol (mužijak, ženka), boja očiju (crna, braon, plava) ili grupe određenih tretmana u eksperimentu (kontrolna grupa i grupa izložena testiranom agensu).

Dužina oklopa šumske kornjače ili težina potočne mrene su promenljive koje mogu imati bilo koju vrednost u dobijenom opsegu, i spadaju u grupu **kontinuiranih promenljivih**. Između dužine karapaksa od 40mm i 41mm, postoji jos beskonačan broj vrednosti (40.1, 40,01, 40,556, 40.6768 itd) koji je ograničen senzitivnošću mernog instrumenta. To znači da između bilo koje dve vrednosti kontinuirane promenljive postoji nova vrednost. Međutim, ukoliko bi pratili apsolutnu brojnost populacije kornjača postojao bi ograničen broj vrednosti u opsegu merenja. Mogla bi da se konsatuje veličina populacije šumske kornjače od 150 ili 151 individue, ali apsolutna brojnost ne bi mogla da iznosi 150.4 individua. Takve promenljive se nazivaju **diskretne ili**

diskontinuirane promenljive (poznate još i kao merističke promenljive). Racionalne, intervalne i ordinalne skale mogu koristiti oba tipa promenljivih (kontinuirane i diskrente) dok je promenljiva na nominalnoj skali po svojoj prirodi diskretna.

2.1 R objekti – tipovi podataka u R programskom jeziku

Merenjem ekoloških parametara dobijaju se kvantitativni podaci koji su organizovani kao **skup podataka**. Uzimajući u obzir različitu prirodu podatak, skupovi podataka su u R programskom jeziku organizovani u obliku: **vektora** (*vectors*), **matrica** (*matrix*), **tabela** (*data. frame*), **nizova** (*array*) i **lista** (*lists*).

Vektori

Više elemenata istog tipa (brojeva, karaktera ili logičkih operatera) čini **vektor** koji predstavlja osnovni tip podataka u R-u. U zavisnosti od toga da li vektor čini niz realnih brojeva, (kontinuirane promenljive), celih brojeva (diskretne promenljive), niz slova ili reči (nominalna promenljiva), R razlikuje sledeće tipove vektora: *numeric vector*, *integer vector*, *character vector* i *factor vector* (Tabela 2.1):

Tablea 2.1 Tipovi promenljivih u R programskom jeziku

Tip promenljive	**skraćenica**	**Opis promenljive**
Integer	In	Koristi cele brojeve (na primer, 1, 2, 3, 4, 5)
Double	Dbl	Koristi realne brojeve (na primer, 1,23. 3.45, 5.67)
Character	Chr	Koristi znakovne nizove (slova ili reči, na primer: mužijak, ženka)
Date-time	Dttm	Koristi datume i vreme (na primer, 07,08,2020 ili 13:20am)
Logical	Lgl	Sadrži samo dve vrednosti TRUE i FALSE i koristi se za logičke vektore
Factors	Fctr	Koristi kategorijske promenljive sa fiksnim mogućim vrednostima
Dates	Date	Koristi datume

```
a=1:10
```

```
a
##[1]  1  2  3  4  5  6  7  8  9 10

b=c(1.2 2.4, 3.7, 5.25)
b
##[1] 1.20 2.40 3.70 5.25

c=c(“plavo”, “zeleno”, “crveno”)
c
##[1] "plavo"  "zeleno" "crveno"
```

Niz karaktera (na primer, mužijak, ženka) je neophodno unositi pod znacima navoda. U suprotnom R će navedeni niz prepoznati kao objekat:

```
b=c (muzijak, zenka)
## [1] Error: object 'muzijak' not found
```

Funkcijama *class(), str (), is()* ili *typeof()* je moguće proveriti tip objekta:

```
> class(a)
## [1] "integer"
> class(b)
## [1] "numeric"
> class(c)
## [1] "character

> str(a)
## [1] int [1:10] 1 2 3 4 5 6 7 8 9 10
> str(b)
## [1] num [1:4] 1.2 2.4 3.7 5.25
> str(c)
## [1] chr [1:2] "muzijak" "zenka"

> typeof(a)
## [1] "integer"
> typeof(b)
```

```
## [1] "double"
> typeof(c)
## [1] "character"

> is.integer(a)
## [1] TRUE
> is.numeric(b)
## [1] TRUE
> is.double(b)
## [1] TRUE
> is.character(c)
[1] TRUE
```

Format vektora je veoma važan prilikom izvršavanja različitih operacija i funkcija na vektorima. Na primer, matematičke operacije poput sabiranja, oduzimanja, množenja, deljenja, stepenovanja i kvadriranja vektora moguće je sprovesti samo na objektima koji su numerički:

```
> a+10
##[1] 11 12 13 14 15 16 17 18 19 20
> b*10
##[1] 12.0 24.0 37.0 52.5
> a^2
##[1]   1   4   9  16  25  36  49  64  81 100
> sqrt(a)
##[1] 1.000000 1.414214 1.732051 2.000000 2.236068 2.449490 2.64
5751 2.828427 3.000000 3.162278
> c + 2
##[1] Error in c + 2 : non-numeric argument to binary operator
```

Ukoliko vektor sadrži samo jedan karakter u nizu a sve ostale brojeve, takav vektor će od strane R-a biti prepoznat kao *character vector*:

```
d= c( a , „plavo“)
d
[1] "1"     "2"     "3"     "4"     "5"     "6"     "7"     "8"
"9"    "10"    "plavo"
```

```
str(d)
 chr [1:11] "1" "2" "3" "4" "5" "6" "7" "8" "9" "10" "plavo"
```

Sprovođenje matematičkih operacija i funkcija sa takvim tipom vektora je takođe nemoguće. Me đutim, funkcijom *as()* je moguće izvršiti transformaciju tipa vektora. Na primer, ukoliko se transf ormiše *character* vektor u *integer* vektor funkcijom *as.integer ()*, R transformiše karaktere u vekt oru u *NA* (nepoznatu vrednost eng. „*Not available*'') i ne dodeljuje mu konkretnu vrednost.

```
e=as.integer(d)
e
 [1]  1  2  3  4  5  6  7  8  9 10 NA
Warning message:
NAs introduced by coercion
```

S obzirom da je promenjena forma vektora u *integer* , ponovo je moguće sprovoditi bilo koje matematiče operacije i funakccije nad vektorom:

```
e+2
[1]  3  4  5  6  7  8  9 10 11 12 NA
```

Iz ovog izraza se vidi da svaka operacija u kojoj učestvuje nepoznata vrednost (*NA*), takođe postaje nepoznata.

Funkcijom *methods (as)* moguće je proveriti sve dostupne tipove funkcija *as()*:

```
Method (as)

[1] as.array                        as.array.default
  [3] as.call                         as.character
  [5] as.character.condition          as.character.Date
  [7] as.character.default            as.character.error
  [9] as.character.factor             as.character.hexmode
 [11] as.character.numeric_version    as.character.octmode
 [13] as.character.POSIXt             as.character.srcref
 [15] as.complex                      as.data.frame
 [17] as.data.frame.array             as.data.frame.AsIs
 [19] as.data.frame.character         as.data.frame.complex
 [21] as.data.frame.data.frame        as.data.frame.Date
 [23] as.data.frame.default           as.data.frame.difftime
 [25] as.data.frame.factor            as.data.frame.integer
 [27] as.data.frame.list              as.data.frame.logical
 [29] as.data.frame.matrix            as.data.frame.model.ma
trix
 [31] as.data.frame.noquote           as.data.frame.numeric
 [33] as.data.frame.numeric_version   as.data.frame.ordered
 [35] as.data.frame.POSIXct           as.data.frame.POSIXlt
 [37] as.data.frame.raw               as.data.frame.table
 [39] as.data.frame.ts                as.data.frame.vector
 [41] as.Date                         as.Date.character
```

Tabele

Tabele su najčešći format podataka koji podrazumeva sekvencu kolona iste dužine, gde svaka kolona može sadržati različite tipove podataka. U ekološkim studijama, skupovi podataka, dobijeni merenjem parametara su najčešće organizovani u vidu tabela različitih ekstenzija (na primer, *.txt*, *.csv*, i *.xcl*) gde kolone predstavljaju registrovane vrste na lokalitetu ili neke metričke parametre (indekse diverziteta i sličnosti) kada je reč o biotičkim podacima, ili sredinske parametre (temperatura, koncentracija kiseonika, koncentracija nitrata) kada je reč o abiotičkim promenljivim. Redove u tako organizovanoj tabeli predstavljaju lokaliteti gde je sprovedeno uzorkovanje biotičkog materijala ili merenje abiotičkih parametara (Table 2.2).

Tabela 2.2 Biotički i abiotički podaci ekološke studije organizovani u formi tabele

	Vrsta 1	Vrsta 2	Temperatura vode (°C)
Lokalitet 1	3	54	12
Lokalitet 2	4	0	14
Lokalitet 3	6	3	13

Format tabele koji koristi R programski jezik je *.dataframe*. Tabele u programskom jeziku se mogu generisati formiranjem i spajanjem vektora, ili pak učitati iz drugih formata. Formiranje tabela spajanjem vektora se izvršava funkcijom *data.frame()* na sledeći način:

```
> Vrsta 1= c(3, 4, 6)
Error: unexpected numeric constant in "Vrsta 1"
> Vrsta_1= c(3, 4, 6)
> Vrsta_2= c(54, 0, 3)
> Temperatura_vode = (12, 14, 13)
Error: unexpected ',' in "Temperatura_vode = (12,"
> Temperatura_vode = c(12, 14, 13)
> Tabela = data.frame(Vrsta_1, Vrsta_2, Temperatura_vode)
> Tabela
  Vrsta_1 Vrsta_2 Temperatura_vode
1       3      54               12
2       4       0               14
3       6       3               13
```

Nakon izvršavanja prethodnog koda, R je generisao tabelu *data.frame* formata što se može proveriti funkcijom *class()*:

```
Class(Tabela)
[1] "data.frame"
```

Osnovne karakteristke tabele kao što su dimenzije (broj redova i kolona) prikaz prvih 6 redova tabele i prikaz poslednjih 6 redova tabele moguće je dobiti pomoću funkcija *dim(), head() i tale()*:

```
dim (Tabela)
##[1] 3 3
head (Tabela)
```

```
  Vrsta_1 Vrsta_2 Temperatura_vode
1       3      54                12
2       4       0                14
3       6       3                13
tail (Tabela)
  Vrsta_1 Vrsta_2 Temperatura_vode
1       3      54                12
2       4       0                14
3       6       3                13
```

Ukoliko je potrebno da se na postojeću tabelu dodaju nove kolone ili redovi, to je moguće učiniti funkcijama *cbind()* i *rbind()*. Na primer, sledećim kodom se dodaje novi parametar koncentracije kiseonika u vodi (O2mg/l) i vrednosti svih merenih parametara za novi lokalitet, *lokalitet_4*:

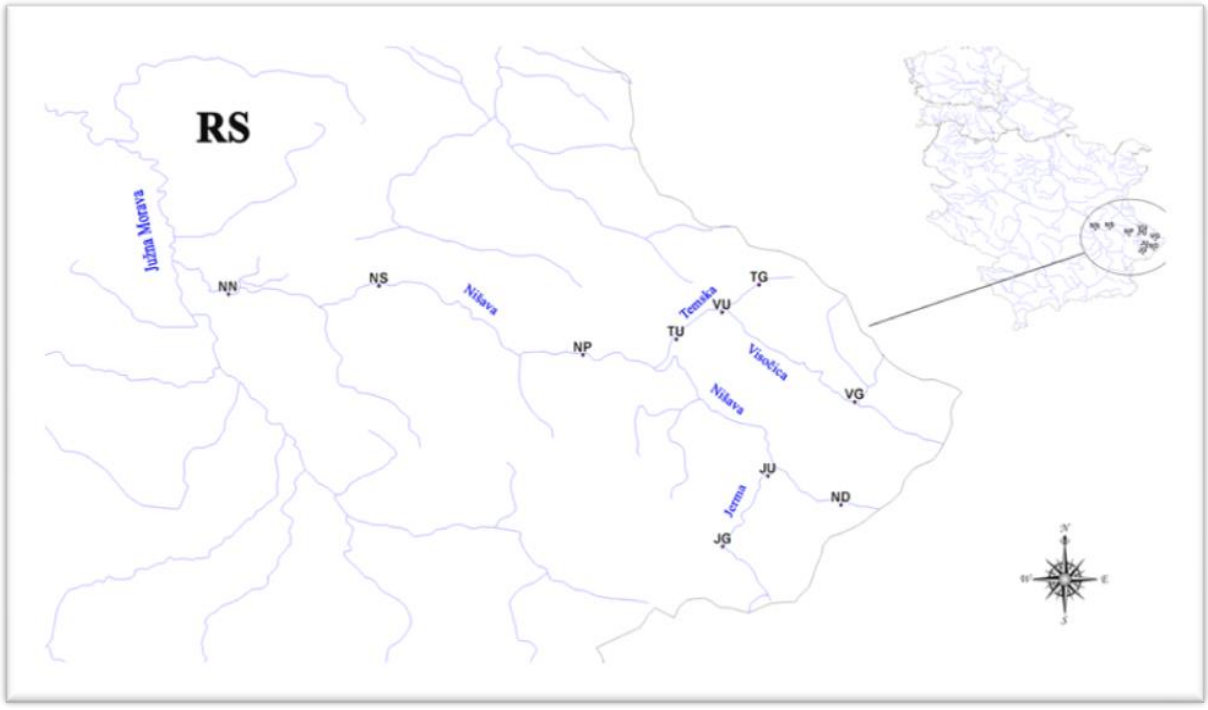
```
Tabela_cbind=cbind(Tabela, O2=c(8.23, 5.35,3.23))
Tabela_cbind
Vrsta_1 Vrsta_2 Temperatura_vode   O2
1       3      54                12 8.23
2       4       0                14 5.35
3       6       3                13 3.23

Tabela_rbind=rbind(Tabela_cbind, c(3, 5,14, 5.32))
Tabela_rbind
Vrsta_1 Vrsta_2 Temperatura_vode   O2
1       3      54                12 8.23
2       4       0                14 5.35
3       6       3                13 3.23
4       3       5                14 5.32
```

Učitavanje već postojećih tabela i njihovo uvoženje u R ambijent se sprovodi funkcijama *read.table()* i *read.csv()*:

```
Tabela_2 = read.table
(c:/ECOBIAS/Numericka_ekologija/ime_fajla.txt, sep=“,“,
header=TRUE)
```

U okviru funkcije *read.table*, prvi argument definiše put do foldera u kome se nalazi tabela za uvoz. Argument *sep* definiše način na koji su razdvojene kolone u dokumentu koji se uvozi u R, tačnije definiše separator koji je najčešće zapeta, dve tačke ili prazan prostor (*space* ili *tab*). Argument *header* definiše postojanje zaglavlja u tabeli, odnosno prvog reda tabele koji nosi imena kolona. Ukoliko treba transformisati prvi red kolone u zaglavlje, vrednost argumenta *header* je *TRUE*, a u suprotnom je *FALSE*.

U ovom polavlju će se kao primer koristiti tabela u *csv* formatu (eng. *Comma-separated Values*, odnosno, zarezima razdvojene vrednosti), sa podacima prikupljenim tokom dvomesešne kampanje uzorkovanja makrobeskičmenjaka i riba na slivu Nišave. Na 10 različitih lokaliteta, distribuiranih duž sliva prikupljani su uzorci bentosnih markobesičmenjaka i riba kvantitativnom metodom (bentosnom mrežom, odnosno elektroagregatom). Svi podaci su organizovani u tabeli sa 128 kolona (vrste makrobeskičmenjaka i riba i sredisnki parametri) i 20 redova (lokaliteti), u *csv* formatu.



Slika XX Prostorni raspored lokaliteta na slivu Nišave

Uvoženje tabele u R se sprovodi funkcijom read.*cvs ()*:

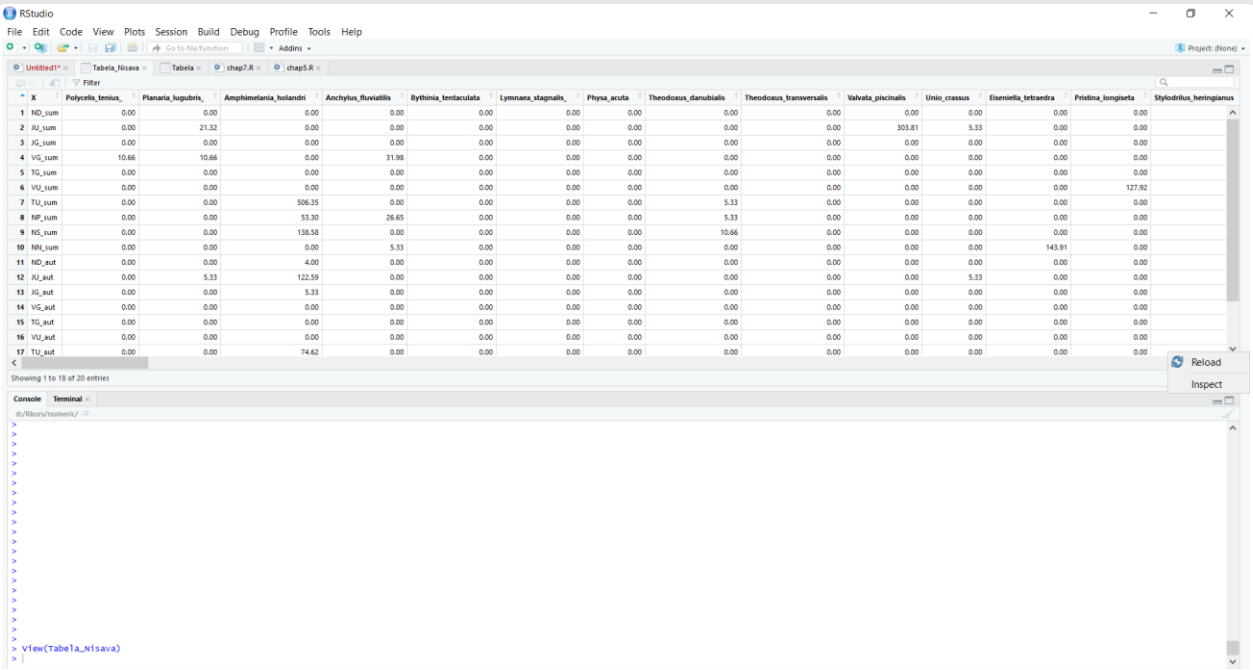
```
Tabela_Nisava= read.csv (“Tabela_Nisava.csv)
```

Uvezena *csv* tabela ima formu *data.frame*:

```
class (Tabela_Nisava)
[1] "data.frame"
```

Za prikazivanje strukture uvezene tabele se obično koristi funkcija *head()* i *tale()*. Kod glomaznih tabela sa velikim brojem parametara (kolona) praktičnije je koristiti funkciju *View ()* kojom je moguće otvoriti posebni prozor u R studiju u kome je prikazana tabela

```
View (Tabela_Nisava
```



U R programskom jeziku je prilično jednostavno manipulisati tabelama kroz različite operacije selekecije i ekstraktcije elemenata tabele. Uglaste zagrade [ ] se koriste za selektovanje redova i kolona u tabeli. U okviru zagrada su definisane dve pozicije odvojene zapetom, prva se odnosi na redove a druga na kolone u tabeli. Ukoliko je potrebno selektovati deo redova u tabeli, moguće je napraviti novi objekat, tabelu koja će sadržati samo selektovane redove:

```
Tabela_5_lokaliteta = Tabela_Nisava [ 1:5, ]
Tabela _5
```

```
      X Polycelis_teniujus_ Planaria_lugubris_ Amphimelania_holandri Anchylu
sf
1 ND_sum              0.00              0.00                0          0.00
2 JU_sum              0.00             21.32                0          0.00
3 JG_sum              0.00              0.00                0          0.00
4 VG_sum             10.66             10.66                0         31.98
5 TG_sum              0.00              0.00                0          0.00
```

Pored redova, uglaste zagrade pružaju mogućnost selektovanja i odgovarajućih kolona u tabeli, na primer moguće je izdvojiti samo abiotičke parametre iz tabele *Tabela_Nisava*:

```
Tabela_5_lokaliteta_abiotički = Tabela_Nisava [1:5 , 115:127]
Tabela_5_lokaliteta_abiotički
t    v  ep   ph  o2mg   o2. bpk5   no3    po4   nh3 tvrdoca nad_vis sirina
1 17.9 0.40 533 6.84  9.65 104.3 3.29 4.296 1.2710 1.705 296.80   461      9
2 17.1 0.74 436 6.94 10.52 113.7 3.94 3.253 0.3680 0.544  206.02  426     15
3 16.2 0.51 431 7.11 11.77 126.9 4.17 4.767 0.1425 0.511  263.51  554     12
4 13.8 0.48 261 7.24 11.69 125.7 3.72 1.538 0.1425 0.640  186.85  694     12
5 23.1 0.33 123 6.85  9.58 119.6 3.00 0.496 0.1637 0.547   74.26  642     10
```

Ukoliko je potrebno, moguće je selektovati i pojedinačne ćelije u tabeli, na primer izmerena temperatura vode (*t*) na lokalitetu broj 5 ili prikazati brojnost vrste potočne pastrme (*Salmo trutta)* na prva tri lokaliteta :

```
Tabela_Nisava [5, 115]
##[1] 23.1
Tabela_Nisava [c(1,2,3), 108]
##[1]  0  4 11
```

Ukoliko su tabele sačinjene iz velikog broja kolona, označiti odgovarajući parametar nije jednostavno i može doći lako do greške, u tom slučaju je jednostavnije da željenu kolonu selektujete prema imenu. Tom prilikom ime kolone označite apostrofima (‘’):

```
Tabela_Nisava [5, ‘t’]
##[1] 23.1
Tabela_Nisava [c(1,2,3), ‘Salmo_trutta’]
##[1]  0  4 11
```

Selektovanje određenih kolona prema imenu se može sprovesti funkcijom (atributom) $:

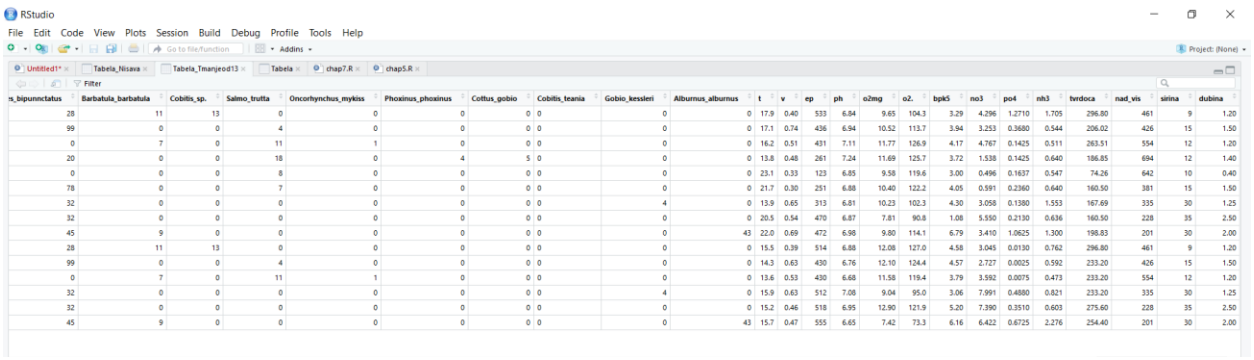
```
Table$t[5]
##[1] 23.1
```

Isključivanje određenih tabela ili redova se sprovodi na isti način kao i selektovanje, preko uglastih zagrada ali sa dodatim prefiksom – (minus). Na primer, ukoliko je potrebno da se iz tabele *Tabela_Nisava* isključe podaci o makrobeskičmenjacima i sredinskim parametrima, takva operacija se izvršava sledećim kodom:

```
Tabela_ribe= Tabela [,-c(1:100, 114:127]
Tabela_ribe
Leuciscus_cephalus Rhodeus_sericeus Gobio_gobio Barbus_balcanicus Alburnoides_bipunnctatus
Barbatula_barbatula Cobitis_sp. Salmo_trutta Oncorhynchus_mykiss Phoxinus_phoxinus Cottus_gobio
Cobitis_teania Gobio_kessleri
1                 84                0           0               122                         28
11          13             0                   0                 0             0              0
0
2                 14                4           9                 35                         99
0           0             4                   0                 0             0              0
0
3                 51                0          15                 47                          0
7           0            11                   1                 0             0              0
0
4                 30                0           0                 40                         20
0           0            18                   0                 4             5              0
0
5                  0                0           0                 96                          0
0           0             8                   0                 0             0              0
0
  Alburnus_alburnus
1                 0
2                 0
3                 0
4                 0
5                 0
```

Funkcija *subset()* se moze koristiti za filtriranje tabela prema određenim kriterijumima. Na primer, ukoliko je potrebno da iz tabele selektujemo sve lokalitete gde je izmerena temperatura vode (*t*) veća od 13 stepeni to se može realizovati sledećim kodom:

```
Tabela_Tmanjeod13=subset(Tabela_Nisava, Tabela_Nisava$t>13)
```



ns_bipunnctatus	Barbatula_barbatula	Cobitis_sp.	Salmo_trutta	Oncorhynchus_mykiss	Phoxinus_phoxinus	Cottus_gobio	Cobitis_teania	Gobio_kessleri	Alburnus_alburnus	t	v	ep	ph	o2mg	o2.	bpk5	no3	po4	nh3	tvrdoca	nad_vis	sirina	dubina
28	11	13	0	0	0	0	0	0	0	17.9	0.40	533	6.84	9.65	104.3	3.29	4.296	1.2710	1.705	296.80	461	9	1.20
99	0	0	4	0	0	0	0	0	0	17.1	0.74	436	6.94	10.52	113.7	3.94	3.253	0.3680	0.544	206.02	426	15	1.50
0	7	0	11	1	0	0	0	0	0	16.2	0.51	431	7.11	11.77	126.9	4.17	4.767	0.1425	0.511	263.51	554	12	1.20
20	0	0	18	0	4	5	0	0	0	13.8	0.48	261	7.24	11.69	125.7	3.72	1.538	0.1425	0.640	186.85	694	12	1.40
0	0	0	8	0	0	0	0	0	0	23.1	0.33	123	6.85	9.58	119.6	3.00	0.496	0.1637	0.547	74.26	642	10	0.40
78	0	0	7	0	0	0	0	0	0	21.7	0.30	251	6.88	10.40	122.2	4.05	0.591	0.2360	0.640	160.50	381	15	1.50
32	0	0	0	0	0	0	0	4	0	13.9	0.65	313	6.81	10.23	102.3	4.30	3.058	0.1380	1.553	167.69	335	30	1.25
32	0	0	0	0	0	0	0	0	0	20.5	0.54	470	6.87	7.81	90.8	1.08	5.550	0.2130	0.636	160.50	228	35	2.50
45	9	0	0	0	0	0	0	0	43	22.0	0.69	472	6.98	9.80	114.1	6.79	3.410	1.0625	1.300	198.83	201	30	2.00
28	11	13	0	0	0	0	0	0	0	15.5	0.39	514	6.88	12.08	127.0	4.58	3.045	0.0130	0.762	296.80	461	9	1.20
99	0	0	4	0	0	0	0	0	0	14.3	0.63	430	6.76	12.10	124.4	4.57	2.727	0.0025	0.592	233.20	426	15	1.50
0	7	0	11	1	0	0	0	0	0	13.6	0.53	430	6.68	11.58	119.4	3.79	3.592	0.0075	0.473	233.20	554	12	1.20
32	0	0	0	0	0	0	0	4	0	15.9	0.63	512	7.08	9.04	95.0	3.06	7.991	0.4880	0.821	233.20	335	30	1.25
32	0	0	0	0	0	0	0	0	0	15.2	0.46	518	6.95	12.90	121.9	5.20	7.390	0.3510	0.603	275.60	228	35	2.50
45	9	0	0	0	0	0	0	0	43	15.7	0.47	555	6.65	7.42	73.3	6.16	6.422	0.6725	2.276	254.40	201	30	2.00

Pored integrisanih funkcija u R-u koje manipulišu podacima, organizovanim u tradicionalnom okviru (data.frame), u poslednje vreme sve popularniji postaje paket *dplyr* i *tibble* koji su deo jezgra paketa *tidyverse*. Pomenuti paketi uvode novine koje olakšavaju programiranje u R-u i čine ga intuitivnijim od prethodnih R-ovih ugrađenih okvira s podacima i funkcijama. Novi okvir podataka se zove *tibl* (eng „tibble“) i za razliku od *data.frame* formata preglednije prikazuje strukturu tabele koja staje na jedan ekran. Prvi red *tibla* troslovnim ili četvoroslovnim skraćenicama pruža informacije o tipu svake promenljive (kolone):

- *int* (*integer*) -celi brojevi
- *dbl* (*double*) – realni brojevi
- *chr* (*character*) – znakovni nizovi
- *dttm* (*date-time*) – datum i vreme
- *lgl* (*logical*) – vektori sa dve vrednost, TRUE ili FALSE
- *fctr* (*factor*) – kategorijska promenljiva
- *date* (*dates*) – datumi

Uvoženje podataka u *cvs* formatu u *tible* okvir se sprovodi funkcijom *read_csv()* koja je jako slična funkciji *read.csv()* koja se koristi za data.frame format. Ukoliko je potrebno, moguće je prevoditi podatke iz jednog formata u drugi funkcijama *as.tibble()* i *as.data.frame()*.

```
> Tabela_Nisava_data_frame=read.csv("Tabela.csv")
> class(Tabela_Nisava_data_frame)
[1] "data.frame"
> Tabela_Nisava_tibble=read_csv("Tabela.csv")
Parsed with column specification:
```

```
cols(
  .default = col_double(),
  Lok = col_character()
)
See spec(...) for full column specifications.
> class(Tabela_Nisava_tibble)
[1] "spec_tbl_df" "tbl_df"      "tbl"         "data.frame"
>
     Tabela_Nisava_tibble_transform=as.tibble(Tabela_Nisava_data
     _frame)
> class(Tabela_Nisava_tibble_transform)
[1] "tbl_df"     "tbl"        "data.frame"
```

Još jedna prednost *tibl*-a je da može sadržati nesintaktička imena kolona poput razmaka ali je u tom slučaju neophodno označiti takva imena obrnutim apostrofima (’’)

Pored navedenih novina u okviru *tibble* paketa, struktura sintakse sa funkcijama *dplyr* je drugačija. Sve *dplyr* funkcije koriste tabele kao prvi argument dok pomoću operatora cevi (%>%) grupiše i sprovodi niz funkcija izbegavajući formiranje međuobjekta. Operator %>% treba čitati kao veznik “onda”. Operatori cevi su jako korisni u kodu ali ih treba izbegavati ukoliko je kod predugačak (više od 10 koraka) ili ima više ulaznih promenljivih. Primer koda sa operatorom %>% gde je krajnji produkt rangirana tabela frekventnosti u opadajućem nizu polne zrelosti riba u plavnim područjima Dunava izgleda ovako:

```
> Baza_Dunav%>%
+   group_by(`polna_zrelost`) %>%
+   summarize(freq=n())%>%
+   arrange (desc(freq))
# A tibble: 4 x 2
  polna_zrelost  freq
          <dbl> <int>
1             3    72
2             4    48
3             1    44
     4             2    23
```

Ukoliko se operatori cevi koje grupišu tri funkcije u naredbodavnu akciju ne bi koristile u kodu, isti produkt bi zahtevao generisanje tri međuprodukta i kod bi bio opširniji:

```
> Baza_Dunav_gr=group_by(Baza_Dunav,`polna_zrelost`)
> Baza_dunav_freq=summarize(Baza_Dunav_gr, freq=n())
> Baza_dunav_freq
# A tibble: 4 x 2
  polna_zrelost  freq
          <dbl> <int>
1             1    44
2             2    23
3             3    72
4             4    48

> Baza_dunav_arrange=arrange(Baza_dunav_freq, desc(freq))
> Baza_dunav_arrange
# A tibble: 4 x 2
  polna_zrelost  freq
          <dbl> <int>
1             3    72
2             4    48
3             1    44
4             2    23
```

U okviru paketa dplyr, postoji niz funkcija koje jednostavno transfromišu tabele i mogu se grupisati u tri kategorije:

Manipulacija redova:

- *filter( )* selektuje redove na osnovu vrednosti kolona,
- *slice( )* selektuje redove na osnovu lokacije u tabeli
- *arrange( )* menja redosled redova u tabeli
- *summarise( )* sažima okvir sa podacima u samo jedan red i obično se koristi zajedno sa funkcijom *group_by*( )

Manipulacija kolona:

- *select( )* selektovanje kolona

- *rename( )* promena imena kolona
- *mutate( )* formiranje novih kolona uz transformaciju postojećih
- *relocate( )* promena redosleda kolona u tabeli

Sve navedene funkcije će biti korišćene u kodovima narednih poglavlja u knjizi.

3. Istraživačka analiza podataka

U ekološkim studijama se prikupljaju podaci i organiziuju u obliku skupova podataka koji sadrže različit broj promenljivih, izmerenih tokom više opservacija (slučaja). Svaki skup podataka u sebi ima skrivenu varijabilnst, tendenciju promene vrednosti merenih parametara kroz set opservacija. Na primer, koncentracija rastvorenog kiseonika u vodi u jednoj reci će na različitm lokalitetima ili čak na istom lokalitetu biti različita prilikom svakog narednog merenja. Obrazac varijabilnosti je informacija koju treba obezbediti iz skupa podataka i koja će pružiti odgovor na postavljena pitanja, neophodna za formulaciju hipoteze. Način na koji se menja koncentracija rastvorenog kiseonika duž longitudinalnog profila reke je informacija koja može delimčno da objasni smenjivanje vrsta makrobeskičmenjaka u bentosnoj zajednici od izvora do ušća.

S obzirom da je svaki ekološki sistem (populacija, zajednica, ekosistem) kompleksna struktura čija dinamika zavisi od velikig broja parametara, pored varijabilnosti unutar promenljive, potencijalna kovarijabilnost koja predstavlja odnos između promenljivih takođe predstavlja veoma značajnu informaciju koju treba ekstrahovati iz skupa podataka. Kovarijabilnost definiše tendenciju da dve ili više promenljivih pokazuju obrazac varijabilnosti koji je povezan. Uokliko se koncentracija rastvorenog kiseonika u vodi duž longitudinalnog gradijenta reke menja zajedno sa promenom temperature vode, gde u hladnijoj vodi su vrednosti rastvorenog kiseonika više, a u toplijoj niže, takvi parametri pokazuju kovarijabilnost i upućuju na njihovu povezanost.

Za objašnjavanje ekoloških fenomena u prirodi nekada je potrebno pratiti zasebne promenljive tokom više opservacija koje pružaju univarijantnu informaciju. Na primer, količina kiseonika u vodi se meri koncentracijom rastovernog kiseonika u miligramima po litru vode, dok se količina rastovrenih soli u vodi prati konduktivitetom (elektroporvodljivost u µS/cm). Metričke osobine poput totalne dužine riba i mase tela, pružaju informaciju o kondicionom stnaju populacije. Međutim, mnogo je češći slušaj da su ekološki sistemi opisani istovremeno velikim brojem promeljivih. Tada je reč o multidiemnzionalnim podacima koji za analizu zahtevaju poseban statistički dizajn. Zajednica makrobeskičmenjaka se sastoji iz određenog broja vrsta, gde je svaka vrsta prisutna sa određenim brojem jedinki. U takvom skupu podataka, svaka vrsta (kolona) predstavlja zasebnu promenljivu, dok svaka opservacija (red) predstavlja uzorkovanu zajednicu, a u ćelijama tabele se nalaze vrednosti brojnosti prisutnih vrsta. Za opisivanje strukture

zajednice kao ekološkog atributa je u tom slučaju neophodno uzeti u obzir sve vrste (promeljive) istovremeno.

Pre odabira adekvatnih metoda za statističku analizu skupa podataka, prvi korak predstavlja prikupljanje inofrmacija o samoj prirodi podataka, glavnim obrascima varijabilnosti i prisustvu uobičajenih i neuobičajenih vrednosti u samom skupu. Ovaj proces se naziva **istraživačka analiza podataka (IAP)**, kada istraživač na svom prvom susretu sa podaacima pomoću tehnika vizuelizacije i jednostavnih statistčkih alata traži odgovore na niz postavljenih pitanja koja ga vode do konačne formulacije hipoteze.

3.1 Deskriptivna statistika

Skupovi podataka u ekološkim istraživanjima čine ponovljena merenja jedne ili više promenljihiv. Najvažnija karakteristika skupa podataka je varijabilnost, zbog čega se podaci statistički opisuju stepenom varijabilnosti čiji su pokazatelji **opseg merenja**, **varijansa**, **standardna devijacija** i **standardna greška**, o čemu će biti reči kasnije.

U tabli Tabela_Nisava.csv, prisutne su vrednosti temperature vode, izmerene na svih 10 lokaliteta tokom 2 sezone. Potskup podataka o temperaturi vode se može izdvojiti sledećim kodom:

```
> Tabela_t=select(Tabela_Nisava, t)
> Tabela_t
# A tibble: 20 x 1
       t
   <dbl>
 1  17.9
 2  17.1
 3  16.2
 4  13.8
 5  23.1
 6   9.5
 7  21.7
 8  13.9
 9  20.5
10  22
11  15.5
12  14.3
```

```
13  13.6
14  10
15  12.2
16   7.3
17  11.8
18  15.9
19  15.2
20  15.7
```

Na osnovu podataka iz tabele se može zaključiti da vrednosti temperature vode nisu slučajno raspoređene već se grupišu oko centralne vrednosti, u ovom primeru oko vrednosti 15, 16 i 17, odnosno nisu nasumično raspoređene od 7.3 do 23.1. Parametri koji opisuju obrazac grupisanja nazivaju se i mere centralne tendencije.

Mera centralne tendencije daje informacije o tome kako su raspoređene vrednosti temperature vode na testiranim lokalitetima. Ona se koristi da bi predstavila neku karakterističnu (prosečnu) vrednost posmatranog uzorka. Mere centralne tendencije koje se najčešće koriste su srednja vrednost, medijana i mod.

Srednja vrednost temperature vode u datom setu lokaliteta predstavlja njenu aritmetičku sredinu. To je suma svih vrednosti datog parametra u setu testranih lokaliteta, podeljena brojem merenja:

$$\textbf{Srednja vrednost} = \overline{X} = \frac{\sum x_i}{N}$$

xi – vrednost temperature vode na bilo kom lokalitetu;

N – ukupan broj merenja (lokaliteta)

Medijana predstavlja centralnu vrednost rangiranih vrednosti temperature vode u ovom skupu podataka, od najnižih do najviših vrednosti ili obrnuto. To je tačka ispod i iznad koje se nalazi po 50% elemenata uzorka. Ona deli osnovni skup podataka na dva podskupa koji imaju isti broj elemenata. Ukoliko postoji paran broj merenja (10 merenja) u skupu podataka, medijana je jednaka srednjoj vrednosti između pete (r_5) i šeste rangirane (r_6) vrednosti:

$$M = \frac{(r_5 + r_6)}{2}$$

gde je M medijana a r_i vrednost iz skupa podataka sa rangom i

U slučaju neparnog broja merenja (na primer 9), ona odgovara centralno rangiranoj vrednosti:

$$M = r_5$$

Na medijanu utiču vrednosti svih elemenata uzorka, ali samo svojim položajem u nizu rangiranih vrenodsti, a ne i svojom vrednošću, što je čini otpornom na uticaj veoma ekstremnih pojedinačnih vrednosti.

Treća mera centralne tendencije je **mod** koji predstavlja vrednost koja se najčešće javlja u skupu podataka. Na primeru temperature vode, mod bi predstavljao vrednost temperature koja je najčešće izmerena tokom kampanje uzorkovanja. Pošto na mod ne utiču sve vrednosti temperature u ovom uzorku već samo one koje se najviše ponavljaju, mod je najmanje pouzdana i najmanje korišćena mera centralne tendencije.

Srednja vrednost je najčešće upotrebljavana mera centralne tendencije, ali se medijana i mod koriste u slučaju kada u uzorku postoje ekstremne vrednosti. Medijana i mod su za razliku od srednje vrednosti jedine mere centralne tendencije pogodne za ordinalne podatke. Primena mera centralne tendencije u sledećoj vežbi otkriva jasnu razliku između njih.

U R programskom jeziku, mere centralne tendecije se jednostavno računaju sledećim funkcijama:

```
> mean(Tabela_Nisava$t) ###srednja vrednost
[1] 15.365
> median(Tabela_Nisava$t)###mediana
[1] 15.5
```

U R-u ne postoji funkcija koja direktno računa mod kao meru centralne tendencije. Međutim, pomoću funkcija *count()* i *sort()* je moguće izračunati mod na sledeći način:

```
Tabela_t %>%
  count(t)%>%
  arrange(desc(n))
# A tibble: 18 x 2
      t     n
   <dbl> <int>
 1  15.5     3
 2   7.3     1
 3   9.5     1
 4  10       1
```

```
      5  11.8     1
      6  12.2     1
      7  13.6     1
      8  13.8     1
      9  13.9     1
     10  14.3     1
     11  15.9     1
     12  16.2     1
     13  17.1     1
     14  17.9     1
     15  20.5     1
     16  21.7     1
     17  22       1
     18  23.1     1
```

Funkcija *arrange()* je prikazala vrednost koja se najčešće javlja u skupu podataka na vrh tabele. Vrednost 15.5 pojavljuje se tri puta i predstavlja mod za dati skup podatakaka

Varijabilnost u skupu podataka

Mere centralne tendencije ne pružaju informaciju o variranju vrednosti unutar skupa podataka. Kao primer, može se navesti potočna mrena (*Barbus balcanicus*), koja je izlovljavana na četiri lokaliteta (reka Sokobanjska Moravica):

```
     Baza_B_balcanicus=read_csv("Baza_Moravica.csv")
     Baza_B_balcanicus
     # A tibble: 4 x 11
       X1    Jedinka1 Jedinka2 Jedinka3 Jedinka4 Jedinka5 Jedinka6
Jedinka7
       <chr>    <dbl>    <dbl>    <dbl>    <dbl>    <dbl>    <dbl>
<dbl>
     1 Lok1        22       24       25       29       30       NA
NA
     2 Lok2         5       10       15       30       70       NA
NA
     3 Lok3        22       24       25       24       27       26
29
```

```
    4 Lok4         5       44       10       60       15       35
30
    # ... with 3 more variables: Jedinka8 <dbl>, Jedinka9 <dbl>,
Jedinka10 <dbl>
```

U uzorku potočne mrene na lokalitetu 2 izmerena težina ribe pokazuje mnogo veću varijabilnost u odnosu na uzorak sa lokaliteta 1, iako njihovi skupovi podataka imaju istu srednju vrednost:

```
    Baza_B_balcanicus %>%
      transmute(X1, Mean = rowMeans(select(., -X1), na.rm=TRUE))
    rowMean(Baza_B_balcanicus)
    # A tibble: 4 x 2
      X1     Mean
      <chr> <dbl>
    1 Lok1   26
    2 Lok2   26
    3 Lok3   26.2
    4 Lok4   34.8
```

S obzirom da na lokalitetu Lok1 i Lok2 u skupu podataka ne postoje vrednosti za svih 10 jedinki, u tim ćelijama tabele stoji vrednost NA. Funkcija *rowMeans()* ne funkcioniše ukoliko u skupu podataka postoje NA vrednosti. Zbog toga se dodaje argument *na.rm=TRUE*, koji definiše odnos funkcije prema NA ignorišći takve vrednosti u skupu podataka i omogućavajući izračunavanje srednju vrednost uzorka. Prethonim kodom je izračunata srednja vrednst lokaliteta Lok1 i Lok2, koja je pored očigledne razlike između dva skupa podataka identična i iznosi 26g. Prema tome, podatke je neophodno statistički opisati **i stepenom varijabilnosti**, koji se predstavlja pomoću nekoliko parametara od kojih su najznačajniji **opseg**, **varijansa**, **standardna devijacija** i **standardna greška**.

Opseg merenja (variranja) na primeru potočne mrene predstavlja razliku između maksimalne i minimalne vrednosti izmerenih težina. Opseg merenja na lokalitetu Lok1 iznosi 8g (22g-30g), dok na lokalitetu Lok2 iznosi 65g (5g-70g). Srednja vrednost težine potočne mrene po uzorku je ista za oba lokaliteta, ali opseg pokazuje mnogo veću varijabilnost podataka u slučaju

uzorka protočne mrene izlovljavane na lokalitetu Lok2. Opseg merenja može prikazati nerealno veću varijabilnost usled prisustva ekstremnih vrednosti u slučaju nereprezentativnog uzorka.

U R-u se opseg merenja računa pomoću funkcija *max()* i *min()* koji izračunavaju maksimalnu i minimalniu vrednost u skupu podataka:

```
row.names(Baza_B_balcanicus) = Baza_B_balcanicus[,1]
Baza_B_balcanicusT=t(Baza_B_balcanicus[,-1])
Baza_B_balcanicusT
Baza_B_balcanicusT2=as.tibble(Baza_B_balcanicusT)
Baza_B_balcanicusT2
# A tibble: 10 x 4
    Lok1  Lok2  Lok3  Lok4
   <int> <int> <int> <int>
 1    22     5    22     5
 2    24    10    24    44
 3    25    15    25    10
 4    29    30    24    60
 5    30    70    27    15
 6    NA    NA    26    35
 7    NA    NA    29    30
 8    NA    NA    30    38
 9    NA    NA    28    70
10    NA    NA    27    41

max(Baza_B_balcanicusT2$Lok1,na.rm=TRUE)-
min(Baza_B_balcanicusT2$Lok1, na.rm=TRUE)
[1] 8
max(Baza_B_balcanicusT2$Lok4) -min(Baza_B_balcanicusT2$Lok4)
[1] 65
```

Pošto funkcije *max()* i *min()* izračunavaju maksimalnu i minimalnu vrednost za kolone, generisana je nova tabela „Baza_B_balcanicusT2" koja je transponovana funkcijom *t()* pa su nakon transformacije, redovi postali kolone i obrnuto. Nakon oduzimanja minimalne vrednosti skupa od maksimalne, dobijeni su opsezi merenja za Lok1 i Lok4 i iznose 8g, odnosno 45g.

Varijansa (σ^2) težine u datom uzorku potočne mrene predstavlja prosečno kvadratno odstupanje pojedinačnih merenja od srednje vrednosti u uzorku. Na taj način se stiče uvid u to kako podaci variraju oko srednje vrednosti. Varijansa pruža mnogo više informacija u odnosu na opseg, a izračunava se veoma jednostavno. Prvi korak je izračunati srednju vrednost ($\overline{\boldsymbol{X}}$). Nakon toga treba izračunati kvadratno odstupanje svakog merenja (X_i) te promenljive od njene srednje vrednosti. Nakon toga sledi sumiranje kvadrata svih odstupanja promenljive od njene srednje vrednosti u datom uzorku (SS):

$$\mathrm{SS} = \sum_{i=1}^{N} \left(x_i - \bar{x}\right)^2$$

Deljenjem sume kvadrata brojem merenja izračunava se vrednost varijanse:

$$\sigma^2 = \frac{\sum_{i=1}^{N} \left(x_i - \bar{x}\right)^2}{N-1}$$

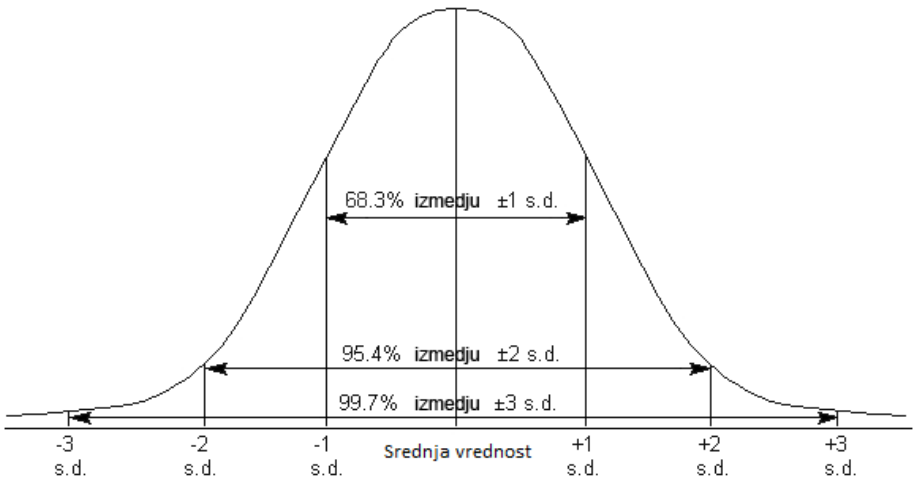
Gde je:N - ukupan broj merenja, $\overline{X}$ - srednja vrednost uzorka za datu promenljivu, X_i – vrednost date promenljive jedinke *i*.

R programski jezik računa varijansu funkcijom *var()*:

```
var(Baza_B_balcanicusT2$Lok1,na.rm=TRUE)
[1] 11.5
var(Baza_B_balcanicusT2$Lok2,na.rm=TRUE)
[1] 692.5
var(Baza_B_balcanicusT2$Lok3)
[1] 6.177778
var(Baza_B_balcanicusT2$Lok4)
[1] 436.1778
```

Varijansa na dobar način predstavlja disperziju (rasipanje) podataka oko srednje vrednosti i koristi se u mnogim univarjintnim statistčkim testovima o čemu će biti reči u narednim polavljima.

Drugi parametar koji prikazuje stepen varijabilnosti težina u datom uzorku potočne mrene i koji je češće u upotrebi u analizama varijabilnosti skupova podataka je **standardna devijacija (SD).** Ona predstavlja vrednost kvadratnog korena varijanse (σ^2). Standardna devijacija nam pomaže da razumemo raspodelu vrednosti u skupu podataka. Za normalnu raspodelu vrednosti, opseg koji obuhvata vrednosti od srednje vrednosti uzorka umanjene za vrednost jedne standardne devijacije do srednje vrednosti uzorka uvećane za vrednost jedne standardne devijacije (±1 SD) uključuje 68% merenja. Opseg od srednje vrednosti uzorka umanjene za dve standardne devijacije do srednje vrednosti uzorka uvećane za dve standardne devijacije ± 2 SD obuhvata 95% merenja (slika 1.2.).



Slika 3.1 Gausova kriva normalne raspodele sa intervalima poverenja

R omogućava integrisanom funkcijom *sd()* jednostavno izračunavanje najčešće korišćenog parametra koji opisuje stepen varijabillnosit:

```
> sd(Baza_B_balcanicusT2$Lok1,na.rm=TRUE)
[1] 3.391165
> sd(Baza_B_balcanicusT2$Lok2,na.rm=TRUE)
[1] 26.31539
> sd(Baza_B_balcanicusT2$Lok3)
[1] 2.485514
> sd(Baza_B_balcanicusT2$Lok4)
[1] 20.88487
```

Još jedan koristan parametar koji prikazuje stepen varijabilnosti podataka je **standardna greška aritmetičke sredine (**$S_{\bar{X}}$**)**. Ona ukazuje na nepreciznost prilikom korišćenja nereprezentativnih uzoraka (uzoraka manjeg obima). Uzorak manjeg obima daće nepreciznije rezultate prilikom uzorkovanja od uzorka većeg obima. Na primer, uzorci potočne mrene od 10 jedinki (Lok 3 i Lok4) vernije će prikazivati stvarnu varijabilnost u datoj populaciji od uzorka koji se sastoji od samo 5 jedinki (Lok1 i Lok2). Uzorci na lokalitetu Lok1 i Lok3 imaju isti opseg vrednosti težina ali različit broj merenih jedinki. Očekuje se da lokalitet Lok3 ima manju standardnu grešku aritmetičke sredine. Formula za izračunavanje standardne greške aritmetičke sredine je:

$$S_{\bar{X}} = \frac{SD}{(n)^{1/2}}$$

Standardna greška aritmetičke sredine je bitna zbog kasnijeg izračunavanja intervala poverenja. Ovaj parametar predstavlja opseg vrednosti u okviru kojih se sa određenom verovatnoćom nalazi prava srednja vrednost populacije. Ukoliko se primeni formula u R-u sa integrisanim funkcijama *sd() sqrt()* i *length()* za izračunavanje standardne devijacije, kvadratnog korena i broja merenja,respektivno dobija se sledeći kod:

```
>sd(Baza_B_balcanicusT2$Lok1,na.rm=TRUE)/sqrt(length(Baza_B
_balcanicusT2$Lok1))
[1] 1.072381
>sd(Baza_B_balcanicusT2$Lok2,na.rm=TRUE)/sqrt(length(Baza_B
_balcanicusT2$Lok2))
[1] 8.321658
```

```
    >                                sd(Baza_B_balcanicusT2$Lok3)/
sqrt(length(Baza_B_balcanicusT2$Lok3))
    [1] 0.7859884
    >                                sd(Baza_B_balcanicusT2$Lok4)/
sqrt(length(Baza_B_balcanicusT2$Lok4))
    [1] 6.604376
```

U R programskom jeziku postoje integrisane funkcije koje istovremeno pružaju veći broj parametara deskriptivne statistke. Na primer funkcija *summary()* izračunava minimalnu vrednost, prvi kvartil, medianu, srednju vrednost, treći kvartil i maksimalnu vrednost za sve varijable u skupu podataka:

```
> summary(Baza_Dunav[,-c(1, 2,4)])
       TL        polna_zrelost
 Min.   : 2.00   Min.   :1.000
 1st Qu.: 6.20   1st Qu.:2.000
 Median :12.00   Median :3.000
 Mean   :12.67   Mean   :2.663
 3rd Qu.:15.85   3rd Qu.:4.000
 Max.   :45.00   Max.   :4.000
```

Ukoliko je potrebno da se opišu varijable u okviru grupa, R koristi funkciju *by().* Na primer moguće je prikazati deskriptivnu statistiku totalne dužine tela i polne zrelosti za svaku vrstu ponaosob:

```
> by(Baza_Dunav[,-c(1, 2,4)], Baza_Dunav$Vrsta, summary)
Baza_Dunav$Vrsta: Abramis brama
       TL        polna_zrelost
 Min.   :12.00   Min.   :3.00
 1st Qu.:13.00   1st Qu.:3.00
 Median :14.10   Median :3.00
```

```
 Mean   :15.47   Mean   :3.25
 3rd Qu.:17.57   3rd Qu.:3.25
 Max.   :20.50   Max.   :4.00
------------------------------------------------------------
------------------
Baza_Dunav$Vrsta: Alburnus alburnus
       TL          polna_zrelost
 Min.   : 5.500   Min.   :2.00
 1st Qu.: 6.250   1st Qu.:2.00
 Median : 7.000   Median :2.00
 Mean   : 7.967   Mean   :2.50
 3rd Qu.: 9.250   3rd Qu.:2.75
 Max.   :12.300   Max.   :4.00
------------------------------------------------------------
------------------
Baza_Dunav$Vrsta: Aspius aspius
       TL          polna_zrelost
 Min.   : 3.000   Min.   :1.000
 1st Qu.: 7.000   1st Qu.:1.000
 Median : 9.200   Median :3.000
 Mean   : 8.478   Mean   :2.333
 3rd Qu.:10.000   3rd Qu.:3.000
 Max.   :12.200   Max.   :3.000
------------------------------------------------------------
------------------
Baza_Dunav$Vrsta: Carassius gibelio
```

```
       TL        polna_zrelost
 Min.   : 4.00   Min.   :1.000
 1st Qu.: 7.40   1st Qu.:2.000
 Median :15.00   Median :3.000
 Mean   :13.44   Mean   :2.836
 3rd Qu.:19.60   3rd Qu.:4.000
 Max.   :24.00   Max.   :4.000
------------------------------------------------------------
------------------
Baza_Dunav$Vrsta: Cyprinus carpio
       TL        polna_zrelost
 Min.   :30.00   Min.   :3
 1st Qu.:30.75   1st Qu.:3
 Median :31.50   Median :3
 Mean   :31.50   Mean   :3
 3rd Qu.:32.25   3rd Qu.:3
 Max.   :33.00   Max.   :3
------------------------------------------------------------
------------------
Baza_Dunav$Vrsta: Esox lucius
       TL        polna_zrelost
 Min.   :10.00   Min.   :2.000
 1st Qu.:19.62   1st Qu.:3.000
 Median :30.00   Median :3.000
 Mean   :26.54   Mean   :3.188
 3rd Qu.:33.50   3rd Qu.:4.000
```

```
 Max.   :45.00   Max.   :4.000
------------------------------------------------------------
------------------
Baza_Dunav$Vrsta: Hypophthalmichthys molitrix
       TL        polna_zrelost
 Min.   :2.000   Min.   :1
 1st Qu.:3.200   1st Qu.:1
 Median :4.200   Median :1
 Mean   :4.284   Mean   :1
 3rd Qu.:5.100   3rd Qu.:1
 Max.   :6.000   Max.   :1
------------------------------------------------------------
------------------
Baza_Dunav$Vrsta: Lepomis gibossus
       TL        polna_zrelost
 Min.   : 4.50   Min.   :1.000
 1st Qu.: 5.50   1st Qu.:1.000
 Median : 7.90   Median :2.500
 Mean   : 8.30   Mean   :2.375
 3rd Qu.:10.62   3rd Qu.:3.250
 Max.   :12.80   Max.   :4.000
------------------------------------------------------------
------------------
Baza_Dunav$Vrsta: Pseudosaspbora parva
       TL    polna_zrelost
 Min.   :4   Min.   :1
```

```
 1st Qu.:4   1st Qu.:1
 Median :4   Median :1
 Mean   :4   Mean   :1
 3rd Qu.:4   3rd Qu.:1
 Max.   :4   Max.   :1
------------------------------------------------------------
------------------
Baza_Dunav$Vrsta: Rutilus rutilus
       TL         polna_zrelost
 Min.   : 4.00   Min.   :1.00
 1st Qu.:11.50   1st Qu.:3.00
 Median :13.50   Median :4.00
 Mean   :12.58   Mean   :3.27
 3rd Qu.:14.60   3rd Qu.:4.00
 Max.   :17.00   Max.   :4.00
------------------------------------------------------------
------------------
Baza_Dunav$Vrsta: Sander lucioperca
       TL    polna_zrelost
 Min.   :6   Min.   :1
 1st Qu.:6   1st Qu.:1
 Median :6   Median :1
 Mean   :6   Mean   :1
 3rd Qu.:6   3rd Qu.:1
 Max.   :6   Max.   :1
```

```
------------------------------------------------------------
------------------
Baza_Dunav$Vrsta: Scardinius erythrophthalmus
       TL        polna_zrelost
 Min.   :10.40   Min.   :3
 1st Qu.:11.80   1st Qu.:3
 Median :12.00   Median :3
 Mean   :12.63   Mean   :3
 3rd Qu.:13.00   3rd Qu.:3
 Max.   :15.50   Max.   :3
```

U okviru paketa *pastecs*, funkcijom *stat.desc()* moguće je izračunati sve prethodno navedene parametre deskriptivne statistke sa dodatkom testa normalnosti o čemu će biti reči kasnije:

```
install.packages("pastecs")
library(pastecs)
> stat.desc(Baza_Dunav[,-c(1, 2,4)], norm = TRUE)
                        TL polna_zrelost
nbr.val      1.870000e+02  1.870000e+02
nbr.null     0.000000e+00  0.000000e+00
nbr.na       0.000000e+00  0.000000e+00
min          2.000000e+00  1.000000e+00
max          4.500000e+01  4.000000e+00
range        4.300000e+01  3.000000e+00
sum          2.368360e+03  4.980000e+02
median       1.200000e+01  3.000000e+00
```

```
mean          1.266503e+01  2.663102e+00
SE.mean       5.647493e-01  8.056771e-02
CI.mean.0.95  1.114137e+00  1.589440e-01
var           5.964210e+01  1.213846e+00
std.dev       7.722830e+00  1.101747e+00
coef.var      6.097760e-01  4.137082e-01
skewness      1.237541e+00 -3.665132e-01
skew.2SE      3.481917e+00 -1.031213e+00
kurtosis      1.895788e+00 -1.202310e+00
kurt.2SE      2.680666e+00 -1.700079e+00
normtest.W    9.037859e-01  8.332498e-01
normtest.p    1.152678e-09  2.316338e-13
```

3.1 Istraživačka analiza univarijantnih skupova podataka u R ambijentu

Obrazac varijabilnosti promenljive u skupu podataka se može najefikasnije sagledati vizuelizacijom raspodele njenih vrednosti. Lista svih izmerenih vrednosti u skupu padataka, kao i učestalost njihovog javljanja može se predstaviti u formi **tabele frekventnosti** (npr. broj uzoraka u kojima se javila potočna mrena određene staorsne kategorije). U okviru paketa *dplyr*, tabela frekventosti se može lako generisati funkcijama *group_by*, *sumarize()* i *arange ()*. U tabeli „Koviljski_Dunavac_ribe.csv“ se nalaze podaci o kondicionom stanju populacije više vrsta riba uzorkovanih na vlažnim staništima duž Dunava. Baza podataka se sastoji od dve varijable koje su kontinuirane (totalna dužina riba (TL u cm) i masa (m u gramima) i četiri kategorijske (jedna ordinalna (polna zrelost) i tri nominalne (lokalitet, vrsta ribe i pol) promenljive. Kako bi primenili istraživačke tehnike na ovaj skup podataka prvo je potrebno da se uveze tabela u *csv* formatu i definiše skup podataka u R-u, u *data.frame* formatu pomoću funkcije *read_csv()*:

```
Baza_Dunav=read_csv("Koviljski_Dunavac_ribe.csv")
Parsed with column specification:
cols(
  Lokalitet = col_double(),
  Vrsta = col_character(),
  TL = col_double(),
  pol = col_character(),
  `polna zrelost` = col_double()
> Baza_Dunav
# A tibble: 187 x 5
   Lokalitet Vrsta              TL pol   `polna zrelost`
       <dbl> <chr>           <dbl> <chr>            <dbl>
 1         1 Carassius gibe~  24   f                    4
 2         1 Carassius gibe~  24   f                    4
 3         1 Carassius gibe~  22   f                    4
 4         1 Carassius gibe~  20   m                    4
 5         1 Carassius gibe~  17.5 m                    4
 6         1 Carassius gibe~  10.2 f                    3
 7         1 Carassius gibe~   7.2 f                    2
 8         1 Carassius gibe~   6   j                    1
 9         1 Rutilus rutilus  13.5 m                    4
10         1 Rutilus rutilus   6   j                    1
# ... with 177 more rows
```

Sledeći kod generiše tabelu frekventosti polne zrelosti potočne mrene:

```
> Baza_Dunav%>%
+   group_by(`polna zrelost`) %>%
+   summarize(freq=n())%>%
+   arrange (desc(freq))
# A tibble: 4 x 2
  `polna zrelost`  freq
            <dbl> <int>
1               3    72
2               4    48
3               1    44
4               2    23
```

Pored tabele frekventnosti moguće je generisati i tabelu relativne frekventnosti (proporciju ukupne frekvence). Ako je ukupna frekventnost u skupu podataka *n*, relativna frekventnost svake klase (npr. jedinke potočne mrene ženskog pola i određene starosne kategorije) se računa kao broj slučajeva konkretne klase podeljen sa *n*. U tom slučaju je suma svih relativnih frekvenci 1.

```
> Baza_Dunav %>%
+   group_by(`polna zrelost`) %>%
+   summarize(freq=n())%>%
+   mutate (rel.freq=freq/sum(freq))
# A tibble: 4 x 3
  `polna zrelost`  freq rel.freq
            <dbl> <int>    <dbl>
1               1    44    0.235
2               2    23    0.123
```

```
3                3    72    0.385
4                4    48    0.257
```

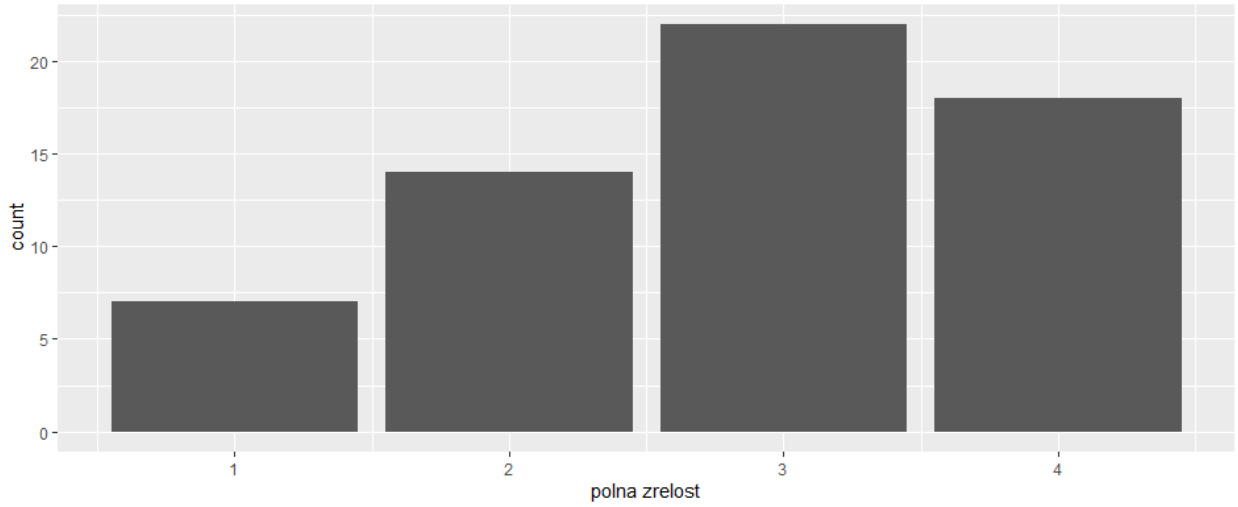
U tabeli frekventnosti se teško mogu doneti zaključci o obrascima raspodele frekvenci, posebno ako se radi o skupovima podataka sa velikim brojem klasa. U tom slučaju se najčešće pribegava vizuelizaciji distribucije frekventnosti. Metoda vizuelizacije raspodele frekventnosti zavisi od tipa podataka. Ukoliko je reč o kategorijskoj promeljivoj, distribucija frekventnosti se efikasno vizualizuje pomoću stubičastih dijagrama (histograma). Za generisanje tog tipa dijagrama, koristi se funkcija *ggplot()* i *geom_bar().* S obzirom da se u bazi podataka nalaze informacije o polnoj zrelosti različitih vrsta riba, prvi korak je generiasti novu bazu sa samo jednom vrstom, konkretno u ovom primeru -babuškom (*Carassius gibelio*):

```
> Baza_Carassius = filter(Baza_Dunav, Vrsta=="Carassius
gibelio")
> Baza_Carassius
# A tibble: 61 x 5
   Lokalitet Vrsta                TL pol   `polna zrelost`
       <dbl> <chr>             <dbl> <chr>           <dbl>
 1         1 Carassius gibelio  24   f                   4
 2         1 Carassius gibelio  24   f                   4
 3         1 Carassius gibelio  22   f                   4
 4         1 Carassius gibelio  20   m                   4
 5         1 Carassius gibelio  17.5 m                   4
 6         1 Carassius gibelio  10.2 f                   3
 7         1 Carassius gibelio   7.2 f                   2
 8         1 Carassius gibelio   6   j                   1
 9         2 Carassius gibelio   7   j                   1
```

```
10         2 Carassius gibelio   7.5 j                    1
# ... with 51 more rows
```

Vizuelni prikaz raspodele polne zrelosti babuške se može dobiti sledećim kodom:

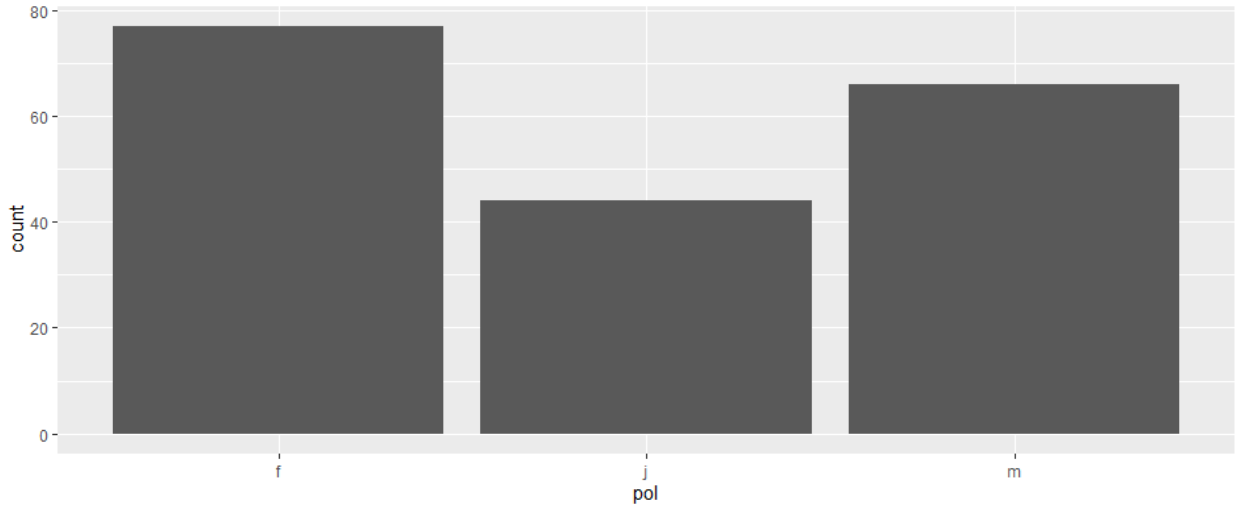
```
ggplot(data=Baza_Carassius)+
   geom_bar(mapping=aes (x=`polna zrelost`))
```

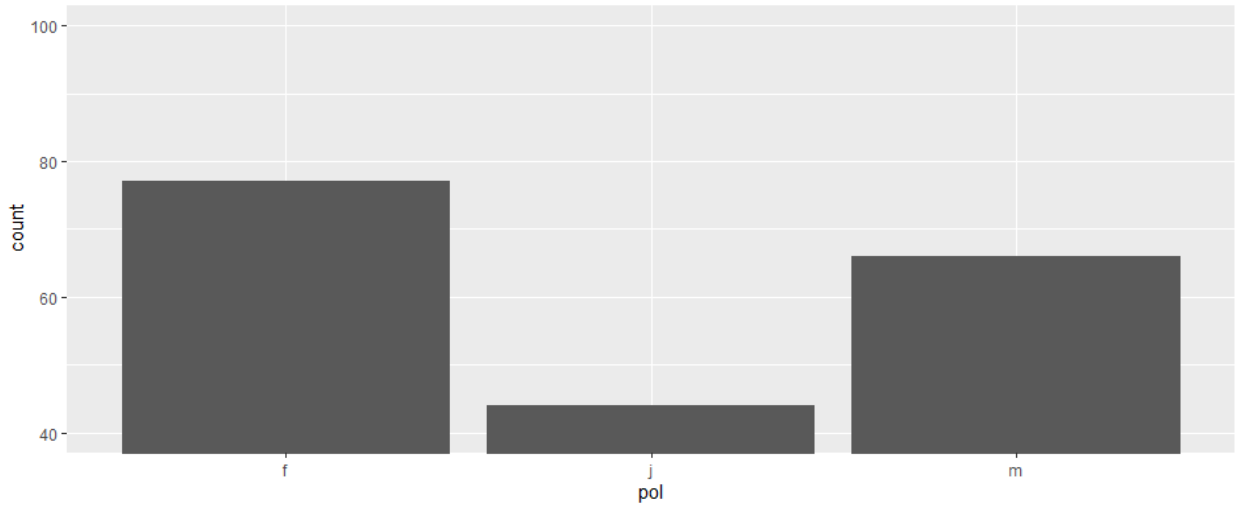


U prvoj liniji prethodnog koda prvi argument *ggplot()* funkcije definiše skup podataka koji će se vizuelizovati pomoću stubičastog dijagrama. Izvršavanjem funkcije *ggplot()*, R crta koordinatni sistem kome se mogu dodavati slojevi. Sledeća funkcija u kodu, *geom_bar()*, dodaje novi sloj postojećem dijagramu u obliku stubova. Agrument *mapping*, uparen sa funkcijom *aes()* definiše promenljivu koja će biti vizuelizovana (u ovom slušaju polna zrelost riba).

Kada su na stubičastom dijagramu vrednosti frekventnosti visoke, a skala na *y* osi počinje od 0, kako bi se bolje uočile razlike između grupa, moguće je zumirati skalu frekventnosti i ograničiti je na određeni opseg. U R-u se takva promena realizuje preko funkcije *coord_cartesian ()*:

```
ggplot(data=Baza_Dunav)+
  geom_bar(mapping=aes (x=pol))+
 coord_cartesian (ylim=c(40,100))
```

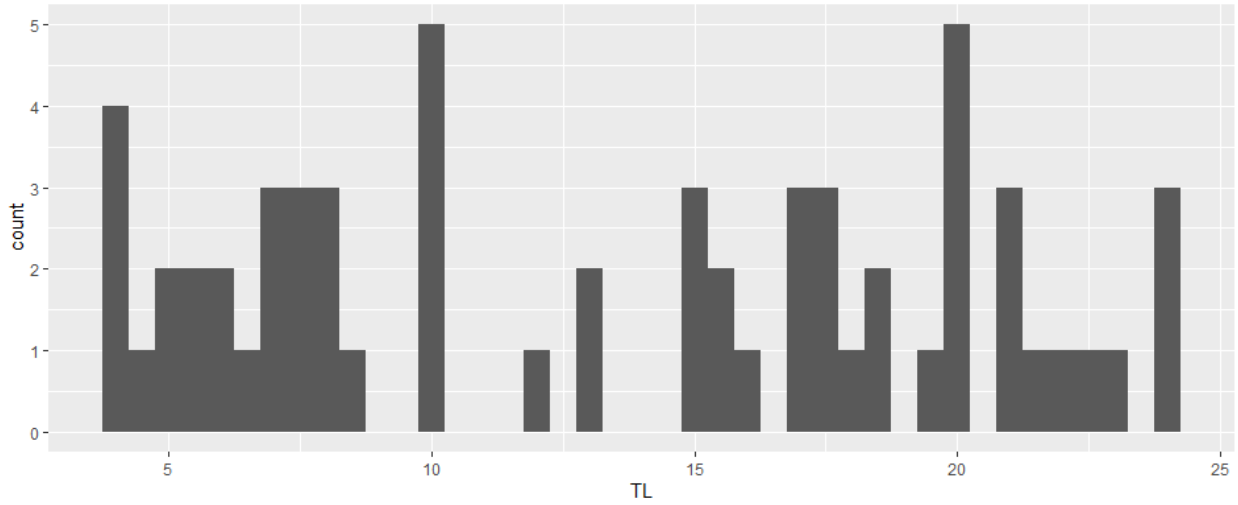




Argument *ylim* je definisao opseg od 40 do 100 slučajeva što je dodatno naglasilo posotjeće razlike između grupa.

Za prikazivanje raspodele frekvetnosti kontinuirane promenljive, koristi se histogram sa definisanim intervalima iste širine (odeljcima) na x-osi. Totalna dužina riba (TL) vrste *Carassius gibelio* predstavlja kontinuiran varijablu:

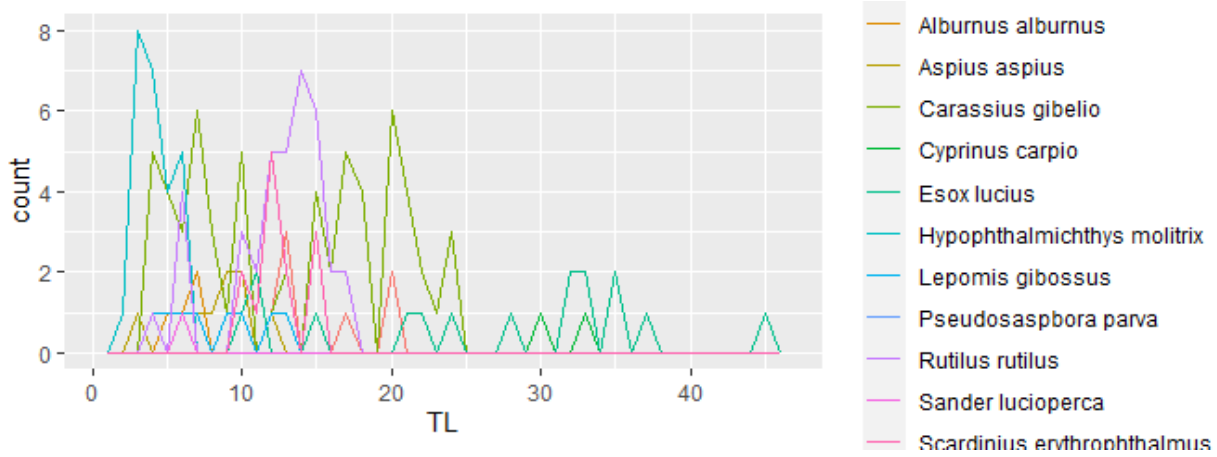
```
ggplot(data=Baza_Carassius) +
  geom_histogram(mapping = aes(x=TL), binwidth=0.5)
```



Argument *binwidth* definiše širinu intervala na x-osi i prikazuje se u jedinicama promenljive *x* (u ovom slučaju u cm). Širina intervala zavisi od varijabilnosti samog skupa te je potrebno ispitati i sagledati različite širine intervala kako neka informacija o obrascu varijabilnost iskupa ne bi ostala skrivena. Prema dobijenom histogramu, najveći broj jedinki u populaciji babuške ima totalnu dužinu od 10cm i 20 cm.

Ukoliko je potrebno istovremeno prikazati raspodelu frekventnosti više promelljivih na istom grafikonu, na primer totalnu dužinu tela (TL) svih prisutnih ribljih vrsta u uzorku, koristi se funkcija *geom_freqpoly ()*:

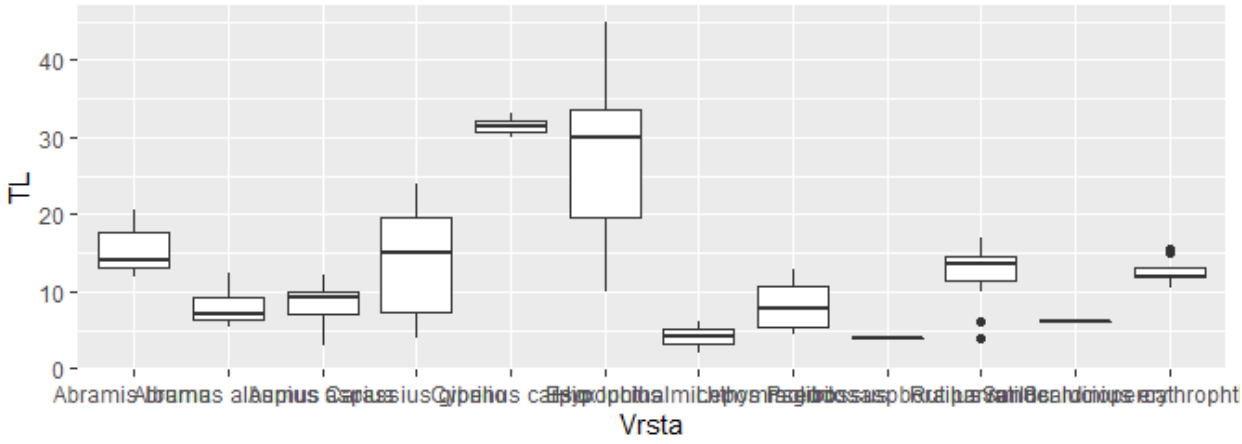
```
ggplot(data=Baza_Dunav, mapping = aes(x=TL, color=Vrsta))+
       geom_freqpoly(binwidth=1.0)
```



Ova funkcija omogučava da se prikaže veza između dve ili više promeljivih i na taj način prikaže obrazac kovarijabilnosti. Meuđutim, pored histograma, mnogo su popularniji **kutijasti dijagrami** ili pravougaono-linijski grafik (eng. box-whisker plot) koji vernije opisuju strukturu podataka koristeći neparametarske mere centralne tendencije (medijana), o kojima će biti reči kasnije. Kutija (pravougaonik) na grafikonu povezuje prvi i treći kvartil skupa rangiranih podoataka, pokrivajući interkvartilni opseg, odnosno rastojanje između 25. i 75. percentrila. Linija u pravouganiku označava medijanu, centralnu vrednost skupa i u kombinaciji sa paralelnim linijama koje čine ivicu kutije pokazuju simetričnost raspodele podataka u odnosnu na medijanu. Pored neparametarskih mera, kutijasti grafikoni mogu da prikazuju i srednju vrednost i standardnu devijaciju. Linije koje polaze od kutije povezuju vrednosti podataka u okviru 1.5 x intevalni opseg.

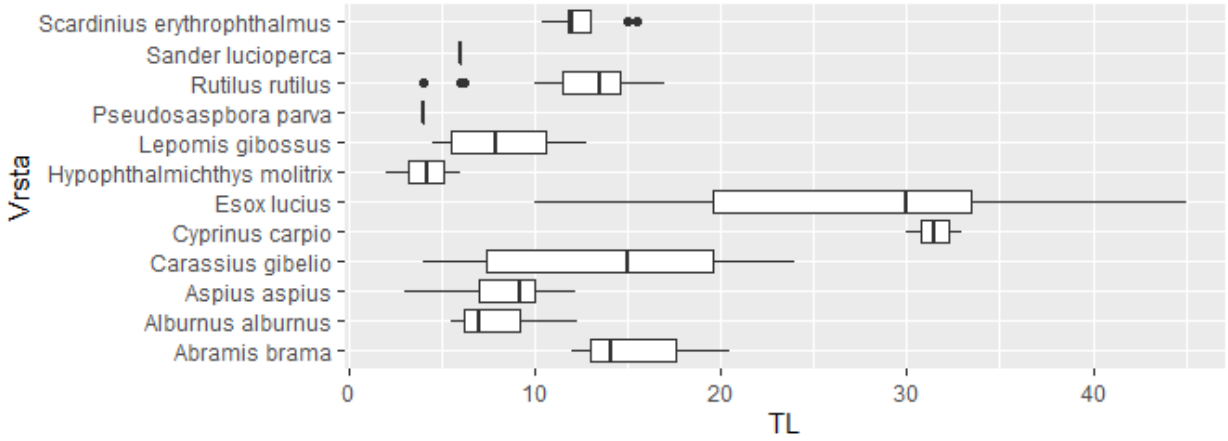
Netipične vrednosti (eng. outlayer) u skupu podataka se često javljaju u ekološkim istraživanjima. Njihovo poreklo može biti dvojako, ili zbog greške istraživača prilikom prikupljanja podatak ili zbog uticaja faktora koji nisu uzeti u obzir tokom studije. Netipične vrednosti ne treba isključivati iz skupa podataka već je potrebno prvo detektovati njihovo prisustvo i ukoliko je dišlo do greške izvršiti korekciju. Ako se ne radi o grešci, netipične vrednosti bi trebalo tretirati adekvatnim statističkim metodama koje bi umanjile njihov uticaj na rezultate analiza. Jedan od načina da se lako detektuje prisustvo netipičnih vrednosti u podacima je i pomoću kutijastih dijagrama. Svi slučajevi (opservacije) čije se vrednosti u vidu tačaka jave van linija (1.5x intervalni opsega) se smatraju netipičnim vrednostima. U R-u se kutijasti dijagrami crtaju pomoću funkcije *geom_boxplot()* u okviru funkcije *ggplot():*

```
ggplot(data=Baza_Dunav, mapping = aes(x=Vrsta, y=TL))+
  geom_boxplot()
```



S obzirom da kategorijska promenljiva *Vrsta* ima veliki broj kategorija sa dugačkim imenima vrsta, grafikon bi bio mnogo pregledniji ukoliko se pomoću funkcije *coord_flip()* zarotira za 90 stepeni:

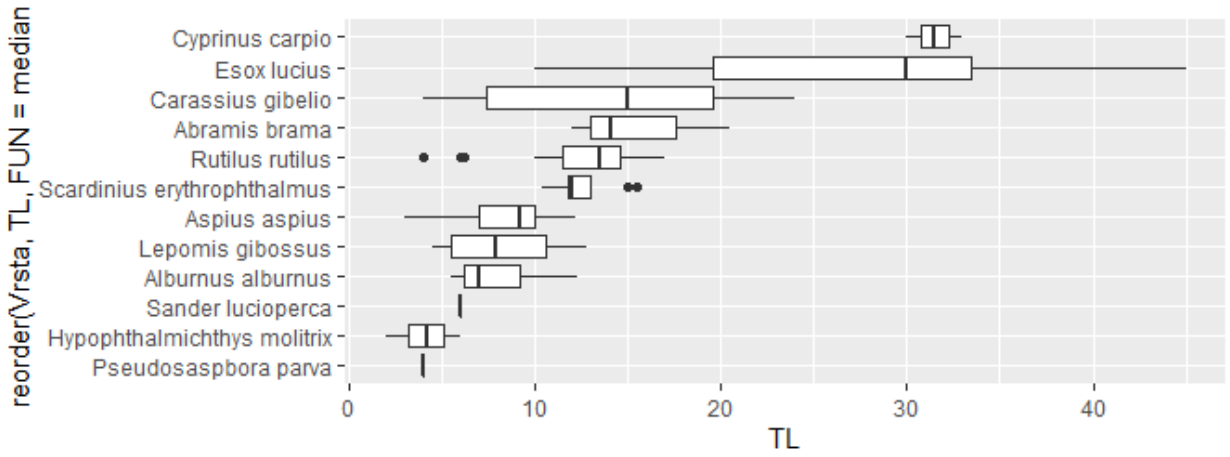
```
ggplot(data=Baza_Dunav, mapping = aes(x=Vrsta, y=TL))+
  geom_boxplot()+
  coord_flip()
```



Kutijasti grafikon prikazuje variranje totalne dužine tela kod svih vrsta u skupu podataka. Sada je moguće pratiti kako dužina tela varira kod različitih vrsta riba. S obzirom da je istovremeno

prikazan veliki broj vrsta na grafikonu, kako bi se poboljšala preglednost moguće je reorganizovati raspored kutija u rastućem nizu na osnovu varijable TL:

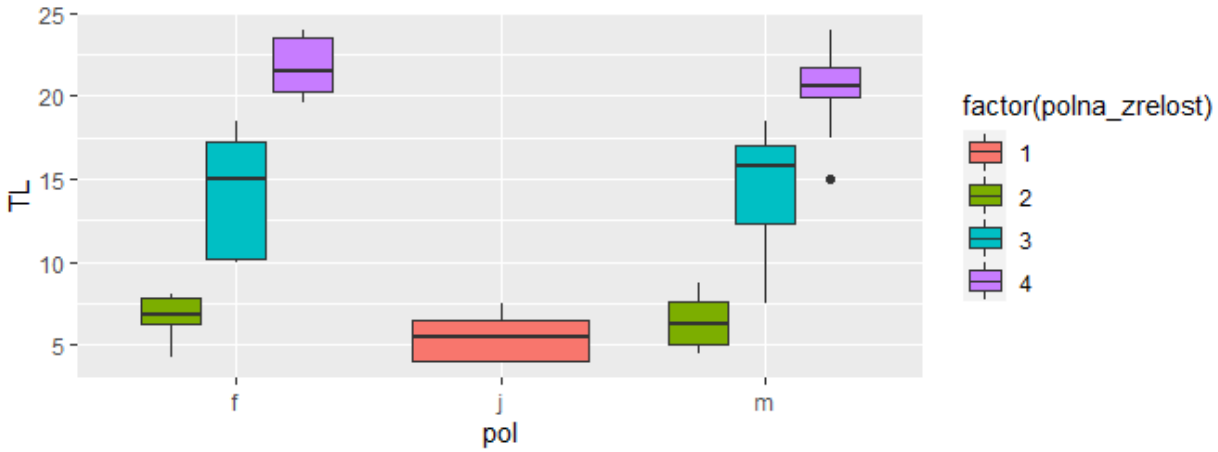
```
ggplot(data=Baza_Dunav, mapping = aes(x=reorder(Vrsta, TL, FUN=median), y=TL))+
  geom_boxplot()
```



Sada se jasno uočava da vrste *C. carpio* i *E. lucius* imaju najduže jedinke dok vrste *P.parva* i *H.molitrix* imaju najkraće jedinke. Netipične dužine tela riba su zabeležene kod vrsta *S. erzthrophthalmus* (netipično duže) i *R. rutilus* (netipično kraće).

Funkcija *geom_boxplot()* omogućava istovremeno uvođenje i drugog atributa, kao što je polna zrelost, pomoću boja:

```
ggplot(data=Baza_Carassius, mapping = aes(x=pol, y=TL, fill = factor(polna_zrelost)))+
  geom_boxplot()
```



Iz grafikona se uočava očekivani obrazac, da totalna dužina tela raste kako jedinke postaju polno zrelije te se polna zrelost može ustanoviti i na osnovu dužine ribe. Ova pojava nije pol- specifična i karakteristična je i za mužijake i za ženke. Iz grafikona se vidi i da je u jednom slučaju jedinka polne zrelosti 4 imala neočekivano malu totalnu dužinu tela.

3.2 Istraživačka analiza multivarijantnih skupova podataka u R ambijentu

Analiza bioloških podataka u hidrobiološkim istraživanjima se najčešće svodi na analizu strukture hidrobiocenoza (zajednica). Takvi multivarijantni podaci, gde svaka vrsta u zajednici predstavlja promenljivu, zahtevaju poseban statistički dizajn. Prvi korak, koji se odnosi na israživačku analizu podataka je određivanje dimenzionalnosti skupa i njegova vizuelizacija duž prostornog (mreža uzorkovanih lokaliteta) ili vremenskog (različite sezone) gradijenta. Najednostavniji način da se opiše zajednica je određivanjem broja vrsta i brojnosti jedinki (abundantnosti), kao i definisanja obrasca varijabilnosti tih promenljivih duž sredinskih gradijenata. Baza podataka „Baza_Nisava“ sadrži informacije o brojnosti vrsta makrobeskičmenjaka i riba na 20 lokaliteta. Broj vrsta na svakom lokalitetu se može izračunati funkcijom *apply()* koja selektuje sve slučajeve gde je vrednost brojnosti jedinki veća od nule i sumira ih po kolonama:

```
> Baza_Nisava=read_csv("Tabela.csv")

> has_rownames(Baza_Nisava)

[1] FALSE

> Baza_Nisava_MZB=column_to_rownames(select(Baza_Nisava, 1:100),
var = "Lok")
```

```
>         Baza_Nisava_RIBE=column_to_rownames(select(Baza_Nisava,
1,101:115), var = "Lok")

> Baza_Nisava_MZB_BV <- apply(Baza_Nisava_MZB > 0, 1, sum)

> Baza_Nisava_MZB_BV

ND_sum  JU_sum  JG_sum  VG_sum  TG_sum  VU_sum  TU_sum  NP_sum  NS_sum
NN_sum ND_aut

    21      26      19      27      21      11      18      26      17
11     10

JU_aut JG_aut VG_aut TG_aut VU_aut TU_aut NP_aut NS_aut NN_aut

    14      12      17      12      10      10      12      14      13

> Baza_Nisava_RIBE_BV <- apply(Baza_Nisava_RIBE > 0, 1, sum)

> Baza_Nisava_RIBE_BV

ND_sum  JU_sum  JG_sum  VG_sum  TG_sum  VU_sum  TU_sum  NP_sum  NS_sum
NN_sum ND_aut

     6       7       7       7       3       2       5       6       5
7      6

JU_aut JG_aut VG_aut TG_aut VU_aut TU_aut NP_aut NS_aut NN_aut

     7       7       7       3       2       5       6       5       7
```

Prvi argument funkcije *apply()* određuje input podatke dok drugi argument, *MARGIN*, definiše da li funkcija procesuira podatke duž kolona (1) ili duž redova (2). Poslednji argumentum (*FUN=sum)* sumira broj slučajeva u koloni koji su bili T*RUE* (*x> 0*).

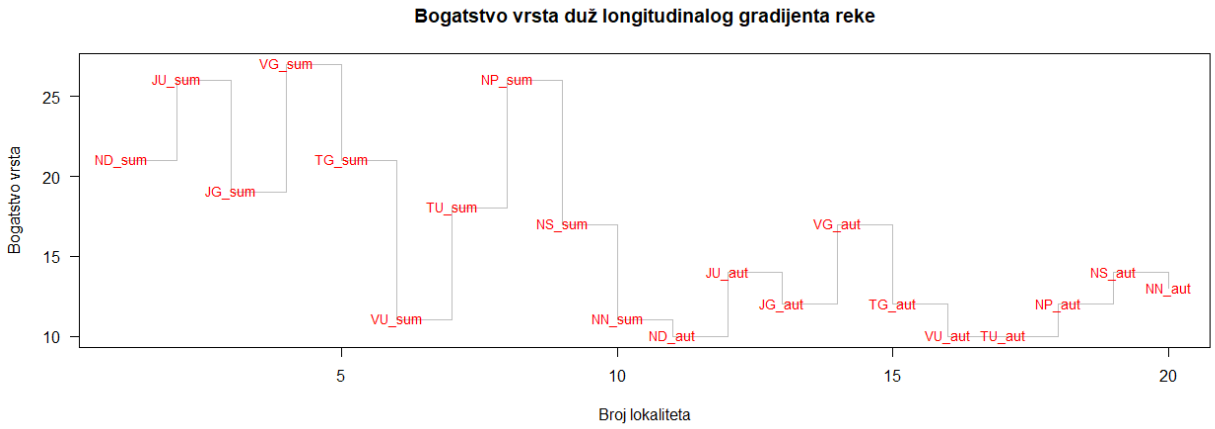
Kako bi distribucija bogatstva vrsta po lokalitetima bila preglednija, funckija *sort()* će poređati lokalitete u rastući niz u odnosu na broj vrsta:

```
> sort(Baza_Nisava_MZB_BV)

ND_aut VU_aut TU_aut VU_sum NN_sum JG_aut TG_aut NP_aut NN_aut
JU_aut NS_aut

    10      10      10      11      11      12      12      12      13
14     14
```

```
NS_sum VG_aut TU_sum JG_sum ND_sum TG_sum JU_sum NP_sum VG_sum
    17     17     18     19     21     21     26     26     27
> sort(Baza_Nisava_RIBE_BV)
VU_sum VU_aut TG_sum TG_aut TU_sum NS_sum TU_aut NS_aut ND_sum
NP_sum ND_aut
     2      2      3      3      5      5      5      5      6
6      6
NP_aut JU_sum JG_sum VG_sum NN_sum JU_aut JG_aut VG_aut NN_aut
     6      7      7      7      7      7      7      7      7
```

Dobijeni rezultat je moguće vizualizovati funkcijom *plot()* koja se koristi u R-u za izradu dijagrama:

```
> plot(Baza_Nisava_MZB_BV,type = "s",
+      las = 1,
+      col = "gray",
+      main = "Bogatstvo vrsta duž longitudinalog gradijenta
reke",
+      xlab = "Broj lokaliteta",
+      ylab = "Bogatstvo vrsta"
+ )
> text(Baza_Nisava_MZB_BV, row.names(Baza_Nisava_MZB),  cex =
.9, col = "red")
```

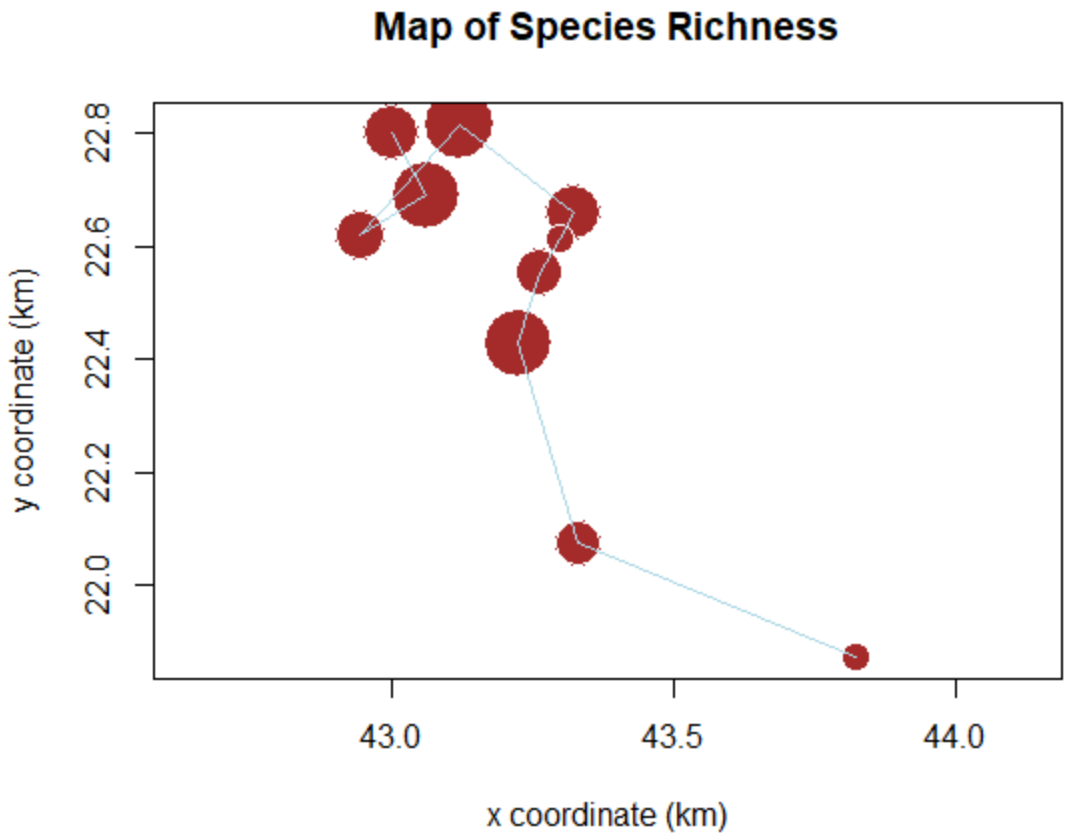


U okviru funkcije *plot()* prvi argument (*x*) definiše coordinate dijagrama dok argument *type* definiše tip dijagrama koji je u ovom slučaju stepeničast ("s"). Iz dobijenog dijagama se uočava pravilnost promene bogatstva vrsta, kako duž prostornog tako i između različitih sezona. Broj vrsta makrobeskičmenjaka je uočljivo već tokom letnje kampanje uzorkovanja, dok se tokom jeseni taj broj smanjuje na svim lokalitetima. Što se tiče prostornog obrasca, lokaliteti JU, VG imaju najveći broj vrsta u obe kampanje, dok je najmanji diverzitet vrsta zabeležen na lokalitetu VU.

Bogatstvo vrsta je mogće predstaviti i preko geografskih koordinata lokaliteta, koristeći dijagram sa mehurićima:

```
spa=read_csv("KoordinateNisava.csv")
plot(spa,
     asp = 1,
     main = "Map of Species Richness",
     pch = 21,
     col = "white",
     bg = "brown",
     cex = 5 * Baza_Nisava_MZB_BV / max(Baza_Nisava_MZB_BV),
     xlab = "x coordinate (km)",
     ylab = "y coordinate (km)"
```

```
)
lines(spa, col = "light blue"))
lines(spa, col = "light blue")
```



4. Osnova statističkog testiranja hipoteze

U osnovi svakog istraživanja je da se pomoću analize prikupljenih podataka izmeri uticaj nezavisne promenljive na zavisnu promenljivu. Po definiciji, **nezavisna promenljiva** predstavlja faktor koji se menja ili je kontrolisana u eksperimentu i čija promena direktno uzrokuje promenu **zavisne promenljive**. Na primer, različiti sektori reke sa različitim kvalitetom staništa predstavljaju nezavisnu promneljivu koja može direktno uticati na kondiciono stanje populacije riba koja naseljava ta staništa i koja je u tom slučaju zavisna promenljiva. Da bi se ustanovio uticaj nezavisne na zavisnu promenljivu, u ekološkim istraživanjima vrlo često se utvrđuje da li postoji razlika između dva skupa podataka koji su pod različitim uticajem nezavisne promenljive. Ako se opisuje kondiciono stanje riba na akvatičnim staništima različitog kvaliteta, potrebno je ustanoviti da li se totalna dužina tela i masa razlikuje između lokaliteta (sektora reke različitog kvaliteta vode). Da bi se odgovorilo na to pitanje, neophodno je definisati termin razlika. Ukoliko se svaki skup podataka sastoji iz više slučajeva (opservacija), najpraktičniji način je porediti njihove parametre centralne tendencije, aritmetičke sredine. Međutim, kako bi se ustanovila razlika, nije dovoljno da se samo srednje vrednosti poređenih skupova razlikuju, jer to razlika može biti slučajna, odnosno posledica prirodne varijabilnosti, a ne uticaja testiranih faktora (u ovom slučaju kvaliteta staišta). Na primer, Fultonov indeks ili kondicioni factor (K), koji izražava masu ribe u kubiku njene dužine i indikuje kondiciono stanje određene populacije riba, bi trebalo da se razlikuje među populacijama koje žive na staništima različitog kvaliteta. Međutim svaki skup podataka pokazuje određeni nivo varijabilnosti gde se merenja u okviru uzorka rasipaju oko srednje vrednosti. Moguće je da jedinke veoma dobrog opšteg stanja sa lokaliteta niskog kvaliteta staništa i jedinke veoma lošeg opšteg stanja sa lokaliteta visokog kvaliteta imaju sličan Fultonov indeks i predstavljaju deo skupova podataka koji se međusobno preklapaju. Intenzitet preklapanja slučajeva između dva seta podataka utiče na njihove razlike. Zbog toga je neophodno utvrditi postojanje značajne razlike između dve srednje vrednosti. Testiranje značajne razlike se sprovodi korišćenjem **statističkih testova**.

Deskriptivna statistika kondicionog faktora populacije riba koja podrazumeva opisivanje varijabilnosti unutar populacije se jedino može predstaviti na nivou uzorka jer je nemoguće izvršiti merenje svake jedinke u okviru populacije datog staništa. Zbog toga se aritmetička sredina parametra *K* riblje populacije može predstaviti jedino preko aritmetičke sredine uzorka. To je upravo glavni cilj statistčke analize da se na osnovu uzorka pouzdano donose zaključci o čitavoj

populaciji. Ako je varijabilnost ispitivane osobine mala, a veličina uzorka velika, onda srednja vrednost uzorka sa velikom preciznošću procenjuje srednju vrednost date osobine u populaciji.

Statističke metode koje testiraju razliku između dva ili više skupova podataka koriste **nultu hipotezu o aritmetičkoj sredini populacije (H_0).** Nulta hipoteza primenjuje koncept nepostojanja razlike, i u slučaju kondicionag stanja karaša *(Carassius carassius)* tvrdi da nema razlike u srednjoj vrednosti parametra *K* između testiranih populacija riba **(H_0:μ=0)**. Međutim da bi se pouzdano odredile verovatnoće odstupanja srednje vrednosti uzoraka od srednje vrednosti populacija i na taj način uspešno koristila metoda uzorka, neophodno je da testirane promenljive zadovolje preduslove u pogledu njihove raspodele frevenci, Oslanjajući se na poznate proporcije normalne raspodele, statistički testovi proračunavaju verovatnoću da uzorak pouzdano oslikava populaciju opisanu nultom hipotezom.

4.1 Normalna raspodela

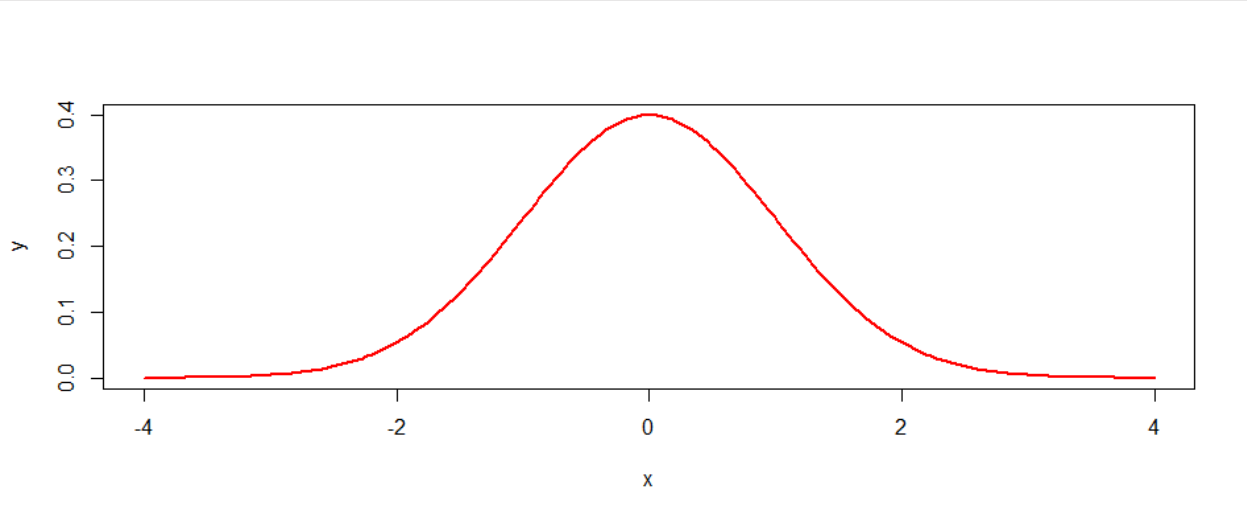
Varijable koje se mere u prirodi, pogotovu kada se radi o velikom skupu podataka obično pokazuju normalnu distribuciju frekvenci, što znači da je veći broj vrednosti skupa skoncentrisan oko aritmetičke sredine dok broj istih progresivno pada ka maksimalnoj i minimalnoj vrednosti. Ukoliko se grafički prikaže, distribucija frekvenci normalnog skupa podataka ima oblik zvona, a kriva takvog oblika se još naziva i *Gausova kriva*. Normalna distribucija je simetrična i poznavajući njene parameter poput srednje vrednosti i standardne devijacije, moguće je proceniti proporcije merenja u svakoj takčki distribucije. Za te potrebe je definisana "z transformacija", čiji prosek i varijansa su jednaki nuli, odnosno jedinici ($\mu=0$ I $\sigma^2=1$):

$$Z = \frac{Xi - \mu}{\sigma}$$

gde je μ aritmetička sredina, a σ standardna devijacija populacije. Inace *Z* transformacija čini osnovu statistčkog testiranja srednjih vrednosti populacija i naziva se još i **standardna normalna raspodela.** U R-u je moguće konstruisati krivu standardne nomralne raspodele funkcijom *dnorm():*

```
x=seq(-4,4,length=200)
> y=dnorm(x,mean=0,sd=1)
```

```
>plot(x,y,type="l",lwd=2,col="red")
```



Prva linija koda sa funkcijom *seq()* generiše vektor *x* od 200 vrednosti koje su jednako udaljene na skali od -4 do 4. Druga linija koda sa funkcijom *dnorm ()* koristi relaciju koja definiše krivu standardne normalne raspodele kada je $\mu = 0$ i $\sigma = 1$.

U **tabeli S4.1** je moguće izračunati proporciju normalne distribucije za bilo koju vrednost Z u skupu. Proporciju, prikazanu veravatnoćom je moguće jednostavno izračunati funkcijom *pnorm()*. S obzirom da je kriva standardne normalne raspodele simetrična, sa obe strane aritmetičke sredine se nalazi po 50% vrednosti, odnosno verovatnoća da se broj nađe sa desne ili leve strane aritmetičke sredine je 0.5 (P(Z>0)=0.5).

```
> pnorm(0, mean=0, sd=1)
[1] 0.5
```

Funkcijom *pnorm()* je moguće izračunati verovatnoću dobijanja vrednosti koja je jednaka ili manja od zadatog broja. Na primer, skup podataka vrste *Carassius carassisus* je sačinjen od 61 merenja totalne dužine riba- Funkijom *stat.desc()* je moguće dobiti prvu sliku o strukturu analiziranog skupa podataka:

```
Baza_Carassius
stat.desc(Baza_Carassius)
nbr.val       61.0000000
nbr.null       0.0000000
```

```
nbr.na          0.0000000
min             4.0000000
max            24.0000000
range          20.0000000
sum           819.7000000
median         15.0000000
mean           13.4377049
SE.mean         0.8283968
CI.mean.0.95    1.6570404
var            41.8607213
std.dev         6.4699862
coef.var        0.4814800
pnorm(20, mean=14.43, sd=6.46)
[1] 0.8057193
```

Funkcija *pnorm()* je izračunala sa zadatim vrednostima aritmetičke sredine i standardne devijacije kao i pod pretpostavkom da skup podataka ima normalnu raspodelu, da je 80% podataka (TL riba) ima vrednost jednaku ili manju od 20cm, odnosno da postoji verovatnoća od 0.8 da se u skupu nađe vrednos manja od 20cm.

Istom funkcijom je moguce izračunati verovatnoću raspodele date vrednosti u okviru jedne, dve ili tri standardne devijacije od aritmetičke sredine koja je jednaka nuli:

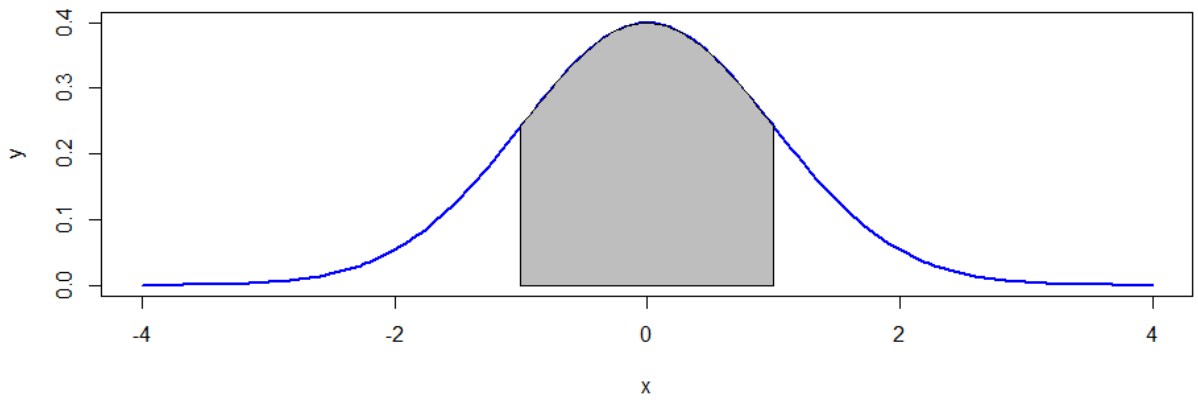
```
> pnorm(1,mean=0,sd=1)-pnorm(-1,mean=0,sd=1)
[1] 0.6826895
pnorm(2,mean=0,sd=1)-pnorm(-2,mean=0,sd=1)
[1] 0.9544997
```

```
> pnorm(3,mean=0,sd=1)-pnorm(-3,mean=0,sd=1)
[1] 0.9973002
```

Prethodni kod otkriva verovatnoću od 68.2%, 95.4% i 99.7% da vrednost iz standardnog normalnog skupa slučajno upadne u opseg između jedne, dve i tri standardne devijacije (SD) respektivno.

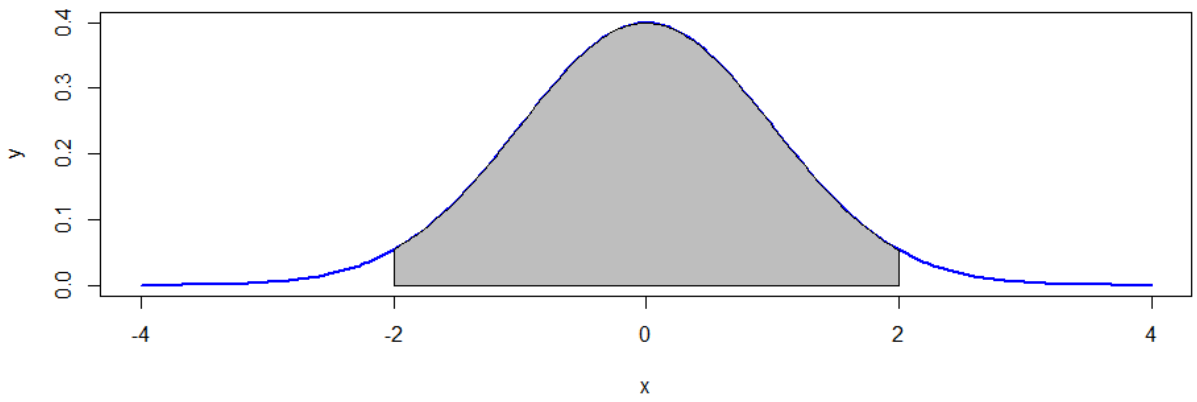
Dobijene verovatnće je moguće vizualizovati funkcijom *plot()* i *polygon ()*

```
x=seq(-4,4,length=200)
y=dnorm(x)
plot(x,y,type="l", lwd=2, col="blue")
x=seq(-1,1,length=100)
y=dnorm(x)
polygon(c(-1,x,1),c(0,y,0),col="gray")
```

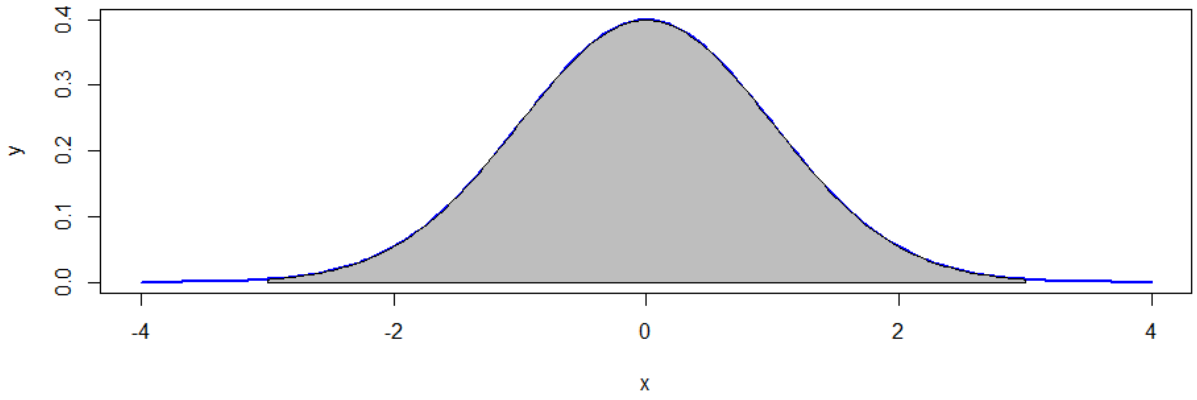



```
x=seq(-4,4,length=200)
y=dnorm(x)
plot(x,y,type="l", lwd=2, col="blue")
x=seq(-2,2,length=200)
y=dnorm(x)
```

```
polygon(c(-2,x,2),c(0,y,0),col="gray")
```




```
x=seq(-4,4,length=200)
y=dnorm(x)
plot(x,y,type="l",lwd=2,col="blue")
x=seq(-3,3,length=200)
y=dnorm(x)
polygon(c(-3,x,3),c(0,y,0),col="gray")
```



4.2 Populacija i slučajni uzorak

U ekologiji, populacija predstavlja skup jedinki iste vrste koji ulaze u različite odnose, pre svega reprodukcije, dele zajednički genofond i nalaze se pod istim ekološkim uslovima. U ekološkim istraživanjima se najčešće statistički obrađuju podaci proistekli iz merenja ili procene ograničenog broja svojstava neke populacije. Na primer, može se posmatrati dužina karapaksa šumske kornjače u Kunovici (Srbija) ili težina vrste *Carrasius carrasius* u Nišavi, ali bi to podrazumevalo veliki broj merenja s obzirom na veliki broj jedinki šumske kornjače u šumi ili sve jedinke ribe u reci. U najvećem broju slučajeva populacija je jako velika, te bi stoga merenje koje bi uključivalo čitavu populaciju bi bilo neizvodljivo. Nemoguće je izračunati dužinu karapaksa sivh kornjača u šumi ali je moguće proceniti dužinu karapaksa na osnovu dela populacije, uzorka koji broji daleko manji broj jedinki od populacije.

Ekolozi nekada formiraju uzorak na osnovu populacije koja ne postoji (u ekotoksikološkim studijama). Ako se u laboratoriji testira uticaj toksičnog agensa na 20 larvi hironomida, to znači da populacija predstavlja grupu hironomida koja se gaji istom hranom u kontrolisanim uslovima. Takve populacije u prirodi ne postoje i nazivaju se imaginarnim ili hipotetičkim.

Uzorci se mogu prikupljati različitim metodama u zavisnosti od cilja istraživanja i odgovarajućeg dizajna uzorkovanja. Međutim kako bi uzorak bio validan i omogućio donošenje zaključaka o populaciji na osnovu njega, statističke tehnike podrazumevaju da je uzorak formiran na osnovu prikupljanja merenja iz populacije po principu slučajnosti. To znači da svako merenje u populaciji ima pođednaku šansu da se nađe u uzorku ali i da selekcija bilo kog člana populacije ne sme da utiče na slekeciju ostalih članova, čime se formira slučajan uzorak.

4.3 Statističko testiranje i verovanoća

Statističko testiranje nulte hipoteze (na primeru *C. carassius* **H_0:μ=0),** uključuje izračunavanje srednje vrednosti slučajnog uzorka (*X*) koji potiče iz populaije. S obzirom da se radi samo o proceni ali ne i tačnoj srednjoj vrednosti populacije, postavlja se pitanje kolika je verovatnoća da uzorak opisan srednjom verdnošću i varijansom potiče iz populacije koja je opisana nultom hipotezom **(H_0:μ=0).** Na primeru praćenja kondicionog stanja populacije karaša (*Carassiuss carassius)* može se sprovesti testiranje hipoteze. Uzorci su prikupljeni pre i posle narušavanja staništa. Promenljiva *X* predstavlja promenu kondicionog faktora godinu dana nakon narušavanja staništa. Pozitivne i negativne vrednosti ukazuju na porast, odnosno pad kondicionog stanja riba. Nulta hipotzea pretpostavlja da narušavanje staništa nije uticalo na promenu kondicionog stanja populacije riba dok alternativna hipoteza (HA) pretpostavlja da je došlo do promene kondicionog staja (HA:μ≠0). Srednja vrednost uzorka (*X*) iznosi 1.29. Iako je varijansa populacije nepoznata u svrhu demonstracije primera σ^2 populacije iznosi 13.4621. Odatle sledi da je standardna greska srednje vrednosti iznosi:

$$\sigma_{\overline{X}} = \sqrt{\frac{\sigma^2}{n}} = \sqrt{\frac{13.4621kg^2}{17}} = \sqrt{0.7919kg^2} = 0.89kg$$

Statistička analiza podataka u ekologiji se bazira na donošenju zaključaka o populaciji na onsovu analize njenih uzoraka. Tačnije statističkom analizom se na osnovu srednje vrednosti uzoraka (*X*) donosi zaključak o jednoj ili više srednjih vrednosti populacije (μ).

Za izračunavanje verovatnoca distribucije mogućih srednjih vrednosti može se koristiti *Z* transformacija (**standardna normalna raspodela**) koja ce u ovom slučaju testiranja hipoteze predstavljati test statistiku. Prema formuli za izračunavanje *Z* vrednosti sledi da je:

$$Z=\frac{\overline{X}-\mu}{\sigma_{\overline{X}}} = \frac{1.29kg-0}{0.89kg}=1.45$$

Na osnovu izračunate *Z* vrednosi u **tabeli S4.1** je moguće proceniti verovatnoću distribucije srednjih vrednosti kada je Z>= 1.45 i ona iznosi 0.0735. S obzirom da nulta hipoteza pretpostavlja da nema odstupanja od srednje vrednosti u oba smera od 0 i da je normalna distribucija simetrična, verovatnoća da je Z<1.45 je takođe 0.0735. Iz toga sledi:

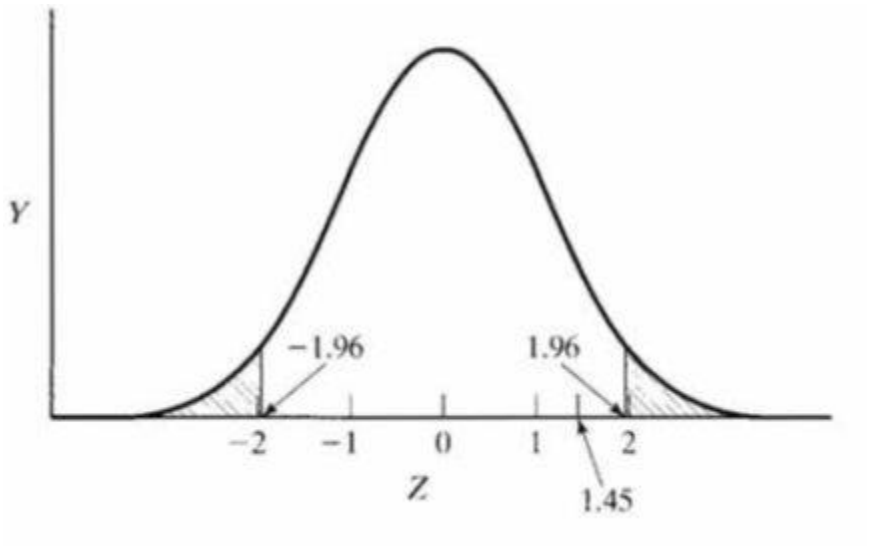
$$P\ (\overline{X} \geq 1.29\text{kg ili } \overline{X} \leq -1.29\text{kg}) =$$

$$P(Z \geq 1.45 \text{ ili } Z \leq -1.45) =$$

$$0.0735 + 0.0735 = 0.1470$$

4.4 Statistčka greška u testiranju hipoteze.

Poželjno je imati objektivan kriterijum u donošenju zaključka o nultoj hipotezi u statističkom testu. U nekim slučajevima je srednja vrednost slučajnog uzorka u velikoj meri udaljena od srednje vrednosti populacija iako je nulta hipoteza tačna. Međutim takva pojava je malo verovatna I što je veća vrednost *Z*, manja je verovatnoća da uzorak potiče iz populacije koja je opisana nultom hipotezom. Zbog toga se postavlja pitanje **koliko mala verovatnoca moze biti, odnosno koliko visoka vrednost *Z* da bi se zakljucilo da nulta hipoteya nije tačna**. Verovatnoća koja se koristi kao kriterijum za odbijanje nulte hipoteze se naziva **nivo značajnosti** i obično je označen malim grčkim slovom alfa (α). Najčešći nivo značjanosti koji se primenjuje u statističkom testiranju hipoteza je α=0.05. *Z* vrednost koja odgovara ovom nivou značajnosti se naziva **kritična vrednost** i na osnovu Tabele S4.1 iznosi P(Z>1.96)=0.025 i P(Z<1.96)=0.025. Vrednost *Z* se moze jos označiti i kao $Z_{0.025(1)}$ =1.96 i $Z_{0.05(2)}$=1.96 gde vrednost u zagradi indikuje da li je reč o jednostranoj ili dvostranoj normalnoj distribuciji. Prema ovome sve vrednosti *Z* statistike veće od 1.96 ili manje od -1.96 se nalaze u "regionu odbijanja" (Slika 4.1, zasenčeni deo *x* ose) su razlog za odbijanje nulte hipoteze. Na primeru *Carassius* vrednost *Z* je iznosila 0.1470 sto je ispod kritične vrednosti i zbog toga se nulta hipoteza ne odbacuje.



Slika 4.1 Standardna normalna raspodela i region odbijanja kada je nivo značajnosti $\alpha<0.05$

Kada nulta hipozea nije odbijena, neki istraživači često koriste reč prihvaćena iako je potrebno da budu obazrivi prilikom statističkog testiranja zbog moguće greške. Nekada se dešava da se hipoteza odbija iako je tačna sto dovodi do greške prilikom procene populacije na osnovu uzorka. Frekventnost pravljenja takve greške je definisana parametrom α (*alfa).* A greska se naziva greska I tipa ili alfa greska. S druge strane statistički test će previdеti da nulta hipoteza nije tačna i neće je odbaciti. Takva greška se naziva greška drugog tipa i verovanoća pravljenja iste je definisana parametrom beta Koristeći parametar b, moguće je definisati snagu statistčkog testa koja je definisana kao 1-b, verovatnoća ispravnog odbijanja nulte hpoteze kada je ona netačna

Tabela 4.1 Greške prilikom statističkog testranja hipoteze

	Ako je H0 tačna	Ako je Ho netačna
H0 se odbacuje	Greska prvog tipa	Nema greske
H0 se ne odbacuje	Nema greske	Greska II tipa

4.5 Testiranje normalne raspodele

Imajući u vidu da se statističko testiranje hipoteze oslanja na normalnu raspodelu promenljive, vrlo je važno pre primene bilo kog testa univarijantne statistike da se ustanovi da li distribucija frekvenci testirane promenljive ima normalnu raspodelu.

Statističko testranje normalne raspodele

Dve osnovne karakteristike distribucije vrednosti skupa su **simetrija** (odnosno asimterija (eng. *skewness*)) i **spljoštenost** (eng. *kurtosis*). Simetrična distribucija ima identičan broj vrednosti distribuiran sa leve i desne strane srednje vrednosti i medijane. U slučaju simetričnih distribucija, medijana i aritmetička sredina imaju istu vrednost. Međutim, simetrija nije dovoljan parametar za testiranje normalnosti jer postoje promenljive koje su približno simetrične ali nemaju normalnu raspodelu. Spljoštenost je komplementaran parametar u odnosu na simetriju koji opisuje oblik distribucije u odnosu na oblik normalne distribucije. Tačnije, spljištenost ukazuje u kojoj meri je pik na poligonu distribucije frekvenci šiljatiji ili spljošteniji u odnosu na normalnu. Ukoliko je mnogo više podataka u odnosu na normalnu raspodelu distribuirano oko srednje vrednosti, u okviru jedne standardne devijacije (X±1SD) i nalazi se u „ramenima“ distribucije tada je reč o *platykurtic* distribuciji. S druge strane promenljiva može imati manje vrednosti u ramenima od normalne distribucije dok je veći broj vrednosti distribuiran ili oko srednje vrednosti ili u repovima distribucije kada je reč o *leptokurtic.* Parametri $\sqrt{b1}$ i *b2* procenjuju simetriju i spljoštenost distribucije populacije, odnosno, na osnovu uzoraka i baziraju se na *k*-tom standardizovanom momentu. Kada je statistika $\sqrt{b1} = 0$, distribucija vrednosti promenljive je simetrična. Ukoliko je vrednost $\sqrt{b1}$ pozitivna (>0), distribucija je pomerena u desno, aritmetička sredina veća od medijane, a desni rep distribucije duži od levog. Ista asimetričnost samo u suprotnom smeru se javlja kod promenljivih koje imaju vrednost $\sqrt{b1}$<0. Parametar *b2*, indikuje spljoštenost i kod normalno distribuiranih promenljivih ima vrednost 3. *Platykurtic* distribucije imaju b2<3 dok kod leptokurtic distribuija b2>3.

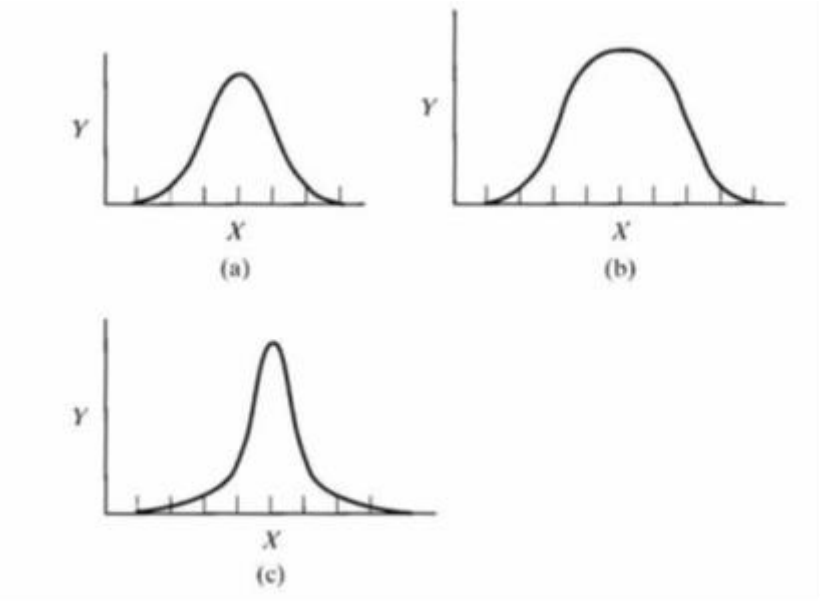
Statistčka metoda koja koristi simultano mere simetrije i spljoštenost se naziva *Shapiro-Wilks* test i često se preporučuje u testiranju normalnosti promenljive. Ova metoda koristi statistiku koja je označena kao W i testira sledeće hipoteze:

H0: Uzorak potiče iz populacija koja nema nomralnu raspodelu

HA Uzorak potiče iz populacije sa normalnom raspodelom vrednosti.

R funkcija *Shapiro.test()* računa vrednost *W* statistike i prikazuje *p* vrednost. Ukoliko je p>0.05, nulta hipoteza se odbacuje dok se alternativna hipoteza (HA) prihvata što znači da testirani skup podataka pokazuje normalnu raspodelu:

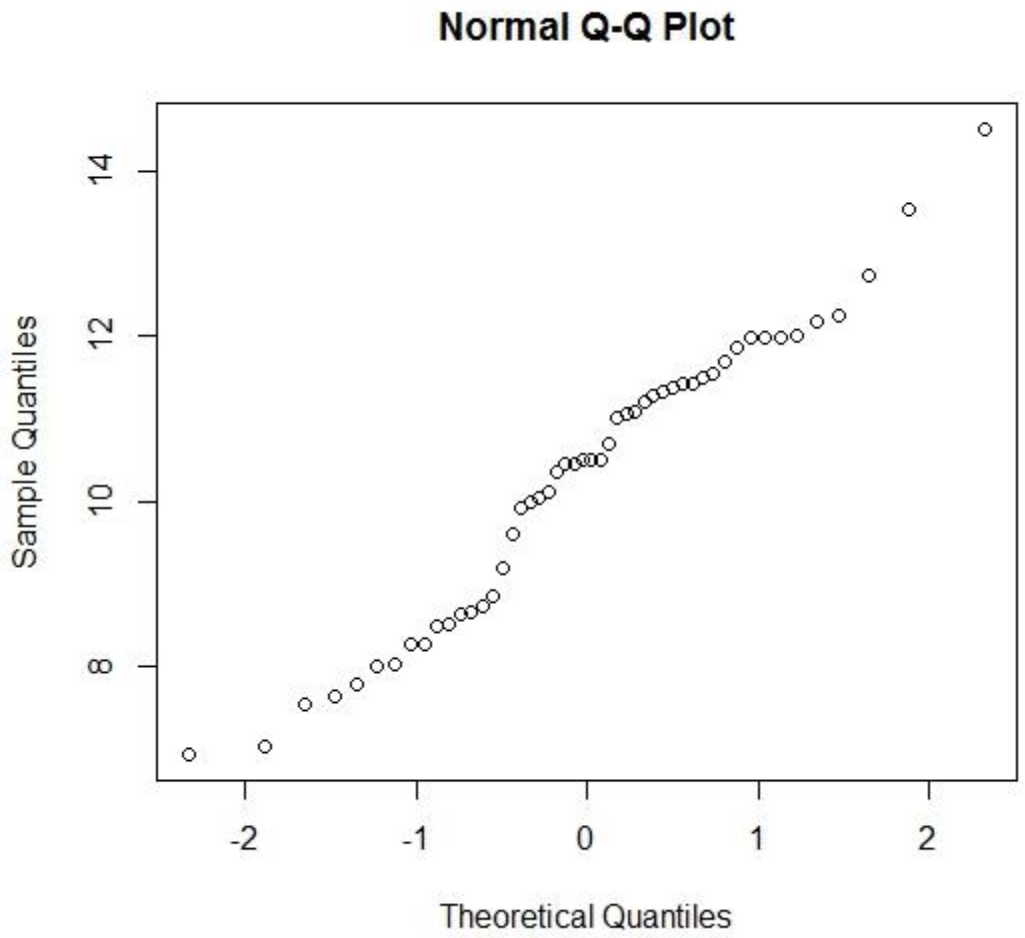
```
install.packages("dplyr")
install.packages("ggpubr")
shapiro.test(my_data$len)
Shapiro-Wilk normality test
data:  my_data$len
W = 0.96743, p-value = 0.1091
```



Slika 4.2 Procena odstupanja od normalne raspodele. Tri tipa spljoštenosti (Kurtosis): a) β2=3 – *mesokurtic* c)β2>3 – *leptokurtic* b)β2<3 – *platykurtic*

Grafičko testiranje normalne raspodele

Grafičko testiranje normalne raspodele se bazira na poređenju distribucije dobijenih vrednosti sa teroijskom distribucijom. Čest metod za ovaj vid testiranja se naziva kvantil-kvantil dijagram (QQ plot). Kvantil je mera koja deli set rangiranih vrednosti na *q* delova gde je *p* jedna od vrednosti (od 1 do q-1). Na pirmer, ukoliko je *q*=4 onda je set podataka podeljen na kvartile (*p*=1,2 i 3). QQ dijagram je grafička metoda koja pomoću dijagrama rasturanja vizualizuje set kvantila teorijske normalne distribucije (*y* osa) i set kvantila testiranog uzorka (*x* osa). Nula na *x* osi predstavlja 50ti percentil sa čije leve i desne strane je distribuirano po polovinu slučajeva iz uzorka standardne normalne distribucije. S druge strane su na *y* osi su rangirane vrednosti test uzorka u rastućem maniru. Ako oba seta kvantila imaju isti tip distribucije (na primer normalnu distribuciju) tačke na dijagramu rasturanja će formirati približno pravu liniju (Slika 4.3).

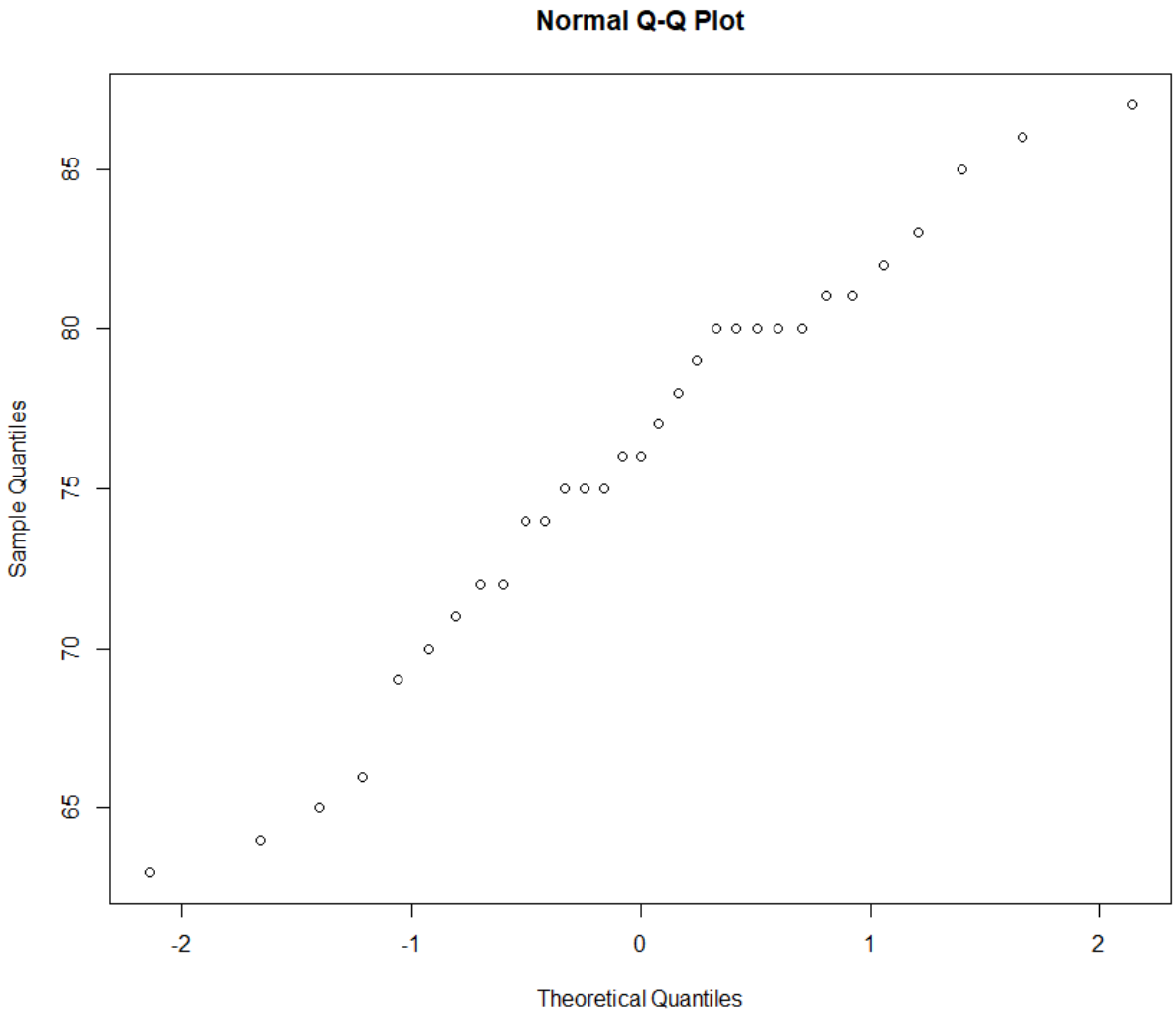


Slika 4.3. Vizualizuja seta kvantila teorijske normalne distribucije (*y* osa) i seta kvantila testiranog uzorka (*x* osa) kod normalne distribucije

Za izradu QQ dijagrama u R-u su dizajnirane funkcije *qqnorm()* I *qqplot()*

Funkcija *qqnorm()* konstruiše dijagram rasturanja testirane varijable u odnosu na teorijsku gde su vrednosti test promenljive rangirani u odnosu na kvantile standardne normalne distribucije. Na primer u okviru integrisanog seta podataka u R-u pod nazivom "*trees*" jedna od promenljivih je i "*height*", visina drveća. Ako se konstruiše QQ dijagram pomoću funkcije *qqnorm* za promenljivu *Height*, moguće je testirati da li ovaj uzorak potiče iz populacije sa nomrlanom distribucijom:

.



Pored funkcije *qqnorm()* u prethodnom kodu se navodi i funkcija *qqline()* koja dodaje referentnu liniju na konstruisani QQ dijagram:

```
qqnorm(trees$Height)
qqline(trees$Height, col = "steelblue", lwd = 2)
```

4.6 Interval poverenja

Ekolozi obično prilikom procenjivanja srednje vrednosti populacije na osnovu uzoraka određuju razumne granice u kojima bi trebalo da se nađe tačna srednja vrednost populacije. Te granice određuju intervali poverenja. Verovatnoća da se na osnovu uzorka dobije interval vrednosti koji stvarno sadrži tačnu srednju vrednost populacije zove se nivo poverenja (*1-α*; *α* je verovatnoća greške tj. verovatnoća da se na osnovu uzorka dobija interval koji ne sadrži tačnu srednju vrednost). Najčešće korišćen nivo poverenja je 95% gde je verovatnoća greške 5%. Srednja vrednost populacije sa intervalom poverenja od 95% se računa kao:

$$\mu = \bar{x} + t_{\alpha}\left(S_{\bar{x}}\right)$$

gde je μ =srednja vrednost populacije, $\overline{X}$ = srednja vrednost uzorka, t_{α} - vrednost sa studentove *t* tabele za *1-α* nivo poverenja, $S_{\overline{X}}$ =standardna greška

Vrednost t_{α} se očitava iz tabele za Studentovu t raspodelu (Tabela S4.2). Određuje se na osnovu stepena slobode (*N-1*) i nivoa poverenja koji iznosi *1-α*.

Iz jednačine se može zaključiti da što je standardna greška uzorka manja ($S_{\overline{X}}$) to je, interval poverenja manji, a samim tim je i procena srednje vrednosti populacije preciznija. S obzirom da se sa povecanjem veličine uzorka (*n*) standardna greška smanjuje, veći uzorci će uvek preciznije procenjivati parametre populacije.

Prilikom statističkog testiranja hipoteze, preporučuje se da pored *p* vrednosti prikaže i *1-α* interval poverenja srednje vrednosti populacije *μ*.

U R-u je moguće izračunati interval poverenja srednje vrednosti populacije sledećim kodom:

```
data("mtcars")


sample.mean <- mean(mtcars$mpg)
print(sample.mean)
```

```
sample.n <- length(mtcars$mpg)
sample.sd <- sd(mtcars$mpg)
sample.se <- sample.sd/sqrt(sample.n)
print(sample.se)

alpha = 0.05
degrees.freedom = sample.n - 1
t.score = qt(p=alpha/2, df=degrees.freedom,lower.tail=F)
print(t.score)

margin.error <- t.score * sample.se
lower.bound <- sample.mean - margin.error
upper.bound <- sample.mean + margin.error
print(c(lower.bound,upper.bound))
```

5 Testiranje hipoteza sa jednim ili dva uzorka.

5.1Testranje hipoteza sa jednim uzorkom

Teorijska osnova

U ekološkim studijama se često javlja potreba poređenja vrednosti merenih promenljivih sa nekom standardnom (teorijskom ili hipotetičkom) vrednošću. Standardna vrednost se obično određuje prethodnim eksperimentima. Na primer, da li se prosečna težina kalifornijske pastrmke (*Oncorhynchus mykiss*) razlikuje od 200g, što je težina dobijena u prethodnom eksperimentu. S druge strane u ekotoksikološkim studijama se uticaj bilo kog agensa u vidu tretmana na model organizam prati u odnosu na kontrolu, gde je testriani agens odsutan. U tom slučaju ako se izrazi promenljiva kao procenat kontrole, moguće je testrirati da li se uslovi u tretmanu razlikuju značajno od 100%.

Navedena pitanja se mogu formulisati u vidu statističke nulte i alternativne hipoteze:

$$H0: \mu=\mu_0$$

$$HA: \mu\neq\mu_0$$

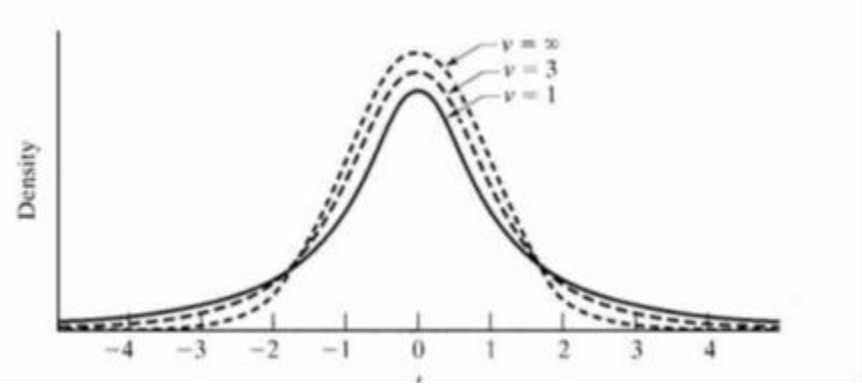
gde je μ_0 standardna vrednost sa kojom se poredi srednja vrednost populacije.

U prethodnom poglavlju je prikazana Z transformacija kojom je moguće testirati hipotezu o srednjoj vrednosti populacje μ, određivanjem verovatnoća dobijanja uzorka srednje vrednosti X Međutim, za izračunavanje statistike Z neophodno je poznavati standardnu grešku populacije, parametar koji obično nije poznat s obzirom da svaka statistička analiza koristi uzorke za testiranje hipoteza. Najbolji način da se proceni standardna greška populacije je pomoću standardne greške uzorka. Zbog toga se umesto Z distribucije koristi alternativna t distribucija koja je napravila najveći uspeh u statistčkoj metodologiji. T statistka se koristi u studentovom t testu i definisana je sledećom formulom:

$$t = \frac{\overline{X} - \mu}{S_{\overline{X}}}$$

gde je $\overline{X}$ srednja vrednost uzorka, μ srednja vrednost populacije i $S_{\overline{X}}$ standardna greška uzorka.

U zavisnosti od veličine uzorka, *t* distribucija može imati različit oblik i to je definisano stepenom slobode (√=*n*+*1*). Kako veličina uzorka raste tako i *t* distribucija menja svoj oblik i približava se normalnoj distribuciji gde konačna za √=∞, *t* distribucija i *Z* distribucija postaju identične.



Slika 5.1 t distribucija i stepeni slobode √. Za √=∞, *t* distribucija je identična normalnoj distribuciji.

Pretpostavke

Kako bi se primenio t test neophodno je da podaci zadovoljavalju sledee pretpostavke:

1. Teorijska osnova *t* testa predpostavlja da uzorak potiče iz populacije sa normalnom raspodelom zbog čega se očekuje i da sam uzorak ima normalnu raspodelu.
2. Podaci su iz slučajnog uzorka. Tačnje uzorak koji se statistički testira se sastoji od ponovljenih merenja gde svako merenje je definisano kao najmanja eksperimentalna jedinica na koju je tretman nezavisno primenjen.

Primer statističkog testiranja hipoteze t-testom za jedno merenje u R-u

Za izvođenje *t* testa za jedan uzorak u R-u koristi se funckija *t.test():*

```
t.test(x, mu = 0, alternative = "two.sided")
```

Prvi argument „*x*“ definiše vektor koji sadrži podatke iz uzorka, u ovom slučaju težinu riba kalifornijske pastrmke. Argument „*mu*“ definiše teorijsku vrednost koja se određuje *a priori* i koja je podešena na 0. Poslednji argument „*alternative*“ definiše alternativnu hipoteuzu, odnosno da li

je srednja vrednost populacije različita „*two sided*“ (opcija koja je predefinisa u podešavanjima), veće („*greater*„) ili manje („*less*“) od μ_0.

U slučaju primera težine riba kalifornijske pastrmke moguće je testirati da li srednja vrednost težine uzorka se značajno razlikuje od teorijske vrednost:

$$H0: \mu=\mu0$$

$$HA: \mu\neq\mu0$$

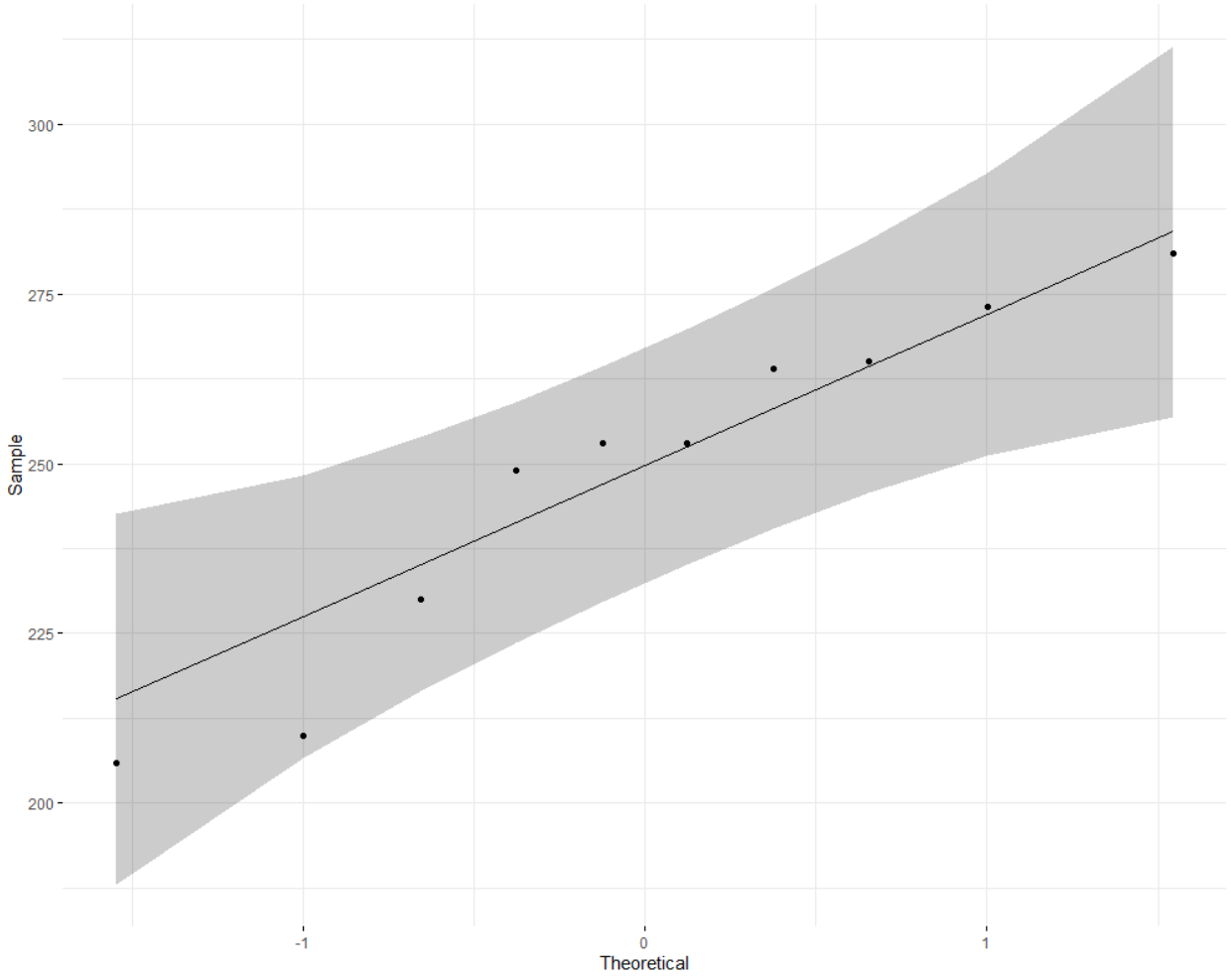
Testirani uzorak se sastoji od 10 merenja (Tabela 5.1) gde je $\mu0$=250g, vrednost koja je dobijena prethodnim eksperimentima.

Tabela 5.1 Težina riba (M1-M10) kalifornijske pastrmke u gramima (g)

Izmerena jedinka	Težina
M1	206g
M2	230g
M3	249g
M4	253g
M5	264g
M6	253
M7	210g
M8	273g
M9	265g
M10	281g

```
> Tezina_riba= data.frame(name=paste0(rep("m_", 10), 1:10),
weight= c(206,230,249,253,264,253,210,273,265,281))

> shapiro.test(Tezina_riba$weight)
```

```
     Shapiro-Wilk normality test

data:  Tezina_riba$weight
W = 0.91488, p-value = 0.3162
```




```
> library("ggpubr")
> ggqqplot(Tezina_riba$weight,
+          ggtheme = theme_minimal())
> rezultat <- t.test(Tezina_riba$weight, mu = 250)
```

```
> rezultat

     One Sample t-test

data:  Tezina_riba$weight

t = -0.19862, df = 9, p-value = 0.847

alternative hypothesis: true mean is not equal to 250

95 percent confidence interval:

 230.1769 266.6231

sample estimates:

mean of x

    248.4

> rezultat2 <-t.test(Tezina_riba$weight, mu = 250,

+        alternative = "less")

> rezultat2

     One Sample t-test

data:  Tezina_riba$weight

t = -0.19862, df = 9, p-value = 0.4235

alternative hypothesis: true mean is less than 250

95 percent confidence interval:

     -Inf 263.1669

sample estimates:

mean of x

    248.4

> rezultat3 <-t.test(Tezina_riba$weight, mu = 250,

+                   alternative = "greater")

> rezultat3
```

```
     One Sample t-test
data:  Tezina_riba$weight
t = -0.19862, df = 9, p-value = 0.5765
alternative hypothesis: true mean is greater than 250
95 percent confidence interval:
 233.6331      Inf
sample estimates:
mean of x
    248.4
```

Rezultati analize prikazuju sledeće statističke indikacije: „*t*“ predstavlja vrednost *t* test statistike (*t*=-0.1986), *df* označava broj stepene slobode (*df*=9), *p* je nivo značajnosti *t* testa (*p*=0.847), *confint* predstavlja 95% interval poverenja (*conf.int* =230.1769, 266.6231) i *mean* predstavlja srednju vrednost uzorka (*mean*=248.4). Na osnovu rezultata testa možemo zaključiti da je *p* vrednost veća od granične (0.85>0.05) i da se na osnovu toga nulta hipoteza (H0) ne odbacuje, odnosno da se srednja vrednost težine riba ne razlikuje od teorijske vrednosti (μ_0=250g).

5.2.Testiranje hipoteza sa dva uzorka

Teorijska osnova

Testiranje hipoteza sa dva uzorka je jedan od najčešćih statističkih dizajna u ekologiji gde je cilj testirati da se utvrdi postojanje značajne razlika određenog parametra između dve populacije. To može biti sprovedeno testiranjem razlika srednjih vrednosti uzoraka koji su prikupljeni po principu slučajnosti iz analiziranih populacija. Ovakav statistički dizajn može biti definisan sledećim hipotezama:

H0: μ1= μ 2

HA: m1≠m2

odnosno da nulta hipoteza pretpostavlja da ne postoji razlika srednje vrednosti parametra između dve populacije dok alternativna hipoteza prepostavlja da se dve srednje vrednosti značajno razlikuju.

Na primer, t test za dva nezavisna uzorka se može primeniti na testiranju koncentracije hlorofila *a* u vodi (mg/m3) koji je dobar pokazatelj eutrofizacije voda. Kako bi se testrialo postojanje značajne razlike količine hlorofila *a* u vodi potrebno je sprovesti *t* statistiku koja je za *t* test za dva nezavisna uzorka formulisan na sledeći način:

$$t = \frac{\overline{X}\,1 - \overline{X}\,2}{S\,\overline{X}\,1 - S\,\overline{X}\,2}$$

gde $\overline{X}\,1 - \overline{X}\,2$ predstavlja razliku između dve srednje vrednosti uzorka, dok je SX1-SX2 standardna greška razlike srednje vrednosti uzoraka i predstavlja meru varijabilnosti podataka u okviru ova dva uzorka. Na taj način metoda poredi razlike između dve srednje vrednosti uzoka i razlike u okviru celog seta podataka. Isti pristup se koristi i u metodama koje testiraju razlike između više od dva uzorka i o tome će biti reči kasnije (videti poglavlje 6, *ANOVA*).

Standardna greška razlike srednje vrednosti uzoraka je izračunata pomoću podataka iz uzoraka, a procenjuje populacione parametre (standardna greška razlike srednje vrednosti populacija). Matematički se može prikazati da varaijansa razlika zimeđu dve nezavisne promenljive je jednaka sumi varijansi te dve varijable tako da je $\sigma^2\,\overline{X}\,1 - \overline{X}\,2 = \sigma^2\,\overline{X}_1 + \sigma^2\,\overline{X}_2$. Pošto je poznato da je standardna greška populacije $\sigma^2\,\overline{X} = \sigma^2/n$ onda je:

$$\sigma^2\,\overline{X}\,1 - \overline{X}\,2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

S obzirom da *t*-test za dva uzorka pretpostavlja da su varijanse dve populacije jednake ($\sigma^2_1 = \sigma^2_2$) onda je

$$\sigma^2\,\overline{X}\,1 - \overline{X}\,2 = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}$$

Pošto se za računanje standardna greške razlike srednje vrednosti populacije ($\sigma^2\,\overline{X}\,1 - \overline{X}\,2$) koristi varijansa populacije (σ^2) i da je ona procenjena varijansom dva uzorka (S^2_1 i S^2_2),

neophodno je definisati puliranu varijansu (S^2_p) uzorka koja će najprihližnije proceniti varijansu populacija σ^2:

$$S^2_p = \frac{SS1+SS2}{v1+v2}$$

Gde je SS, suma razlike kvadratnog odstupanja dobijenih vrednosti od srednje rednosti ($SS=\sum(Xi - \overline{X})^2$, parametar koji se koristi za izračunavanje varijanse.

Ukoliko se standardna greška uzorka primeni u glavnoj formuli za izračunavanje standardne greške razlike srednje vrednosti uzorka, dobija se sledeća formula:

$$S^2_{\overline{X}1-\overline{X}2} = \frac{S_p^2}{n_1} + \frac{S_p^2}{n_1}$$

Odnosno

$$S_{\overline{X}1-\overline{X}2} = \sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_1}}$$

Ukoliko se ovaj izraz primeni u formuli za izračunavanje *t* statistke dobija se:

$$t = \frac{\overline{X}1 - \overline{X}2}{\sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_1}}}$$

U slučaju da se varijanse dva uzorka razlikuju onda je nemoguće izračunati puliranu standardnu grešku niti primeniti *t* test za statističko testiranje hipoteze. Zbog toga je razvijen alternativni test koji je poznat kao *Welch aproksimacija*. Ova modifikovana verzija *t*-testa računa standardnu grešku razlike srednje vrednosti uzorka pomoću dve odvojene varijanse:

$$S_{P'}\,\overline{X}_1 - \overline{X}_2 = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_1}}$$

S obzirom da je standardna greška jednog uzorka $S_{\overline{X}_i} = S^2_i/n_i$ onda se standardna greška razlike srednjih vrednosti uzoraka sa različitim varijansama može prikazati sledećom formulom:

$$S_{P'}\,\overline{X}_1 - \overline{X}_2 = \sqrt{S^2_{\overline{X}_1} + S^2_{\overline{X}_2}}$$

Ako se ova modifikacija primeni u formuli za izračunavanje *t* statistike dobija se:

$$t' = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{S^2_{\overline{X}_1} + S^2_{\overline{X}_2}}}$$

Pretpostavke za *t*-test sa dva uzorka

1. Dva slučajna uzorka dolaze iz populacija sa normalnom raspodelom
2. Homogenost varijanse, dva uzorka imaju istu varijansu ($\sigma^2_1 = \sigma^2_2$)

Primer statističkog testiranja hipoteze *t*-testom za jedno merenje u R-u

Pomoću funkcije *t_test* u R-u je moguće testirati razlike između dva nezavisna uzorka. Kao primer koristi se uzorak vode sa dva različita jezera sa različitim nivoom trofičnosti. Kao indikator trofičnosti se koristi koncentracija hlorofila *a* u vodi izražena kao mg/cm2. Uzorkovanje je sprovedeno na dest mernih mesta u dva tako da svaki od dva uzorka sadrži deset merenja. Pre primene *t* testa za statistčko testiranje postavljenih hipoteza, potrebno je uztvrditi da li su zadovljene pretpostavke o normalnoj raspodeli i homogenosti varijansi (poglavlje 5.1.2). Za procenu normalnnosti promenljive za dva uzorka, koristiće se statistički test (*Shapiro-Wilk*-ov test) i vizuelna procena (*QQ* dijagram).

Prvi korak je konstruisati ulaznu matricu koja će sadržati podatke o hlorofilu *a* za dva vodna tela:

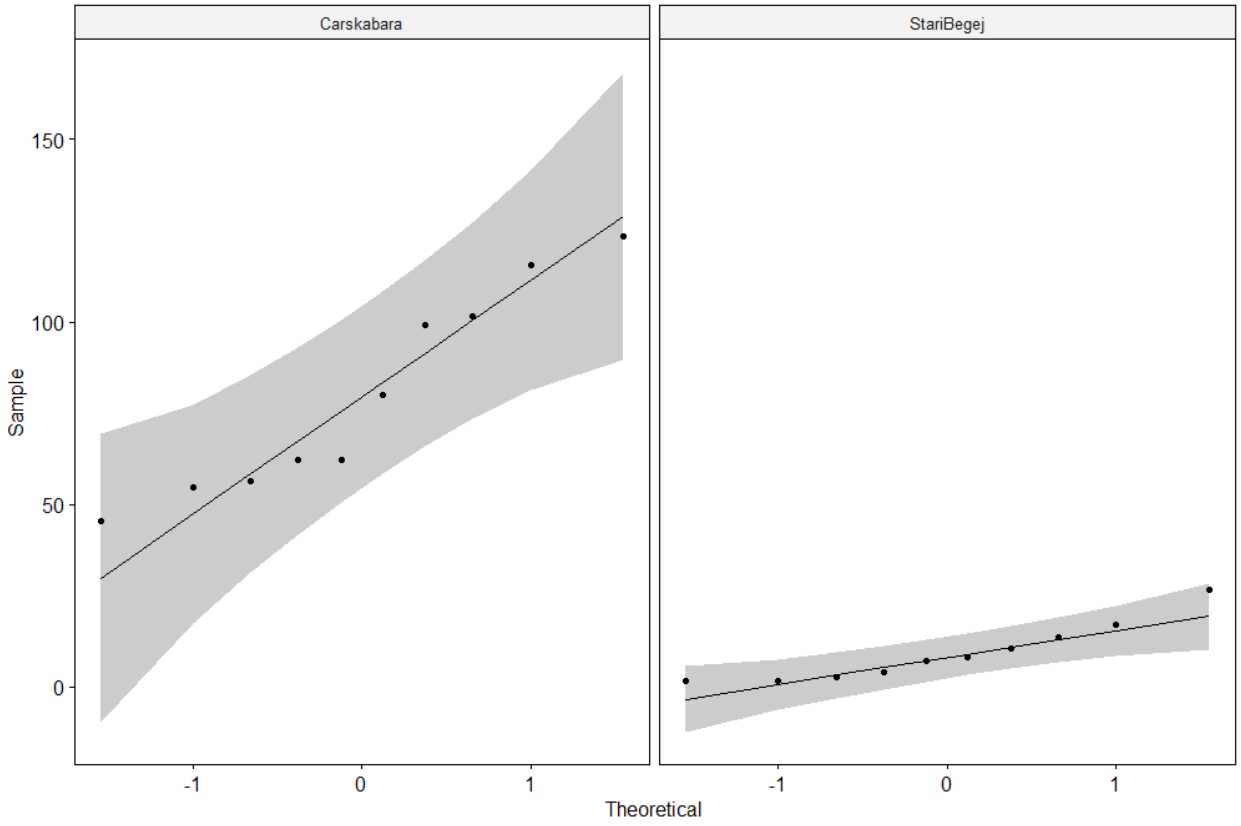
```
> Carskabara <- c(80.1, 62.2, 115.6, 123.5, 45.4, 56.3, 62.3, 54.7, 101.6, 99.3)

> StariBegej <- c(10.7, 26.7, 8.1, 4.1, 2.7, 13.7, 16.9, 1.7, 7.2, 1.8)

> my_data <- data.frame( lokalitet = rep(c("Carskabara", "StariBegej"), each = 10), hlorofil = c(Carskabara_hlorofil, StariBegej_hlorofil) )

> library(dplyr)

> group_by(my_data, lokalitet) %>% summarise( count = n(), mean = mean(hlorofil, na.rm = TRUE), sd = sd(hlorofil, na.rm = TRUE) )

`summarise()` ungrouping output (override with `.groups` argument)

# A tibble: 2 x 4

  lokalitet  count  mean    sd

  <chr>      <int> <dbl> <dbl>

1 Carskabara    10 80.1  27.9

2 StariBegej    10  9.36  7.97
```

Nakon što je matrica formirana, normalnost podataka se testira funkcijama *shapiro_test ()* i *ggqqplot()*:

```
> #provera normalnosti varijabli

> # Shapiro-Wilk normality test for Men's weights with my_data

> install.packages("rstatix")

Error in install.packages : Updating loaded packages

> library(rstatix)

> my_data=as.tibble(my_data)
```

```
> my_data
# A tibble: 20 x 2
   lokalitet  hlorofil
   <chr>         <dbl>
 1 Carskabara     80.1
 2 Carskabara     62.2
 3 Carskabara    116.
 4 Carskabara    124.
 5 Carskabara     45.4
 6 Carskabara     56.3
 7 Carskabara     62.3
 8 Carskabara     54.7
 9 Carskabara    102.
10 Carskabara     99.3
11 StariBegej     10.7
12 StariBegej     26.7
13 StariBegej      8.1
14 StariBegej      4.1
15 StariBegej      2.7
16 StariBegej     13.7
17 StariBegej     16.9
18 StariBegej      1.7
19 StariBegej      7.2
20 StariBegej      1.8
> my_data%>%
```

```
+   group_by(lokalitet)%>%
+   shapiro_test(hlorofil)
# A tibble: 2 x 4
  lokalitet  variable statistic     p
  <chr>      <chr>        <dbl> <dbl>
1 Carskabara hlorofil     0.907 0.258
2 StariBegej hlorofil     0.886 0.152
> #Napraviti qqplot
> ggqqplot (my_data, x="hlorofil", facet.by = "lokalitet")
```



Dobijen rezultat (*p*=0.25 za lokalitet Carska bara i *p*=0.152 za Stari Begej) ukazuje na noramlnu distribuciju vrednosti hlorofila *a* na dva analizirana vodna tela.

Druga pretpostavka o jednakosti varijansi se može testirati levenovim testom. Ovaj statistčki metod koristi *t* statistiku i testira sledeće hipoteze:

H0: $\sigma^2_1 = \sigma^2_2$

HA: $\sigma^2_1 \neq \sigma^2_2$

To znači da za ispunjavanje pretpostavke o homogenosti varijansi, potrebno je prihvatiti nultu hipotezu, odnosno da $p>0.05$. U R-u se homogenost varijansi dva uzorka može testirati funkcijom *levene_test*():

```
> #provera jednakosti varijansi)
> my_data %>% levene_test(hlorofil~lokalitet)
# A tibble: 1 x 4
    df1   df2 statistic       p
  <int> <int>     <dbl>   <dbl>
1     1    18      12.5 0.00239
```

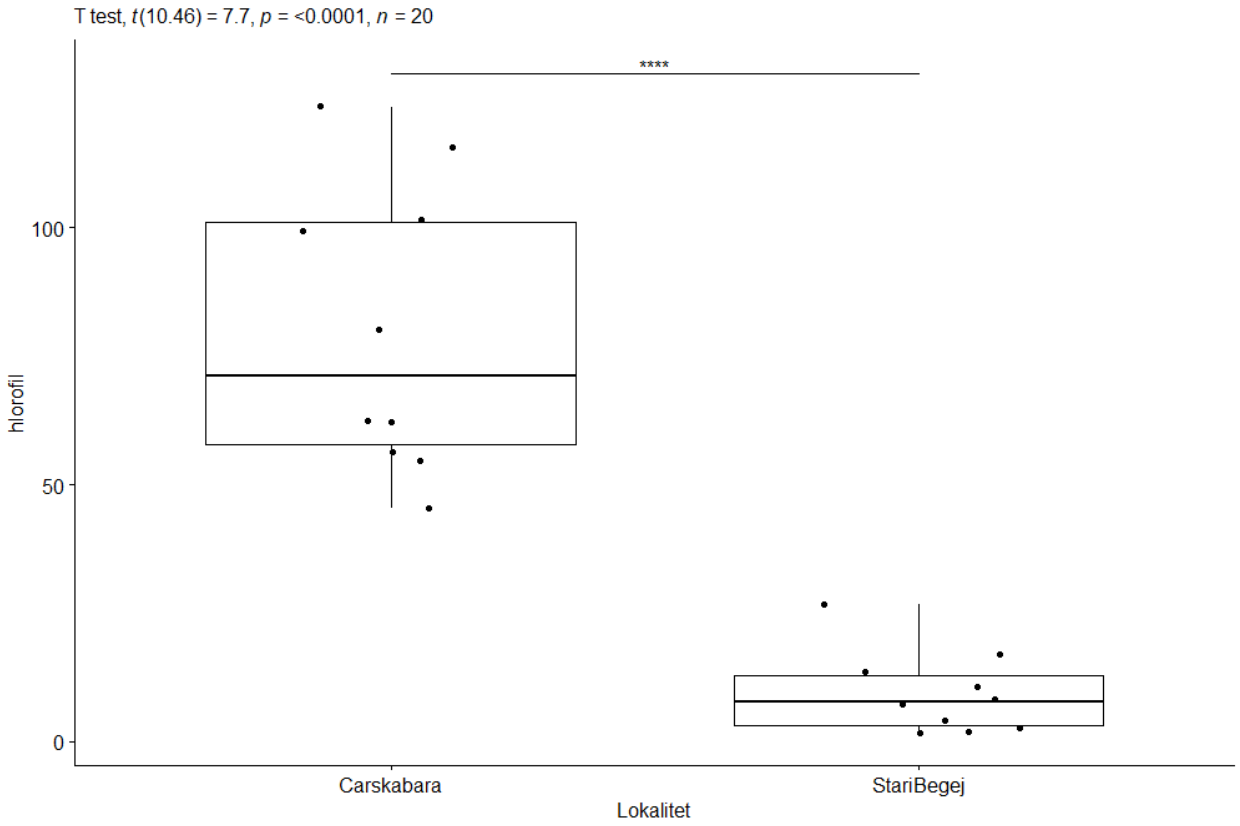
S obzirom da je prema rezultatim Leveneovog testa $p<0.05$ (p=0,00239) nulta hipoteza se odbacuje i prihvata alternativna hipoteza (HA: $\sigma^2_1 \neq \sigma^2_2$), što znači da se varijanse ova dva uzorka značajno razlikuju. Zbog toga je neophodno da se prilikom testiranja srednjih vrednosti uzoraka primeni *t* test sa *Welwich* aproksimacijom. Funkcija *t_test()* sadrži argument *var.equal* koji definiše da li se prilikom testiranja koristi standardni *t* test (*var.equal = TRUE*) ili *Welwich* aproksimacija (kao predefinisano podešavanje stoji *var.equal = FALSE*).

```
> statisticki.test= my_data %>% t_test(hlorofil ~ lokalitet)%>%
+   add_significance()
> statisticki.test
# A tibble: 1 x 9
  .y.      group1     group2        n1    n2 statistic    df
p p.signif
  <chr>    <chr>      <chr>      <int> <int>     <dbl> <dbl>
<dbl> <chr>
1 hlorofil Carskabara StariBegej    10    10      7.70  10.5
0.0000127 ****
```

```
> statisticki.test2= my_data %>% t_test(hlorofil ~ lokalitet,
var.equal = TRUE) %>%

+   add_significance()

> statisticki.test2

# A tibble: 1 x 9

  .y.      group1     group2        n1    n2 statistic    df
p p.signif

  <chr>    <chr>      <chr>      <int> <int>     <dbl> <dbl>
<dbl> <chr>

1 hlorofil Carskabara StariBegej    10    10      7.70    18
0.000000422 ****
```

Prema rezultatima *t* testa, gde je $p<0.05$ ($p=0.000000442$) nulta hipoteza se odbacuje i prihvata alternativna, da se srednje vrednosti dva uzorka hlorofila značajno razlikuju. Tačnije da je koncentracija hlorofila *a* u Starom Begeju značjano veća od one u Carskoj Bari. Kako bi se rezultati prikazali na što jasniji način, najbolje je kombinovati najznačajnije statistčke indikacije testa i vizuelnu prezentaciju rezultata. Tako nešto je moguće generisati u R-u sledećim kodom:

```
> staisticki.test<- statisticki.test %>% add_xy_position(x=
"group")

> bxp=ggboxplot(my_data, x="lokalitet", y="hlorofil", ylab =
"hlorofil", xlab = "Lokalitet", add="jitter")

> bxp+

+ stat_pvalue_manual(statisticki.test, tip.length = 0,
y.position = 130)+

+ labs(subtitle = get_test_label(statisticki.test, detailed=
TRUE))
```


T test, t(10.46) = 7.7, p = <0.0001, n = 20
****
100
50
0
hlorofil
Carskabara
StariBegej
Lokalitet

5.3 Neparametarsko testiranje hipozeza za dva uzorka

Neparametarski statistički testovi ne koriste procene populacionih parametara (μ i σ) dok su hipoteze tako koncipirane da se u pretpostavkama ne pominju populacioni parametri. Ova grupa statistčkih testova je suprotna parametarskim testovima kao što je *t*-test koji se bazira na proceni populacionih parametara dok hipoteze pretpostavljaju da se srednje vrednosti populacija ne razlikuju (H0: μ1= μ2) , odnosno razlikuju (HA: μ1≠ μ2). Zbog ove činjenice, neparametarski testovi se koriste kao alternativa parametarskim kada su pretpostavke o normalnoj raspodeli i varijansi testiranih promenljivih ozbiljno narušene.

Neparametarski testovi koriste rangirane vrednosti umesto dobijenih in ta taj način isključuju uticaj ekstremnih vrednosti u uzorku koje remete normalnu raspodelu skupa podataka- S druge strane ovakvom transformacijom gde se vrednosti poređaju u rastući ili opadajući niz, gubi se određeni broj informacija iz skupa podataka što predstavlja glavni nedostatak ove metode.

Teorijska osnova

U ekologiji se načešće javlja potreba za statističkim testiranjem značajne razlike između centralnih tendencja dva uzorka koja nemaju normalnu raspodelu i za tu svrhu se koristi *Mann-Whitney*-ev test. Na primer u ekotoksikološkim analizama veliki broj biomarkera široko varira u replikama i promenljiva nema normalnu raspodelu dok se varijanse između tretmana značajno razlikuju. U tom slučaju nije moguće koristiti t-test već se često primenjuje neparametarska alternativa kao što je *Mann-Whitney*-ev (MW) test. Ulazni podaci za ovaj test su rangirane vrednosti skupa podataka koje mogu biti raspoređene u nizu od najmanje do najveće vrednosti ili obrnuto. *Mann-Whitney* test koristi *U* statistiku:

$$U=n1n2 + \frac{n1(n1+1)}{2}+R1$$

gde su n_1 i n_2 broj obzervacija u uzorku 1 i 2, respektivno, a R1 suma rangova uzorka 1. Ukoliko se U statistika izrazi preko R2, sume rangova uzorka 2, statistička interpretacija poprima sledeći izgled:

$$U'=n2n1 + \frac{n2(n2+1)}{2}+R2$$

Označavanje uzoraka je arbitrarno, pa je zbog toga moguće koristiti bilo koju od ove dve jednačine koje se lako mogu transformisati jedna u drugu:

$$U' = n1n2 - U, \text{ odnosno } U = n1n2 - U'$$

S obzirom da MW test ne pripada grupi parametarskih testova koji procenjuju parametere populacije (μ i σ), hipoteze koje se testiraju manje su specifične i porede medijane populacija, otkrivajući da li su distribucije podataka dve populacije iste. Na primer, ukoliko se testira kako cijanotoksin utiče na genotoksičnost larvi hironomida (Diptera: Chironomidae), prateći nivo oštećenja DNK koji je kvantifikovan dužinom repa DNK, postavljaju se sledeće hipoteze:

H0: Dužina repa DNK hironomida kontrole se ne razlikuje od dužine repa DNK hironomida koje su bile tretirane cijanotoksinom Mikrocistin LR.

H0 Dužina repa DNK hironomida kontrole se razlikuje od dužine repa DNK hironomida koje su bile tretirane cijanotoksinom Mikrocistin LR.

Pretpostavke

Pretpostavke za korišćenje neparamterarskih testova su manje striktne nego kod parametarskih testova jer se ne oslanjaju na parametre populacija (μ i σ):

1. Slučajnost uzorka
2. Nezavisnost opservacija

Primer statističkog testiranja hipoteze Mann-Wtthney testom u R-u

Funkcijom *wilcox_test ()* se u R-u koristi za testiranje hipoteza MW testom. Eskprimentom u kontrolisanim uslovima testiran je toksični efekat cijanotoksina Mikrocistin LR na larve familije hironomida. Kao biomarker je korišćen nivo oštećenosti DNK lanca (dužina repa DNK) čije su vrednosti predstavljene u Tabeli 5.2. Kako bi se podaci pripremili za statističko testiranje, neophodno je generisati ulaznu matricu:

Tabela 5.2 Dužina repa DNK jeidnki koje su bile u kontrolnim uslovima i one u tretmanu, tretirane cijanotoksinom Mikrocistin LR

Konrola (H_2O)	Tretman (Mikrocistin LR 1mg/l)
1.8434	24.4429

19.3141	25.1643
37.3159	33.4689
1.16004	34.9058
1.68472	29.0558
5.14486	30.1883
1.67865	39.1202
0.81668	2.04586
0.85719	23.9191
1.2763	12.9059

```
> Kontrola <- c(1.843399, 19.314072, 37.315882, 1.160038, 1.684717, 5.144862, 1.678647, 0.816683, 0.857185, 1.276296)

> Cijanotoksin <- c(24.442907, 25.164253, 33.468926, 34.905848, 29.055759, 30.188337, 39.120239, 2.045863, 23.919067, 12.905896)

> my_data <- data.frame( Tretman = rep(c("Kontrola", "Cijanotoksin"), each = 10), Duzina_repa = c(Kontrola, Cijanotoksin) )

> group_by(my_data, Tretman) %>% summarise(count = n(), mean = mean(Duzina_repa, na.rm = TRUE), sd = sd(Duzina_repa, na.rm = TRUE), M = median(Duzina_repa, na.rm = TRUE) )

summarise()` ungrouping output (override with `.groups` argument)

# A tibble: 2 x 5

  Tretman      count  mean    sd     M

  <chr>        <int> <dbl> <dbl> <dbl>

1 Cijanotoksin    10 25.5   11.0 27.1

2 Kontrola        10  7.11  12.0  1.68

> post

# A tibble: 20 x 2
```

```
   Tretman      Duzina_repa
   <chr>              <dbl>
 1 Kontrola           1.84
 2 Kontrola          19.3
 3 Kontrola          37.3
 4 Kontrola           1.16
 5 Kontrola           1.68
 6 Kontrola           5.14
 7 Kontrola           1.68
 8 Kontrola           0.817
 9 Kontrola           0.857
10 Kontrola           1.28
11 Cijanotoksin      24.4
12 Cijanotoksin      25.2
13 Cijanotoksin      33.5
14 Cijanotoksin      34.9
15 Cijanotoksin      29.1
16 Cijanotoksin      30.2
17 Cijanotoksin      39.1
18 Cijanotoksin       2.05
19 Cijanotoksin      23.9
20 Cijanotoksin      12.9
> my_data%>%
+   group_by(Tretman)%>%
+   shapiro_test(Duzina_repa)
```

```
# A tibble: 2 x 4
  Tretman      variable    statistic         p
  <chr>        <chr>           <dbl>     <dbl>
1 Cijanotoksin Duzina_repa     0.910 0.281
2 Kontrola     Duzina_repa     0.600 0.0000558
```

Na primeru genotoksičnog efekta cijanotoksina na larve hironomida je moguće primeniti MW s obzirom da promenljiva dužina repa DNK nema normalnu raspodelu za kontrolni uzorka (p=0.281):

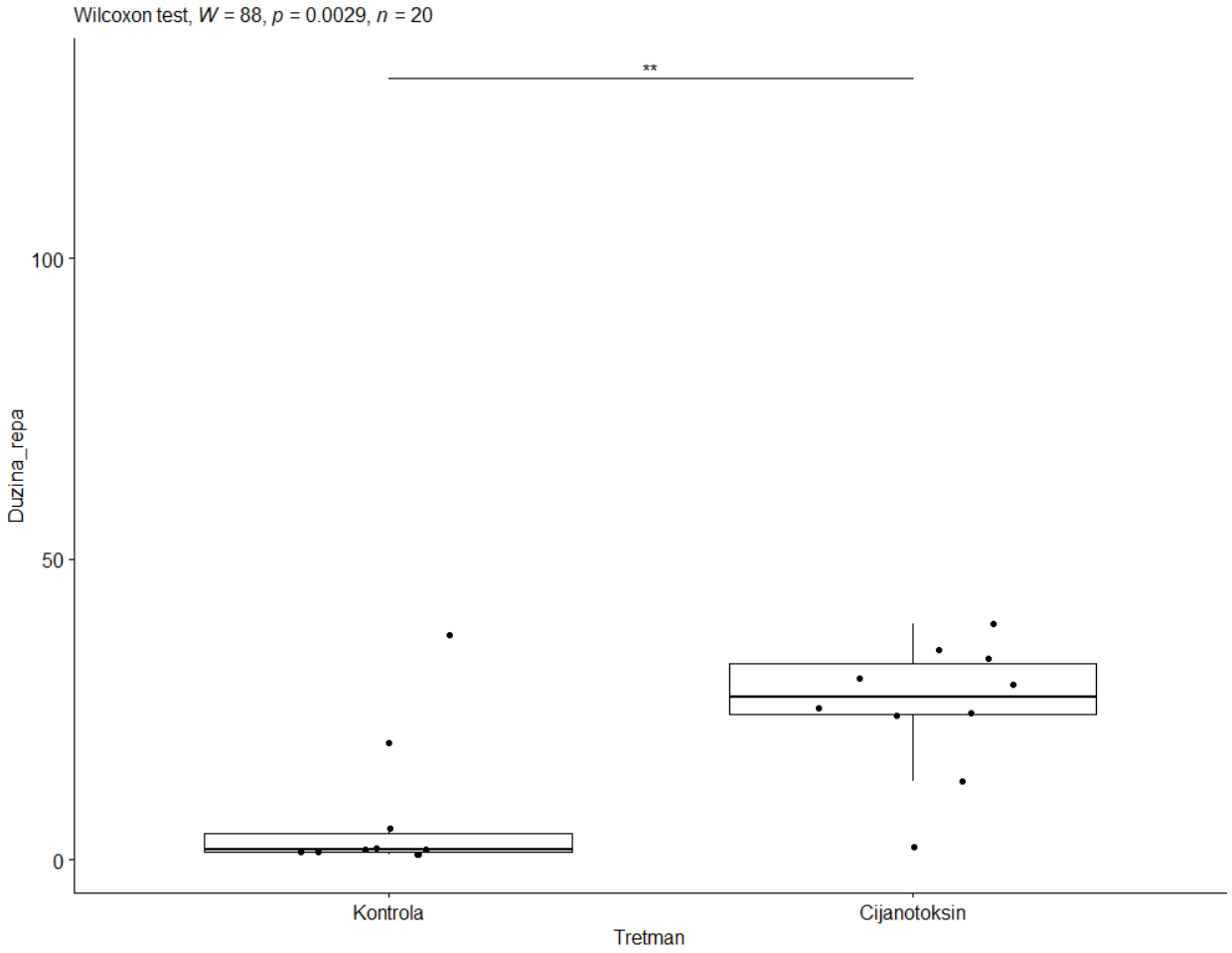
```
> statisticki.test<- my_data %>%
+   wilcox_test(Duzina_repa ~ Tretman)%>%
+   add_significance()
> statisticki.test
# A tibble: 1 x 8
  .y.         group1       group2      n1    n2 statistic
p p.signif
  <chr>       <chr>        <chr>    <int> <int>     <dbl>
<dbl> <chr>
1 Duzina_repa Cijanotoksin Kontrola    10    10        88
0.00288 **
> effect.size<- my_data %>%
+   wilcox_effsize(Duzina_repa ~ Tretman)
> effect.size
# A tibble: 1 x 7
```

```
  .y.          group1       group2   effsize    n1    n2
magnitude
* <chr>       <chr>        <chr>      <dbl> <int> <int> <ord>
1 Duzina_repa Cijanotoksin Kontrola   0.642    10    10 large
```

Rezultati MW testa pokazuju da se duzina repa u tretmanu statistčki značajno razlikuje od džine repa DNK u jedinkama kontrolnog uzorka (p=0,00288). Funkcijom *wilcox_effsize()* je testirana veličina uticaja faktora (tretmana) na promenljivu koja pokazuje visok nivo uticaja (r=0.642).

Kako bi se rezultat statističkog testa prikazao na efikasan način, rezultatima u tekstu se obično pridružuje vizuelni prikaz rezultata. Za ovaj statistčki dizajn je najefikasnije koristiti kutijaste dijagrame (eng. *Box plot*):

```
> staisticki.test<- statisticki.test %>% add_xy_position(x=
"Tretman")
> bxp=ggboxplot(my_data, x="Tretman", y="Duzina_repa", ylab =
"Duzina_repa", xlab = "Tretman", add="jitter")
> bxp+
+   stat_pvalue_manual(statisticki.test, tip.length = 0,
y.position = 130)+
+   labs(subtitle = get_test_label(statisticki.test, detailed=
TRUE))
```


Wilcoxon test, W = 88, p = 0.0029, n = 20
**
100
50
0
Duzina_repa
Kontrola
Cijanotoksin
Tretman

6. Testiranje hipoteze sa više uzoraka

6.1Parametarsko testiranje hipoteza sa više uzoraka i analiza varijansi ANOVA

Kada studija uključuje testiranje razlike između dva uzorka, takav statistički dizajn može sadržati metode koje se pominju u poglavlju 4 i 5 (*t*- test i *Mann Whitney* test, respektivno). Međutim u ekološkim studijama su često ulazni podaci sačinjeni od tri ili više uzoraka. Kada bi se na takve podatke formulisne kroz hipotezu H0: $\mu 1 = \mu 2 = \mu 3$, za svaki mogući par uzoraka, u nizu primenili statistički testovi za dva uzorka H0: $\mu 1 = \mu 2$, H0: $\mu 1 = \mu 3$. Ho: $\mu 2 = \mu 3$ došlo bi do greške. Kada se primenjuje t test za dva uzorka, pretpostavlja se da uzorci potiču iz iste populacije (ili dve populacije sa identičnom srednjom vrednošću) sa verovatnoćom greške te pretpostavke od 5% (α=0.05). Ukoliko bi se isti statistički dizjan upotrebio za tri uzorka računanjem tri *t* test u nizu, verovatnoća pogrešnog zaključivanja za sva tri uzorka potiču iz iste populacije bi iznosio 14%. Ukoliko bi se isto testiranje sprovelo na 4 uzorak, greška bi skočila na čak 26% gde se nivo greške prvog reda sa porastom broja uzorka računa kao:

$$1-(1-\alpha)^c$$

gde je *c* maksialni mogući broj kombinacija *k* uzoraka. S obzirom da se greška u testiranju većeg broja uzoraka brzo povećava sa povećanjem broja uzoraka, *t* test za dva uzorka nije odgovarajuća metoda za hipoteze za tri i više uzoraka (Tabela 6.1).

Tabela 6.1 Promena nivo greške prvog reda α sa porastom broja uzorka. *C* predstravlja višestruku primenu *t* test između k grupa (C=k(k-1)/2)

k	C	0.10	0.05	0.01	0.005	0.001
2	1	0.10	0.05	0.01	0.005	0.001
3	3	0.27	0.14	0.03	0.015	0.003
4	6	0.47	0.26	0.06	0.030	0.006
5	10	0.65	0.40	0.10	0.049	0.010
6	15	0.79	0.54	0.14	0.072	0.015
10	45	0.99	0.90	0.36	0.202	0.044
	∞	1.00	1.00	1.00	1.00	1.00

Teorijska osnova

Kako bi se prevazišao problem greške u testiranju hipoteza sa većim brojem uzoraka, dizajnirana je metoda pod imenom jednofaktorska analiza varijanse (*ANOVA*) ili jednofaktorska *ANOVA* koja prestavlja modifikaciju *t* testa za dva uzorka. Ulazni podaci za jednofaktrosku *ANOVA-u* su organizovani u više grupa koje su definisane jednom promenljivom koja se naziva faktorska promenljiva. Na primer, na slivu Južne Morave je merena koncentracija rastvorenog kiseonika u vodi, abiotički parametar koji je odlican pokazatelj saprobnosti vode. Na tri različite reke je sprovedeno po deset merenja. Broj grupa je *k*=3 dok svaka grupa ima *n1= n2= n3= 10* merenja. Ukupran broj merenja u sve tri grupe je:

$$N=\sum_{i=1}^{k} ni$$

Naziv statističkog testa *ANOVA*-leži u njenoj matematičkoj osnovi. Ova metoda istražuje nekoliko izvora varijabilnosti u okviru testiranih podataka u eksperimentu. Za tu svrhu koristi se suma kvadrata odstupanja dobijenih vrednosti od srednje vrednosti (*SS*), parametar koji je predstavljen u deskriptivnoj statistici (poglavlje 3) i opisuje varijabilnost u skupu podataka. Ukupna varijabilnost testiranih podataka (*N*) se može predstaviti sledećim izrazom:

$$\text{Ukupna SS}=\sum_{i=1}^{k}\left[\sum_{j=1}^{ni}(X_{ij}-\overline{X}\quad\right]^2$$

Gde je Xij, merenje j u grupi (lokalitetu) i dok je $\overline{X}$ srednja vrednost svih merenja N. Stepeni slobode vezani za ukupnu varijabilnost podataka se računaju kao:

$$\text{Ukupna DF= N} - 1$$

Na primeru koncentracije rastvorenog kiseonika u tri različite reke Sliva Južne Morave, ukupni broj stepena slobode iznosi 30-1=29.

Od ukupne varijabilnosti u skupu podataka N, deo se odnosi na varijabilnost između testiranih grupa, definisanu kao razliku između srednjih vrednosti k grupa:

$$\text{Međugrupna SS}= \sum_{i=1}^{k} ni(\overline{X}\,\text{i} - \overline{X})^2$$

Gde je $\overline{X}$ i srednja vrednost uzorka n_i, $\overline{X}$ srednja vrednost svih podataka iz *N* skupa. Broj stepena slobode za međugrupnu varijaiblnost se računa kao:

$$\textit{Međugrupna DF = k-1}$$

Na primeru koncentracije rastovrenog kisenoika u vodi, međugrupna *DF* iznosi 3-1=2

Deo ukupne varijabilnosti koja nije opisana razlikama između grupa je varijabilnost unutar grupa:

$$\text{Unutargrupna SS} = \sum_{i=1}^{k} \left[\sum_{j=1}^{ni}(X_{ij} - \overline{X}_{\,i}\right]^2$$

Ukoliko se međugrupna i unutargrupna *SS* podele odgovarajućim *DF*, dobija se parametar srednje kvadratno odstupanje od srednje vrednosti (*MS*), neophodan za računanje statistike koju koristi *ANOVA*:

$$\text{Međugrupna MS} = \frac{međugrupna\ SS}{međugrupna\ DF};\ \text{Unutargrupna MS} = \frac{unutargrupna\ SS}{unutargrupna\ DF}$$

Za testiranje hipoteza za tri i više groupa jednofaktorska *ANOVA* kotisti *F* statistiku koja ja kao i u slučaju *t* testa bazirana na poređenju varijabilnosti između grupa (koja je objašnjena faktorskom

promenljivom) i varijabilnosti unutar groupa (koja je okarakterisana kao prirodna varijabilnost i posledica je slučajnosti):

$$F=\frac{me]ugrupna\ MS}{unutargrupna\ MS}$$

Kada se srednje vrednosti testiranih grupa ($\overline{X}_i$) značajno razlikuju varijaiblnost između grupa je veća od unutargupne varijaiblnosti, pa je vrednost *F* statistike veća od 1. U slučaju da ne postoji značajna razlika između grupa, unutargrupna varijabilnost će biti veća od međugrupne te je *F* vrednost manja od 0. Za testiranje nulte hipoteze važno je definisati F distribuciju koja predstavlja kontinuiranu distribuciju verovatnća i opisana je sa dva parametra međugrupni *DF* (*df1*) i unutargrupni *DF* (*df2*) odakle se računa kritična vrednost *Fα(df1)(df2)*.

Odbacivanje nulte hipoteze (H0: μ1=μ2=μ3) i prihvatanje HA hipoteze ne znači da se srednje vrednosti sva tri uzorka statistički značajno razlikuju. Na primer, ukoliko se odbaci H0 hipoteza koja pretpostavlja da se koncentracija rastvorenog kiseonika u vodi ne razlikuje značajno, i dalje je nepoznato između kojih uzoraka (reka) postoji značajna razlika u nivou rastvorenog kisenika u vodi. Da bi se ustanovilo između kojih uzoraka postoji statistčki značajna razlika srednjih vrednosti potrebno je sprovesti naknadni test (eng. *post-hoc* test). Na početku poglavlja je objašnjeno da bi višestruka primena *t* test za dva uzorka prilikom testiranja svakog mogućeg para uzoraka vodila ka značajnom porastu greške prvog reda. Zbog toga je razvijena grupa metoda za višestruko poređenje uzoraka od kojih se *Tukey HSD* test najčešće korist. Ova metoda računa novu kritičnu vrednost statistke *q* koja se koristi prilikom testiranja hipoteze između bilo koja dva para od *k* uzoraka. *Q* statistka je definisana sledećom formulom:

$$q=\frac{\overline{X}_A-\overline{X}_B}{SE}$$

gde je $\overline{X}_A-\overline{X}_B$ razlika srednje rednosti para testiranih uzoraka dok je *SE* količnik unutargrupne *MS* i broja merenja po uzorku:

$$SE=\sqrt{\frac{unutargrupna\ MS}{n}}$$

Nakon određivanja nove kritične vrednosti *Tukey's HSD* testa ($q_{\alpha,\sqrt{,}k,}$) razlike srednjih vrednosti svih mogućih parova se porede sa kritičnom vrednošću i ukoliko je veća vrednost *q* statistke od

kritične vrednosti, hipoteza H0: $\overline{X}\,\mathrm{A} = \overline{X}\,\mathrm{B}$ se odbacuje a između srednjih vrednosti testiranog para uzoraka postoji statistčki značajna razlika.

Pretpostake

Pretpostavke za primenu *Tukey's HSD* testa su iste kao i za *ANOVA*-u:

1. Uzorci potiču iz populacija koje imaju normalnu raspodelu
2. Homogenost varijansi, sve grupe nemaju značajno različitu varijansu populacija ($\sigma_1^2 = \sigma_2^2 = \sigma_3^2$)

Primer statističkog testiranja hipoteze ANOVA testom u R-u

Funkcija *anova_test()* primenjuje jednofaktorsku *ANOVA* u testiranju hipoteze sa tri i više uzoraka u R-u. Međutim, pre primene ove funkcije neophodno je formatirati ulazne podatke i testirati pretpostavke metode (normalnost promenljive i homogenost varijansi). U tabeli 6.1 prikazani su rezultati merenja rastvorenog kiseonika u na tri različite reke u okviru juznomoravskog sliva.

Tabela 6.1 Koncentracija rastvorenog kiseonika (O2 mg/l) izmerena na tri lokaliteta.

Merenje/lokalitet	Temska	Visočica	Veternica
1	8.91	6.67	4.21
2	8.21	7.35	3.65
3	7.12	7.56	3.41
4	7.49	9.21	3.26
5	8.22	6.45	3.65
6	8.45	6.87	5.21
7	9.00	7.77	5.23
8	7.89	8.86	4.32
9	9.12	8.43	6.12
10	6.80	8-32	4.21

Prvi korak je generisati ulaznu matricu na osnovu podataka iz tabele 6.1:

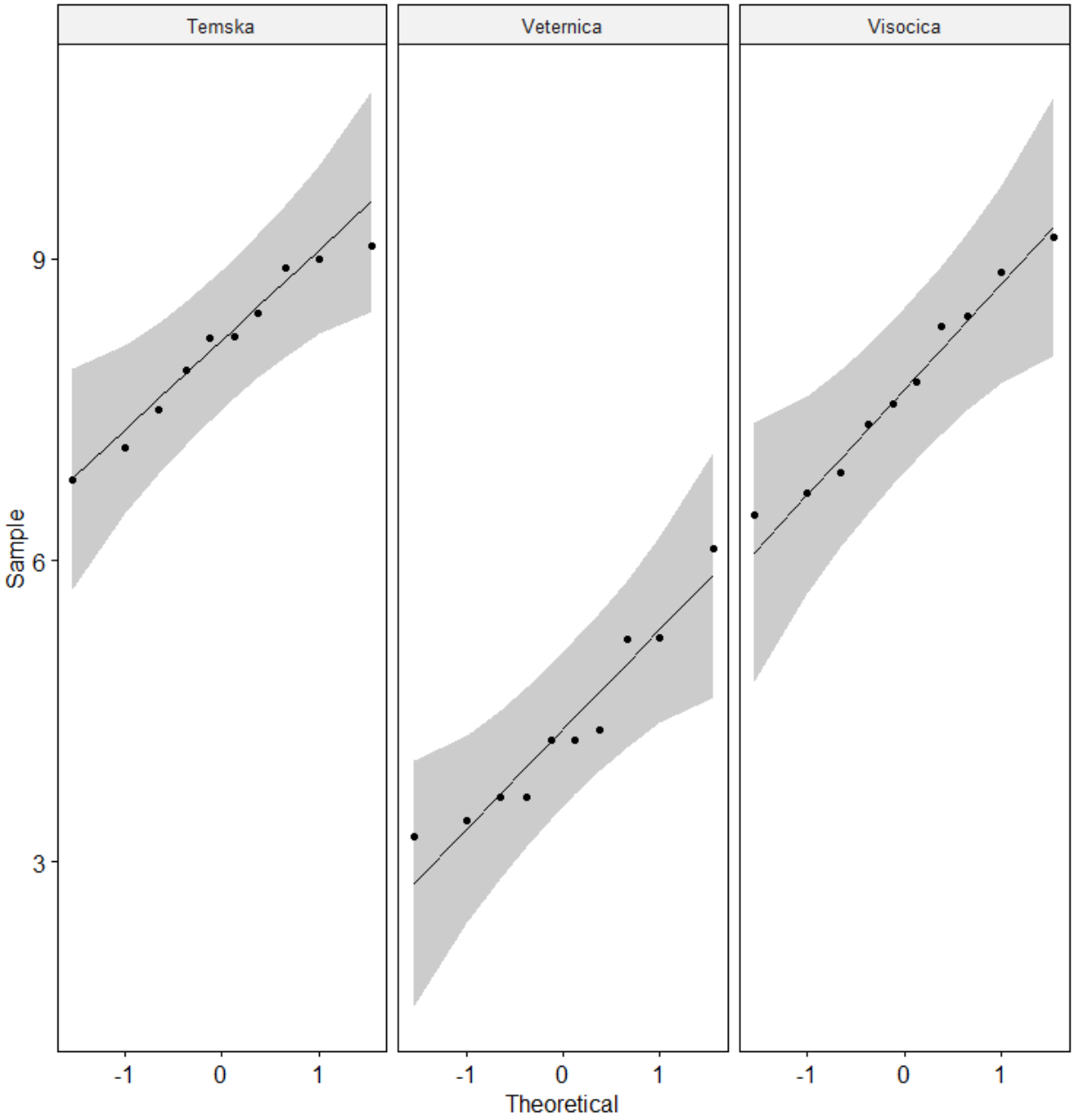
```
> Temska <- c(8.91, 8.21, 7.12, 7.49, 8.22, 8.45, 9.00, 7.89, 9.12, 6.80)

> Visocica <- c(6.67, 7.35, 7.56, 9.21, 6.45, 6.87, 7.77, 8.86, 8.43, 8.32)

> Veternica <- c(4.21, 3.65, 3.41, 3.26, 3.65, 5.21, 5.23, 4.32, 6.12, 4.21)

> my_data <- data.frame( Lokalitet = rep(c("Temska", "Visocica", "Veternica"), each = 10), koncentracija_kiseonika = c(Temska, Visocica, Veternica) )

> group_by(my_data, Lokalitet) %>% summarise(count = n(), mean = mean(koncentracija_kiseonika, na.rm = TRUE), sd = sd(koncentracija_kiseonika, na.rm = TRUE))

`summarise()` ungrouping output (override with `.groups` argument)

# A tibble: 3 x 4

  Lokalitet count  mean    sd

  <chr>     <int> <dbl> <dbl>

1 Temska       10  8.12 0.797

2 Veternica    10  4.33 0.927

3 Visocica     10  7.75 0.942
```

Pre testiranja hipoteze, neophodno je potvrditi pretpostavke za jednofaktorsku ANOVA-u, normalnost promenljive u svakoj grupi i homogenost varijansi:

```
> #provera normalnosti varijabli

> # Shapiro-Wilk normality test for Men's weights with my_data

> my_data=as_tibble(my_data)

> my_data
```

```
# A tibble: 30 x 2
   Lokalitet koncentracija_kiseonika
   <chr>                       <dbl>
 1 Temska                       8.91
 2 Temska                       8.21
 3 Temska                       7.12
 4 Temska                       7.49
 5 Temska                       8.22
 6 Temska                       8.45
 7 Temska                       9
 8 Temska                       7.89
 9 Temska                       9.12
10 Temska                       6.8
# ... with 20 more rows
> my_data%>%
+   group_by(Lokalitet)%>%
+   shapiro_test(koncentracija_kiseonika)
# A tibble: 3 x 4
  Lokalitet variable                statistic     p
  <chr>     <chr>                       <dbl> <dbl>
1 Temska    koncentracija_kiseonika     0.944 0.596
2 Veternica koncentracija_kiseonika     0.912 0.297
3 Visocica  koncentracija_kiseonika     0.959 0.772
> ggqqplot(my_data, "koncentracija_kiseonika" , facet.by =
"Lokalitet")
```

```
> #homogenost varijansi - Levene's test
> my_data%>%
+   levene_test(koncentracija_kiseonika~Lokalitet)
# A tibble: 1 x 4
    df1   df2 statistic     p
  <int> <int>     <dbl> <dbl>
1     2    27     0.211 0.811
```



Na osnovu rezultata *Shapiro-Wilikov* i *Levene*-ovog testa (p>0.05), promenljiva rastvoreni kiseonik pokazuje normalnu raspodelu dok su varijanse između grupa (reka), jednake. S obzirom da su obe pretpostake jednofaktorske *ANOVA* zadovoljene, sledeći korak je sprovođenje *ANOVA* testa pomoću funkcije *anova_test()*:

```
> #jednofaktorska ANOVA
```

```
> res.ANOVA<- my_data %>% anova_test(koncentracija_kiseonika ~
Lokalitet)

Coefficient covariances computed by hccm()

> res.ANOVA

ANOVA Table (type II tests)


     Effect DFn DFd      F        p p<.05   ges
1 Lokalitet   2  27 55.108 2.94e-10     * 0.803
```

Funkcija *anova_test()* generiše tabelu sa informacijama o F distribuciji (numerator i denominator), *F* statistici, *p* vrednosti i „ges“ parametru koji indikuje veličinu uticaja (eng. *effect size*), odnosno količinu varijabilnosti (*u* %) koju objašnjava kategorijska promenljiva (različite reke). Rezultati testa pokazuju da se srednje vrednosti grupa statistički značajno razlikuju (F=55.10, p<0.001).

S obzirom da je *F* statistika veća od kritične vrednosti, H0 hipoteza se odbacuje i prihvata hipoteza HA. Međutim, na osnovu toga se ne može zaključiti između kojih reka postoji značajna razlika u količini rastovernog kiseonika u vodi. Da bi se to utvrdilo potrebno je primeniti *posthoc* test za višestruko testiranje hipoteza. U R-u postoji funkcija *tukey_hsd()* koja može sprovesti fišestruko testiranje hipoteza između svih mogućih parova uzoraka:

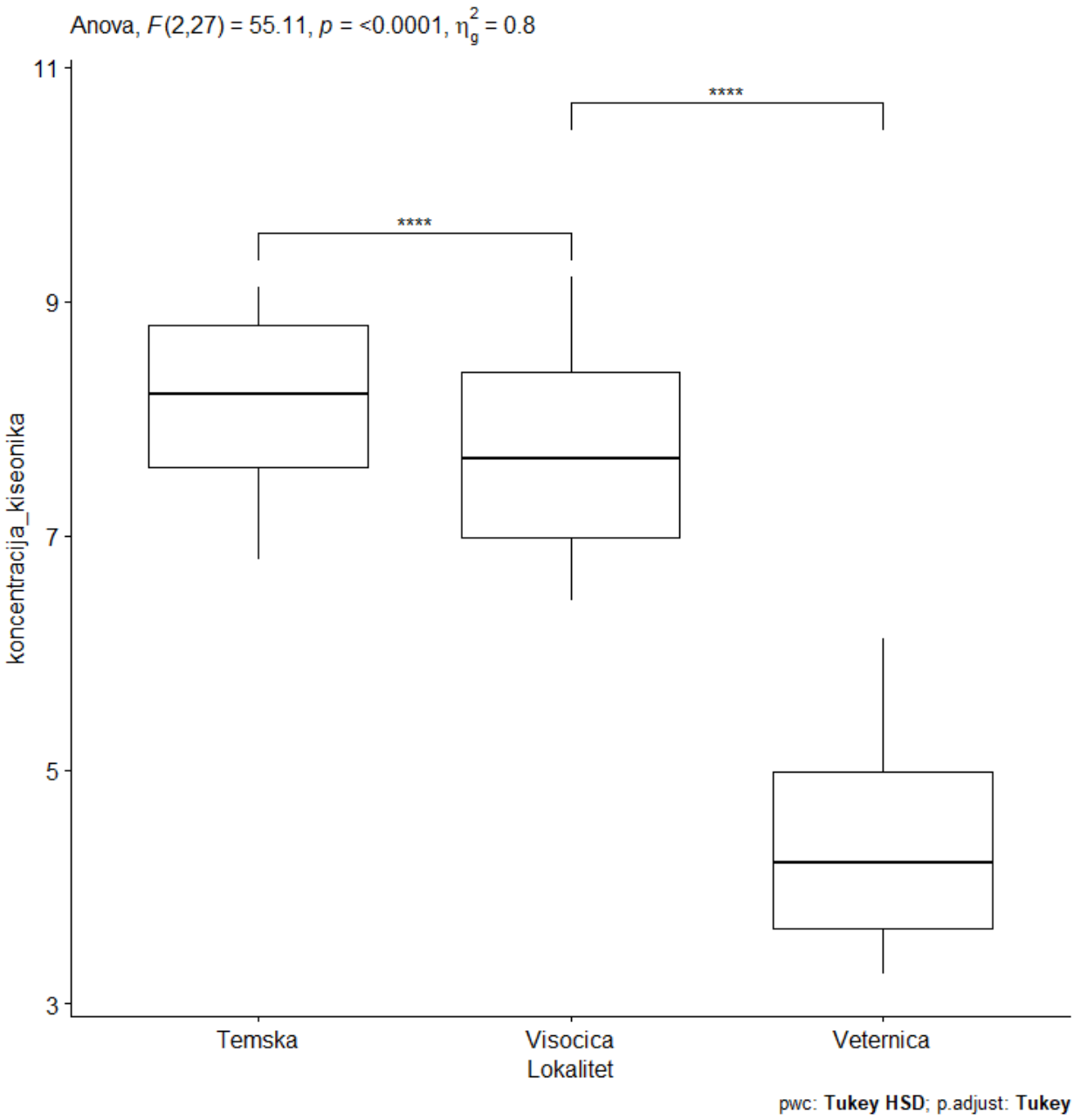
```
> #Post-hoc test

> pwc<- my_data %>%tukey_hsd(koncentracija_kiseonika ~
Lokalitet)

> pwc

# A tibble: 3 x 9

  term      group1    group2    null.value estimate conf.low
conf.high         p.adj p.adj.signif

* <chr>     <chr>     <chr>          <dbl>    <dbl>    <dbl>
<dbl>         <dbl> <chr>
```

```
1 Lokalitet Temska    Veternica          0   -3.79     -4.78
-2.81  0.00000000119 ****

2 Lokalitet Temska    Visocica           0   -0.372    -1.36
0.616 0.624         ns

3 Lokalitet Veternica Visocica           0    3.42      2.43
4.41  0.00000000987 ****
```

Izlazni podaci pokazuju informaciju o testiranim parovima, 95% intervala poverenja i modifikovanoj *p* vrednosti. Na osnovu rezultata višestrukog testiranja hipoteze, rastvoreni kisenik u vodi se jedino ne razlikuje značajno između lokaliteta na Temskoj i Veternici, dok je u svim ostalim parovima detektovana statistčka značajnost koja potvrđuje razliku između izmerenih koncentracija O_2 u vodi.

Konačno, rezultati jednofaktorske *ANOVA* se mogu i grafički prikazati preko boks dijagrama sledećim R kodom:

```
#graficki prikaz rezultata

pwc<- pwc %>% add_xy_position(x= "Lokalitet")

ggboxplot(my_data, x="Lokalitet", y="koncentracija_kiseonika") +

  stat_pvalue_manual(pwc, hide.ns = TRUE)+

  labs(subtitle = get_test_label(res.ANOVA, detailed= TRUE),
caption = get_pwc_label(pwc))
```


Anova, F(2,27) = 55.11, p = <0.0001, η²g = 0.8
11
9
7
5
3
koncentracija_kiseonika
****
****
Temska
Visocica
Veternica
Lokalitet
pwc: Tukey HSD; p.adjust: Tukey

6.2 Neparametarsko testiranje Hipoteze sa više uzoraka

Teorijska osnova

Kada je testirani uzorak takav da pretpostavke *ANOVA* testa (normalna raspodela uzoraka i homogenost varijansi, videti poglavlje 6.1) nisu zadovoljene onda se hipoteza testira neparametarskom alternativom analize varijansi, ***Kruskal-Wallis-ov* testom**,

Kao i u slučaju neparametarskih testova za testoranje hipozeze za dva uzorka, neparametarska *ANOVA, Kruskal-Wallis-ov* test ne koristi populacione parametre u formulaciji hipoteze, kao ni parametre uzorka. *Kruskal-Walliso-va* test statistka *H* se računa kao:

$$\mathrm{H}=\frac{12}{N(N+1)}\sum_{i=1}^{k}\frac{R_i^2}{n_i}-3(N+1)$$

Gde je n_i broj merenja u uzorku *i*, $\mathrm{N}=\sum_{i=1}^{k} n_i$ je ukupan broj merenja u svim grupama *k*, *Ri* je suma rangova *ni* merenja u uzorku *i*. Procedura rangiranja vrednosti je prikazana u poglavlju 11 (*Mann-Whitney* test). Dobar način da se proveri da li su rangovi dodeljeni na ispravan način je da se proveri da li suma je svih rankova jednaka *N(N+1)/2*.

Kao i u slučaju *ANOVA* testa, i u okviru neparametarske varijante je moguće izračunati veličinu uticaja pomoću eta^2 parametra. Ovaj parametar definiše procenat varijabilnsti zavisne promenljive koji je objašnjen nezavisnom promenljivom (kategorijskom promenljivom, brojem grupa). Eta^2 se računa sledećom formulom:

$$\mathrm{Eta}^2=\frac{(\mathrm{H}-\mathrm{k}+1)}{(n-K)}$$

Gde je *H* statistika *KS* testa, *k* broj grupa i *n* ukupan broj merenja. Vrednost Eta^2 se kreće između 0 i 1 a vrednosti se mogu interpretirati na sledeći način: 0.01-0.06 (mali efekat), 0.06-0.14 (srednji efekat) i >0.14 (veliki efekat).

Kako bi se ustanovilo između kojih parova grupa postoji značjan razlika testirane promenljive, naknadni test (eng. *post-hoc* test) za višestruko testiranje se sprovodi *Mann-Whitney* test-om. Zbog višestrukog testiranja, kako bi se izbegla greška prvog tipa, potrebno je sprovesti **Bonferonijevu korekciju alfa vrednosti**. Bonefornijevo prilagođavanje znači da se alfa brednost 0.05 podeli

brojem testova koji je planiran priliko naknadnog testiranja i potom upotrebiti tako revidirani alfa nivo kao kriterijum za testirnjae hipoteza.

Pretpostake

Pretpostavke za korišćenje neparamterskog *ANOVA* testa su iste kao i za *Mann-Whitney* test i ne oslanjaju se na parametre populacija (μ i σ):

1. Slučajnost uzorka
2. Nezavisnost opservacija

Primer statističkog testiranja hipoteze Kruskal-Wallisov testom u R-u

U poglavlju 5 je *Mann-Whitney* testom testiran toksični efekat cijanotoksina Mikrocistin LR upoređivanjem stepena oštećenja DNK između kontrole i tretmana. Međutim u ekotoksikologiji je eksperimentalni dizajn takav da često postoji više tretmana, a samim tim i više grupa. Na primer, toksični efekat cijanotoksina Mikrocistin LR na larve familije hironomida je ispitivan preko dve koncentracije toksina, sredisnke relevantne i deset puta povećane u odnosu na sredinsku. Kao biomarker je korišćen nivo oštećenosti DNK lanca (dužina repa DNK) čije su vrednosti prdstavljene u tabeli6.1.

Konrola (H_2O)	Tretman (Mikrocistin LR 1mg/l)	Tretman (Mikrocistin LR 10mg/l)
1.8434	24.4429	28.424
19.3141	25.1643	1.234
37.3159	33.4689	45.345
1.16004	34.9058	32.345
1.68472	29.0558	45.345
5.14486	30.1883	13.463
1.67865	39.1202	28.345
0.81668	2.04586	34.234
0.85719	23.9191	2.345
1.2763	12.9059	45.567

Prvi korak je generisti tabelu koja sadrži zavisnu promenljivu (Dužina repa DNK) i nezavisnu promeljuvu (grupna varijabla-tretmani)

```
> library(tidyverse)
> library(ggpubr)
> library(rstatix)
> Kontrola <- c(1.843399, 19.314072, 37.315882, 1.160038, 1.684717, 5.144862, 1.678647, 0.816683, 0.857185, 1.276296)
> Cijanotoksin_env <- c(24.442907, 25.164253, 33.468926, 34.905848, 29.055759, 30.188337, 39.120239, 2.045863, 23.919067, 12.905896)
> Cijanotoksin_10Xenv <- c(28.424, 1.234, 45.345, 32.345, 45.345, 13.463, 28.345, 34.234, 2.345, 45.567)
> my_data <- data.frame( Tretman = rep(c("Kontrola", "Cijanotoksin_env", "Cijanotoksin_10Xenv"), each = 10), Duzina_repa = c(Kontrola, Cijanotoksin_env, Cijanotoksin_10Xenv) )
> group_by(my_data, Tretman) %>% summarise(count = n(), mean = mean(Duzina_repa, na.rm = TRUE), sd = sd(Duzina_repa, na.rm = TRUE), M = median(Duzina_repa, na.rm = TRUE) )
`summarise()` ungrouping output (override with `.groups` argument)
# A tibble: 3 x 5
  Tretman             count  mean    sd     M
  <chr>               <int> <dbl> <dbl> <dbl>
1 Cijanotoksin_10Xenv    10 27.7   16.8 30.4
2 Cijanotoksin_env       10 25.5   11.0 27.1
3 Kontrola               10  7.11  12.0  1.68
> my_data=as_tibble(my_data)
> my_data
# A tibble: 30 x 2
   Tretman  Duzina_repa
   <chr>          <dbl>
 1 Kontrola       1.84
 2 Kontrola      19.3
 3 Kontrola      37.3
```

```
 4 Kontrola       1.16
 5 Kontrola       1.68
 6 Kontrola       5.14
 7 Kontrola       1.68
 8 Kontrola       0.817
 9 Kontrola       0.857
10 Kontrola       1.28
# ... with 20 more rows
```

Kako bi se ustanovilo da li promenljiva u okviru grupa zadovoljava pretpostavke parametarskih testova, moguće je vizuelizovati varijabilnost Dužine repa DNK box plotovima.

```
> #vizuelizacija
> ggboxplot(my_data, x= "Tretman", y="Duzina_repa")
```

S obzirom da normalna distribucija i homogenost varijabli kao preduslov nije zadovoljen, značajna razlika u dužini repa DNK između kontrole i tretmana se može testirati neparametarskom alternativom pomoću funkcije *test.kruskal()*:

```
> #Kruskal-Wallis test
> test.kruskal <- my_data %>% kruskal_test(Duzina_repa ~ Tretman)
> test.kruskal
# A tibble: 1 x 6
  .y.             n statistic    df       p method
* <chr>       <int>     <dbl> <int>   <dbl> <chr>
1 Duzina_repa    30      10.1     2 0.00632 Kruskal-Wallis
```

Rezultati testa pokazuju da se dužina repa DNK značajno razlikuje (p=0.00632) između larvi kontrole i tretmana.

Veličina uticaja tretmana se može testirati funkcijom *kruskal_effsize()*:

```
> #effect size
> my_data %>% kruskal_effsize(Duzina_repa ~ Tretman)
# A tibble: 1 x 5
  .y.             n effsize method  magnitude
* <chr>       <int>   <dbl> <chr>   <ord>
1 Duzina_repa    30   0.301 eta2[H] large
```

Vrednost parametra *eta2* koji kvantifikuje veličinu uticaja grupa na zavisnu promenljivu iznosi *eta2*=0.301, što se tumači kao veliki uticaj (vide poglavlje 6.1.1)

Kako bi se ustanovilo između kojih grupa postoje značajne razlike u dužini repa DNK, potrebno je sprovesti dodatno (*post-hoc*) testiranje *Mann-Whitney* testom, pomoću funkcije *wilcox.test()* gde se prilagođena alfa vrednost definiše argumentom *p.adjust.method*

```
> #post-hoc test
> pht <- my_data %>%
+       wilcox_test(Duzina_repa  ~  Tretman,  p.adjust.method  = "bonferroni" )
> pht
# A tibble: 3 x 9
  .y.          group1               group2              n1    n2 statistic     p p.adj p.adj.signif
* <chr>        <chr>                <chr>            <int> <int>     <dbl> <dbl> <dbl> <chr>
1 Duzina_repa Cijanotoksin_10Xenv Cijanotoksin_env     10     10        58 0.571 1     ns
2 Duzina_repa Cijanotoksin_10Xenv Kontrola             10     10        84 0.011 0.034 *
3 Duzina_repa Cijanotoksin_env    Kontrola             10     10        88 0.003 0.009 **
```

Rezultati post-hoc testa pokazuju da se tretmani značajno razlikuju od kontrole ali da ne postoji značajna razlika dužine repa DNK između tretmana. To znači da desetostruka vrednost sredinski relevantne koncentracije mikrocistina LR ne izaziva značajno povećanje subletalnog efekta na larvama hironomida.

7 Korelacija i regresija

7.1 Pearson-ov koeficijent korelacije

Teorijska osnova

Linearna korelacija predstavlja linearnu vezu između dve promenljive, pod pretpostavkom da funkcionalna zavisnost među njima ne postoji. Iako se promenjive uglavnom obeležavaju sa *X* i *Y*, kao i kod regresione analize, u korelaciji među njima nema zavisnosti tako da se isti rezultat dobija bez obzira na to koja promenljiva je označena sa X a koja sa Y. Pearson-ov koeficijent korelacije r se računa pomoću sledeće formule:

$$r = \frac{\sum xy}{\sqrt{\sum x^2 * \sum y^2}}$$

Gde su $\sum x^2$, $\sum y^2$ sume kvadrata, $\sum x^2 = \sum (Xi-X)^2$ i $\sum xy$ suma odstupanja od srednjih vrednosti:

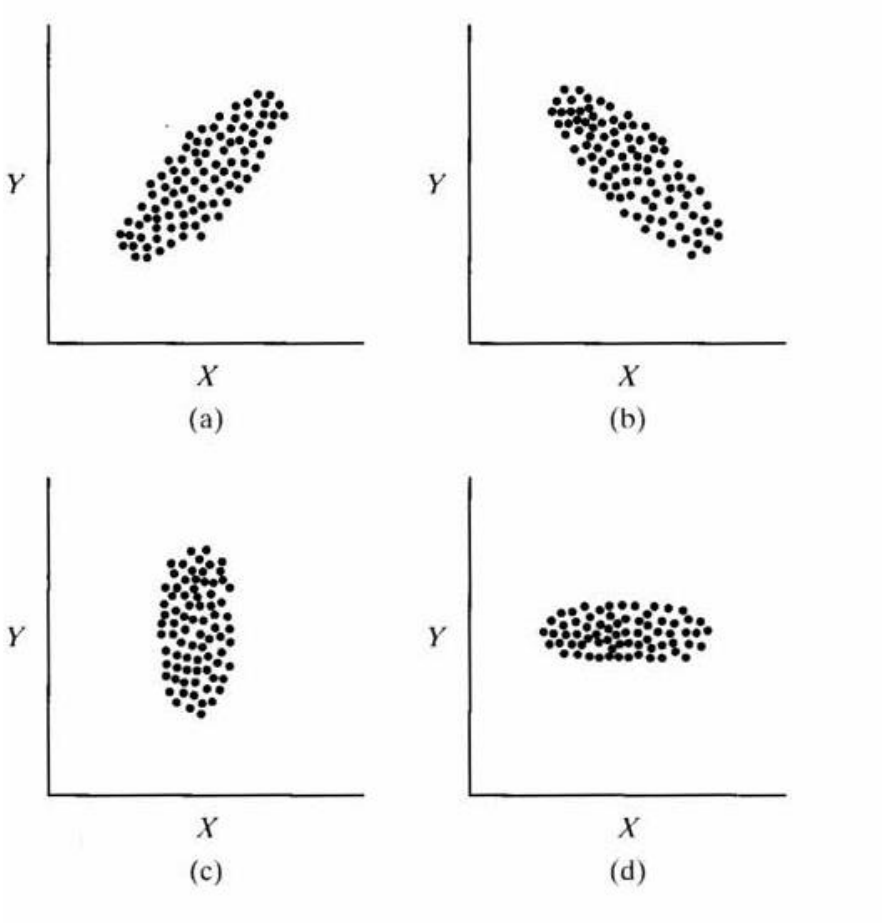
$\sum xy = \sum (Xi-X)(Yi-Y)$

Pearson-ov koeficijent korelacije se kreće u intervalu od -1 do 1.

(-1) ukazuje na jaku negativnu korelaciju. To znači da svaki put kada se X poveća, Y se smanjuje (Slika 7.1b)

(0) znači da ne postoji veza između dve promenljive (*X* i *Y*) (Slika 7.1c i d)

(1) ukazuje na jaku pozitivnu korelaciju. To znači da se *Y* povećava sa povečanjem *X* (Slika 7.1a)



Slika 7.1. Prikaz korelacije dve varijable (*X, Y*): (a) Pozitivna korelacija. (b) Negativna korelacija. (c) Nema korelacije. (d) Nema korelacije.

Pretpostavke

Iako ne postoje preduslovi koje treba ispoštovati da bi se izračunao koeficijent korelacije, postoje određene pretpostavke koje moraju biti zadovoljene prilikom testiranja hipoteza o i određivanja intervala poverenja za koeficijent korelacije. Kako bi se primenio *Pearson*-ov koeficijent korelacije neophodno je da podaci zadovoljavalju slede pretpostavke:

1. *X* i *Y* vrednosti potiču iz populacije sa normalnom raspodelom
2. Obe promenljive koje se koriste u analizi imaju noromalnu raspodelu (pretpostavka bivarijantne normalnosti)

Primer Pearson-ov koeficijent korelacije u R-u

U ekološkim istraživanjima se često se ispituje uticaj sredinskih parametara na strukturu hidrobiocenoze (zajednice). U tom slučaju uglavnom ispituje korelisanost odabranog sredinskog faktora na sastav i strukturu zajednice predstavljenu odgovarajućom metričkom osobinom (indeksom). Takođe, ne retko se dešava da su sredinski paramentri međusobno povezani i uslovljeni, pa se kao takvi neki od njih iskuljučuju iz daljih analiza uticaja na zajednicu ukoliko se ispostavi da pokazuju veliki stepen korelacije. S toga, jedna od prvih analiza koja se sprovodi je ispitivanje korelisanosti između promenljivih.

Baza podataka „Tabela_Nisava“ sadrži informacije o brojnosti vrsta makrobeskičmenjaka i sredindinskim parametrima.

Želimo da ispitamo korelisanost između dve sredinske preomeljive: koncentracija kiseonika (*omg*) i nadmorska visina (*nad_vis*). Nakon učitavanja podataka prvo treba ispitati hipotezu o normalnosti promenljivih koje su uključene u analizu.

```
> library(ggplot2)
> library(dplyr)
> library(ggpubr)
> Table_Nisava=read.csv("Tabela.csv")
> Table_env=Table_Nisava[,115:127]
> Table_env

#Normality test
shapiro.test(Table_env$omg)
shapiro.test(Table_env$nad_vis)

Rezultat:
Shapiro-Wilk normality test

data:  Table_env$nad_vis
W = 0.93342, p-value = 0.1797
```

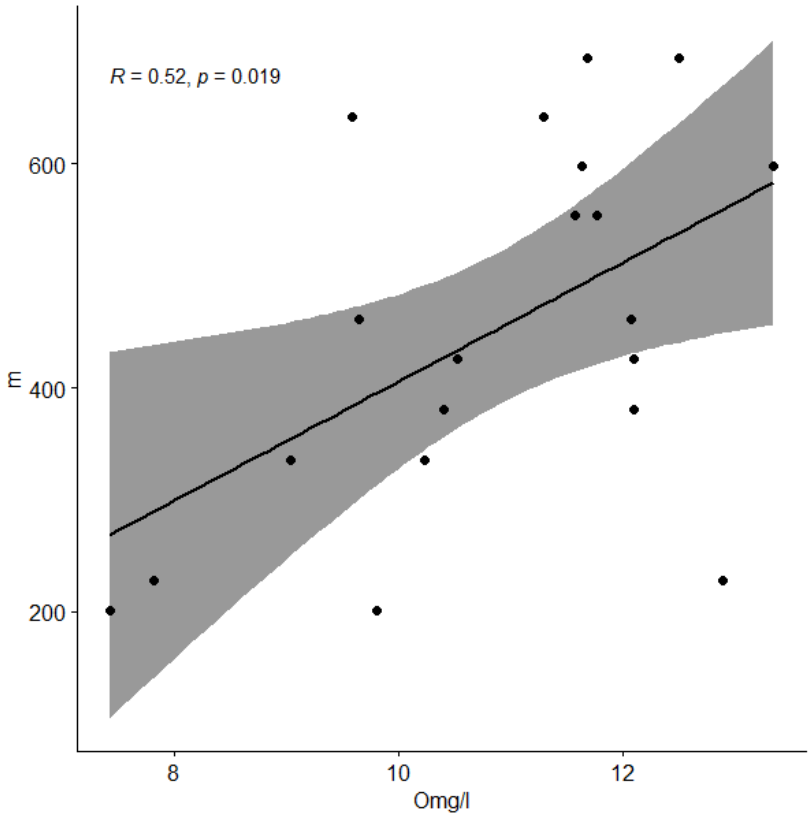
```
> shapiro.test(Table_env$omg)

Shapiro-Wilk normality test

data:  Table_env$omg
W = 0.94745, p-value = 0.3301
```

Ukoliko su obe *p* vrednosti veće od intervala poverenja (0.05), smatra se da raspodela podataka nije statistički značajno različita od normalne raspodele, odnosno važi pretpostavka da oba obeležja imaju normalnu raspodelu. Na osnovu rezultata testa normalnosti u ovom primeru zaključujemo da oba obeležja imaju normalnu raspodelu što znači da za ispitivanje linearne korelacije među njima koristimo *Pearsonov* koeficijen korelacije.

```
res <- cor.test(Table_env$omg, Table_env$nad_vis, method = "pearson")
res

Rezultat:
Pearson's product-moment correlation

data:  Table_env$omg and Table_env$nad_vis
t = 2.5845, df = 18, p-value = 0.0187
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1009702 0.7825973
sample estimates:
      cor
0.5202457
```

Takođe, uz pomoć funkcije *ggscatter* moguće je grafički predstaviti zavisnost između dve promenljive (Slika 7.2).

```
ggscatter(Table_env, x = "omg", y = "nad_vis", add = "reg.line",
conf.int = TRUE,

          cor.coef = TRUE, cor.method = "pearson",

          xlab = "Omg/l", ylab = "m")
```



Slika 7.2. Korelacija nadmorske visine i koncentracije rastvorenog kiseonika

7.2 Spearman-ov koeficijent korelacije ranga

Ukoliko imamo skupove podataka čija obeležja nemaju normalnu raspodelu, onda *Pearson*-ov koeficijent korelacije nije primenljiv. U tom slučaju se ispitivanje povezanosti dve promenljive koristi *Spearman-ov* koeficijent korelacije, koji je neparametrijska mera korelacije ranga. Prvi korak prilikom izračunavanja je rangiranje svake mere promenljive. *Spearmano*-vim

koeficijentom r_s (rho_s) se meri snaga i pravac veze između dve rangirane promenljive i izračunava se po sledećoj formuli:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} di^2}{n^3 - n},$$

Gde je d_i razlika između X i Y ranga: $d_i = rank\ of\ X_i - rang\ Y_i$

Vrednosti *Spearman*-ovog koeficijenta r_s (rho_s) kreću se u rasponu od -1 do +1.

Primer *Spearman*-ovog testa u R-u

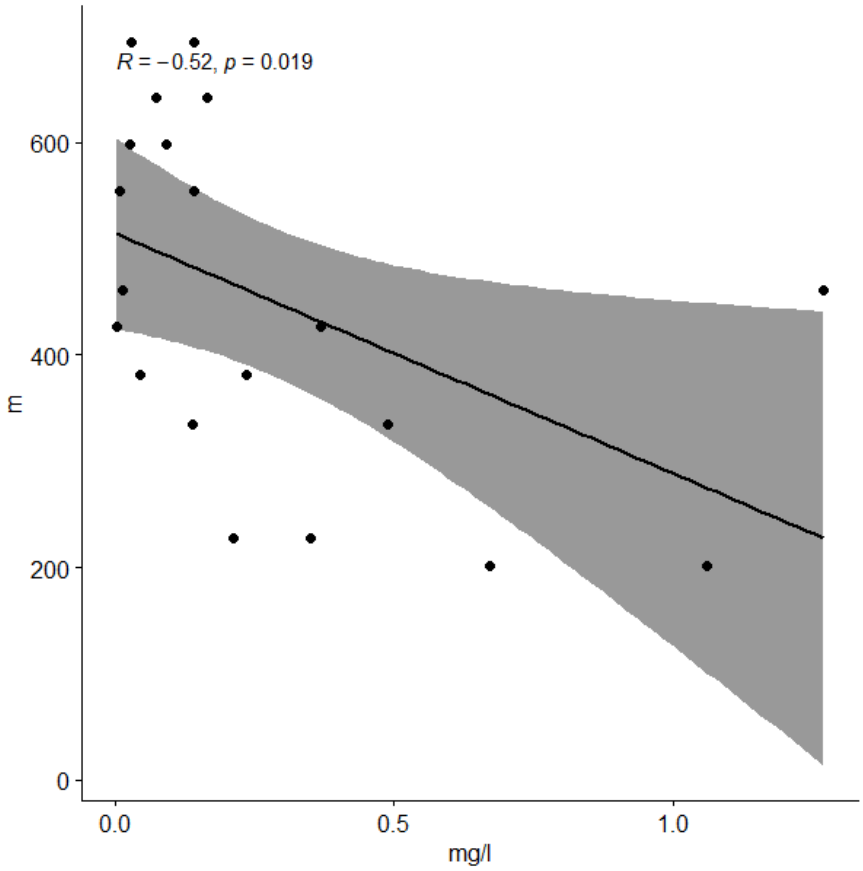
Ukoliko makar jedno od obeležja, čija se korelacija ispituje, nema normalnu raspodelu pribegava se računanju *Spearman*-ovog koefinijenta korelacije ranga. U ovom primeru ispituje se korelisanost između koncentracije ortofosfata u vodi (*po*) i nadmorske visine (*nad_vis*).

```
Normality test
shapiro.test(Table_env$po)
shapiro.test(Table_env$nad_vis)

Rezultat:
Shapiro-Wilk normality test

data:  Table_env$nad_vis
W = 0.93342, p-value = 0.1797

Shapiro-Wilk normality test

data:  Table_env$po
W = 0.74678, p-value = 0.0001537

res2 <-cor.test(Table_env$po, Table_env$nad_vis, method = "spearman", exact = FALSE)
res2
```

```
Rezultat:
Spearman's rank correlation rho

data:  Table_env$po and Table_env$nad_vis
S = 2018.9, p-value = 0.01932
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.5179343
```

Isto kao i u prethodnom primeru, uz pomoć funkcije *ggscatter* moguće je grafički predstaviti zavisnost između dve promenljive (Slika 7.3).

```
ggscatter(Table_env, x = "po", y = "nad_vis", add = "reg.line",
conf.int = TRUE,

          cor.coef = TRUE, cor.method = "spearman",

          xlab = "mg/l", ylab = "m")
```



Slika 7.3. Korelacija izme]u nadmorske visine i koncentracije ortofosfata u vodi

7.2 Regresija

Teorijska osnova

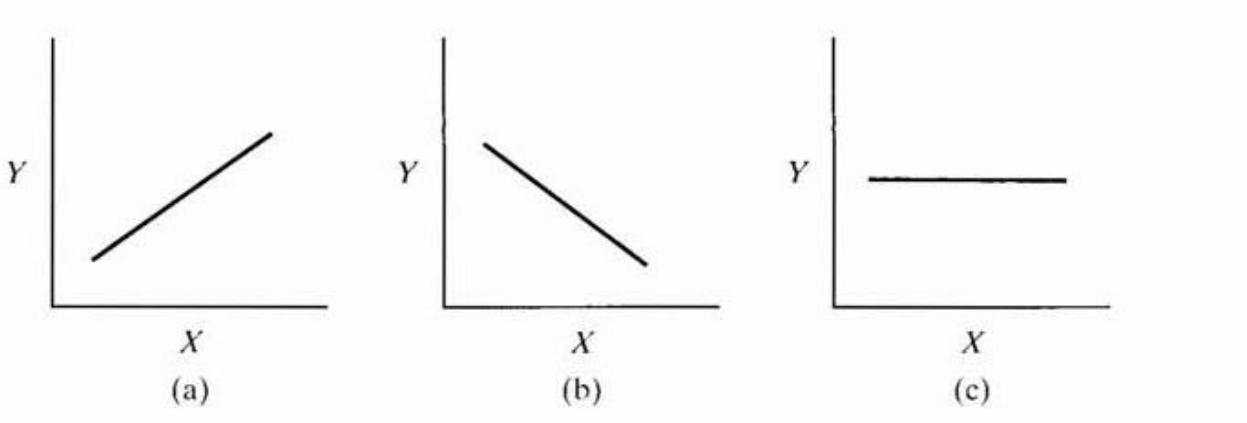
Linearnom regresijom opisuje se odnos između dve promenljive, za koje je karakteristično da svakom jediničnom porastu vrednosti jedne promenljive (nezavisna promenljiva) odgovara približno jednaka linearna promena druge (zavisna promenljiva). Nezavisna promenljiva (X) se drugačije može nazvati i prediktor jer se koristi prilikom predviđanja vrednosti zavisne promenljive ali i objašnjavajuća (eksplanatorna promenljiva) jer se na nju pozivamo prilikom objašnjavanja nekih pojava ili rezultata. Zavisna promenljiva (Y) se može nazvati i odgovor jer se menja kao odgovor na promenu vrednosti nezavisne promenljive ili ishod, jer predstavlja krajnji ishod koje želimo izmeriti. Prosta linearna regresija ima za cilj opisivanje i kvantifikovanje veze između dva obeležja i predstavlja najjednostavniji funkcionalni odnos jedne promenljive prema drugoj u populaciji:

$$Yi=\alpha + \beta Xi,$$

Gde su α i β parametri populacije (a samim tim i konstante) a sam matematički izraz predstavlja opštu jednačinu za pravu liniju. Parameter α zove se *intercept* (odsečak) a β *slope* (nagib) (Slika 7.4).

$$Yi = \alpha + \beta Xi + \varepsilon i,$$

Gde εi naziva "greška" ili "rezidual" i predsavlja slučajnu promenljivu koja dodaje šum u linearnu relaciju između zavisne i nezavisne promenljive. Kada je odnos među promenljivim funkcionalan, svaka vrednost promenljive εi je jednaka nuli, što znači da sve tačke sa koordinatama (xi, yi), i = 1,2,...,n leže na istom pravcu.



Slika 7.4. Nagib regresione krive (a) pozitivan, (b) negativan, (c) nula

Jednačina pravca određena je ako su poznati parametri a i b. Parametar b je regresioni koeficijent koji pokazuje za koliko se u proseku menja vrijednost zavisne promenljive Y za jediničnu promenu vrednosti nezavisne promenljive X. Parametar a je konstanta ili intercept i predstavlja tačku na Y osi kada je $X=0$.

$$b = \frac{\sum xy}{\sum x^2} = \frac{\sum (Xi - \bar{X})\,(Yi - \bar{Y})}{\sum (Xi - \bar{X})^2} = \frac{\sum XiYi - \frac{\sum Xi \sum Yi}{n}}{\sum Xi^2 - \frac{(\sum Xi)^2}{n}}$$

$$a = \overline{Y - b\bar{X}}$$

Specifičan pokazatelj reprezentativnosti regresije jest koeficijent determinacije R^2 (*R squared*). Koeficijent determinacije predstavlja procenat ukupne varijabilnosti zavisne promenljive (Y) koji

se može objasniti dobijenim modelom. Model je reprezentativniji što je koeficijent determinacije bliži 1.

Prilikom definisanja regresione krive i izračunavanja regresionog koeficijenta, sledeće pretpostavke/ uslovi moraju biti zadovoljeni:

1. Svakoj vrednosti *X*, nasumično su dodeljene vrednosti *Y*, nezavisne jedna od druge.
2. Za svaku vrednost *X* u populaciji postoji normalna raspodela *Y* vrednosti.
3. Homogenost varijanse u populaciji
4. Postoji linearna veza između promnenljivih *X* i*Y*
5. Vrednosti promenljive X su dobijene bez greške

Primer regresije u R-u

Jednačinom linearne regresije mogu se prikazati dužisko-maseni odnosi kod riba. Dužinsko-maseni odnosi su veoma korisni za razumevanje bioloških promena u ribljem fondu, i imaju značajnu ulogu u predviđanju stanja, reproduktivne istorije, istorije života ribljih vrsta, kao i morfoloških odnosa vrsta i populacija. Kao podaci koriste se podaci o totalnoj dužini i masi jednki različitog uzrasta. Dužinsko-maseni odnos utvrđuje se je prema eksponencijalnoj funkciji $W = aL^b$, koja u transformisanoj logaritamskoj formi ima oblik:

$$LogW = Loga + bLogL \ ,$$

gde je W = težina u gramima, L = totalna dužina jedinke u centimetrima, a = regresiona konstanta, b = koeficijent regresije, odnosno

$$y = a + bx \ ,$$

što predstavlja jednačinu linearne regresije, gde je:

y – zavisna promenljiva, x - nezavisna promenljiva, a – regresiona konstanta, b – koeficijent regresije.

Koeficijent regresije „b“ određuje nagib regresione prave. Sa aspekta dužinsko-masenih odnosa, koeficijent b može imati sledeće vrednosti:

- b=3, izometrijski rast
- b>3, pozitivan alometrijski
- b<3, negativan alometrijski

Kada riba zadrži isti oblik, njena specifična težina ostaje nepromenjena tokom njenog životnog veka i koeficijent b bi tada bio tačno 3.0 ukazujući na izometrijski rast. Vrednost znatno veća ili manja vrednost od 3.0 ukazuje na alometrijski rast. Vrednost manja od 3.0 pokazuje da riba postaje

lakša (negativni alometrijski rast) dok veća od 3.0 pokazuje da riba postaje teža (pozitivni alometrijski rast) za određenu dužinu.

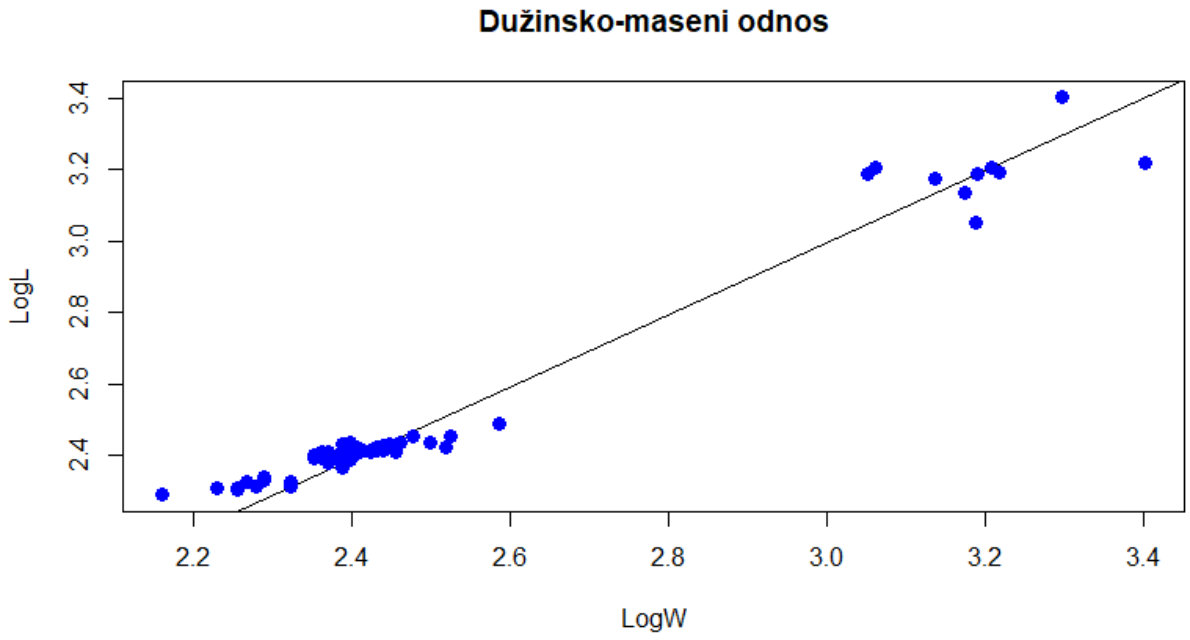
```
library("broom")
Regresija=read_csv("Tabela_reg.csv")
Regresija
# Apply the lm() function.
relation <- lm(Regresija, formula = Regresija$LogW~Regresija$LogL)
glance(relation)


Rezultat:
r.squared adj.r.squared  sigma statistic  p.value
      <dbl>         <dbl>  <dbl>     <dbl>    <dbl>
1     0.966         0.966 0.0541      958. 7.44e-52



print(relation)
Rezultat:
lm(formula = Regresija$LogW ~ Regresija$LogL, data = Regresija)

Coefficients:
   (Intercept)  Regresija$LogL
        -0.03171         1.00954
```

Vrednost koeficijenta I u ovom primeru iznosi 1.00954 što znači da jedinke teže izometrijskom rastu. Izračunavanjem koeficijenata a i b može se napisati jednačina dužinsko-masenih odnosa ispitivane populacije riba:

$$LogW = -0.03171 + 1.00954LogL,$$

Regresiona kriva koja opisuje ispitivani odnos (Slika 7.5), može se nacrtati na sledeći način:

```
# Give the chart file a name.
 png(file = "linearregression.png")
# Plot the chart.
 plot(Regresija$LogW,Regresija$LogL,col    =    "blue",main    =
"Dužinsko-maseni odnos",
     abline(lm(Regresija$LogW~Regresija$LogL)),cex  =  1.3,pch  =
16,xlab = "LogW",ylab = "LogL")
```



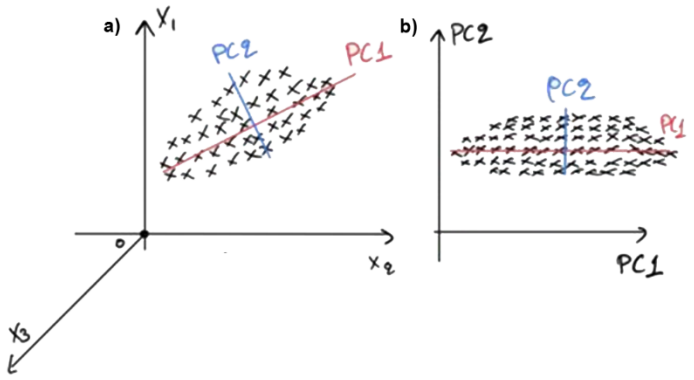
Slika 7.5. Regresiona kriva dužinsko-masenog odnosa kod riba

8. Multivarijantne tehnike u ekologiji – Analiza glavnih komponenti (PCA)

U ekologiji multivarijantni set podataka predstavlja skup uzoraka (lokaliteta) raspoređenih u prostoru u okviru kojih je uzorkovano/mereno više promenljivih (pogledati poglavlje 3.2). U takvom skupu podataka svaka promenljiva predstavlja jednu dimenziju, pa broj je izmerenih promeljivih jednak broju dimenzija. Na primer, tokom hidorbiološke studije najčešće se istovremeno uzorkuju biotički parametri (različite hidrobiocenoze) i mere abitički parametri (hidromorfologija staništa i fiziko-hemija vode). Kao rezultat konstruiše se tabela sa multivarijantnim podacima gde redovi predstavljaju uzorke/lokalitete (*n*), a kolone identifikovane vrste i izmerene abiotičke parametre (*x*). Broj identifikovanih vrsta i abiotičkih parametara predstavlja dimenzionalnost datog seta podataka.

Kako bi se analizirala struktura podataka, najefikasniji način je preko dvodimenzionalnog grafikona rasipanja gde ose predstavljaju dimenzije (promenljive), a tačke na grafikonu uzorke (lokalitete). S obzirom da ekološki podaci najčešće sadrže više od dve promelnljive, bilo bi neefikasno analizirati svaki mogući par promenljivih preko dvodimenzionalnog grafikona rasipanja. Na primer, ukoliko zajednica riba u jednoj studiji broji 10 vrsta, ukupan broj parova iznosi 10x9/2=45, odnosno 45 grafikona rasipanja. Iz takve serije grafikona ne bi bilo moguće otkriti glavne trendove u varijabilnosti podataka kao ni definisati odnose izmešu ispitivanih promenljivih (vrsta riba u navedenom primeru).

Kao rešenje u analizi multivarijantnih skupova podataka u ekologiji se predlaže redukcija dimenzionalnosti kao tehnika koja uspešno analizira multidimenzionalne podatke. Set podataka sa *x* varijabli (vrste ili sredinski parametri) i *n* slučajeva (uzorci/lokaliteti) se u prostoru može prikazati u vidu klastera *n* tačaka u *x*-dimenzionalnom prostoru. Takav klaster obično nema pravialn sferični oblik već je izdužen u nekim pravcima dok je u drugim spljošten (Slika 8.1). Pravac gde je klaster u najvećoj meri izdušen odrogava pravcu duž koga se javlja najveća varijabilnost u okviru klastera tačaka (uzoraka/lokaliteta), te predstavlja najduži gradijent u okviru datog skupa podataka. Projekcija osa na mesto pomenutih pravaca predstaljva glavni operacioni princip analize glavnih komponenti (eng *Principal Component Analyis*; *PCA*), najpopularnije vizuelizacione tehnike u multivarijantnoj statistici.



Slika 8.1 Ordinacija uzoraka u a) 3D prostoru formiranom varijablama x1-x3 i 2D prostoru formiranom glavnim komponentama (*Principal Component PC* osama)

Teorijska osnova analize glavnih komponenti

PCA kao vizuelizaciona tehnika izračunava i konstruiše ose (dimenzije) koje pokrivaju najveći deo varijabilnosti u skupu podataka i označava je kao prvu glavnu komponentu (eng *First Principal Component*, Slika 8.1b). Osa koja je ortogonalna (linearno nezavisna) u odnosu na prvu osu je označena kao druga glavna osa, gde je konačan broj PC osa odgovara broju promeljivih u skupu podataka (*x*). S obzirom da se u skupu podataka obično nalazi nekoliko glavnih trendova varijabilnosti (gradijenata), nakon projekcije osa, značajno manji broj dimenzija (najčešće prve dve) će pokrivati veći deo varijabilnosti podataka. U tome i leži glavna ideja *PCA*, redukovati broj promenljivih (dimenzija) u setu podataka, a i dalje sačuvati glavne informacije koje nose podaci. Distanca između uzoraka/lokaliteta u novom redukovanom prostoru opisanom *PC* osama (najčešće dvodimenzionalnom *PC1* i *PC2*) je relativno sličan distancama lokaliteta u početnom multidimenzionalnom prostoru opisanom *x* brojem promenljivih.

PCA definiše nove *PC* ose kao linearnu kombinaciju inicijalnih promeljivih (dimenzija). Za taj proces koristi matricu kovarijansi koje prikazuju varijansu i kovarijansu svih mogućih parova promenljivih. Na primer za trodimenzionalni set podataka koji sadrži tri pomenljive (*x*,*y* i *z*), matrica kovarijansi je matrica dimenzija 3x3:

$$\begin{matrix} Cov\,(x,x) & Cov\,(x,y) & Cov\,(x,z) \\ Cov\,(y,x) & Cov\,(y,y) & Cov\,(y,z) \\ Cov\,(z,x) & Cov\,(z,y) & Cob\,(z,z) \end{matrix}$$

Iz matrica kovarijansi se računaju karakteristični vektori (eng. *eigenvectors*) i karakteristične vrednosti (eng. *eigenvalues*) koje su neophodne za definisanje *PC* osa. Karakteristični vektori matrica kovarijansi su zapravo pravci *PC* osa u kome se nalazi najveća varijabilnost podataka. Karakteristične vrednosti predstavljaju koeficijente karakterističnih vektora koji daju informaciju o količini varijabilnosti pokrivene od strane odgoarajuće *PC* ose. Rangiranjem karakterističnih vrednosti karakterističnih vektora, od najvećeg do najmanjeg dobija se prikaz *PC* osa prema značajnosti u pogledu varijabilnosti. Kako bi se u potpunosti razumeli detalji funkcionisanja *PCA*, potrebno je dobro poznavanje linearne algebre što u ovom slučaju prevazilazi okvire ovog udzbenika. Zbog toga će na dalje biti glavni fokus na grafičkoj interpretaciji rezultata *PCA*.

Glavni rezultat analize glavnih komponenti je *PCA* ordinacioni dijagram, koji predstavlja klasični dijagram rasturanja u Dekartovom pravouglom koordinatnom sistemu gde su slučajevi (uzorci/lokaliteti) predstavljeni tačkama, a promenljive strelicama (vektorima). Ordinacija u ovom slučaju znači raspređivanje slučajeva (uzoraka/lokaliteta) u koordinatnom sistemu na osnovu sličnosti/razlike kompleksa merenih promenljivih.

Pretpostake

PCA je multivarijantna metoda koja je osetljiva na varijansu ulaznih promeljivih. S obzirom da se za analizu koriste multivarijantni podaci, takvi setovi podataka sadrže promenljive sa različitim ospezima, pa bi u tom slučaju, one sa većim ospegom (na primer, od 0 do 100) bile dominantije u modelu u odnosu na promenljive sa manjim ospezima (od 0 do 1). Zbog toga se prilikom korišćenja ove metoda, kao prvi korak, primenjuje transformacija inicijalih promenljivih na uporedive skale (na primer, transformacija svih varijabli na skalu od 0 do 1, *Z*=vrednost varijable -srednja vrednost/standardna devijacija).

Da bi se podaci koristili kao inptu za *PCA* neophodno je da zadovoljavaju sledeće pretpostvake:

1. Za PCA metodu se mogu koristiti samo kontinuirane varijable (na racionalnim i interval sakalama).
2. Neophodno je da postoji linearna veza između varijabli jer *PCA* koristi *Pirsonov* korelacioni koeficijent kako bi opisao vezu između promenljivih.

3. Podaci ne smeju sadržati značjan broj neočekivanih promeljivih (kada su vrednosti veće od 3 standardne devijacije), pa se pretpostavlja da promeljive imaju normalnu raspodelu.

Primer analize multivarijantnog seta podataka pomoću analize glavnih komponenti u R-u.

Analiza kvaliteta vode na većem broju lokaliteta, opisanom sa većim brojem promenljivih u okviru jednog rečnog sliva, može se sprovesti metodom analize glavnih komponenti (PCA). Kao ulaznu matricu koristiće se rezultati merenja fizičko hemijskog kvatiteta vode na dvadeset različitih lokaliteta duž Sliva Nišave. U okviru kampanje mereno je 13 promenljivih:

```
> library(vegan)
Loading required package: permute
Loading required package: lattice
This is vegan 2.5-6
> Tabela = read.table("Tabela.csv", sep=',' , header=TRUE)
> env = Tabela [,115:127]
> env
      t    v  ep   ph   omg     O  bpk    no     po    nh
tvrdoca nad_vis sirina
1  17.9 0.40 533 6.84  9.65 104.3 3.29 4.296 1.2710 1.705
296.80     461      9
2  17.1 0.74 436 6.94 10.52 113.7 3.94 3.253 0.3680 0.544
206.02     426     15
3  16.2 0.51 431 7.11 11.77 126.9 4.17 4.767 0.1425 0.511
263.51     554     12
4  13.8 0.48 261 7.24 11.69 125.7 3.72 1.538 0.1425 0.640
186.85     694     12
5  23.1 0.33 123 6.85  9.58 119.6 3.00 0.496 0.1637 0.547
74.26     642     10
6   9.5 0.25 286 6.87 11.64 121.2 3.58 1.340 0.0925 0.537
215.60     598      6
7  21.7 0.30 251 6.88 10.40 122.2 4.05 0.591 0.2360 0.640
160.50     381     15
8  13.9 0.65 313 6.81 10.23 102.3 4.30 3.058 0.1380 1.553
167.69     335     30
9  20.5 0.54 470 6.87  7.81  90.8 1.08 5.550 0.2130 0.636
160.50     228     35
```

```
10 22.0 0.69 472 6.98  9.80 114.1 6.79 3.410 1.0625 1.300
198.83     201     30

11 15.5 0.39 514 6.88 12.08 127.0 4.58 3.045 0.0130 0.762
296.80     461      9

12 14.3 0.63 430 6.76 12.10 124.4 4.57 2.727 0.0025 0.592
233.20     426     15

13 13.6 0.53 430 6.68 11.58 119.4 3.79 3.592 0.0075 0.473
233.20     554     12

14 10.0 0.44 261 6.80 12.50 120.0 4.06 1.292 0.0280 0.533
127.20     694     12

15 12.2 0.30 124 7.31 11.29 113.3 3.33 0.523 0.0725 0.400
84.80     642     10

16  7.3 0.40 282 7.00 13.35 115.8 4.47 1.257 0.0250 0.552
169.60     598      6

17 11.8 0.37 255 6.83 12.10 117.3 4.08 0.771 0.0450 0.651
148.40     381     15

18 15.9 0.63 512 7.08  9.04  95.0 3.06 7.991 0.4880 0.821
233.20     335     30

19 15.5 0.46 518 6.95 12.90 121.9 5.20 7.390 0.3510 0.603
275.60     228     35

20 15.5 0.47 555 6.65  7.42  73.3 6.16 6.422 0.6725 2.276
254.40     201     30
```

U okviru R biblioteke *Vegan*, konstruisana je funkcija *rda* pomoću koje je moguće sprovesti analizu glavnih komponenti (*PCA*). Prvi argument funkcije određuje ulaznu matricu za analizu dok drugi argument *scale* definiše transformaciju podataka pre primene *PCA*, standardizujući promenljive različitih mernih skala.

```
> env.pca <- rda(env, scale=TRUE)
> env.pca
Call: rda(X = env, scale = TRUE)

              Inertia Rank
Total              13
```

```
Unconstrained      13   13

Inertia is correlations


Eigenvalues for unconstrained axes:

  PC1   PC2   PC3   PC4   PC5   PC6   PC7   PC8   PC9  PC10
PC11  PC12  PC13

5.743 1.962 1.334 1.055 0.982 0.849 0.550 0.210 0.146 0.113
0.029 0.024 0.003


> summary(env.pca) # Default scaling 2


Call:

rda(X = env, scale = TRUE)


Partitioning of correlations:

              Inertia Proportion

Total              13          1

Unconstrained      13          1


Eigenvalues, and their contribution to the correlations


Importance of components:

                          PC1    PC2    PC3     PC4     PC5
PC6     PC7     PC8     PC9     PC10     PC11

Eigenvalue            5.7435 1.9618 1.3339 1.05496 0.98168
0.84910 0.55014 0.20968 0.14567 0.113395 0.029038

Proportion Explained  0.4418 0.1509 0.1026 0.08115 0.07551
0.06532 0.04232 0.01613 0.01121 0.008723 0.002234

Cumulative Proportion 0.4418 0.5927 0.6953 0.77647 0.85199
0.91730 0.95962 0.97575 0.98696 0.995679 0.997913

                           PC12      PC13

Eigenvalue            0.023955 0.0031804

Proportion Explained  0.001843 0.0002446
```

```
Cumulative Proportion 0.999755 1.0000000


Scaling 2 for species and site scores
* Species are scaled proportional to eigenvalues
* Sites are unscaled: weighted dispersion equal on all
dimensions
* General scaling constant of scores:  3.964371



Species scores


             PC1      PC2       PC3       PC4      PC5       PC6
t       -0.5059  0.49964 -0.169195  0.587302 -0.03529  0.546184
v       -0.6071 -0.18534 -0.472542  0.048548  0.40967 -0.005805
ep      -0.9005 -0.50817 -0.153512 -0.070160 -0.26211  0.118101
ph       0.3752  0.08547 -0.542726  0.521669 -0.14864 -0.662024
omg      0.7632 -0.72150 -0.076623  0.009174  0.17126 -0.028478
O        0.7957 -0.47519 -0.176339  0.406806  0.07438  0.349673
bpk     -0.3193 -0.58988  0.472948  0.365314  0.56022 -0.152277
no      -0.8943 -0.24340 -0.409793 -0.185637 -0.21853 -0.118967
po      -0.7719  0.07465  0.283042  0.547033 -0.27075 -0.148432
nh      -0.8218  0.07313  0.626676  0.022104 -0.01625 -0.238674
tvrdoca -0.5842 -0.79002 -0.007756  0.016251 -0.44505  0.108523
nad_vis  0.9488 -0.01448  0.099099  0.027661 -0.30254 -0.122383
sirina  -0.8586  0.18257 -0.362365 -0.129573  0.40900 -0.098798



Site scores (weighted sums of species scores)


            PC1      PC2      PC3      PC4       PC5       PC6
sit1  -0.90818 -0.05699  1.30580  0.87612 -2.500476 -0.001247
```

```
sit2  -0.23571 -0.13432 -0.84549  0.38663  0.305354  0.336617
sit3   0.23036 -0.89692 -0.95285  0.69437 -0.727488 -0.085696
sit4   0.86320 -0.02944 -0.55678  0.88420 -0.309274 -1.126809
sit5   0.84227  1.90484  0.42358  0.71688  0.009917  1.360560
sit6   0.91200 -0.32376  0.77239 -0.69195 -0.872230 -0.056619
sit7   0.36706  0.84942  0.47038  0.91296  0.230387  1.302578
sit8  -0.54357  0.43646  0.28744 -0.83231  1.486425 -0.498807
sit9  -0.93732  1.80942 -1.61919 -1.28287 -0.530271  0.692766
sit10 -1.23648 -0.11839  0.25934  2.36319  1.558977 -0.144678
sit11  0.16367 -1.30356  0.16316  0.08031 -0.733558  0.902174
sit12  0.09783 -0.97315 -0.24256 -0.38127  0.753893  1.029712
sit13  0.22233 -0.70791 -0.06441 -0.99120 -0.180332  1.183370
sit14  0.96098 -0.12619  0.49855 -0.79445  0.575305 -0.101567
sit15  1.29506  1.10320 -0.24706  0.46867 -0.077455 -1.874170
sit16  1.02089 -0.67517  0.45778 -0.55859  0.229186 -1.166260
sit17  0.61068  0.08449  0.55641 -0.61119  0.825630  0.141517
sit18 -1.05881  0.30547 -1.56235 -0.41635 -0.721339 -1.055802
sit19 -0.66913 -1.39874 -1.06020  0.13752  0.498087  0.018232
sit20 -1.99712  0.25123  1.95605 -0.96069  0.179262 -0.855871

> summary(env.pca, scaling=1)

Call:
rda(X = env, scale = TRUE)

Partitioning of correlations:
              Inertia Proportion
Total              13          1
Unconstrained      13          1

Eigenvalues, and their contribution to the correlations
```

```
Importance of components:
                          PC1    PC2    PC3     PC4     PC5
PC6     PC7     PC8     PC9     PC10     PC11
Eigenvalue            5.7435 1.9618 1.3339 1.05496 0.98168
0.84910 0.55014 0.20968 0.14567 0.113395 0.029038
Proportion Explained  0.4418 0.1509 0.1026 0.08115 0.07551
0.06532 0.04232 0.01613 0.01121 0.008723 0.002234
Cumulative Proportion 0.4418 0.5927 0.6953 0.77647 0.85199
0.91730 0.95962 0.97575 0.98696 0.995679 0.997913
                           PC12      PC13
Eigenvalue            0.023955 0.0031804
Proportion Explained  0.001843 0.0002446
Cumulative Proportion 0.999755 1.0000000


Scaling 1 for species and site scores
* Sites are scaled proportional to eigenvalues
* Species are unscaled: weighted dispersion equal on all
dimensions
* General scaling constant of scores:  3.964371



Species scores


            PC1      PC2      PC3      PC4      PC5      PC6
t       -0.7611  1.28617 -0.52819  2.06165 -0.12843  2.13713
v       -0.9134 -0.47710 -1.47518  0.17042  1.49081 -0.02272
ep      -1.3547 -1.30815 -0.47923 -0.24629 -0.95382  0.46211
ph       0.5645  0.22002 -1.69428  1.83125 -0.54089 -2.59039
omg      1.1482 -1.85729 -0.23920  0.03221  0.62324 -0.11143
O        1.1971 -1.22324 -0.55049  1.42804  0.27069  1.36821
bpk     -0.4803 -1.51847  1.47645  1.28239  2.03867 -0.59584
no      -1.3455 -0.62655 -1.27929 -0.65165 -0.79524 -0.46550
```

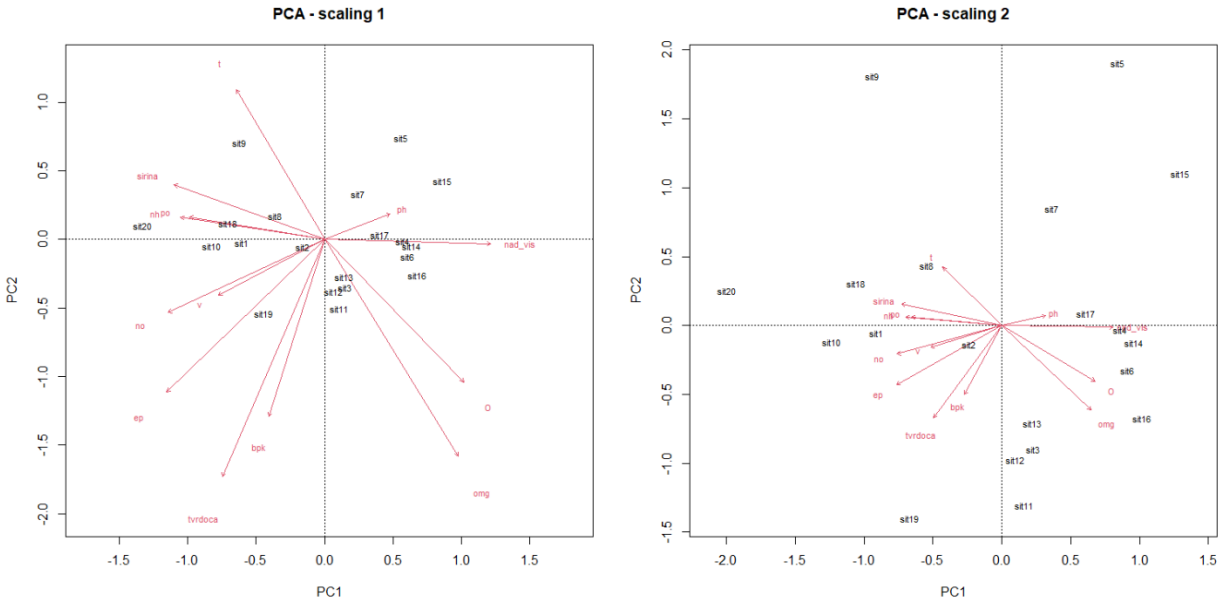
```
po      -1.1613  0.19216  0.88360  1.92029 -0.98528 -0.58079
nh      -1.2364  0.18826  1.95636  0.07759 -0.05915 -0.93390
tvrdoca -0.8790 -2.03369 -0.02421  0.05705 -1.61957  0.42463
nad_vis  1.4274 -0.03727  0.30937  0.09710 -1.10095 -0.47887
sirina  -1.2918  0.46999 -1.13123 -0.45485  1.48837 -0.38658


Site scores (weighted sums of species scores)

            PC1      PC2      PC3      PC4       PC5         PC6
sit1  -0.60366 -0.02214  0.41828  0.24958 -0.687126 -0.0003187
sit2  -0.15667 -0.05218 -0.27083  0.11014  0.083911  0.0860288
sit3   0.15312 -0.34842 -0.30522  0.19781 -0.199912 -0.0219012
sit4   0.57376 -0.01144 -0.17835  0.25188 -0.084988 -0.2879769
sit5   0.55984  0.73997  0.13569  0.20422  0.002725  0.3477164
sit6   0.60619 -0.12577  0.24742 -0.19711 -0.239687 -0.0144701
sit7   0.24398  0.32997  0.15067  0.26007  0.063310  0.3328982
sit8  -0.36130  0.16955  0.09207 -0.23710  0.408467 -0.1274795
sit9  -0.62302  0.70290 -0.51867 -0.36545 -0.145718  0.1770492
sit10 -0.82187 -0.04599  0.08308  0.67320  0.428404 -0.0369751
sit11  0.10879 -0.50639  0.05226  0.02288 -0.201580  0.2305674
sit12  0.06502 -0.37804 -0.07770 -0.10861  0.207168  0.2631621
sit13  0.14778 -0.27500 -0.02063 -0.28236 -0.049555  0.3024322
sit14  0.63875 -0.04902  0.15970 -0.22631  0.158093 -0.0259573
sit15  0.86080  0.42856 -0.07914  0.13351 -0.021285 -0.4789792
sit16  0.67857 -0.26228  0.14664 -0.15912  0.062980 -0.2980596
sit17  0.40591  0.03282  0.17823 -0.17411  0.226881  0.0361674
sit18 -0.70377  0.11867 -0.50046 -0.11860 -0.198223 -0.2698298
sit19 -0.44476 -0.54337 -0.33961  0.03917  0.136873  0.0046595
sit20 -1.32746  0.09759  0.62658 -0.27367  0.049261 -0.2187339
```

Funkcijom *summery* moguće je prikazati rezultate *PCA*, nakon čega R konzola izlista veliki broj tabela i vrednosti. Karakteristična vrednost prikazuje značaj *PC* osa u pogledu varijabilnosti i može biti prikazana preko proporcije objašnjene varijabilnosti (*Proportions Explained*). Ova vrednost se dobija tako što se karakteristična vrednost ose podeli sa sumom varijabilnosti svih PC osa (eng. *total interia*) i pomnoži sa 100. PCA ne spada u grupu statističkih testova i cilj joj je da redukuje broj dimenzija u multidimenzionalnom setu podataka i da najefikasniji način prikaže obrasce varijabilnosti u prostoru sa redukovanim brojem dimenzija/varijabli. Karakteristična vrednost se koristi kao parametar za odabir broja *PC* osa koje će se koristiti prilikom vizuelizacije varijabilnosti multivarijantnog seta podataka. Odluka može biti u potpunosti arbitrarna tako da broj *PC* osa pokriva određeni procenat varijabilnosti u setu podataka (na primer, >50%). Na primeru kvaliteta vode Sliva Nišave, prve dve *PC* ose pokrivaju 59% varijabilnosti podataka (*Cumulative Proportion*: 0.5927) što je dovoljno da se uzme u obzir prilikom ordinacije lokaliteta u prostoru koje formiraju prve dve *PC* ose.

Vizuelna interpretacija rezultata u R-u je moguća pomoću *biplot* funkcije:

```
> # Plots using biplot.rda
> dev.new(width=12, height=6, title="PCA biplots - environmental variables - biplot.rda")
NULL
> par(mfrow=c(1,2))
> biplot(env.pca, scaling=1, main="PCA - scaling 1")
> biplot(env.pca, main="PCA - scaling 2")  # Default scaling 2
```

Rezultat funkcije *biplot* je PCA ordinacioni grafikon koji prikazuje ordinaciju lokaliteta na osnovu kvaliteta vode (Slika 8.2). *PCA* grafikon prikazuje dva tipa rezultata, ordinaciju lokaliteta i korelaciju varijabli. Funkcija *biplot* sadrži argument *scaling* koji definiše način prikazivanja ordinacionih rezultata u redukovanom prostoru. Kada je argument *scaling*=1, biplot funckija generiše grafikon distanci (*Scaling 1*) koji prikazuje distancu između lokaliteta, definisanu euklidovom distancom u multidimenzionalnom prostoru. U tom slučaju ugao imzeđu vektora ne prikazuje korelaciju ulaznih varijabli. S druge strane korelacioni grafikon (*Scaling 2*) vizualizuje korelaciju između varijabli koja je proporcionalna uglovima koji formiraju vektori dok distanca između tačaka ne aproksimira njihovu sličnost.



Slika 8.2 PCA ordinacioni grafikoni. Tčke predstavljaju lokalitete, a vektori promeljive. *Scaling 1* prikazuje sličnost lokaliteta pomoću euklidove distance u multidimenzionalnom prostoru. *Scaling 2* prikazuje odnos promenljivih gde korelacija proporcionalna uglu između vektora.

Prvi korak u interpretaciji rezultata je definisati varijabilnost koju pokrivaju *PC* ose. U slučaju kvaliteta vode na Slivu Nišave, prve dve ose pokivaju 59% varijabilnsoti (*PC1*=0.44 i *PC2*=0.15), što je u slučjau ekoloških podataka gde veliki broj varijabli utiče na varijabilnost dovoljno da se pouzdano intrpretiraju podaci u okvriu prvog para *PC* osa. Na *PCA* grafikonu se može prepoznati tendencija grupisanja lokaliteta duž prve *PC* ose gde se grupa lokaliteta na desnoj strani (sit 4-7, 14-17) nalazi na velikoj nadmorskoj visini, dok su lokaliteti (sit 8-10, 18 i 20) na levoj strani ose na najnižim nadmorskim visinama sa visokim koncentracijama soli (No3-N i PO4-P). Na osnovu korelacionog grafikona (*Scaling 2*), prva *PC* osa se može interpretirati kao gradijent nadmorskih visina, gde sa porastom nadmorske visine raste i koncentracija rastvorednog kiseonika u vodi (O2 i O%). S druge strane kako nadmorska visina opada, tako raste temperatura vode, količina rastvorenih soli u vodi (ep No3-N i PO4-P, NH4-N) i biološka potrošnja kiseonika (BPK5).

9.Prilog

Tabela S4.1 Proporcije normalne distribucije (jednostrane).

Z	0	1	2	3	4	5	6	7	8	9	**Z**
0	0.5	0.496	0.492	0.488	0.484	0.4801	0.4761	0.4721	0.4681	0.4641	0
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247	0.1
0.2	0.4207	0.4168	0.4129	0.409	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859	0.2
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.352	0.3483	0.3
0.4	0.3446	0.3409	0.3372	0.3336	0.33	0.3264	0.3228	0.3192	0.3156	0.3121	0.4
0.5	0.3085	0.305	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.281	0.2776	0.5
0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451	0.6
0.7	0.242	0.2389	0.2358	0.2327	0.2297	0.2266	0.2236	0.2207	0.2177	0.2148	0.7
0.8	0.2119	0.209	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867	0.8
0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.166	0.1635	0.1611	0.9
1	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379	1
1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.123	0.121	0.119	0.117	1.1
1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.102	0.1003	0.0985	1.2
1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823	1.3
1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681	1.4
1.5	0.0668	0.0655	0.0643	0.063	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559	1.5
1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455	1.6
1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367	1.7
1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294	1.8
1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.025	0.0244	0.0239	0.0233	1.9
2	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183	2
2.1	0.0179	0.0174	0.017	0.0166	0.0162	0.0158	0.0154	0.015	0.0146	0.0143	2.1
2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.011	2.2
2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084	2.3
2.4	0.0082	0.008	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064	2.4
2.5	0.0062	0.006	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048	2.5
2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.004	0.0039	0.0038	0.0037	0.0036	2.6
2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.003	0.0029	0.0028	0.0027	0.0026	2.7
2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.002	0.0019	2.8
2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014	2.9
3	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.001	0.001	3
3.1	0.001	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007	3.1
3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005	3.2
3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003	3.3
3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002	3.4
3.5	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	3.5
3.6	0.0002	0.0002	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	3.6
3.7	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	3.7
3.8	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	3.8

Tabela SX.X Studentova t-raspodela. N predstavlja broj stepena slobode.

N \ α	0.1	0.05	0.025	0.01	0.005	0.001	0.0005
1	3.078	6.314	12.076	31.821	63.657	318.310	636.620
2	1.886	2.920	4.303	6.965	9.925	22.326	31.598
3	1.638	2.353	3.182	4.541	5.841	10.213	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221

14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.485	3.767
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	1.296	1.671	2.000	2.390	2.660	3.232	3.460
120	1.289	1.658	1.980	2.358	2.617	3.160	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.090	3.291

Erasmus + Project No ECOBIAS_609967-EPP-1-2019-1-RS-EPPKA2-CBHE-JP
Development of master curricula in ecological monitoring and aquatic bioassessment for Western Balkans HEIs

Co-funded by the Erasmus+ Programme of the European Union