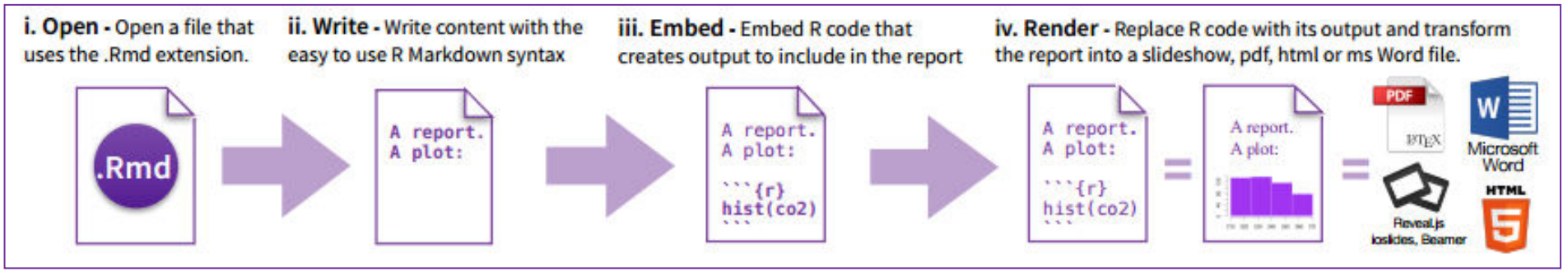
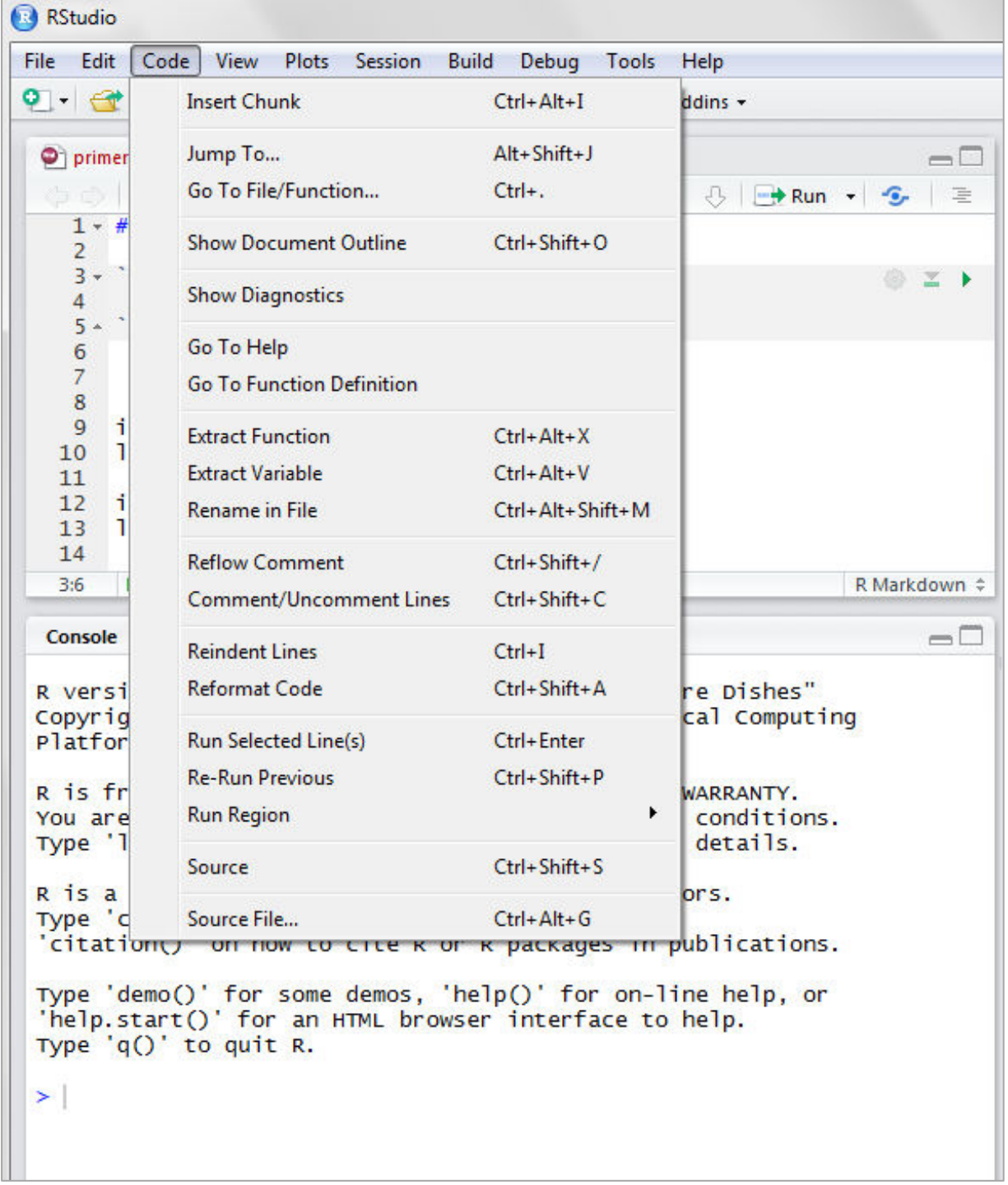
Univerzitet u Beogradu – Elektrotehnički fakultet (ETF)
Katedra za signale i sisteme

Tehnike obrade biomedicinskih signala 13M051TOBS



Dr Nadica Miljković, vanredni profesor
kabinet 68, nadica.miljkovic@etf.bg.ac.rs
Slika na naslovnoj strani je preuzeta iz: https://www.rstudio.com/wp-content/uploads/2015/02/rmarkdown-cheatsheet.pdf, *Fair Use*.

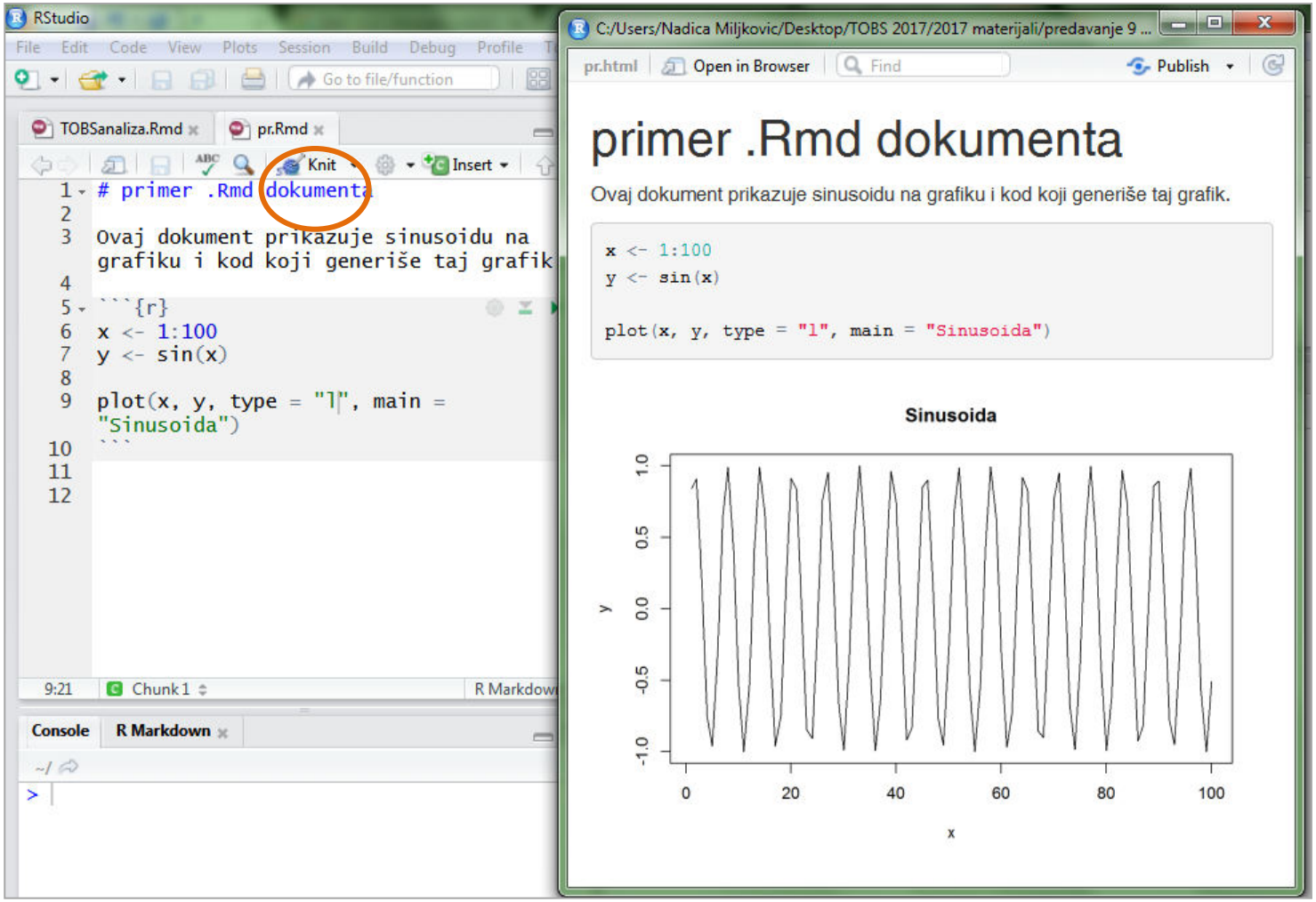
PISANJE I GENERISANJE IZVEŠTAJA



RStudio okruženje

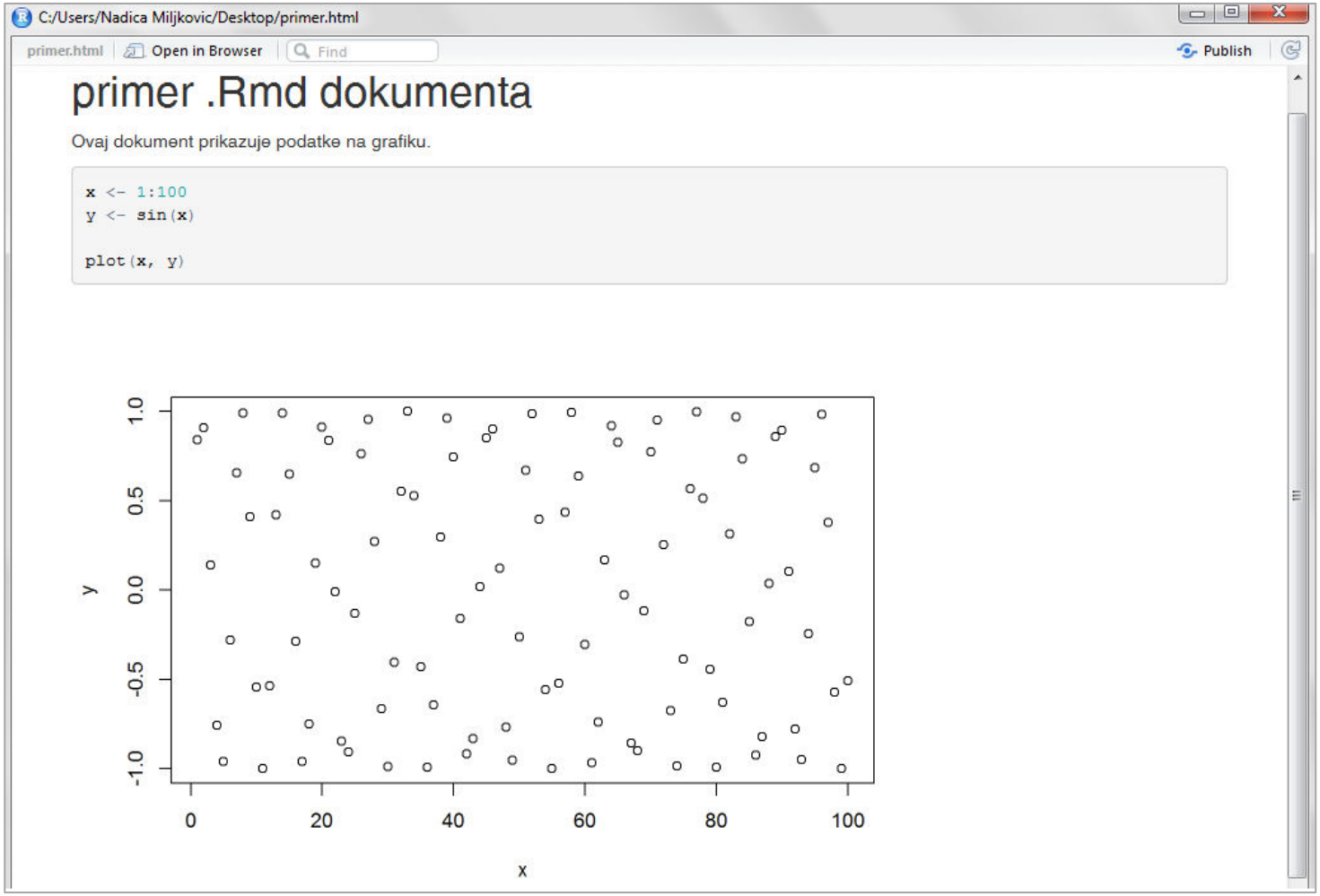
- Skriptu je moguće osim sa ekstenzijom .R snimiti i sa .Rmd (eng. *R markdown*) ekstenzijom.
  - Pomoću opcije iz padajućeg menija moguće je odabrati *Insert Chunk* (dodavanje delova) – služi da se odvoje tekst i kod (kod je unutar *Chunk*-a osenčen)
- R *markdown* služi za istovremeno pisanje koda i generisanje izveštaja (automatski izveštaji).

“Rmd” dokument



- Odabirom opcije *Knit* moguće je pokrenuti konverziju koda u npr. HTML dokument, gde se tekst iako nije označen komentarom, zbog postojanja *Chunk*-a, prikazuje zajedno sa rezultatom koda (grafikom) kao na slici.
- Za opciju *Knit* moguće je odabrati *HTML*, *pdf* ili neku drugu ekstenziju dokumenta za čuvanje automatski generisanog izveštaja.

Rmd HTML, može i ovako



- Rezultat prikaza koda sa komentarima i rezultatima u HTML dokumentu dat je na slici.
- NAPOMENA: za korišćenje ove opcije, potrebno je instalirati paket rmarkdown.
- Primetiti da je naslov dokumenta zapravo prva linija koda pod komentarom tj. posle znaka “#”.

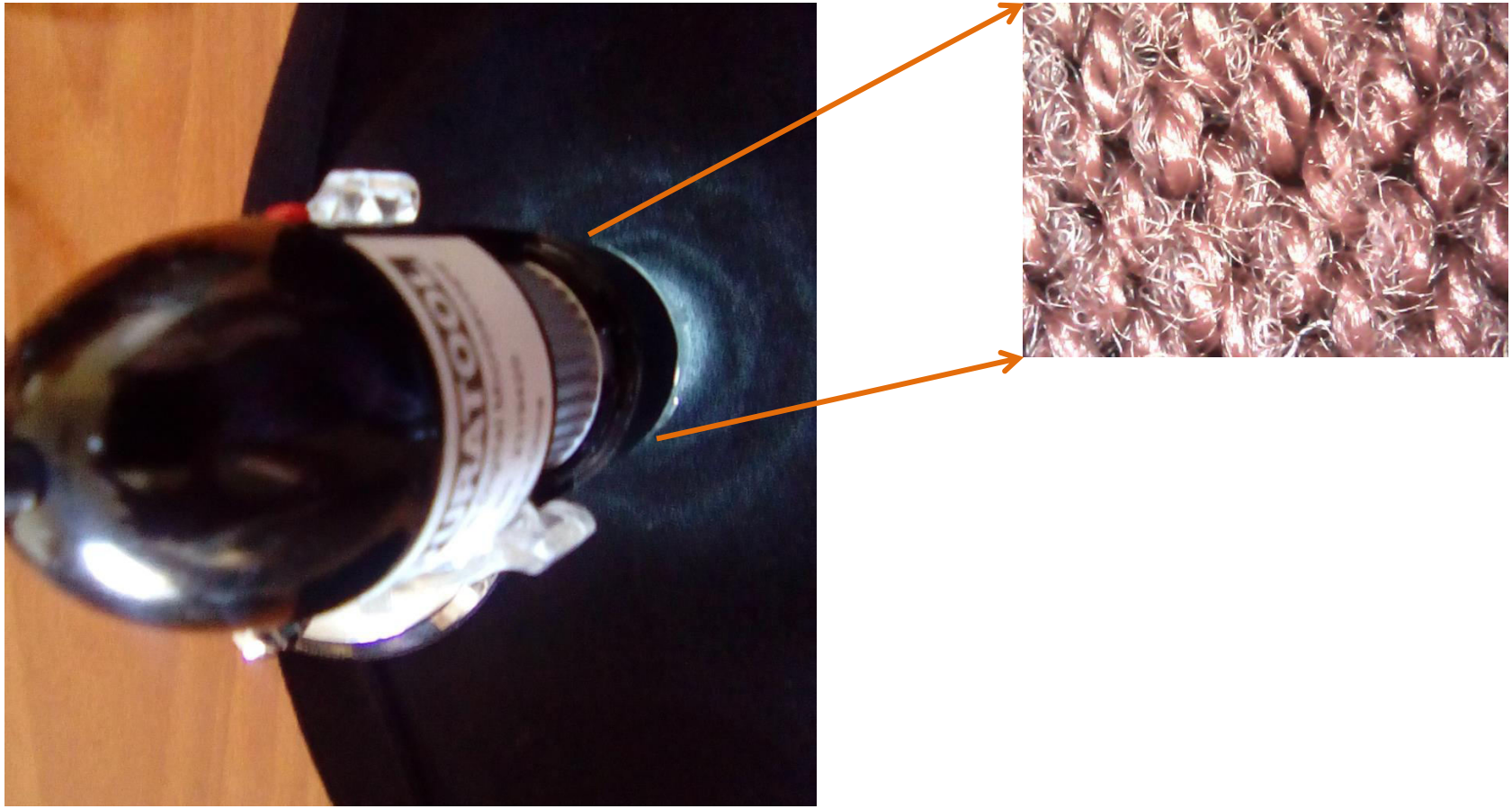
R *Markdown*

- U jednoj datoteci postižete dve funkcije: pišete **kod (obrada signala)** i generišete **izveštaj (vizuelizacija rezultata)**.
- Izlaz (izveštaj) moguće je odabrati da bude u nekom od sledećih formata: HTML, pdf, Word, *slide show*, *book*, *handout*, *dashboard*, *website* i *interactive app*.
- R Markdown se sastoji iz tri dela:
  - *YAML metadata* (komande koje definišu izgled generisanog izveštaja tzv. konfiguracioni kod),
  - *Text* (tekst koji želite da prikažete u Vašem izveštaju) i
  - *Code chunks* (delovi .Rmd dokumenta koji sadrže R kod za obradu signala).
- Svakom delu koda (*Chunk*) moguće je pristupiti pojedinačno i pokrenuti taj kod opcijama koje su dostupne u ".Rmd" datoteci u R Studio interfejsu.

R *Markdown*

- Moguće je i testirati kod i proveriti rezultate u konzoli i u kartici *Plots*.
- Ovo je odlična alatka za distribuiranje izveštaja.
- Velika udobnost za korisnika/icu, jer sve poslove završava u jednom programu pomoću alatke *knitr* (https://yihui.name/knitr/).
- **Rezultat koda će biti prikazan u finalnom automatski generisanom izveštaju. Po želji može biti prikazan i sam kod. Zavisi kome je izvešaj namenjen.**
- Ovo je odlično za dijagnostiku i terapiju u medicini. Svaki uređaj mora da sadrži i odgovarajuć izveštaj za medicinsko osoblje, za arhivu, kliničke inženjere...
- **Ovi izveštaji se ponekad nazivaju i dinamički izveštaji.**
- Modifikovana slika od Hello I'm Nik na Unsplash

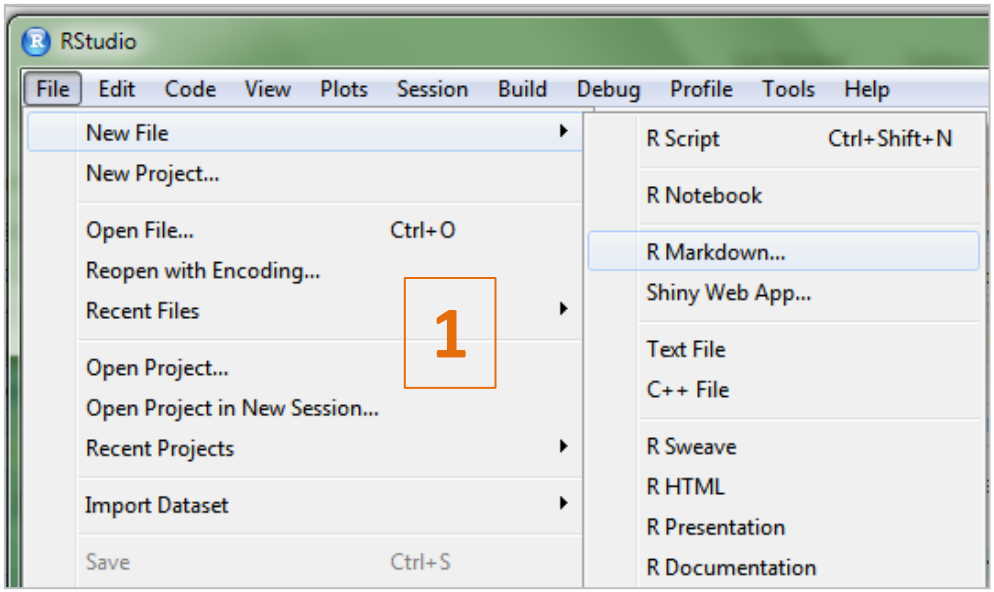
Knitting?

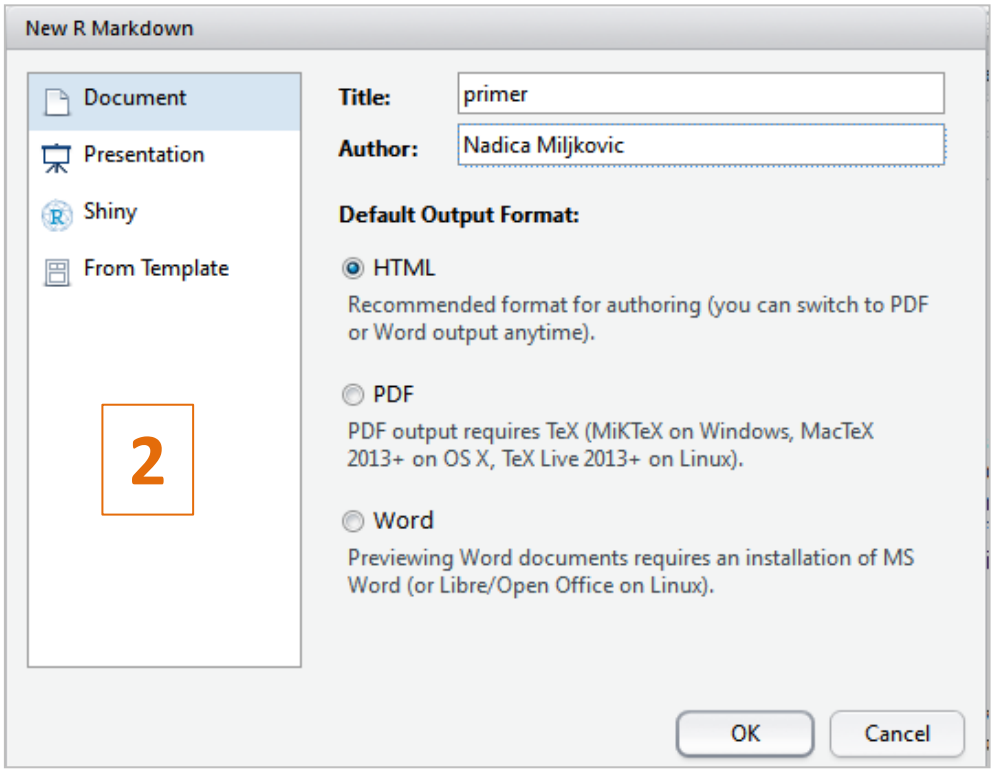


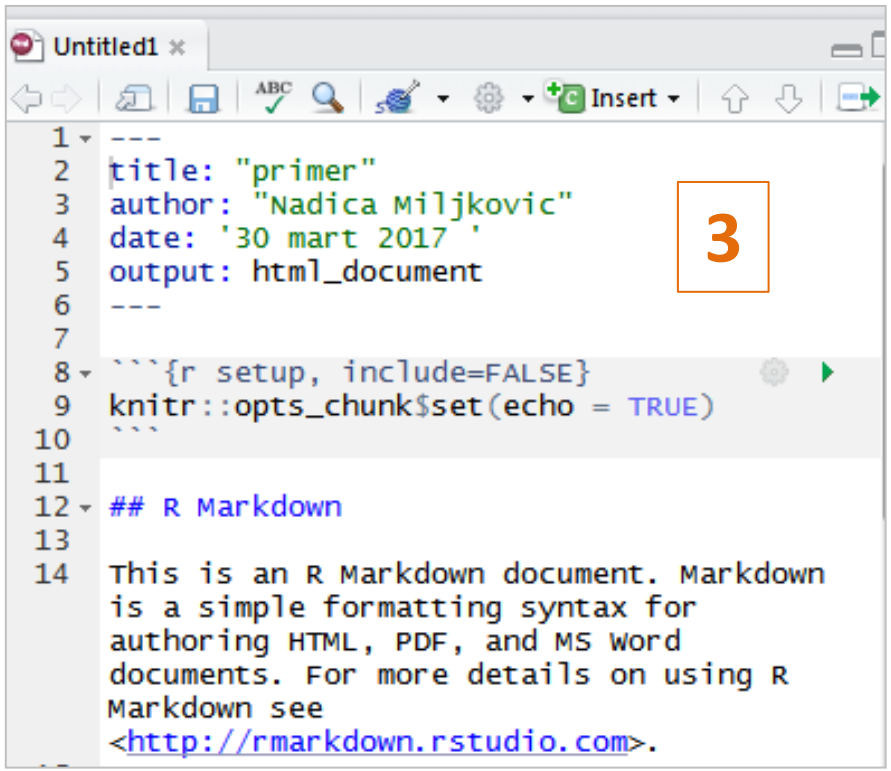
- Zašto crno nije crno?
- Fotografije su snimljene sa digitalnim mikroskopom BW-788 (Duratool, China) sa manuelnim uvećanjem od 25 do 200 puta na Elektrotehničkom fakultetu u Beogradu (mart 2017. godine).

Kako se generiše izveštaj?

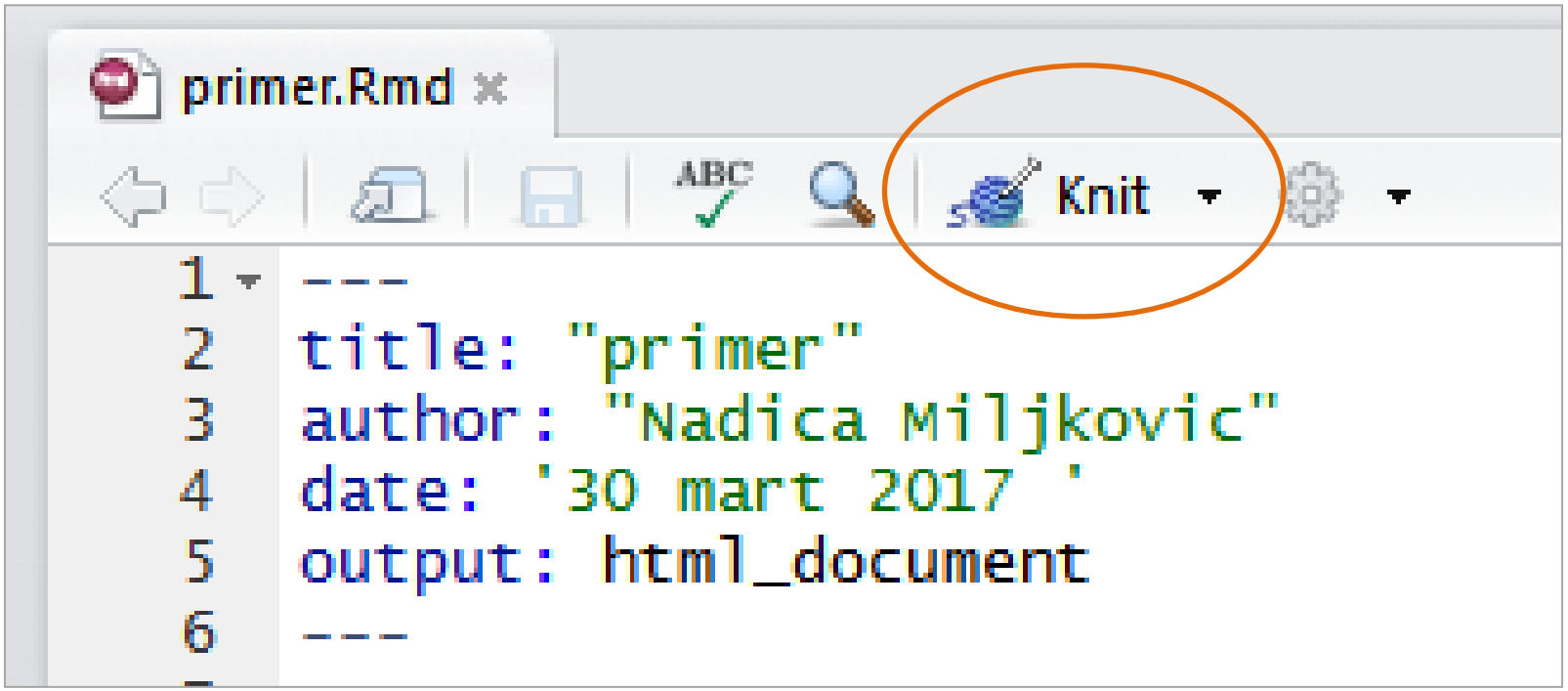
- *File/New File/R Markdown* (kao na slici na gornjem panelu).
- Potom se bira naslov (ne naziv dokumenta) i unosi se npr. ime autora dokumenta. Takođe, bira se podrazumevani format izveštaja (kao na slici na donjem levom panelu).
- Otvara se primer dokumenta (kao na slici na donjem desnom panelu).





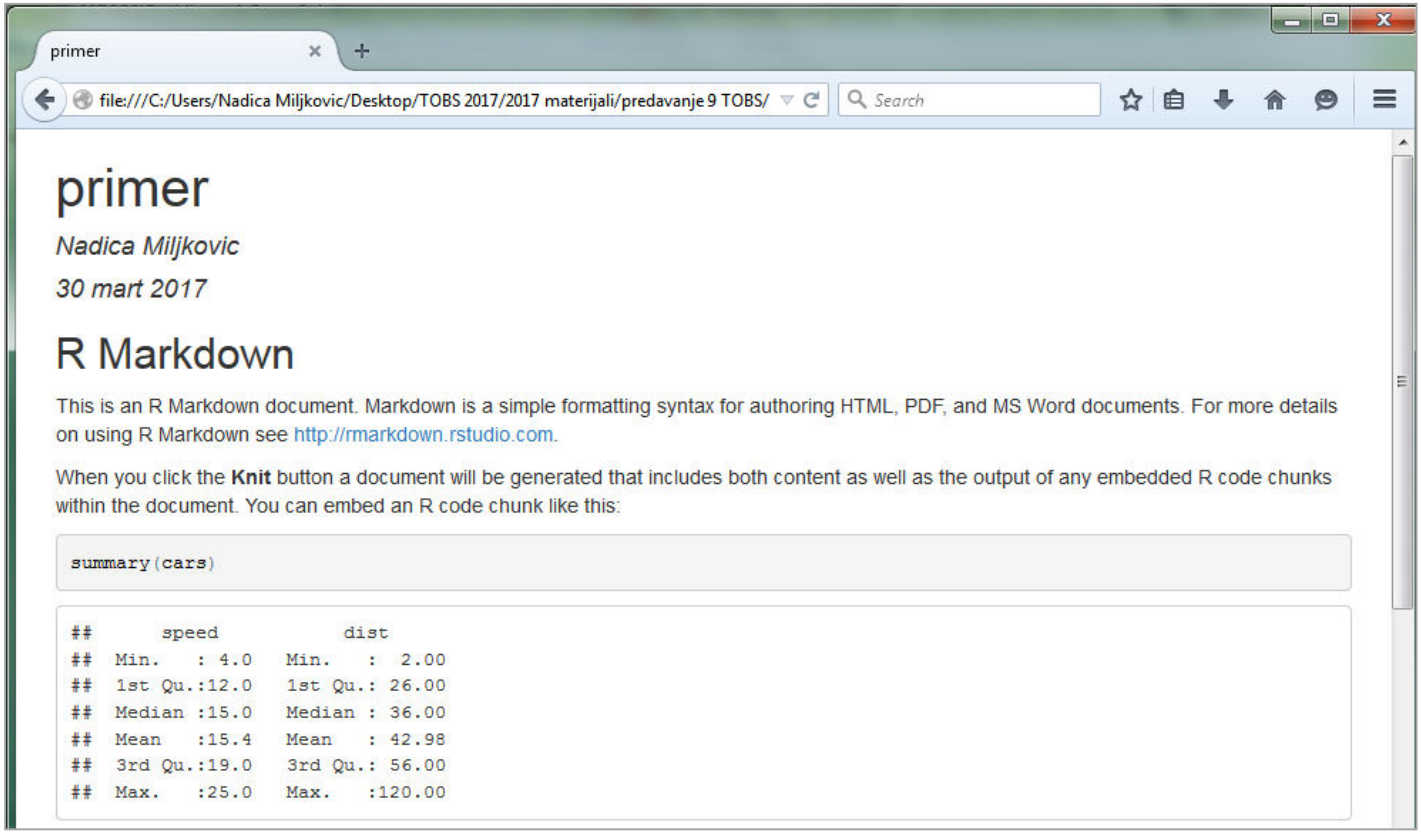


I to je sve!



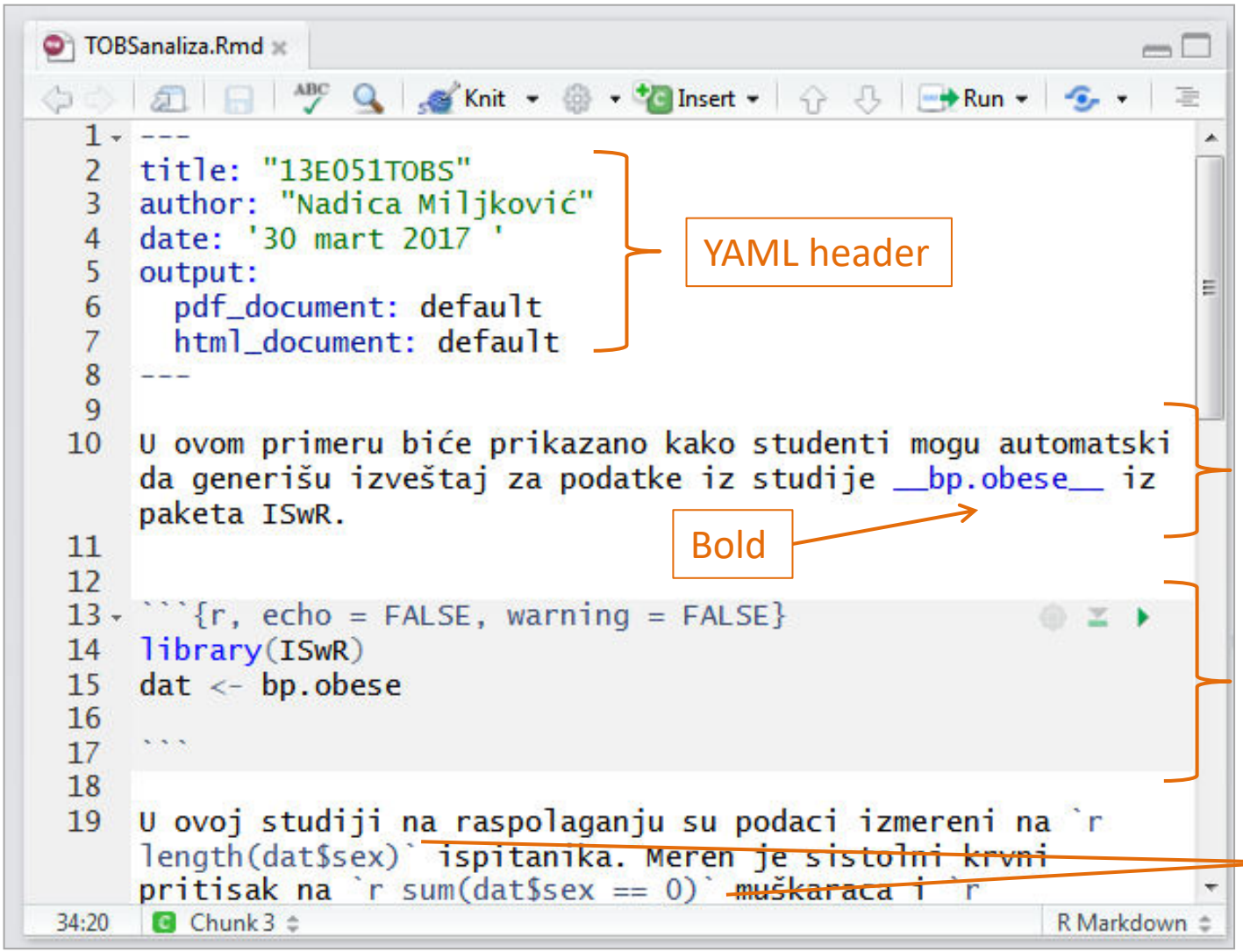
- Još samo jedan korak ... pritiskom na “Knit” dugme kao na slici dobija se HTML izveštaj.
- Ovaj primer je sadržao generički kod i tekst.
- Hajde da vidimo neke jednostavne primere (od nule) ... ali pre toga rezultat!

Rezultat?



- Generisan je “primer.HTML” koji je moguće otvoriti u pretraživaču po želji.
- Na slici je prikazan generički primer otvoren u Mozzila Firefox *browser*-u (koji je kao i R slobodan softver).

Primeri generisanja izveštaja za obradu biosignala



- Nakon kreiranja odgovarajuće *Rmd* datoteke, potrebno je dodati *Chunk* (iz padajućeg menija kao na slajdu ranije *Code* / *Insert Chunk*).
- Potom je potrebno učitati podatke i pokazati ih pomoću “ggplot2” funkcija.

Primeri generisanja izveštaja za obradu biosignala

````
```{r}
# Funkcija summary() se koristi da bi se prikazali osnovni
parametri za skup podataka.
# Sadrži parametre deskriptivne statistike kao što su
minimalna vrednost i dr.
summary(dat)

```


BMI (eng. _Body Mass Index_) se izražava u jedinicama
$kg/m^{2}$.
Zavisnost BMI od sistolnog krvnog pritiska prikazana je na
sledećem grafiku:

```{r, echo = FALSE, warning = FALSE, message = FALSE,
fig.height = 3, fig.width = 4}
library(ggplot2)
p <- ggplot(dat, aes(x = bp, y = obese)) +
        geom_smooth() +
        ggtitle("Zavisnost krvnog pritiska od BMI") +
        xlab("sistolni krvni pritisak [mmHg]") +
        ylab("BMI [kg / m^2]")
print(p)

```


Ovaj izveštaj je u pdf-u zajedno sa Rmd skriptom dostupan
studentima na sajtu predmeta:
[http://automatika.etf.rs/sr/13m051tobs](http://automatika.
etf.rs/sr/13m051tobs).
````

Još jedan *Chunk*, ima ih dva na ovom slajdu.

Italic

formula

Zaokružena je strelica za testiranje *Chunk*-a.

Link

- Obratiti pažnju na oznake koje se koriste i posebno na parametre zaglavlja *Chunk*-a.
- Od velike koristi bi trebalo da bude *Rmd Reference Card* na linku: https://www.rstudio.com/wp-content/uploads/2015/02/rmarkdown-cheatsheet.pdf.

TOBSanaliza.pdf

Page: 1 /1 Find:

13E051TOBS

Nadica Miljković

30 mart 2017

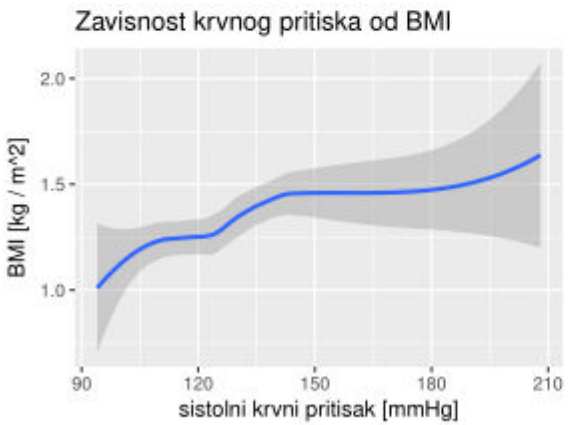
U ovom primeru biće prikazano kako studenti mogu automatski da generišu izveštaj za podatke iz studije **bp.obese** iz paketa ISwR.

U ovoj studiji na raspolaganju su podaci izmereni na 102 ispitanika. Meren je sistolni krvni pritisak na 44 muškaraca i 58 žena. Osnovni elementi ovih podataka su prikazani u tabeli:

```
# Funkcija summary() se koristi da bi se prikazali osnovni parametri za skup podataka.
# Sadrži parametre deskriptivne statistike kao što su minimalna vrednost i dr.
summary(dat)
```

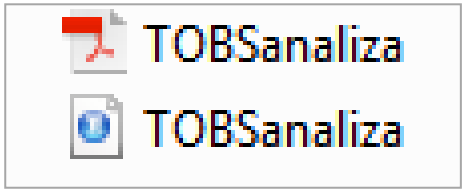
```
##       sex            obese             bp
##  Min.   :0.0000   Min.   :0.810   Min.   : 94.0
##  1st Qu.:0.0000   1st Qu.:1.143   1st Qu.:116.0
##  Median :1.0000   Median :1.285   Median :124.0
##  Mean   :0.5686   Mean   :1.313   Mean   :127.0
##  3rd Qu.:1.0000   3rd Qu.:1.430   3rd Qu.:137.5
##  Max.   :1.0000   Max.   :2.390   Max.   :208.0
```

BMI (eng. *Body Mass Index*) se izražava u jedinicama kg/m^2. Zavisnost BMI od sistolnog krvnog pritiska prikazana je na sledećem grafiku:



Ovaj izveštaj je u pdf-u zajedno sa Rmd skriptom dostupan studentima na sajtu predmeta: http://automatika.etf.rs/sr/13m051tobs.

Izveštaj u pdf-u



Da li primećujete deo koji je pogrešno napisan u izveštaju?

Jedinice se ne pišu kurzivom tj. *Italic*.

Formatiranje teksta

- Osim pokazanih par primera, gde je korisnik/ca mogao/la posebnim komandama da generiše podebljan tekst (eng. *Bold*), kurziv (eng. *Italic*), da ispiše formule, postoji i niz drugih opcija.
- Te opcije omogućavaju podelu izveštaja na paragrafe, definisanje naslova i podnaslova, unos slika iz postojećih datoteka u dokument (npr. fotografija, šema i sl).
- Moguće je definisati i prikaz tabele na način koji korisniku/ci odgovara.
- Za interaktivni dokument, potrebno je iskoristiti *Shiny* paket – o njemu kasnije u ovom semestru.
- Kome se ovo svidi, može komotno sa MS Word-a (https://www.office.com/) da pređe na LATEX (https://www.latex-project.org/).

```
BMI (eng. _Body Mass Index_) se izražava u jedinicama $kg/m^{2}$.
Zavisnost BMI od sistolnog krvnog pritiska prikazana je na sledećem grafiku:
```

A korisnička interakcija?

```
---
title: "prikaz signala iz .txt fajla"
author: "Nadica Miljkovic"
date: '01 april 2017 '
output:
  html_document: default
params:
  filename: "emg1.txt"
---

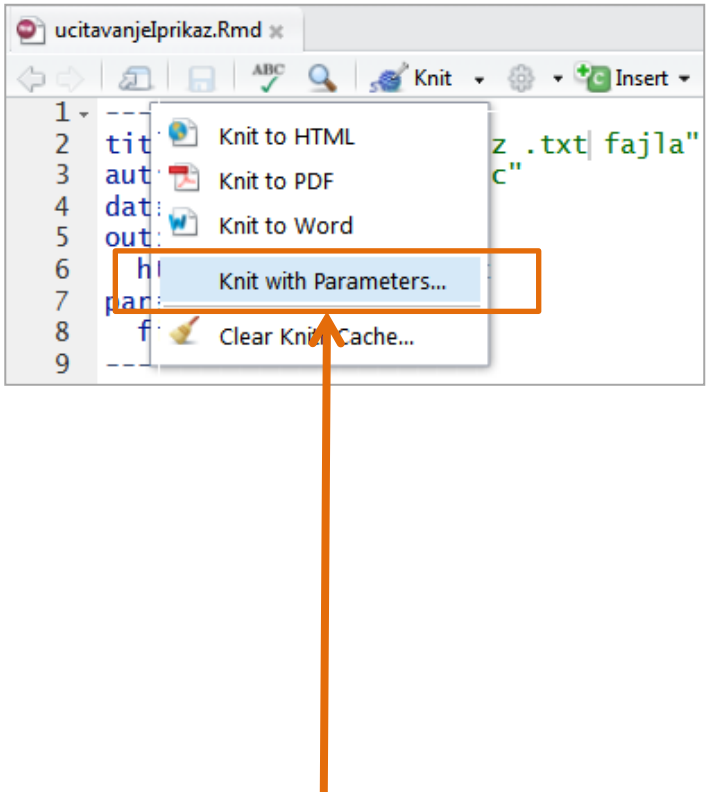
```{r}
dat <- read.table(params$filename, dec = ",")
```

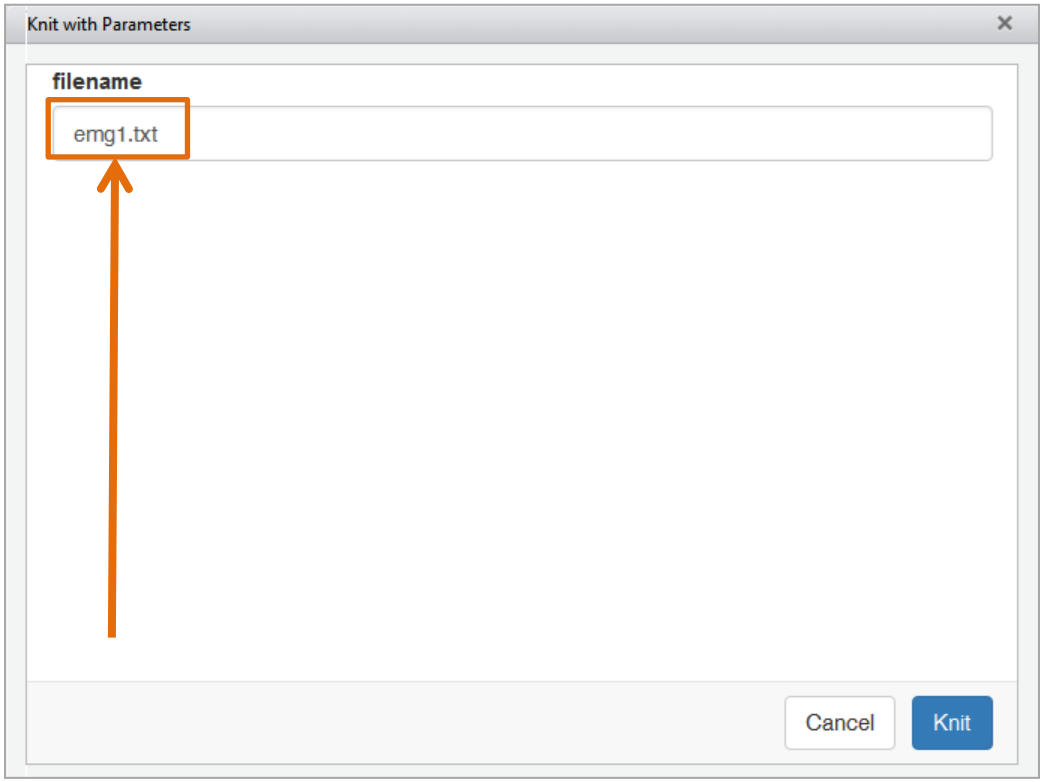
Srednja vrednost učitanog EMG signala iz fajla "`r
params$filename`" je `r mean(dat$V1)`.

```{r}
fs <- 1000
timeA <- seq(0, length(dat$V1)/fs - 1/fs, by = 1/fs)
plot(timeA, dat$V1, type = "l",
     main = "EMG signal",
     xlab = "vreme [s]",
     ylab = "napon [mV]",
     col = "grey50")
grid()
```
```

- U padajućem Knit meniju u .Rmd dokumentu, postoji i opcija “Knit with Parameters ...”.
- Na primer, ako se za potrebe generisanja izveštaja učitava datoteka, ova opcija omogućava odabir datoteke.
- Ovo je samo jedan primer interakcije.
- Prilikom prvog korišćenja opcije “Knit with Parameters ...” potrebno je otvoriti i zatvoriti tj. resetovati R Studio.
- Tada će R ponuditi instalaciju dodatnih paketa što treba odobriti.

Knit with parameters



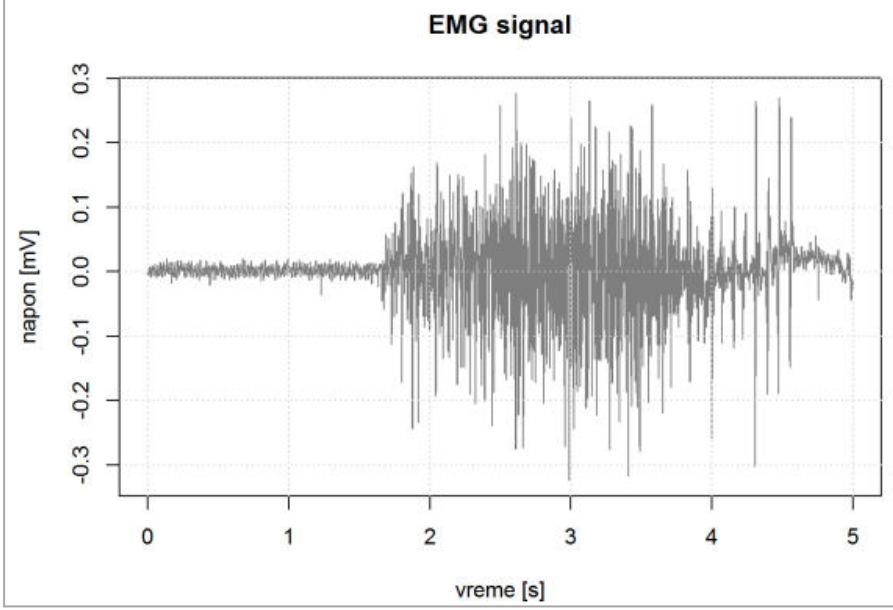


- Prilikom pokretanja ove opcije (kao na slici levo), otvara se prozor kao na slici desno.
- Moguće je odabrati datoteku po izboru i na osnovu odabrane datoteke generiše se odgovarajući izveštaj.

Rezultat(i)

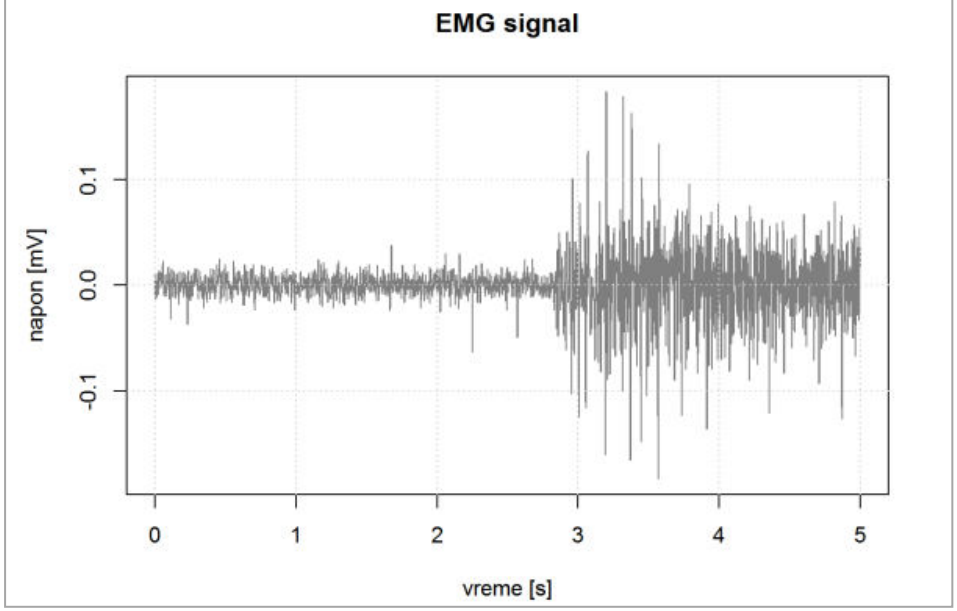
Srednja vrednost učitanog EMG signala iz fajla “emg1.txt” je 0.0034512.

```
fs <- 1000
timeA <- seq(0, length(dat$V1)/fs - 1/fs, by = 1/fs)
plot(timeA, dat$V1, type = "l",
     main = "EMG signal",
     xlab = "vreme [s]",
     ylab = "napon [mV]",
     col = "grey50")
grid()
```



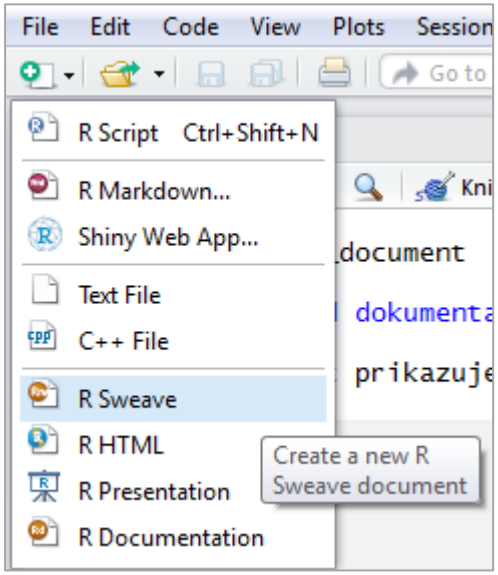
Srednja vrednost učitanog EMG signala iz fajla “emg3.txt” je 0.0010596.

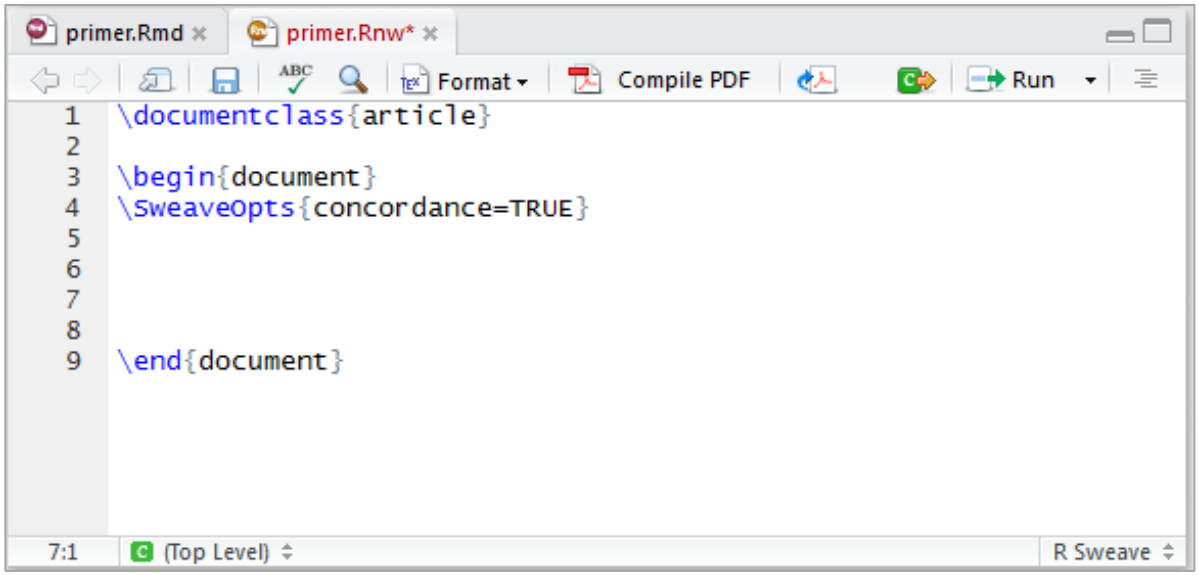
```
fs <- 1000
timeA <- seq(0, length(dat$V1)/fs - 1/fs, by = 1/fs)
plot(timeA, dat$V1, type = "l",
     main = "EMG signal",
     xlab = "vreme [s]",
     ylab = "napon [mV]",
     col = "grey50")
grid()
```



JOŠ NEŠTO O IZVEŠTAJIMA

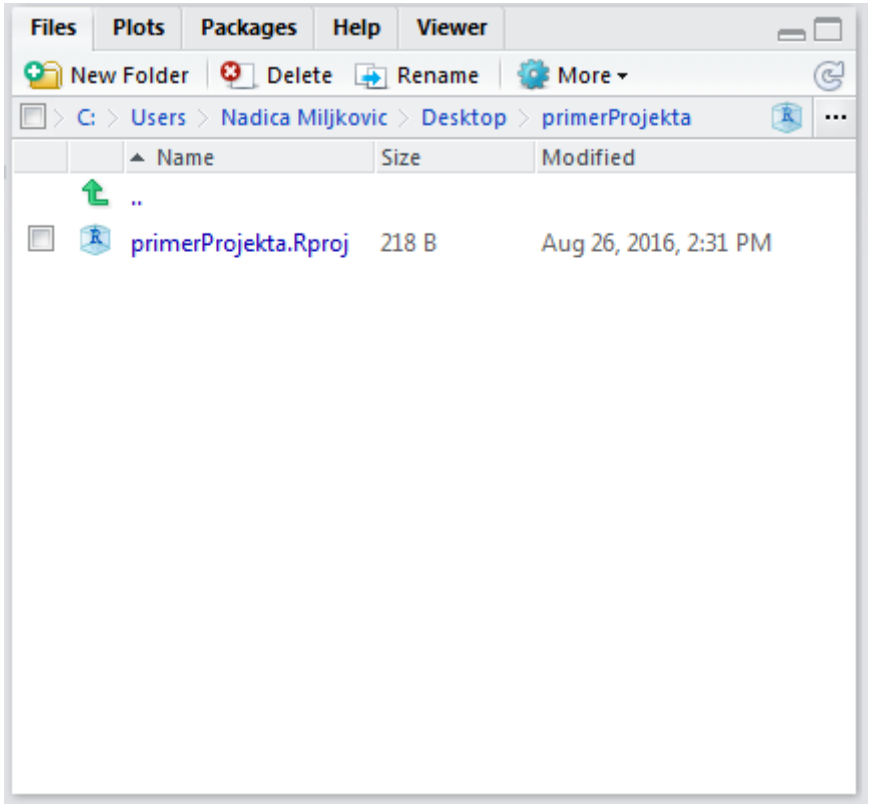
R Sweave dokument





- Odgovaraju Latex formatu (https://www.latex-project.org/) i jednostavno je generisati .pdf dokumenta pozivom komande *Compile pdf*.
- Takođe, poseduju (dodaju se) odgovarajuće celine (*Chunk*) i u dokumentu se nalaze naslov, komentari, kod i rezultat.
- Ekstenzija im je ".Rnw".
- *Rmd* i *Sweave* omogućavaju da se rezultat direktno čuva u izveštaju – posebno je zgodno za deljenje koda i rezultata.

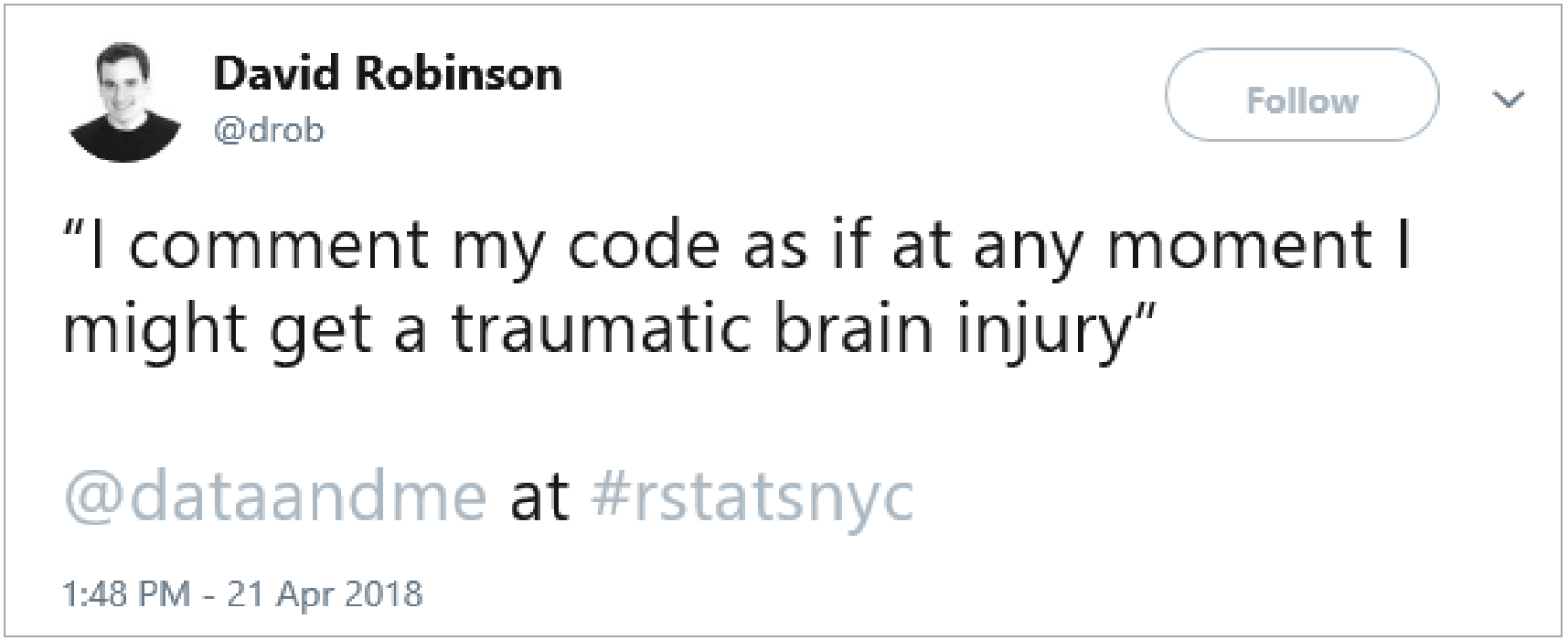
Organizacija podataka i koda u R-u



- Definisanje projekta: *File*/*New Project*/*New Directory*/*Empty Project*.
- Kada se kreira projekat, kreira se i folder/fascikla na računaru gde se nalaze sve informacije od značaja (.R, .Rmd, .Rnw, .txt, .csv, ...). Trebalo bi da se jednostavno (klikom računarskog miša) pregleda sadržaj neke datoteke iz projekta.
- U meniju *File* u R Studio interfejsu, može se videti projekat, kao i sve datoteke koje su vezane za tekući projekat.

Saveti/preporuke

- Jedan “dobar” izveštaj sadrži:
  - ime autora/ke,
  - datum poslednje izmene/generisanja izveštaja,
  - sve relevantne podatke koji omogućavaju da se na osnovu potrebnih znanja i sredstava ponovi merenje/analiza/prikaz rezultata,
  - strukturu
  - i dr.
- Prilikom obrade podataka ne može se na početku definisati izgled finalnog izveštaja, ali je dobro uvek:
  - pisati dnevnik (to je Vaš interni dokument, niko ga ne gleda, pa ne mora formalno, što više podataka to bolje po Vas),
  - sačuvati međuverzije i rezultate obrade podataka i dobro ih organizovati (ako Vi ne umete da se snađete u Vašem kodu ili projektu, onda se neće snaći niko),
  - preuzeti što je više moguće informacija o merenjima,
  - ...



Izvor: https://twitter.com/drob/status/987795355659112453

Daleko bilo!

Šta nikako ne činiti/raditi?

- Postoje neke osnovne razlike u pisanju tehničkog izveštaja, laboratorijskog izveštaja, naučnog izveštaja, ... Potrebno je pratiti uputstva i pravila za pisanje konkretnih primera.
- Nemojte pisati izveštaje koje ni sami ne razumete ili ne možete da ocenite.
- Ne komplikujte i nemojte proširivati temu. Izveštaj koji sadrži > 25-30% uvodnih informacija nije izveštaj, nego pregled teme ...
- Brzo pređite na suštinu.
- Čuvajte se zaključaka! Ako niste obradili podatke na dovoljno velikom statističkom uzorku, možete da pričate isključivo o preporukama, a ne o sigurnim i potvrđenim činjenicama.
- Menjati i modernizovati strukturu izveštaja je OK, ali anarhija u pisanju nije.
- Nemojte preterivati u neozbiljnosti. *Catchy* naslovi su OK, ali šaljive fotografije nisu, osim ako niste nastavnik/ca kome/joj u publici sede studenti/kinje.
- Nemojte mešati stilove. Ako nemate Vaš sopstveni, ugledajte se na nekoga. Minimalizam je uvek dobro rešenje.
- Proverite datume, godine, gramatiku, ... svaka nepreciznost umanjuje značaj Vašeg izveštaja.
- Ne kitite se tuđim perjem ... Pohvalite saradnike/ce.
- Kada počinjete, najvažnije je da se jednog dana setite da ste i Vi bili početnik/ca!
- ...

Na sajtu predmeta

Kako pripremiti i držati prezentaciju

prema Ted VanderNoot, Victoria VanderNoot:
Skills For Maximising Your Graduate Experience
Cognitrix Ltd, London, UK, 2012, ISBN 978-0-9562487-4-9

Korisni linkovi i literatura

- http://rmarkdown.rstudio.com/authoring_quick_tour.html (Uputstvo za pisanje R *markdown* dokumenata, pristupljeno marta 2017. godine)
- http://rmarkdown.rstudio.com/ (Osnovne informacije o ovom paketu sa primerima, pristupljeno marta 2017. godine)
- https://www.rstudio.com/wp-content/uploads/2015/02/rmarkdown-cheatsheet.pdf (Čuvamo šume i recikliramo papir, ali ovaj dokument ima smisla odštampati, pristupljeno marta 2017. godine)
- http://rmarkdown.rstudio.com/articles_intro.html (Kratko uputstvo, pristupljeno marta 2017. godine)
- http://rmarkdown.rstudio.com/gallery.html (Galerija dokumenata i kodova i spisak knjiga, pristupljeno marta 2017. godine)
- https://shiny.rstudio.com/articles/rmarkdown.html (*Shiny* uvod u R *markdown*, pristupljeno aprila 2017. godine)
- http://rmarkdown.rstudio.com/lesson-1.html (R *markdown* kroz *online* lekcije, pristupljeno aprila 2017. godine)

STATISTIKA,
DA, DOBRO STE PROČITALI!

Šta znači rezultat prikazan na slici?

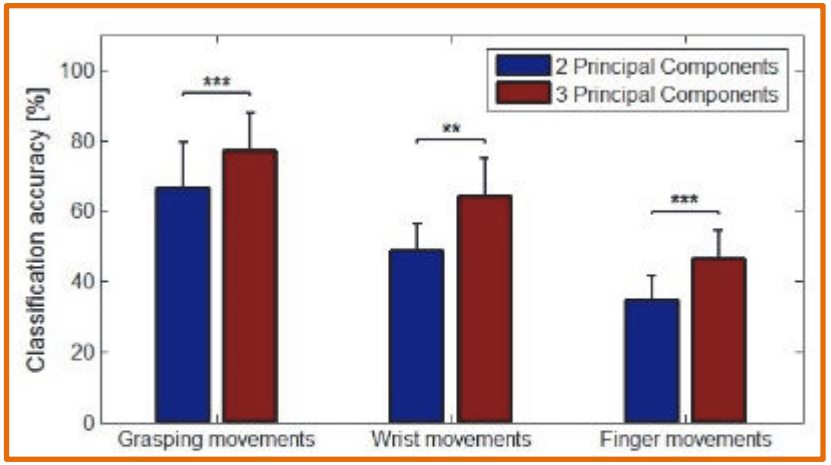


Fig. 5. The classification accuracy for the three sets of movements when using two and three PCs. Graphs report the mean and the standard deviation over all 27 subjects. Horizontal bars with asterisks indicate the statistically significant difference in mean classification accuracy between different numbers of PCs used as classification features. (*, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$).

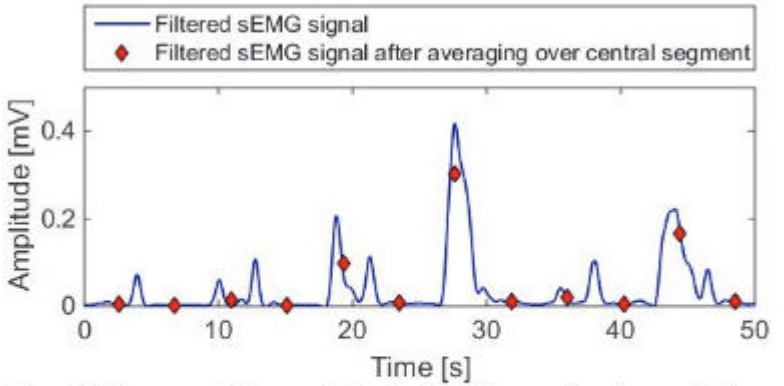
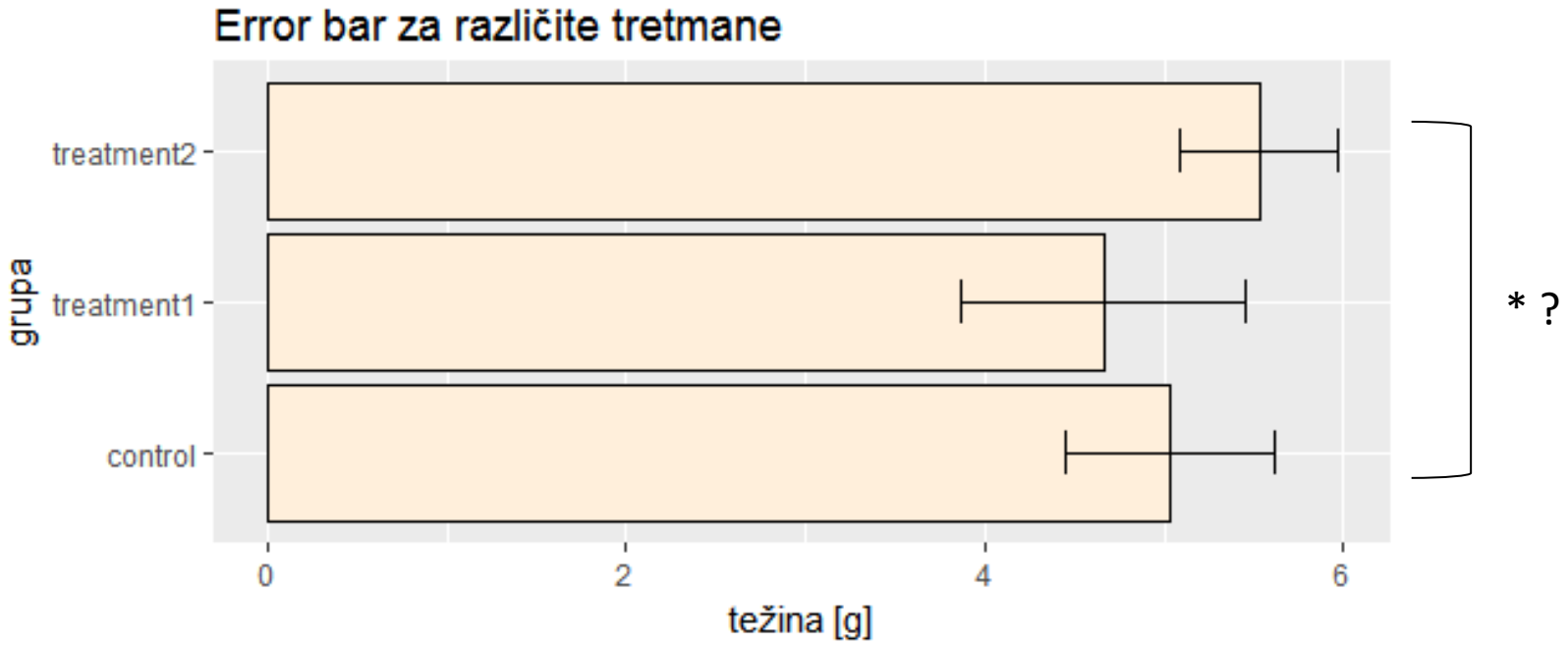


Fig. 6. Six repetitions of the index finger flexion and the following rest for subject no. 1 and one out of 10 electrodes, demonstrating intra-subject variability of sEMG signal.

The selection of sEMG signal preprocessing technique, especially movement segmentation, largely affects the final results of PCA and the classification. The suggested

- Slika je iz rada:
  - M. S. Isaković, N. Miljković, M.B. Popović. Classifying sEMG-based hand movements by means of principal component analysis, *TELFOR JOURNAL*, 7(1): 26-30, 2015, ISSN: 1821 -3251, doi: 10.5937/telfor1501026I, http://journal.telfor.rs/Published/Vol7No1/Vol7No1_A5.pdf.
- Šta znače rezultati označeni sa "**"? To je cilj sledećeg dela lekcije ...

Primer studije



- Primer *PlantGrowth* podataka iz "ISwR" paketa.
- Ako uporedimo srednje vrednosti i standardne devijacije težine biljaka za tri grupe dobija se rezultat kao na slici.
- Kako kvanitativno porediti ove srednje vrednosti? Odnosno, kako odrediti da li tretman #2 utiče **značajno** na povećanje težine biljaka ili je razlika koja se uočava posledica slučajnosti?

Kod?

```
library(ISwR)
dat <- PlantGrowth
head(dat)
summary(dat)

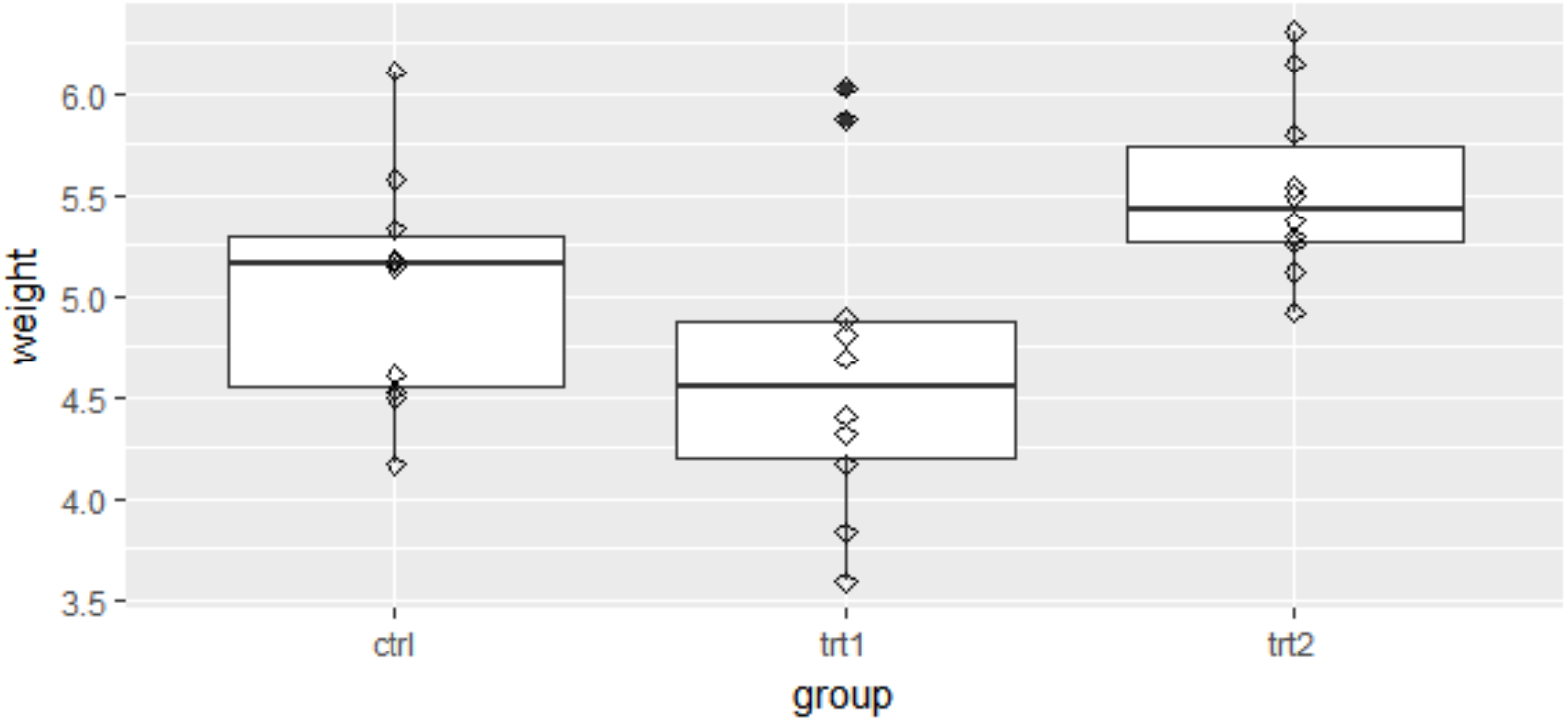
library(dplyr)
ctDat <- filter(dat, group == "ctrl")
t1Dat <- filter(dat, group == "trt1")
t2Dat <- filter(dat, group == "trt2")

pod <- list("numeric")
pod$mean <- c(mean(ctDat$weight), mean(t1Dat$weight),
              mean(t2Dat$weight))
pod$group <- c("control", "treatment1", "treatment2")
pod$sd <- c(sd(ctDat$weight), sd(t1Dat$weight),
            sd(t2Dat$weight))

pod <- as.data.frame(pod)

library(ggplot2)
barE <- ggplot(pod, aes(x = group, y = mean)) +
        ggtitle("Error bar za različite tretmane") +
        xlab("grupa") + ylab("težina [g]") +
        geom_bar(stat = "identity",
                 fill = "antiquewhite1", col = "black")
barE + geom_errorbar(aes(ymin = mean - sd, ymax = mean + sd),
                     width = 0.3) + coord_flip()
```

A pojedinačne realizacije?



- Podaci su u formi pojedinačnih realizacija prikazani romboidima i u formi *box plot*-a. Bolje?
- Za tretman #2 postoje pretpostavke da pozitivno utiče na povećanje težine biljaka u odnosu na kontrolnu grupu, ali bi to trebalo proveriti. Obrnuto važi za tretman #1. Primetiti da ovo ne važi za sve biljke!

Kod?

```
# razlika srednjih vrednosti
mean(ctDat$weight) - mean(t2Dat$weight)

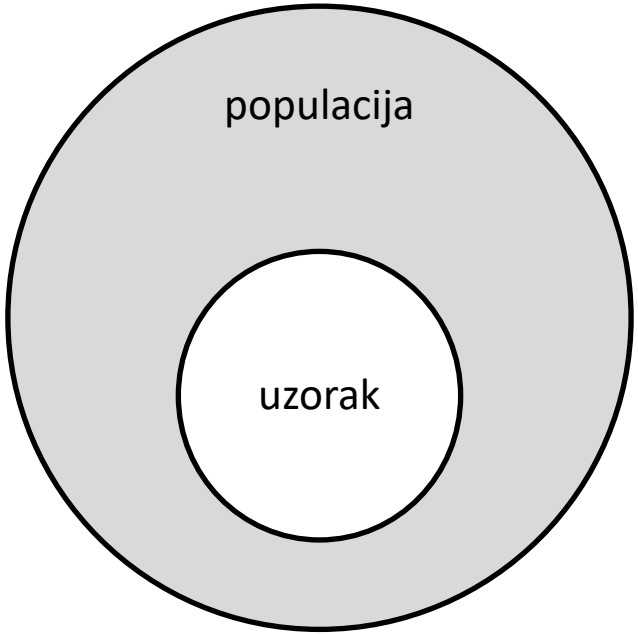

# bar plot i pojedinačne vrednosti
graf <- ggplot(dat,  aes(x = group, y = weight))

graf + geom_boxplot() + geom_point(shape = 5)
```

0,494 g

- Dobijena razlika predstavlja razliku između srednjih vrednosti dva uzorka.
- I srednje vrednosti uzoraka i njihove razlike su slučajne promenljive.

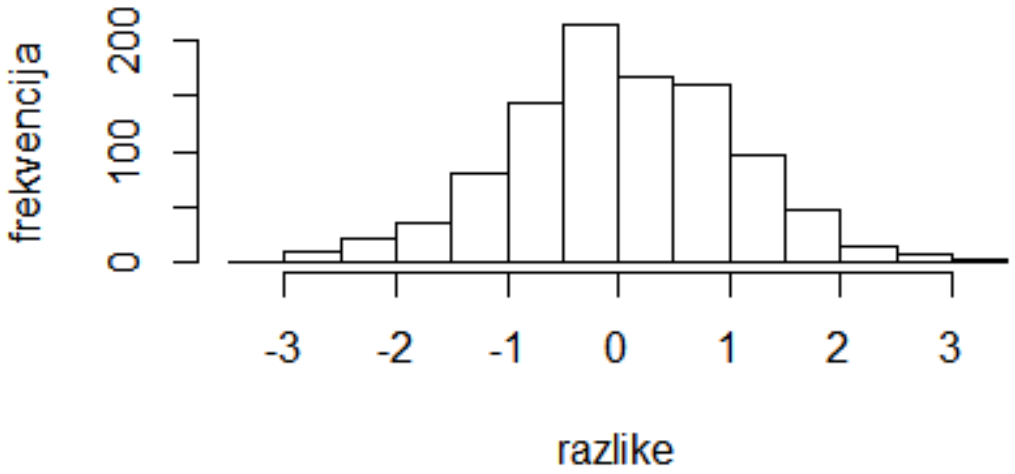
Uzorak *vs*. populacija



- Uzorak je deo populacije koji je dostupan i koji služi da se opiše populacija.
- Populacija nije dostupna (najčešće!)
- Poželjno je da je uzorak što veći.
- Zamislite da su skupovi od po 10 biljaka tretirani kontrolnim i tretmanom #2 zapravo populacije.
- Ako treba da uzmemo uzorak od 5 biljaka iz svake grupe, njihove razlike u srednjim vrednostima će biti kao na slici dole za različita pokretanja koda.
- Da li je razlika sa prethodnim slajdova slučajna ili rezultat tretmana?

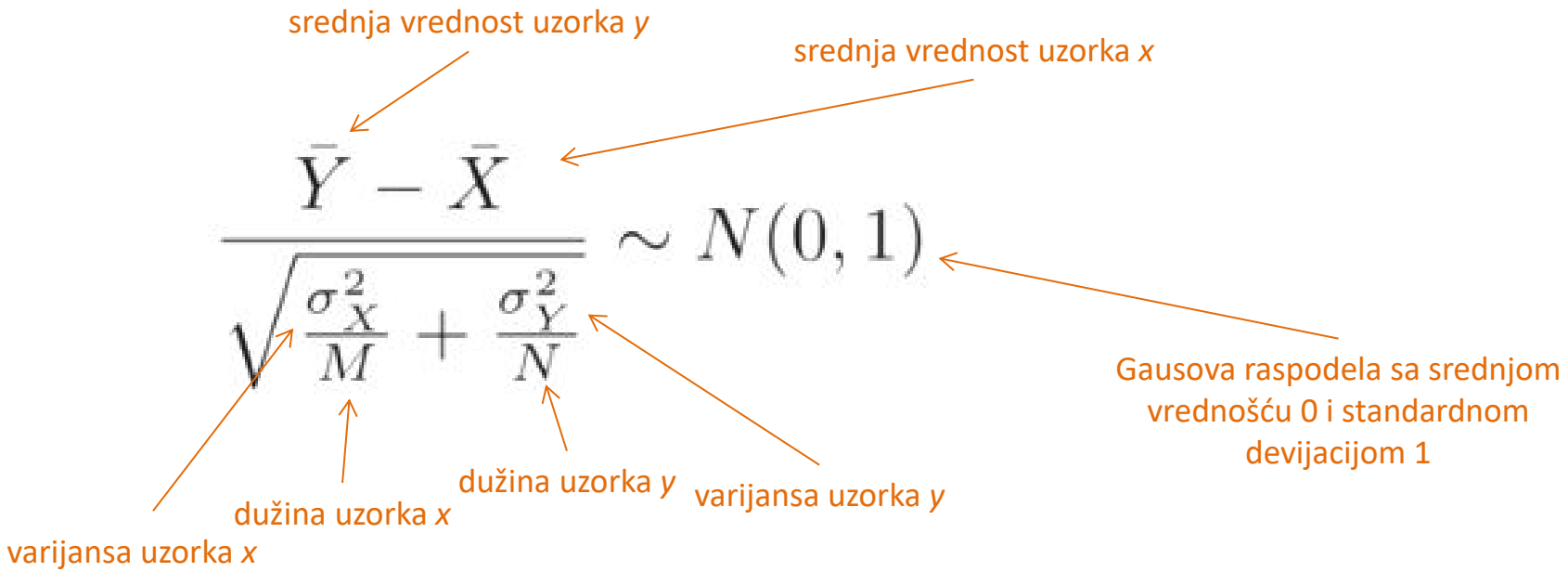
```
> mean(sample(ctDat$weight,5)) - mean(sample(t2Dat$weight,5))
[1] -0.538
> mean(sample(ctDat$weight,5)) - mean(sample(t2Dat$weight,5))
[1] 0.07
> mean(sample(ctDat$weight,5)) - mean(sample(t2Dat$weight,5))
[1] -0.75
> mean(sample(ctDat$weight,5)) - mean(sample(t2Dat$weight,5))
[1] -0.468
```

Testiranje hipoteze



- Postavimo hipotezu:
  - 0-ta hipoteza kaže da ne postoje razlike između težine kontrolnih biljaka i težine biljaka koje su prošle tretman #2, odnosno kada bi smo računali razlike srednjih vrednosti za ove dve grupe (za različite uzorke), onda bi te razlike imale Gausovu raspodelu sa srednjom vrednošću 0: CGT (Centralna Granična Teorema). Što je uzorak veći, to je bolja aproksimacija.
- Pa proveravamo hipotezu.

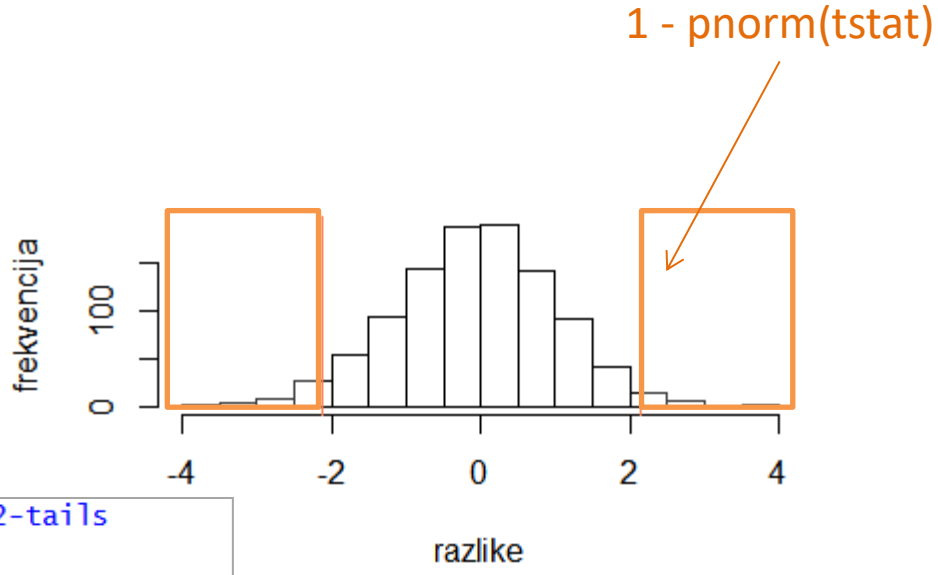
Provera hipoteze




```
> op <- mean(t2Dat$weight) - mean(ctDat$weight)
> N <- length(ctDat$weight)
> sg <- sqrt(
+         var(ctDat$weight) / N +
+         var(t2Dat$weight) / N
+ )
> sg
[1] 0.2314879
```

- Kako se razlikuju srednje vrednosti uzorka i populacije?
- Srednja vrednost uzorka ima normalnu raspodelu, ako je populacija normalno raspodeljena i ako je uzorak dovoljno veliki.
- Na slici je dat kod za računanje **standardne greške razlike dva uzorka**.

Kako dalje?




```
> tstat <- op / sg
> tstat
[1] 2.13402
```

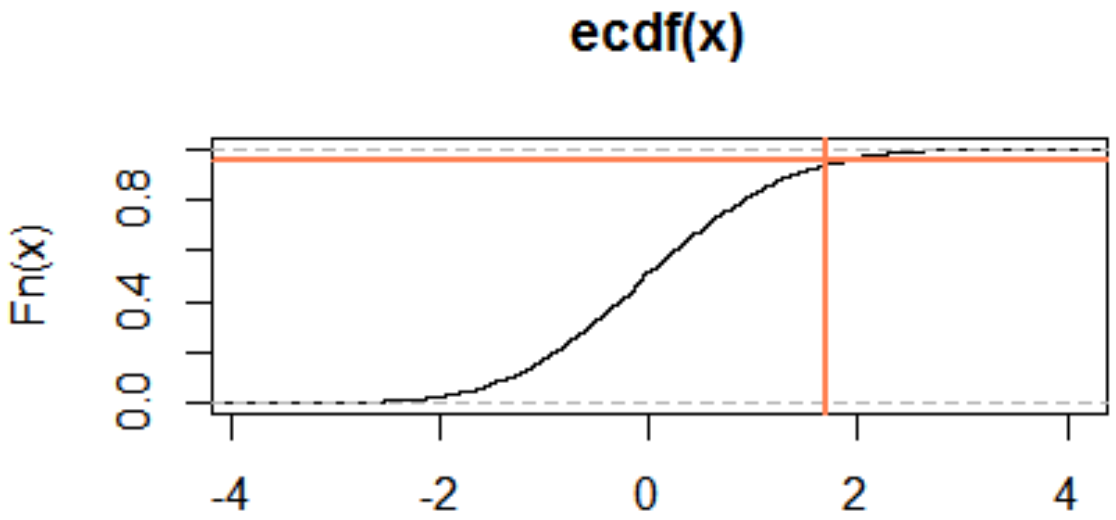
```
> p <- 2 * ( 1 - pnorm(tstat) ) # mnozi se sa 2, 2-tails
> p
[1] 0.03284111
> p < 0.05 # ako je tačno, onda je statistički značajna razlika
[1] TRUE
```

- Ako se dobijena razlika srednjih vrednosti kontrole i tretmana podeli sa standardnom greškom sa prethodnog slajda, dobija se t-statistika (kompletna formula je data na prethodnom slajdu).
- Za ponovljen broj merenja (uzorkovanja) *tstat* promenljiva bi trebalo da ima srednju vrednost 0 i standardnu devijaciju 1, u slučaju da nulta hipoteza jeste tačna.
  - Kada nulta hipoteza nije tačna, onda postoji razlika između kontrole i uzoraka.
- **Da bi se dobila vrednost *p*, potrebno je proveriti kolika je verovatnoća da promenljiva ima vrednost veću od *tstat***. Potrebno je koristiti *pnorm()* funkciju.
- CGT i t-statistika se koriste da se dobije *p* vrednost kada merenja na populaciji nisu na raspolaganju i prethodne formule je moguće koristiti ako uzorak koji je na raspolaganju ima normalnu raspodelu.
- *p* vrednost je verovatnoća da je nulta hipoteza ostvariva, što je ovde 3.28%.

Funkcija raspodele verovatnoće

$$\Pr(a \leq X \leq b) = F(b) - F(a)$$

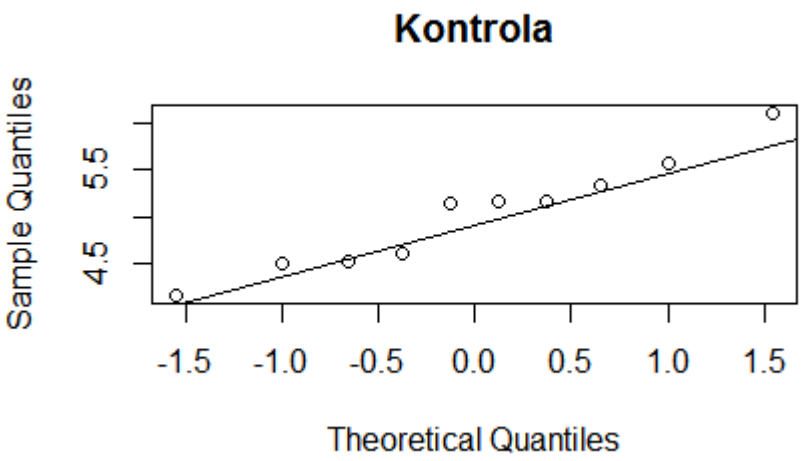
$$\Pr(X \leq b) = F(b)$$

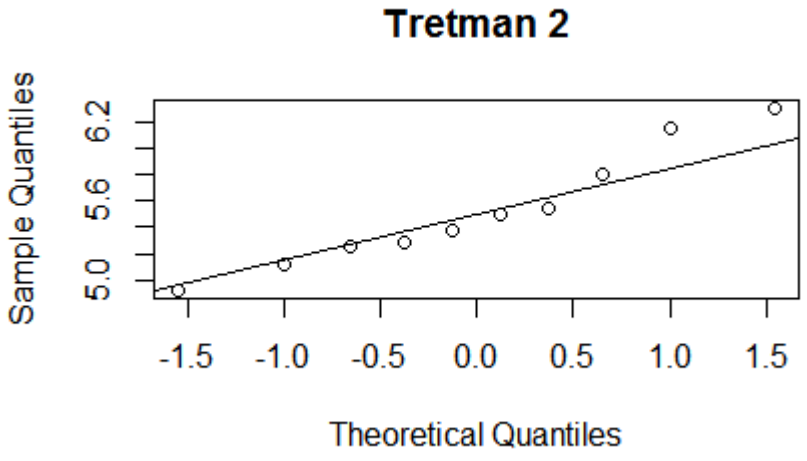


- Verovatnoća da će se slučajna promenljiva *X* biti manja ili jednaka od *b* se izražava preko *F*.
- Pogledati grafik i R kod! ecdf je od eng. *Empirical Cumulative Density Function*.
- *pnorm(b)* = 0.955

```
x <- rnorm(1000)
Fa <- ecdf(x)
plot(Fa)
b <- 1.7
abline(v = b, col = "coral", lwd = 2)
abline(h = pnorm(b), col = "coral", lwd = 2)
```

Da li naši uzorci imaju G. r.?





- Na slici su prikazan grafički način za proveru na osnovu *qqplot*-a.
- Da li su raspodele normalne?

Šta ako nisu Gausove raspodele?

```
> t.test(t2Dat$weight, ctDat$weight)

        Welch Two Sample t-test

data:  t2Dat$weight and ctDat$weight
t = 2.134, df = 16.786, p-value = 0.0479
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.00512787 0.98287213
sample estimates:
mean of x mean of y
    5.526     5.032
```

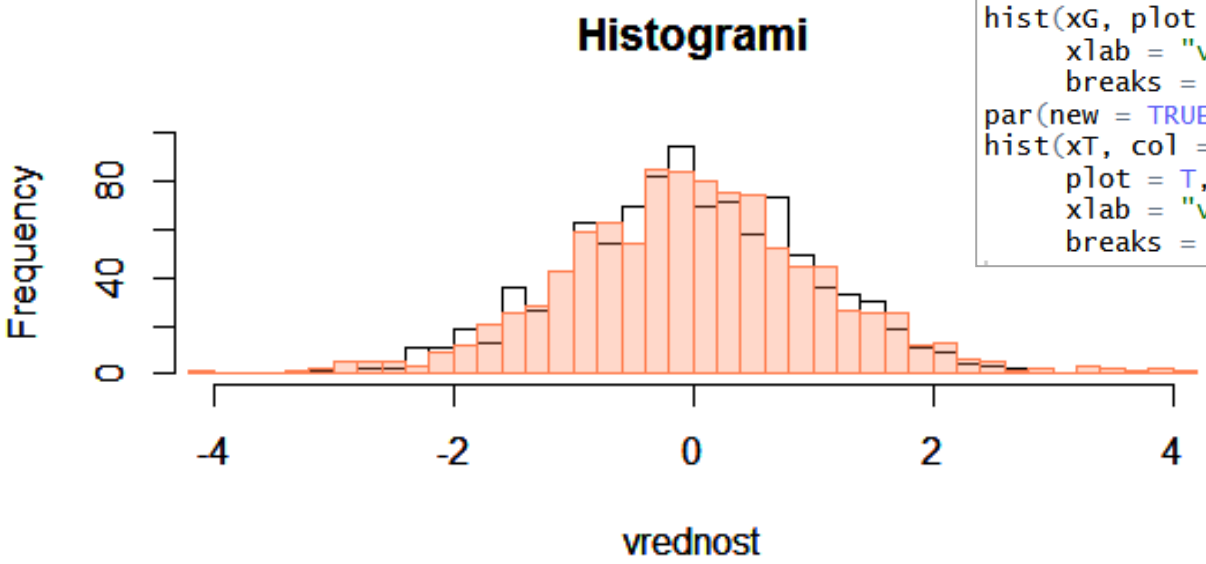
- Ako raspodela nije Gausova, ali "liči" na Gausovu raspodelu, najčešće se aproksimira t-raspodelom tj. Studentovom raspodelom.
- Sa QQ grafika na prethodnom slajdu se vidi da se može raspodela aproksimirati Studentovom t-raspodelom (https://en.wikipedia.org/wiki/Student%27s_t-distribution).
- Tada postoji i gotova funkcija u R-u koja se može koristiti.
- Manuelni račun t-statistike za Studentovu raspodelu je nešto kompleksniji pa neće biti objašnjen. Primetiti da je *t* vrednost ista, ali da je *p* vrednost veća, što je posledica pretpostavke Studentove raspodele.
- **Dakle, tretman 2 ima značajan efekat na porast težine biljaka!**

t-test

Language/Program	Function
Microsoft Excel pre 2010	`TTEST(array1, array2, tails, type)`
Microsoft Excel 2010 and later	`T.TEST(array1, array2, tails, type)`
LibreOffice	`TTEST(Data1; Data2; Mode; Type)`
Google Sheets	`TTEST(range1, range2, tails, type)`
Python	`scipy.stats.ttest_ind(a, b, axis=0, equal_var=True)`
Matlab	`ttest(data1, data2)`
Mathematica	`TTest[{data1,data2}]`
R	`t.test(data1, data2, var.equal=TRUE)`
SAS	`PROC TTEST`
Java	`tTest(sample1, sample2)`
Julia	`EqualVarianceTTest(sample1, sample2)`
Stata	`ttest data1 == data2`

- Realizacija i funkcije za t-test u programskim jezicima.
- Tabela je preuzeta sa sajta Vikipedije: https://en.wikipedia.org/wiki/Student%27s_t-test.

Studentova t-raspodela

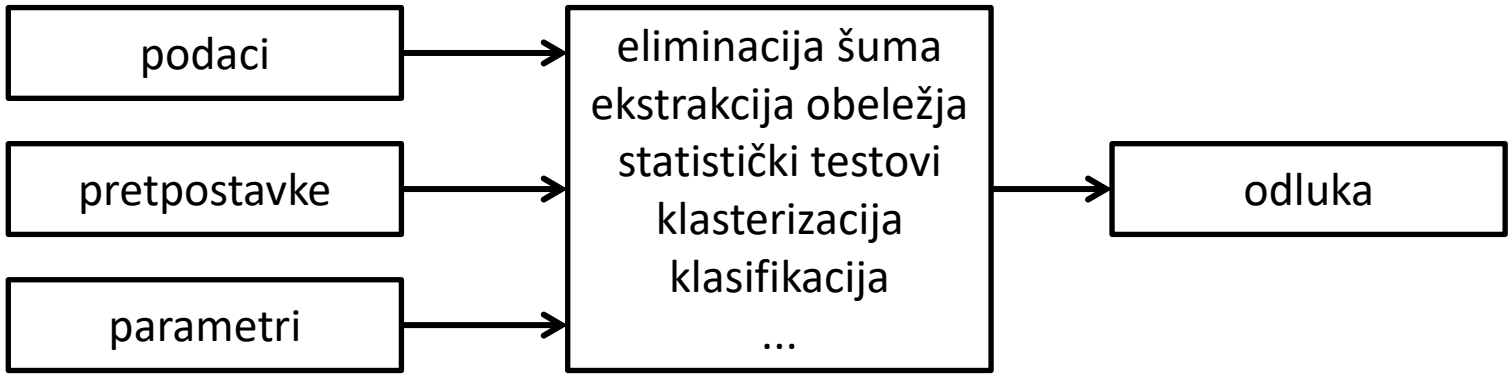



```r
xG <- rnorm(1000)
xT <- rt(1000, df = 10) # df = degrees of freedom

hist(xG, plot = T, ylim = c(0, 100), xlim = c(-4, 4),
     xlab = "vrednost", main = "Histogrami",
     breaks = 33)
par(new = TRUE)
hist(xT, col = alpha("coral", 0.3), border = "coral",
     plot = T, ylim = c(0, 100), xlim = c(-4, 4),
     xlab = "vrednost", main = "Histogrami",
     breaks = 33)
```

- U verovatnoći i statistici, Studentova t-raspodela se koristi kada ne postoji dovoljno veliki broj merenja da bi mogla da se pretpostavi Gausova raspodela tj. kada raspodela merenih podataka (uzorka) podseća na Gausovu raspodelu (https://en.wikipedia.org/wiki/Student%27s_t-distribution).
- Iz ovoga sledi da je za veliki uzorak (veliki broj ponovljenih merenja) Studentova raspodela jednaka Gausovoj raspodelu (CGT).
- Ovu raspodelu je uveo Vilijam Goset (https://en.wikipedia.org/wiki/William_Sealy_Gosset) koji je koristio pseudonim "Student".
- Primetiti za koji broj uzoraka su nacrtani histogrami i uporediti ih.

Obrada podataka



- Potrebno je postaviti hipoteze/pretpostavke.
- Kratak priručnik može se naći i na sajtu: http://kodu.ut.ee/~swen/courses/data-mining-lectures/bioinf-pvalue/index.html.
- Deo ovog predavanja je inspirisan *online* kursom "Data Analysis for Life Science: Statistics and R", profesora Rafaela Irizarrija sa Harvard Univerziteta (https://www.edx.org/course/statistics-r-harvardx-ph525-1x).
- Možete pogledati i materijale u knjizi: Rafael A. Irizarry, Michael I. Love, data Analysis for the Life Sciences, 2016, by Chapman and Hall/CRC, pp. 354, http://genomicsclass.github.io/book/.
- Nastavićemo na sledećem času ...

REZIME

Rezime

Take-home messages

- Automatski izveštaji su korisni.
  - U R-u se najčešće koristi R *markdown* za automatsko generisanje izveštaja.
- Vrlo često se koristi t-test u radu sa biomedicinskim podacima. U R-u postoji funkcija *t.test()*.
- Modifikovana slika od Alice Dietrich na Unsplash