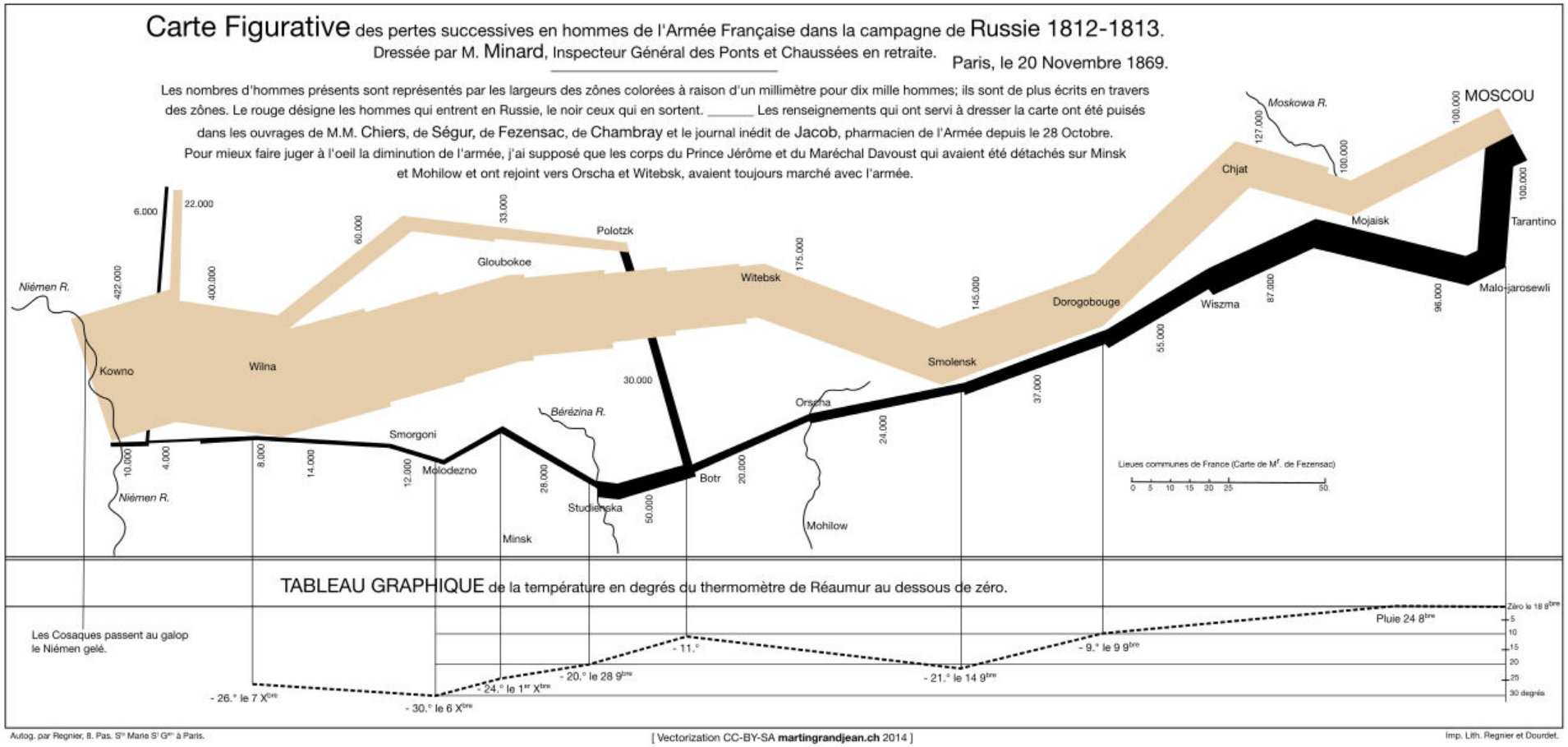
Tehnike obrade biomedicinskih signala 13M051TOBS



Dr Nadica Miljković, vanredni profesor, kabinet 68, nadica.miljkovic@etf.bg.ac.rs

T-TEST SA PRIMERIMA - PODSETNIK

Šta znači rezultat prikazan na slici?

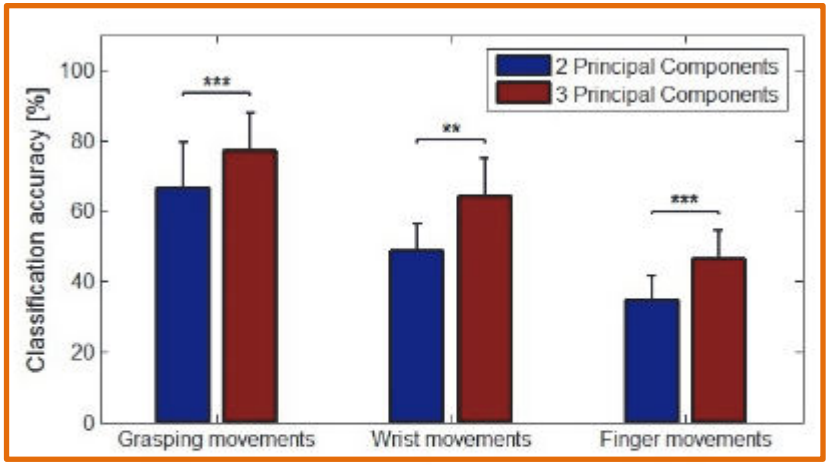


Fig. 5. The classification accuracy for the three sets of movements when using two and three PCs. Graphs report the mean and the standard deviation over all 27 subjects. Horizontal bars with asterisks indicate the statistically significant difference in mean classification accuracy between different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

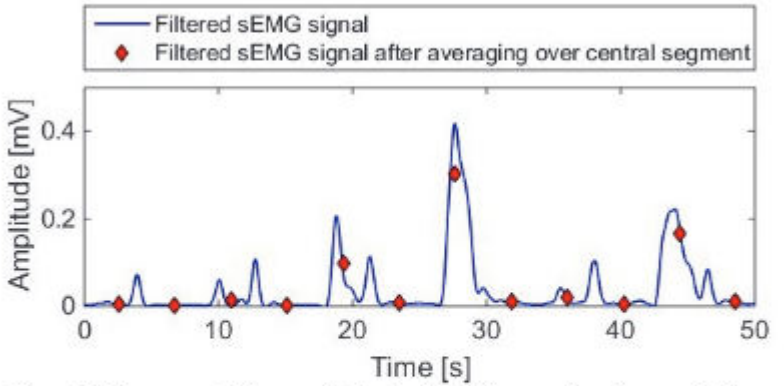
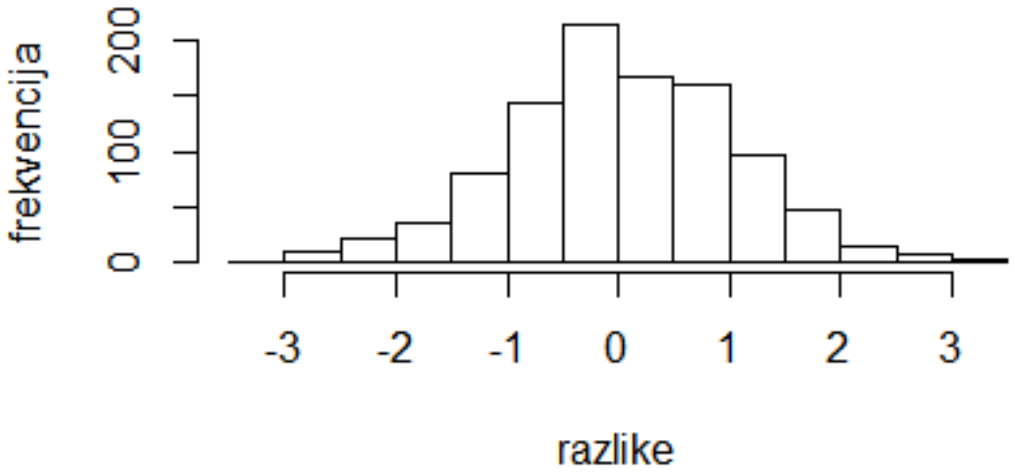


Fig. 6. Six repetitions of the index finger flexion and the following rest for subject no. 1 and one out of 10 electrodes, demonstrating intra-subject variability of sEMG signal.

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

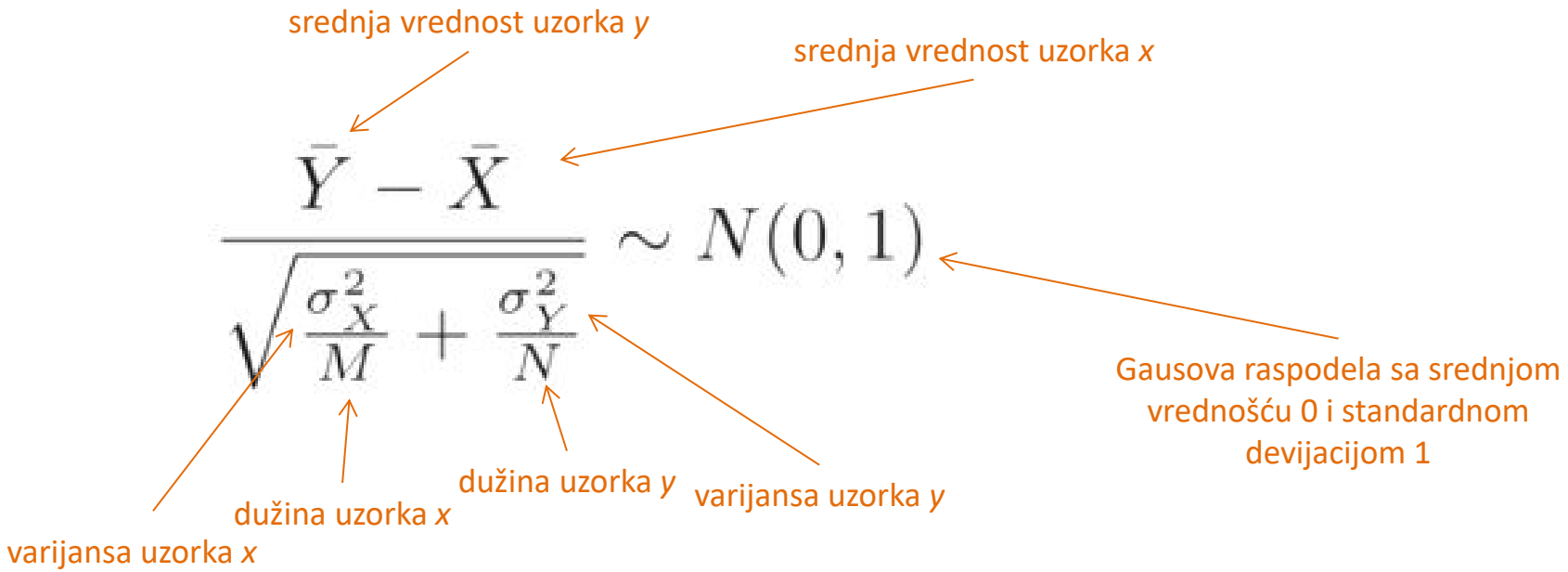
- Slika je iz rada:
  - M. S. Isaković, N. Miljković, M.B. Popović. Classifying sEMG-based hand movements by means of principal component analysis, *TELFOR JOURNAL*, 7(1): 26-30, 2015, ISSN: 1821 -3251, doi: 10.5937/telfor1501026I, http://journal.telfor.rs/Published/Vol7No1/Vol7No1_A5.pdf.
- Šta znače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

Testiranje hipoteze – podsetnik



- 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.

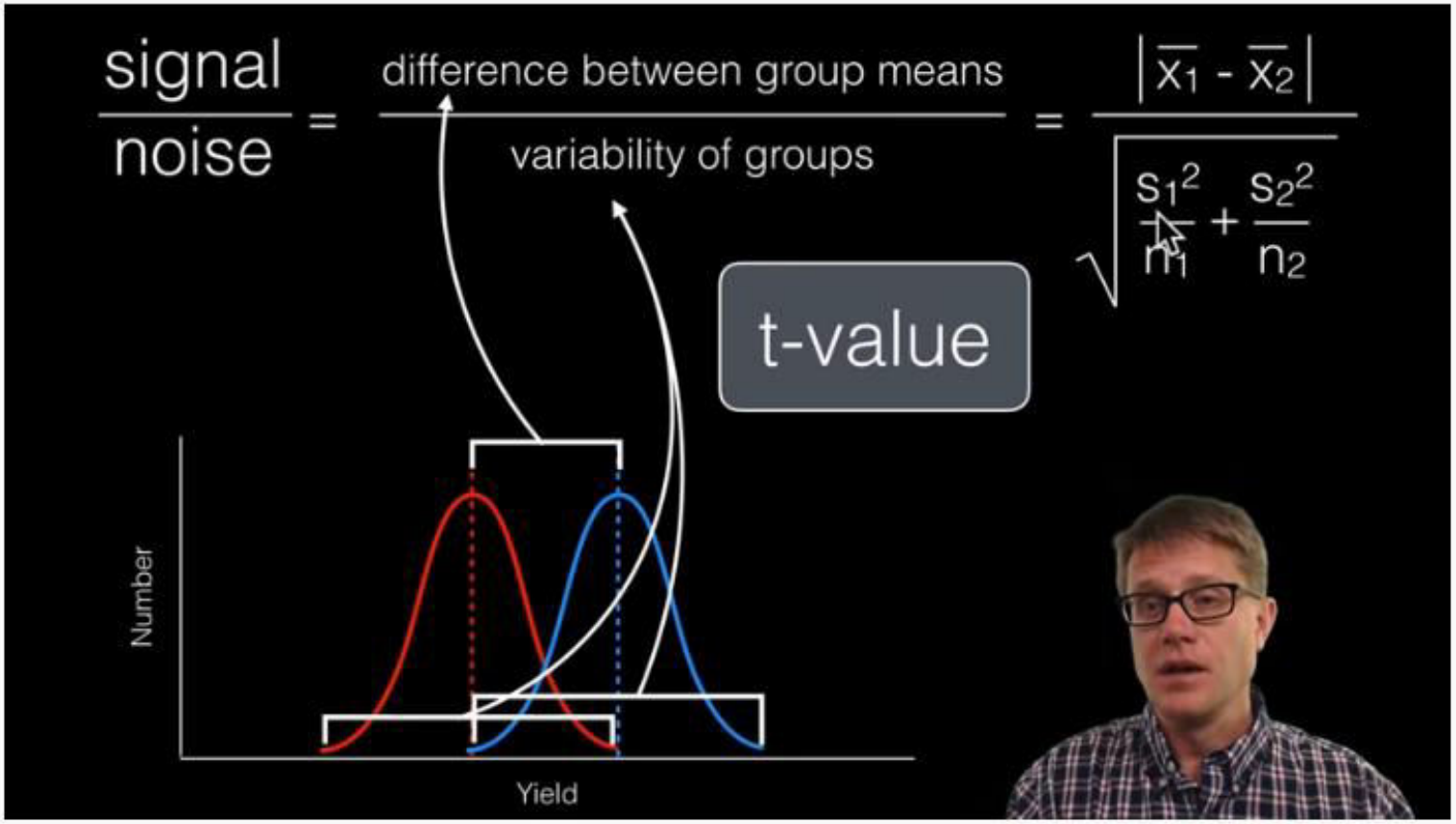
Provera hipoteze – podsetnik




```
> 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
```

- Standardni test t-test značajnosti za proveru nulte hipoteze.
- Na slici je dat način za računanje **standardne greške razlike dva uzorka**.

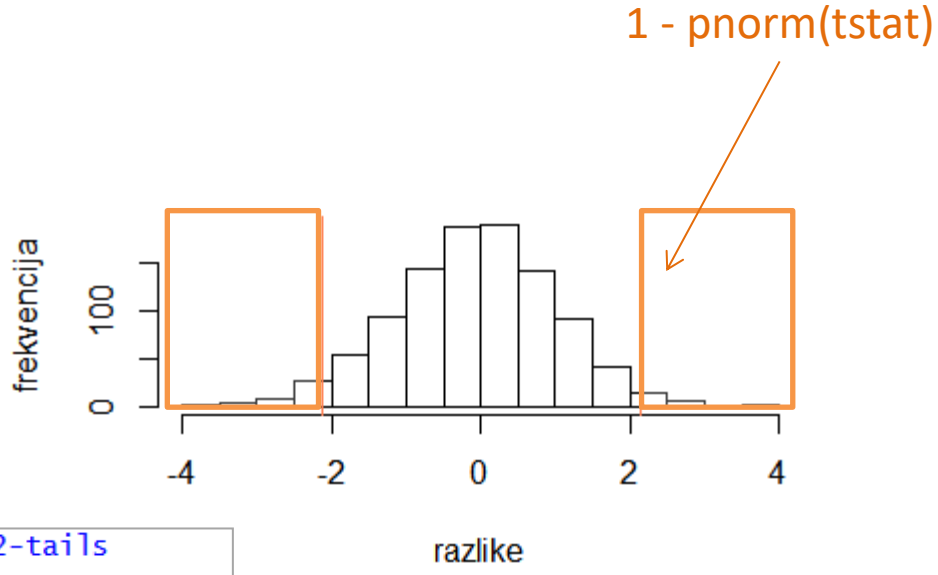
Provera hipoteze – podsetnik



- Odlično video uputstvo
- I objašnjenje formule
- Slika: Student's t-test od Bozeman Science, 2016, https://youtu.be/pTmLQvMM-1M, *Fair Use*
- Primetite kako na izraz za t vrednost utiče povećanje uzorka na slici

Podsetnik

```
> 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
```

- Za ponovljen broj merenja (uzorkovanja) *t* vrednosti 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* tj. od *t*. Potrebno je koristiti *pnorm()* funkciju.
- CPG 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%.

Šta ako nije G. r.? Podsetnik

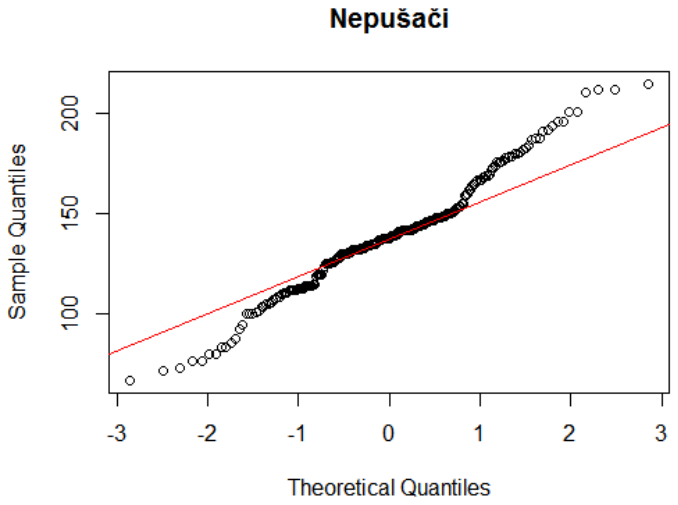
```
> t.test(t2Dat$weight, ctDat$weight)

        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
```

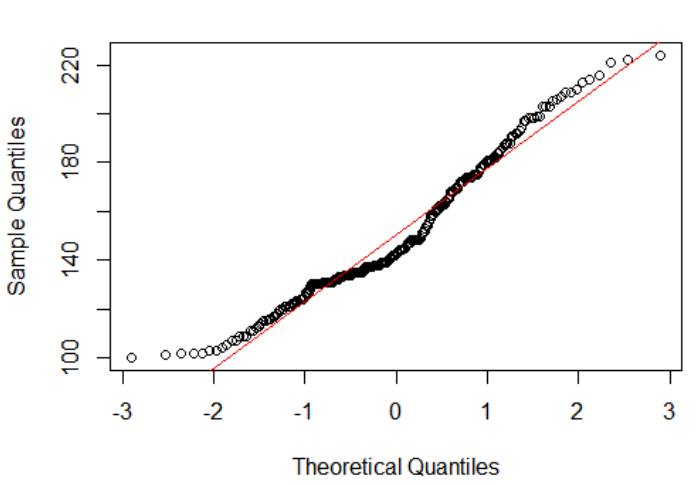
- Ako raspodela nije Gausova, ali "liči" na Gausovu raspodelu, najčešće se aproksimira t-raspodelom.
- Sa QQ grafika se može videti da je najbolje ovu raspodelu aproksimirati Studentovom t-raspodelom (https://en.wikipedia.org/wiki/Student%27s_t-distribution).
- Tada postoji i gotova funkcija u R-u koja 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!**

Razlika srednjih vrednosti od 7.8%.

Podaci iz 8. vežbe

```
> dat <- read.csv("Blood1.csv") # signali iz vežbe 8
> head(dat)
  X SystolicBP Smoke Overwt
1 1        133     0      2
2 2        115     1      0
3 3        140     1      1
4 4        132     0      2
5 5        133     0      1
6 6        138     0      1
>
> # da li pušenje ima uticaja na sistolni krvni pritisak?
> datS <- dat %>% filter(Smoke == 1) %>% select(SystolicBP)
> datNS <- dat %>% filter(Smoke == 0) %>% select(SystolicBP)
>
> qqnorm(datS$SystolicBP, main = "Pušači")
> qqline(datS$SystolicBP, col = "red")
>
> qqnorm(datNS$SystolicBP, main = "Nepušači")
> qqline(datNS$SystolicBP, col = "red")
>
> t <- t.test(datS$SystolicBP, datNS$SystolicBP)
> t$p.value
[1] 1.329196e-05
```



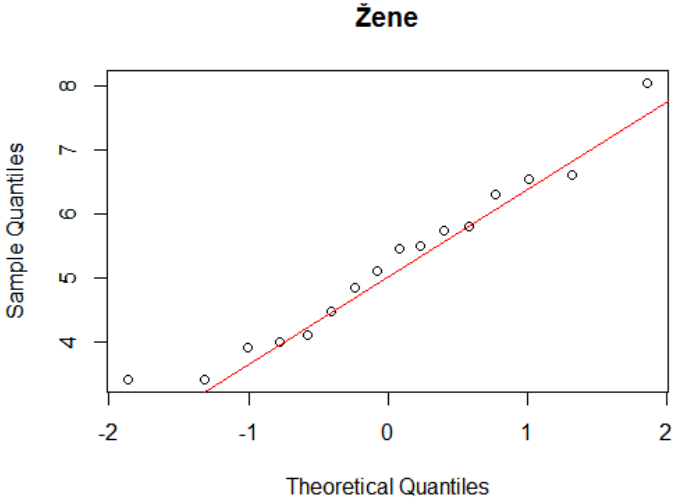


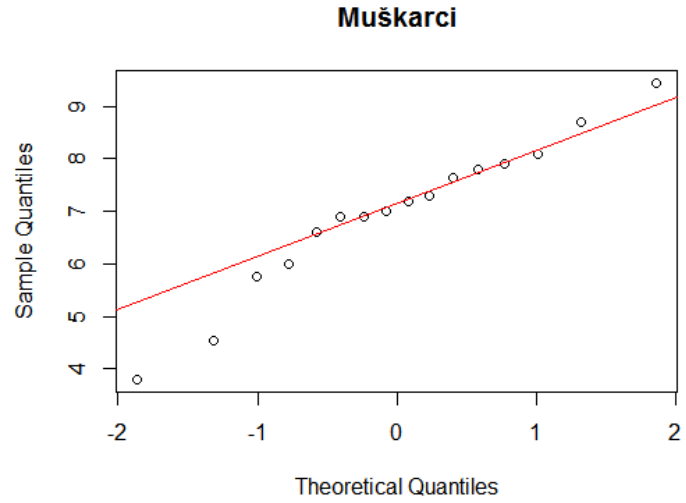
- U podacima koji su korišćeni u ovoj vežbi iz datoteke “Blood1.csv”, potrebno je proveriti da li se sistolni krvni pritisak razlikuje kod pušača i nepušača.
- Kod i rezultat su dati na slici.
- Da li postoji razlika? Da li raspodela podataka prati Gausovu raspodelu?

Razlika srednjih vrednosti od 34%.

Podaci iz 7. vežbe

```
> library(ISwR)
> dat2 <- tlc # podaci korišćeni u vežbi 7
> head(dat2)
  age sex height  tlc
1  35   1    149 3.40
2  11   1    138 3.41
3  12   2    148 3.80
4  16   1    156 3.90
5  32   1    152 4.00
6  16   1    157 4.10
>
> # Da li se razlikuje tlc kod muškaraca i žena u ovoj studiji
> dat2F <- dat2 %>% filter(sex == 1) %>% select(tlc)
> dat2M <- dat2 %>% filter(sex == 2) %>% select(tlc)
>
> qqnorm(dat2F$tlc, main = "Žene")
> qqline(dat2F$tlc, col = "red")
>
> qqnorm(dat2M$tlc, main = "Muškarci")
> qqline(dat2M$tlc, col = "red")
>
> t2 <- t.test(dat2F$tlc, dat2M$tlc)
> t2$p.value
[1] 0.0009492535
```



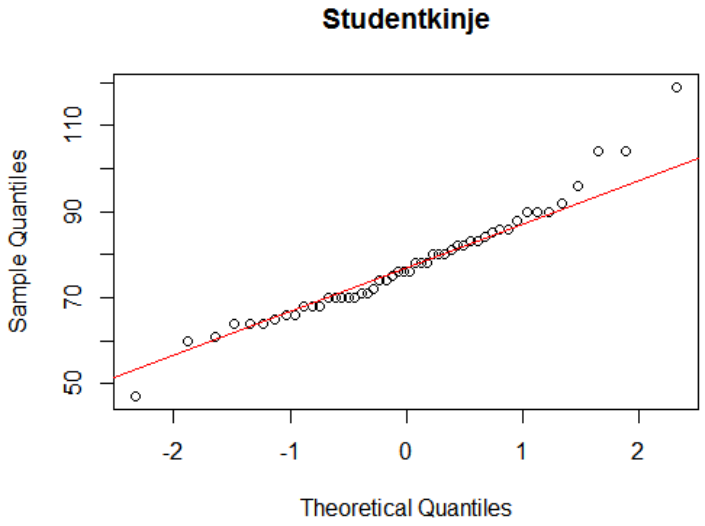


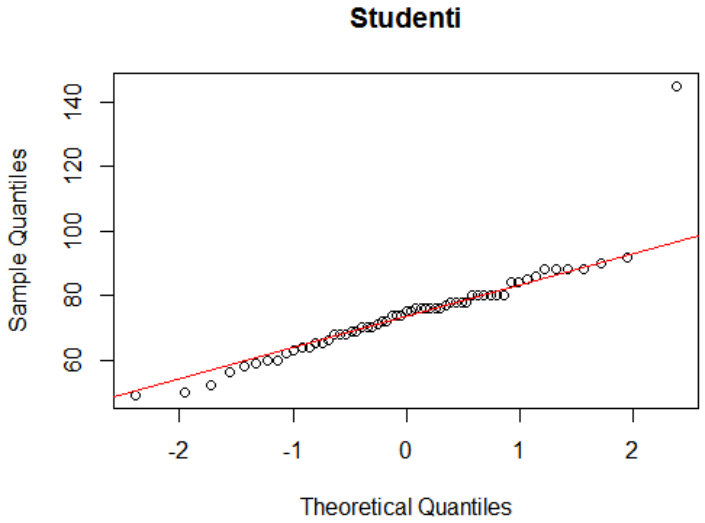
- Korišćeni su podaci u "tlc" bazi iz paketa "ISwR"koji su mereni na kandidatima za transplantaciju srca i pluća. Potrebno je proveriti da li se razlikuje tlc (eng. *total lung capacity*) kod žena i muškaraca na ovom skupu ispitanika.
- Kod i rezultati su dati na slici.
- Da li postoji razlika? Da li raspodela podataka prati Gausovu raspodelu?

Razlika srednjih vrednosti od 4.5%.

Podaci iz 5. vežbe

```
> dat3 <- read.table(url("http://www.statsci.org/data/oz/ms212.txt"),
+                    header = TRUE) # podaci iz vezbe 5
> head(dat3)
  Height Weight Age Gender Smokes Alcohol Exercise Ran Pulse1 Pulse2 Year
1    173     57  18      2      2       1        2   2     86     88   93
2    179     58  19      2      2       1        2   1     82    150   93
3    167     62  18      2      2       1        1   1     96    176   93
4    195     84  18      1      2       1        1   2     71     73   93
5    173     64  18      2      2       1        3   2     90     88   93
6    184     74  22      1      2       1        3   1     78    141   93
>
> # Da li se puls razlikuje kod studenata i studentkinja?
> dat3M <- dat3 %>% filter(Gender == 1) %>% select(Pulse1)
> dat3F <- dat3 %>% filter(Gender == 2) %>% select(Pulse1)
>
> qqnorm(dat3M$Pulse1, main = "Studenti")
> qqline(dat3M$Pulse1, col = "red")
>
> qqnorm(dat3F$Pulse1, main = "Studentkinje")
> qqline(dat3F$Pulse1, col = "red")
>
> t3 <- t.test(dat3M$Pulse1, dat3F$Pulse1)
> t3$p.value
[1] 0.1885446
```

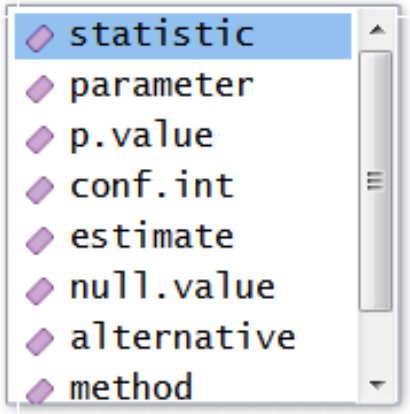




- Ovde su korišćeni podaci sa sajta: http://www.statsci.org/data/oz/ms212.txt koji su mereni na studentima. Potrebno je odrediti da li postoji razlika u pulsu (promenljiva *Pulse1*) kod studenata i studentkinja.
- Kod i rezultati su dati na slici.
- Da li postoji razlika? Da li raspodela podataka prati Gausovu raspodelu?

t.test() u R-u

```
        Welch Two Sample t-test

data:  datS$SystolicBP and datNS$SystolicBP
t = 4.4, df = 489.88, p-value = 1.329e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  6.000844 15.684266
sample estimates:
mean of x mean of y
 150.0263  139.1838
```



- Vrednost funkcije *t.test()* se može dodeliti nekoj promenljivoj.
- Tada postoji niz parametara koji se mogu dobiti iz nove promenljive (jedan od njih je *p* vrednost).
- Ostali parametri su dati na slici. Svi parametri osim intervala poverenja su do sada objašnjeni.

Šta znači interval poverenja?

```
> t$conf.int
[1]  6.000844 15.684266
attr(,"conf.level")
[1] 0.95
```

```
> mean(datS$SystolicBP) - mean(datNS$SystolicBP)
[1] 10.84256
```

- Interval poverenja je slučajan interval sa verovatnoćom od 95% da će "tačna" tj. deterministička razlika koju posmatramo biti u tom intervalu (u ovom slučaju od 6,00 do 15,68 mmHg). Razlika srednjih vrednosti sistolnog pritiska pušača i nepušača koja je dobijena je 10,84 mmHg.
- Interval poverenja je statistički način da se predstavi rezultat koji predstavlja varijabilnost slučajne promenljive (razlike srednje vrednosti dva pritiska iz dva uzorka).
- NAPOMENA: U R-u se može navesti i drugi željeni interval poverenja, koji ne mora da bude 95% (pogledati ulazne parametre *t.test()* funkcije). Naravno, sa promenom verovatnoće, menja se i interval: kada je verovatnoća manja interval je uži.

Značaj intervala poverenja

- Neki autori smatraju, sa opravdanim razlozima, da *p* vrednosti ne bi trebalo koristiti u praksi i da bi ih trebalo zameniti intervalima poverenja.
  - Pogledati više na: R. A. Irizarry, M. I. Love, Data Analysis for the Life Sciences, Leanpub, 2016, https://leanpub.com/dataanalysisforthelifesciences.
- Takođe, preporuka je da *p* vrednost ne bi trebalo samostalno objavljivati iz jednostavnog razloga: STATISTIČKI ZNAČAJAN REZULTAT NE ZNAČI NUŽNO I NAUČNO/TEHNIČKI/TEHNOLOŠKI ZNAČAJAN REZULTAT. Najveći razlog tome je što postoji efekat veličine uzorka na *p* vrednost.
  - Da li statistički značajna promena od 0.1 mmHg sistolnog pritiska ima značaja za zdravlje čoveka koji koristi terapiju sa ovim efektom?
- ZAKLJUČAK: uključite se, proverite razlike (apsolutne, relativne, ...), razmislite o praktičnom značaju rezultata i njihovom smislu, posavetujte se sa kolegama i koleginicama ...

Dodatno

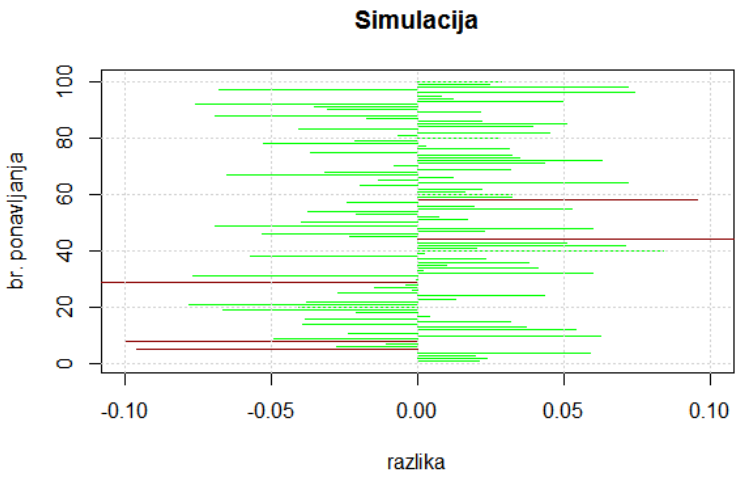
- *Independent/unpaired* tj. nezavisni t-test, odnosi se na odbirke koji se ispituju (eng. *samples*) i koji su dobijene iz nezavisnih opservacija
  - Poređenje krvnog pritiska kod pušača i nepušača
  - Poređenje kapaciteta pluća kod žena i muškaraca
- *Dependent/paired* tj. zavisni t-test, odnosi se na odbirke koji se ispituju (eng. *samples*) i koji su dobijeni na istoj grupi ispitanika
  - Mera ocene oporavka pre i posle terapije kod pacijenata
  - Ovde se očekuje da je isti broj odbiraka dat u obe grupe koje se porede
- *Two-tailed* t-test kada nije poznato da li je promena varijanse pozitivna ili negativna
- *One-tailed* t-test kada je poznat znak promene
- Pogledaćemo još dva primera sa interneta:
  - Kada postoji više grupa koje se porede, http://www.sthda.com/english/wiki/one-way-anova-test-in-r
  - Kada je potrebno grafički predstaviti rezultat, https://rpkgs.datanovia.com/ggpubr/
  - Kada je potrebno primeniti i druge testove, https://cran.r-project.org/web/packages/effectsize/vignettes/effectsize.html
  - Kada raspodela nije Gausova i kako je ispitati

Simulacija

```
> set.seed(3)
> p1 <- rnorm(1000)
> set.seed(7)
> p2 <- rnorm(1000)
>
> t4 <- t.test(p1, p2)
> t4$p.value
[1] 0.9397395
```

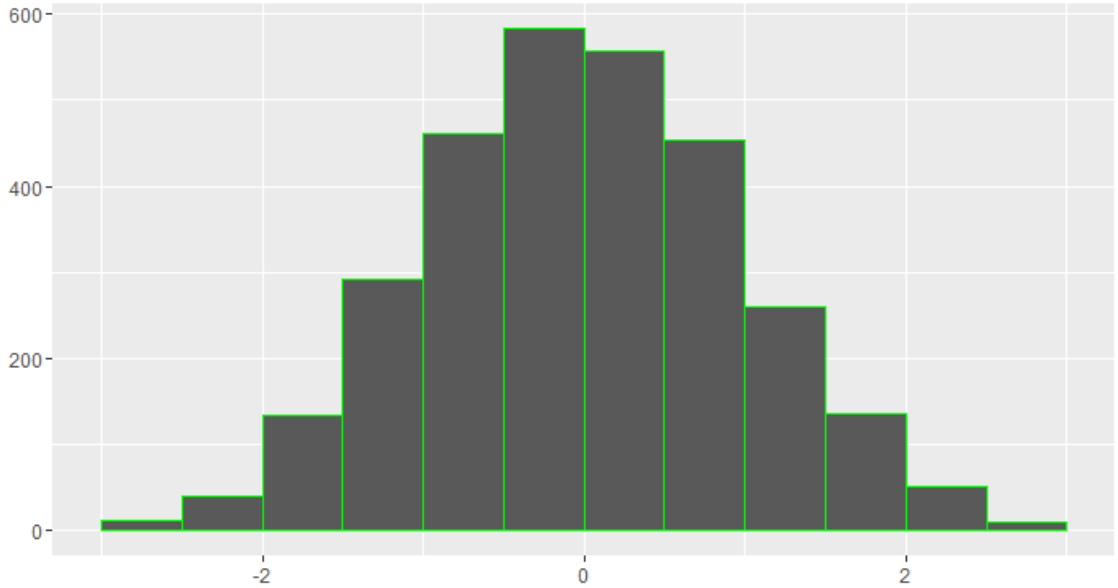
Monte Karlo simulacija?

```
> int <- 100
> plot(0, 0, type = "l", ylim = c(1,100), xlim = c(-0.1, 0.1),
+      xlab = "razlika", ylab = "br. ponavljanja",
+      main = "Simulacija")
> ip <- 0
> for (interval in 1:int) {
+         p1 <- rnorm(1000)
+         p2 <- rnorm(1000)
+         df <- mean(p1) - mean(p2)
+         t <- t.test(p1, p2)
+         color <- ifelse(t$p.value < 0.05, "darkred", "green")
+         ip <- ifelse(color == "darkred", ip <- ip + 1, ip <- ip)
+         lines(c(0, df), c(interval, interval), col = color)
+ }
> grid()
> ip
[1] 5
```



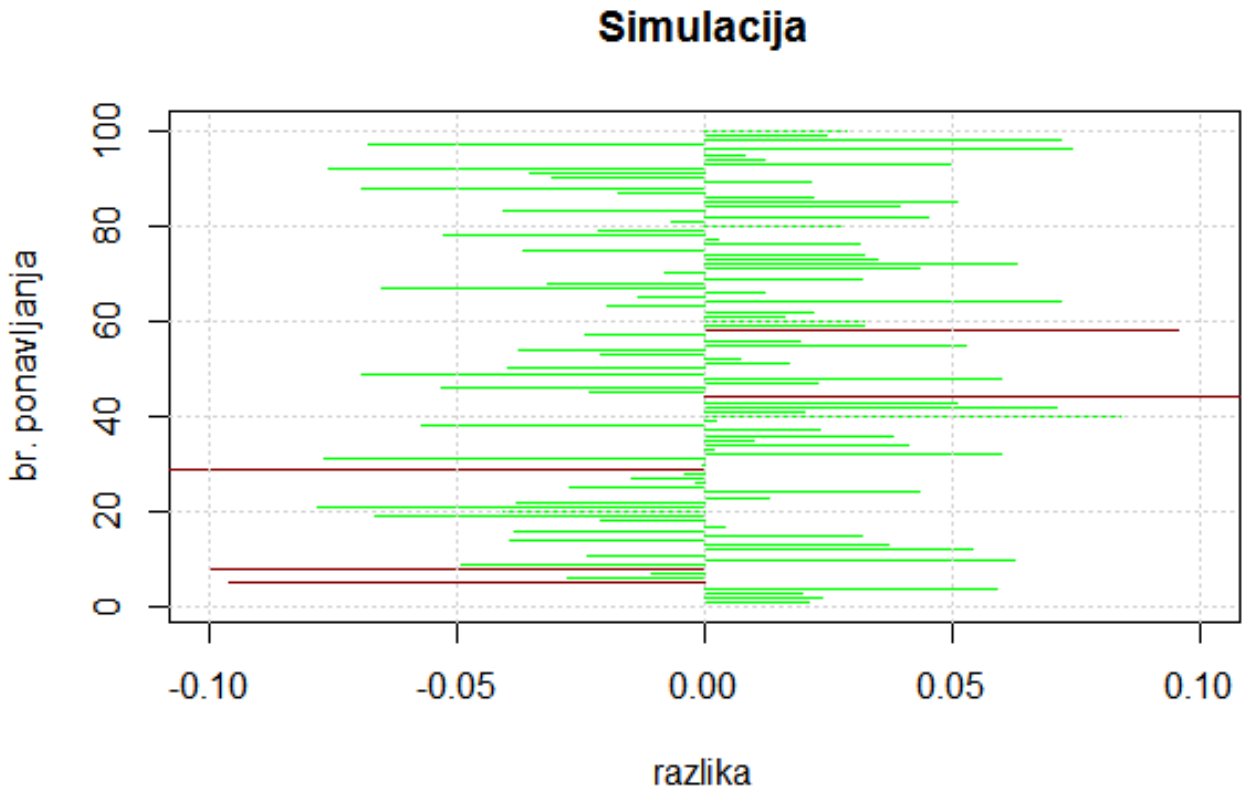
- Da li su razlike koje su dobijene ponovljive? Odnosno, da li bi smo dobili iste zaključke kada bi smo ponovili ova merenja?
- Može se vrlo jednostavno realizovati simulacija koja to pokazuje. Primetiti da je $\alpha = 0.05$.
- Na primer, koristimo *t.test()* za normalnu raspodelu da se pokaže da li postoji značajna razlika između neka dva broja. To kasnije ponavljamo 100 puta.

Monte Karlo simulacija



- Monte Karlo simulacija koristi mogućnost računara da se generišu **pseudo**-slučajni brojevi.
- Ona je veoma korisna da se ispitaju ideje, metode i hipoteze, odnosno da se izvrši simulacija merenja.
- Vrlo često se u praksi koristi **parametarska Monte Karlo simulacija**. To je simulacija koja ne uzima podrazumevane parametre, već parametre (npr. standardnu devijaciju i srednju vrednost) koji su dobijeni na relativno malom uzorku (realnim podacima), kako bi se npr. isplaniralo kompletno merenje. T-test je, takođe, parametarski test.
- Više na: https://en.wikipedia.org/wiki/Monte_Carlo_method.

Da se vratimo na rezultat simulacije



- Da li je ovaj rezultat očekivan?
- Postoje dva tipa greške u statistici: Tip II greške (kada postoji razlika, a naša analiza daje suprotan rezultat) i Tip I greške (kada ne postoji razlika, a naša analiza detektuje značajnu razliku).
- Na slici je crvenom bojom označen Tip I greške.
- Greške se događaju iz dva razloga: 1) radi se o slučajnim podacima i 2) ne postoji dovoljno veliki uzorak. Češći je drugi uzrok greške!

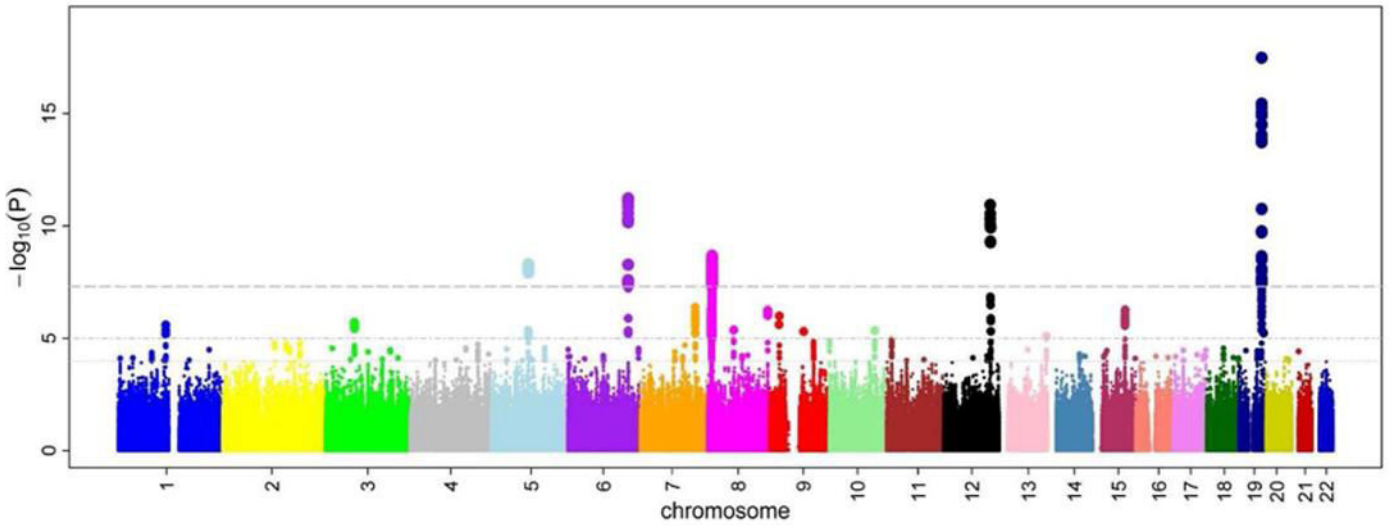
Snaga testa

- Svaki put kada se uradi statistički test, postoji verovatnoća da je došlo do greške! Kada je $p = 0,01$, to će se desiti jedanput u 100 puta (Tip I greške).
  - Ne postoji neki jasno definisan razlog zašto se koristi α koje je jednako 0,01 ili 0,05, osim što su to često korišćene vrednosti u literaturi. Ne postoji ispravno i neispravno α!
- Snaga testa je parametar koji pomaže da se odrediti dovoljan broj uzoraka kako bi se izbegao Tip II greške.
  - U R-u postoji paket "pwr" na CRAN-u koji omogućava da se izračuna snaga testa (https://cran.r-project.org/web/packages/pwr/index.html). O tome ovde neće biti reči.
- Snaga testa je verovatnoća o odbijanju 0-te hipoteze kada je alternativna hipoteza tačna. Što je veći uzorak to je veća snaga testa.
- Za greške Tipa I na prethodnom slajdu snaga testa je 95%.
- Računanje snage testa, najčešće se obavlja pre početka studije iz dva razloga: 1) etički i 2) radi uštede vremena i resursa.
- Snaga testa, razumljivo, raste sa povećanjem uzorka i sa povećanjem α.
  - Otud, manja *p* vrednost može da znači da je uzet relativno veliki uzorak, a ne nužno da je pokazana praktično značajna razlika.

JOŠ STATISTIČKIH TESTOVA I GRAFIKA

Manhattan plot

```
> -log10(0.01)
[1] 2
> -log10(0.05)
[1] 1.30103
```



By M. Kamran Ikram et al - Ikram MK et al (2010) Four Novel Loci (19q13, 6q24, 12q24, and 5q14) Influence the Microcirculation In Vivo. PLoS Genet. 2010 Oct 28;6(10):e1001184. doi:10.1371/journal.pgen.1001184.g001, CC BY 2.5, https://commons.wikimedia.org/w/index.php?curid=18056138

By DigbyDalton - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=56674407

- Ovo je vrsta grafika u kome je omogućena vizuelizacija velike količine podataka i njihovih *p* vrednosti (https://en.wikipedia.org/wiki/Manhattan_plot).
- Kako bi se izdvojile najznačajnije vrednosti na relativno velikom skupu gena (kao što je prikazano na slici) koristi se negativni logaritam sa osnovom 10.
- Grafik je ime dobio po slici prikazanoj na donjem panelu.

U našoj varijanti

Čaj?

By Vanderdecken - Author's original own work., Public Domain, https://commons.wikimedia.org/w/index.php?curid=1613904.

- **Povod:** *Team Building Tea Party*: *Lady Tasting Tea*
- **Mesto:** Cambridge, England
- **Vreme:** 1920.
- **Glavni učesnici:** Muriel Bristol (https://en.wikipedia.org/wiki/Muriel_Bristol) i Ronald Fisher (https://en.wikipedia.org/wiki/Ronald_Fisher)
- **Tema:** Šta je prvo sipano u šolju?
- **Hipoteza:** Da li je to što se dogodilo slučajnost ili ne?
- **NULL hipoteza:** dama samo nagađa

Rešenje? Test asocijacije?

Tea-Tasting Distribution

Success count	Permutations of selection	Number of permutations
0	oooo	1 × 1 = 1
1	ooox, ooxo, oxoo, xooo	4 × 4 = 16
2	ooxx, oxox, oxxo, xoxo, xxoo, xoox	6 × 6 = 36
3	oxxx, xoxx, xxox, xxxo	4 × 4 = 16
4	xxxx	1 × 1 = 1
Total		70

Tabela je preuzeta sa: https://en.wikipedia.org/wiki/Lady_tasting_tea, *Fiar Use*.

- 8 šoljica čaja (4 gde je prvo sipan čaj i 4 gde je prvo sipano mleko) i šoljice su slučajno raspoređene.
- Uspela je da pogodi! To je bilo moguće u 1 od 70 slučajeva (p = 1/70).
- Tako je nastao *Fisher's exact* test, odnosno način kako da se izračuna p vrednost.
  - Primetiti da ovaj test nema veliku snagu, jer je uzorak od 8 šoljica relativno mali.
- U praksi se često koriste testovi slični ovom testu.
- Gde bi ovakav ili sličan test mogao da se koristi u biomedicinskom inženjerstvu? Koje druge testove asocijacije znate?

tip testa	kada se koristi?
korelacija	Kada se prverava asocijacija dve promenljive.
Pirsona korelacija	Kada se proverava asocijacija normalno raspodeljenih promenljivih. Linearna relacija.
Spirmanova korelacija	Kada se proverava asocijacija dve promenljive koje ne moraju da imaju G. r. i koje ne moraju da budu linearno zavisne.
Chi-square test	Za proveru asocijacije kategoričkih promenljivih.
poređenje sr. vr.	**poređenje sr. vr. za dve grupe podataka**
paired t-test	Provera zavisnosti dve zavisne promenljive.
nezavisni t-test	Provera zavisnosti dve nezavisne promenljive.
ANOVA	Provera zavisnosti sr. vr. u grupi.
regresija	**proverava se da li promena jedne utiče na promenu druge promenljive**
jednostavna regresija	kako se promena u prediktoru odražava na izlaznu promenljivu
multiple regresija	kako se kombinovana promena u prediktoru odražava na izlaznu promenljivu
neparametarski testovi	**primenjuje se na podacima na kojima se ne mogu primeniti parametarski testovi**
Wilicoxon rank-sum test	proverava se razlika dve nezavisne promenljive
Wilcoxon sign-rank test	proverava se razlika dve zavisne promenljive
Sign test	proverava se da li se dve zavisne promenljive razlikuju

Više na https://cyfar.org/types-statistical-tests.

EDA

- EDA (eng. *Exploratory Data Analysis*) je način analize podataka i sumiranja osnovnih karakteristika najčešće na vizuelni način (https://en.wikipedia.org/wiki/Exploratory_data_analysis).
  - EDA je definisana 1961. godine i poslužila je kao inspiracija za nastanak programskog jezika S.
  - *Outlier*-i se mogu najčešće relativno jednostavno prikazati u ovoj fazi.
  - Histogram je osnovni alat u EDA i koristi se kao vizuelni deskriptor podataka. Pored histograma i funkcije raspodele verovatnoće, postoji i vizuelizacija aproksimacije sa normalnom raspodelom.
    - Na TOBS-u je rađen i *qqplot*, ali i druge metode provere Gausove raspodele.
    - Kada podaci nemaju Gausovu raspodelu, koristi se *boxplot*, a kada imaju koristi se *errorbar*.
- Još dva važna pojma:
  - *Statistical inference* – bavi se osobinama populacije iz koje su dobijeni podaci (https://en.wikipedia.org/wiki/Statistical_inference)
  - *Descriptive statistics* – bavi se osobinama izmerenih podataka (https://en.wikipedia.org/wiki/Descriptive_statistics)
- Dalja statistička analiza prevazilazi gradivo predviđeno na ovom predmetu.

CASE STUDY

Analiza podataka

- Započinje posmatranjem signala (odbiraka).
- Nastavlja se vizuelizacijom signala.
  - Poželjno je na početku definisanja metode za obradu biomedicinskih signala, vizuelizovati rezultate u raznim fazama manipulacije i obrade biomedicinskih podataka.
- Hipoteza:
  - Koje informacije su od značaja?
  - Na koja pitanja treba dati odgovor?
- U ovom primeru:
  - Puls meren na jednom zdravom ispitaniku Elektrotehničkom fakultetu u Beogradu sa IR (eng. *infra red*) senzora i sa EKG-om (elektrokardiografija).
  - Potrebno je proveriti pouzdanost IR senzora za merenje pulsa ispitanika i proveriti zavisnost ovog merenja u odnosu na standardnu EKG metodu merenja.

Mereni podaci

puls - Notepad

File Edit Format View Help

```
# U prvoj koloni se nalaze EKG podaci mereni na grudnom košu ispitanika. Pojačanje signala je bilo 5000, a frekvencija odabiranja 1000 Hz.
# U drugoj koloni se nalaze podaci mereni sa IR senzorom na kažiprstu leve ruke ispitanika. Frekvencija odabiranja je 1000 Hz.
# Ostale kolone u ovom fajlu (3-16) su posledica projektovanog akvizicionog softvera i ne koriste se.
# Subjekat je tokom merenja bio postavljen u sedeći položaj.
-0.033  1.781   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000
-0.026  1.781   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000
-0.030  1.781   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000
-0.041  1.780   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000
-0.040  1.780   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000
-0.016  1.780   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000
-0.022  1.781   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000
0.012   1.781   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000
-0.003  1.780   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000
0.010   1.782   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000
0.010   1.781   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000
0.019   1.780   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000
0.013   1.781   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000
0.018   1.782   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000
0.022   1.781   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000
0.031   1.781   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000
0.030   1.781   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000
0.028   1.781   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000
0.016   1.781   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000
0.024   1.781   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000
0.011   1.782   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000
0.013   1.781   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000
```

Ln 1, Col 1

- Obzirom da je postavljena hipoteza (prvi korak), potrebno je proveriti strukturu podatka (drugi korak u analizi).
- Na slici je prikazan izgled tekstulane datoteke sa podacima.

EKG i IR za merenje pulsa

```
> dat <- read.table("puls.txt")
> head(dat)
      V1    V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14 V15 V16
1 -0.033 1.781  0  0  0  0  0  0  0   0   0   0   0   0   0   0
2 -0.026 1.781  0  0  0  0  0  0  0   0   0   0   0   0   0   0
3 -0.030 1.781  0  0  0  0  0  0  0   0   0   0   0   0   0   0
4 -0.041 1.780  0  0  0  0  0  0  0   0   0   0   0   0   0   0
5 -0.040 1.780  0  0  0  0  0  0  0   0   0   0   0   0   0   0
6 -0.016 1.780  0  0  0  0  0  0  0   0   0   0   0   0   0   0
> fs <- 1000
> length(dat$V1)/fs
[1] 62.476
>
```

- Treći korak: učitavanje podataka.
- Podaci iz datoteke su smešteni u promenljivu *dat*.
- Potrebno je proveriti u kom vremenskom intervalu su mereni podaci i izvršiti kratak pregled sadržaja promenljive *dat*.

EKG i IR za merenje pulsa

```
> EKG <- dat$V1 / 5 # zbog pojacanja, da bi signal bio u mV
> IR <- dat$V2
> library(signal)
> f1 <- butter(2, 0.3/(fs/2), "high")
> f2 <- butter(3, 40/(fs/2), "low")
> EKGf <- filtfilt(f1$b, f1$a, filtfilt(f2$b, f2$a, EKG))
> IRf <- filtfilt(f1$b, f1$a, filtfilt(f2$b, f2$a, IR))
>
```

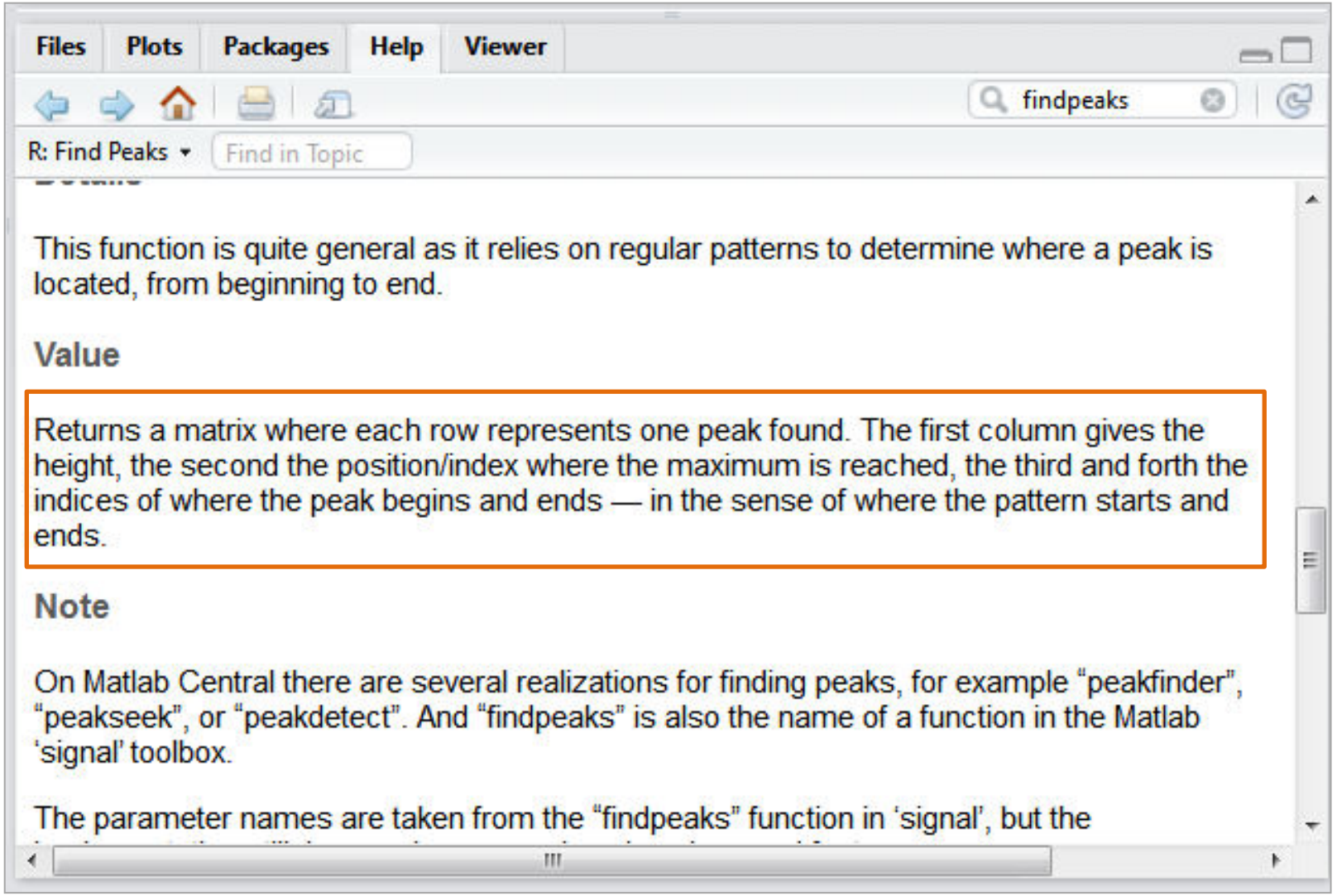
- Četvrti korak: dodeljivanje imena učitanim podacima i po potrebi smeštanje u različite promenljive.
- Ovaj korak uključuje i filtriranje tj. pretprocesiranje podataka kako bi se otklonio artifakt merenja.

EKG i IR za merenje pulsa

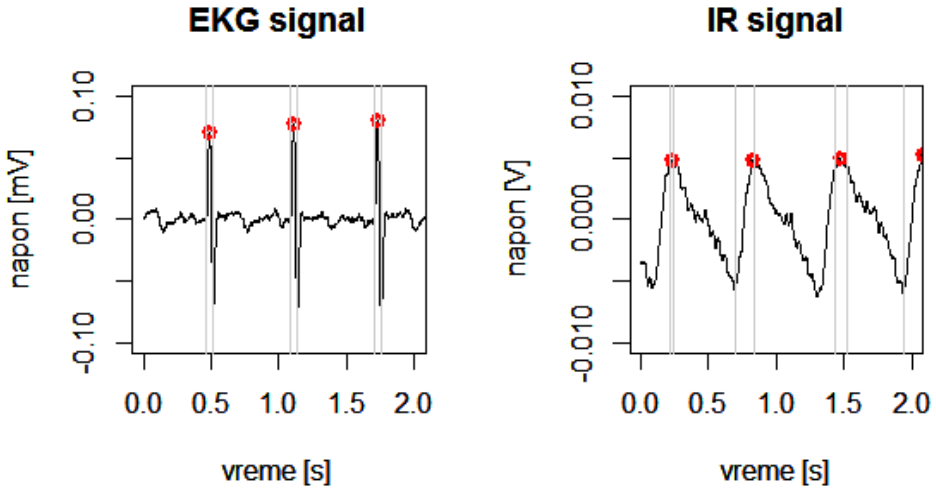
```
> sum(is.na(EKGf))
[1] 0
> sum(is.na(IRf))
[1] 0
> summary(EKGf)
      Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
-9.438e-02 -2.658e-03  6.471e-04  6.780e-06  3.285e-03  8.755e-02
> summary(IRf)
      Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
-1.0480000 -0.0026890  0.0002708  0.0054870  0.0027290  0.9800000
> library(pracma)
> EKGf <- EKGf[(30*fs):(42*fs)]
> leEKG <- findpeaks(EKGf, minpeakheight = 0.05, minpeakdistance = fs/4)
> head(leEKG, 3)
           [,1]  [,2]  [,3]  [,4]
[1,] 0.08495983 10176 10150 10203
[2,] 0.08482469  4756  4730  4783
[3,] 0.08342679  2354  2327  2380
> IRf <- IRf[(30*fs):(42*fs)]
> leIR <- findpeaks(IRf, minpeakheight = 0.0025, minpeakdistance = fs/4)
> head(leIR, 3)
            [,1]  [,2]  [,3]  [,4]
[1,] 0.005771482 10528 10499 10543
[2,] 0.005636414  9282  9154  9299
[3,] 0.005569005  7491  7465  7526
```

- Peti korak: računanje vrednosti od interesa.
- Koristi se “pracma” paket. Primetiti da se radi samo na delu snimljenih podataka.
- Ovaj korak bi uključio i manipulaciju nedostajućim vrednostima. U ovom skupu podataka nedostajuće vrednosti ne postoje.
- Međukoraci uključuju i pregled deskriptivnih statističkih parametara za izmerene (realne) podatke.

findpeaks() funkcija



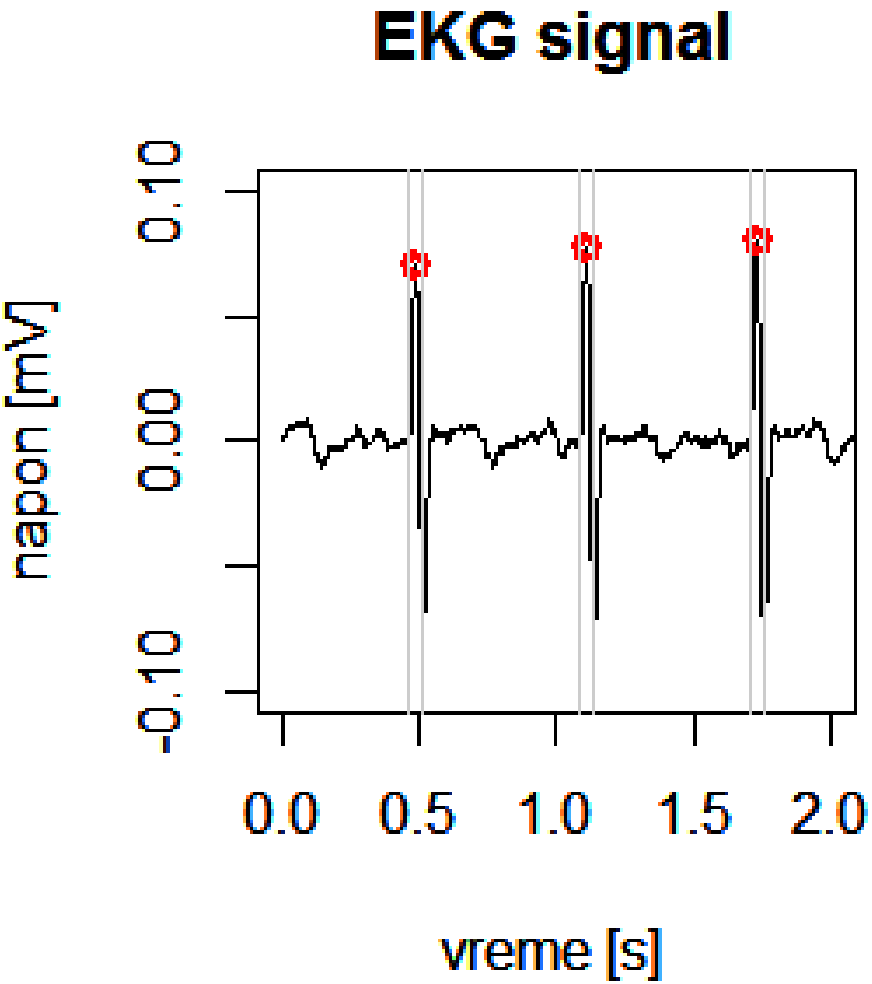
EKG i IR za merenje pulsa

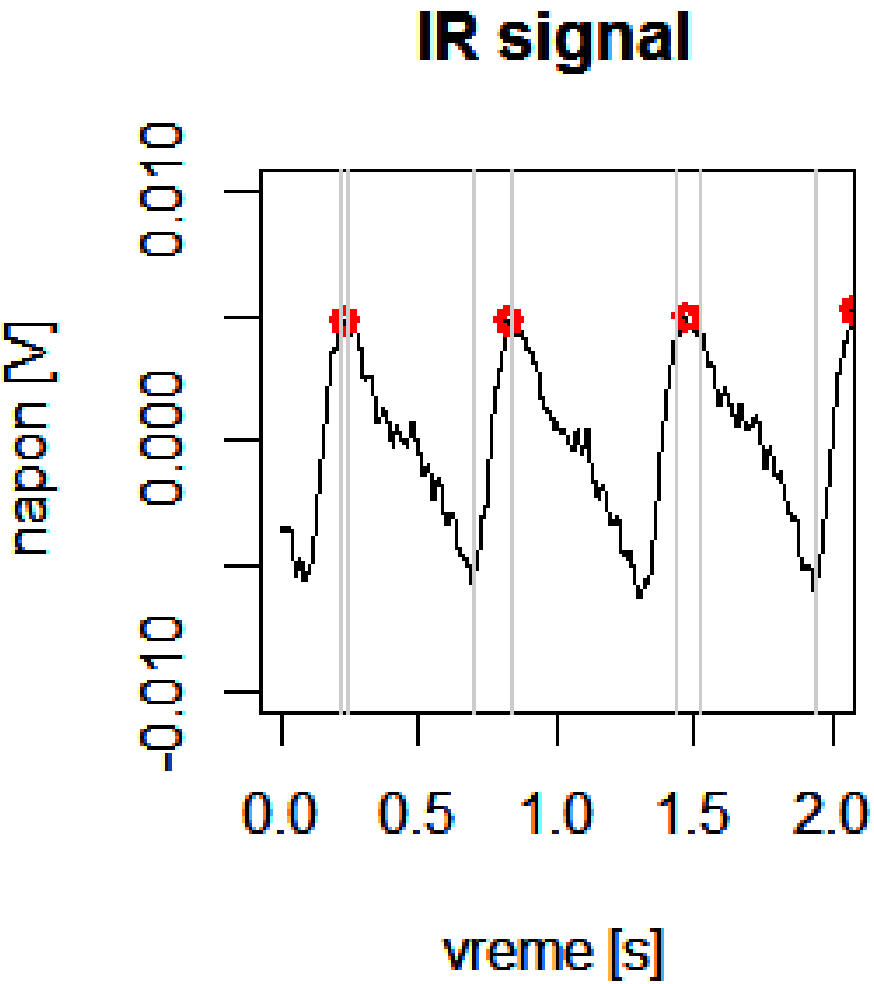



```
> vreme <- seq(0, (length(EKGf)/fs - 1/fs), by = 1 / fs)
> par(mfrow = c(1, 2))
> plot(vreme, EKGf, main = "EKG signal", xlab = "vreme [s]",
+      ylab = "napon [mV]", type = "l",
+      ylim = c(-0.1, 0.1), xlim = c(0, 2))
> par(new = TRUE)
> plot(leEKG[,2]/fs, leEKG[,1], main = "EKG signal", xlab = "vreme [s]",
+      ylab = "napon [mV]", col = "red",
+      ylim = c(-0.1, 0.1), xlim = c(0, 2), lwd = 3)
> abline(v = leEKG[,4]/fs, col = "grey80", pch = 2)
> abline(v = leEKG[,3]/fs, col = "grey80", pch = 2)
>
>
> plot(vreme, IRf, main = "IR signal", xlab = "vreme [s]",
+      ylab = "napon [V]", type = "l",
+      ylim = c(-0.01, 0.01), xlim = c(0, 2))
> par(new = TRUE)
> plot(leIR[,2]/fs, leIR[,1], main = "IR signal", xlab = "vreme [s]",
+      ylab = "napon [V]", col = "red",
+      ylim = c(-0.01, 0.01), xlim = c(0, 2), lwd = 3)
> abline(v = leIR[,4]/fs, col = "grey80", pch = 2)
> abline(v = leIR[,3]/fs, col = "grey80", pch = 2)
```

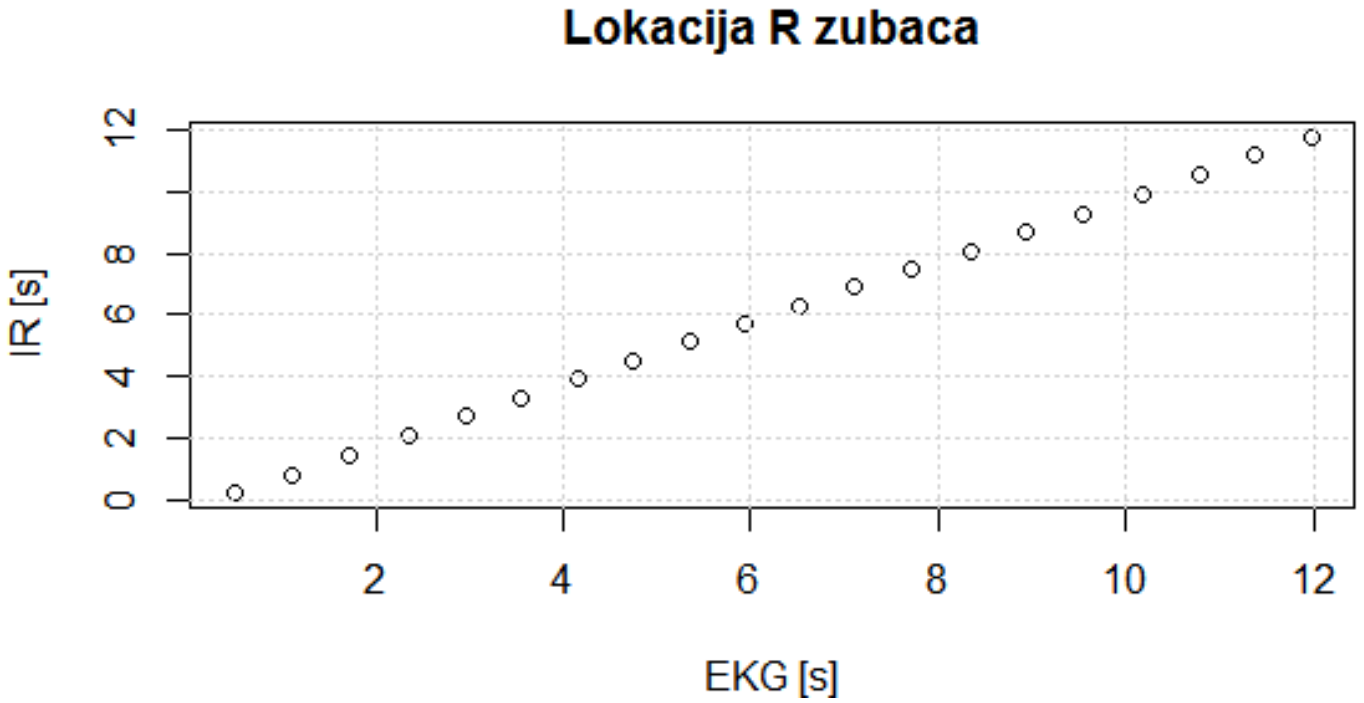
- Šesti korak: vizuelizacija podataka i izračunatih parametara.
- Šta se može zaključiti iz priloženih grafika?

EKG i IR za merenje pulsa





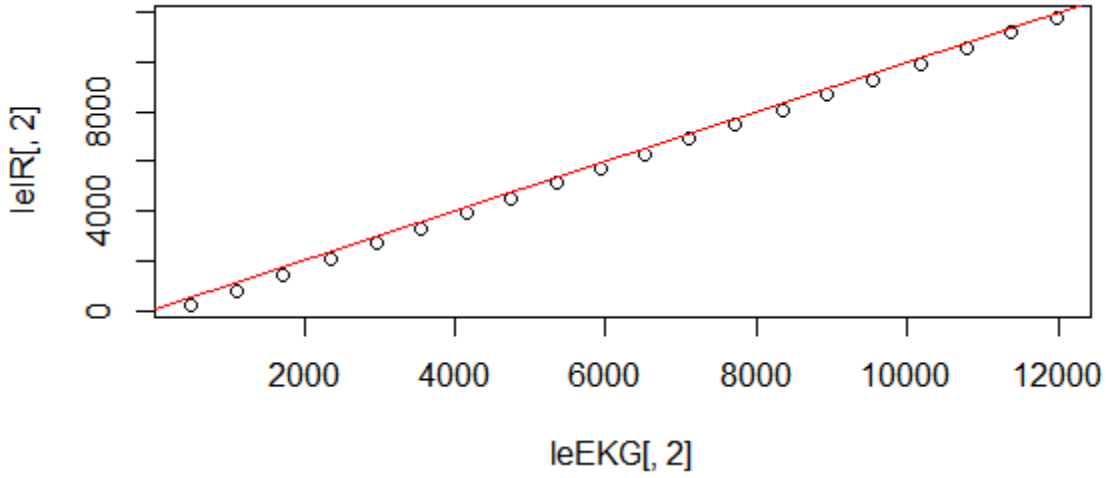
EKG i IR za merenje pulsa



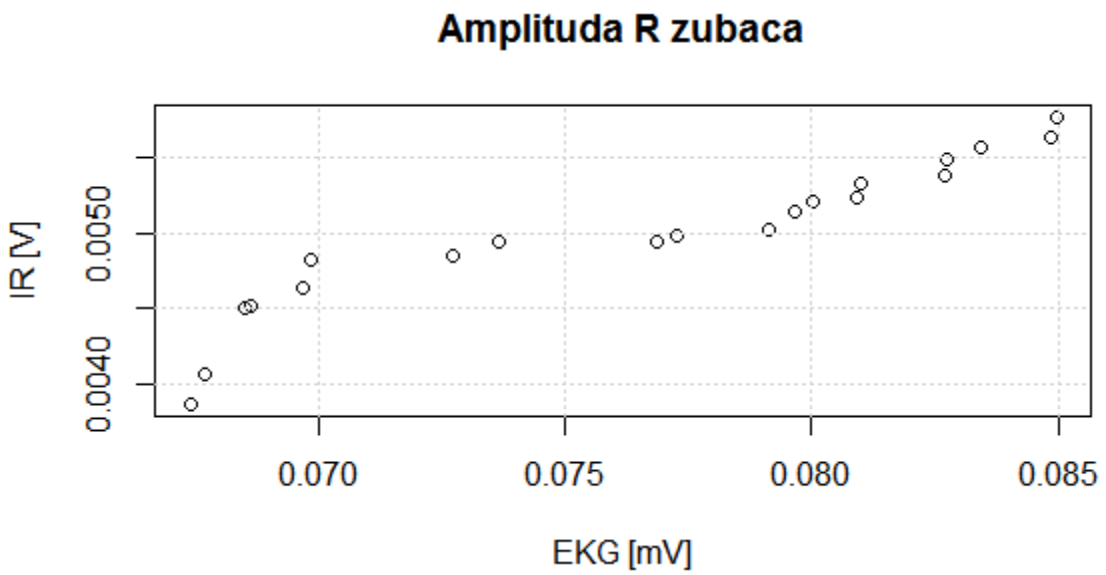

```
> leEKG[,2] <- sort(leEKG[,2])
> leIR[,2] <- sort(leIR[,2])
> plot(leEKG[,2]/fs, leIR[,2]/fs, main = "Lokacija R zubaca",
+      xlab = "EKG [s]", ylab = "IR [s]")
> grid()
```

- Sedmi korak: još vizuelizacije.
- Zaključci?

EKG i IR za merenje pulsa




```
> qqplot(leEKG[,2], leIR[,2])
> abline(1, 1, col = "red")
> cor(leEKG[,2], leIR[,2])
[1] 0.9999858
```



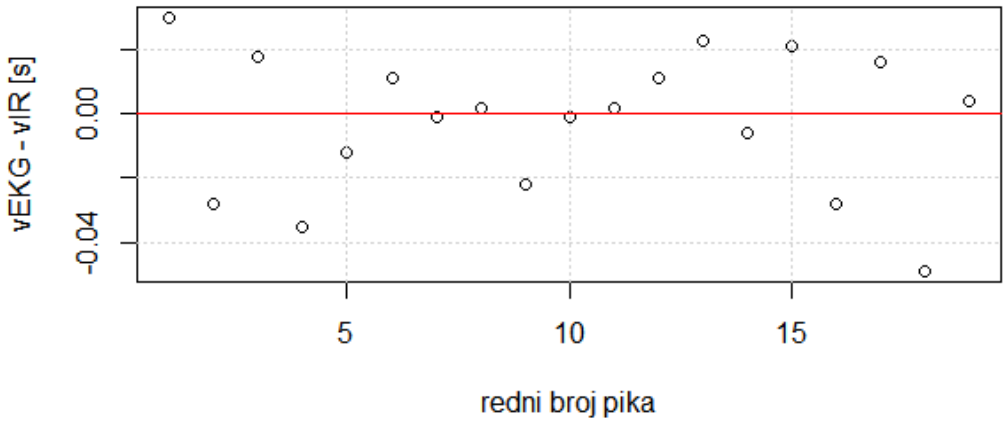

```
> plot(leEKG[,1], leIR[,1], main
 = "Amplituda R zubaca",
+      xlab = "EKG [mV]", ylab =
 "IR [V]")
> grid()
> cor(leEKG[,1], leIR[,1])
[1] 0.9396907
```

Još slika i poređenja?

```
> pulsEKG <- 60 * length(leEKG[,1]) / max(vreme)
> pulsIR <- 60 * length(leIR[,1]) / max(vreme)
> pulsEKG
[1] 100
> pulsIR
[1] 100
```

```
> vEKG <- vector("numeric", le = length(leEKG[,2]) - 1)
> vIR <- vEKG
> for (ind in 1:(length(leEKG[,2]) - 1)) {
+         vEKG[ind] <- leEKG[ind + 1, 2] - leEKG[ind, 2]
+         vIR[ind] <- leIR[ind + 1, 2] - leIR[ind, 2]
+ }
> vEKG <- vEKG / fs # da bi bilo u s
> vIR <- vIR / fs
> cor(vEKG, vIR)
[1] 0.1879677
> mean(vEKG)
[1] 0.6048947
> mean(vIR)
[1] 0.6072105
> sd(vEKG)
[1] 0.01208256
> sd(vIR)
[1] 0.02043575
```

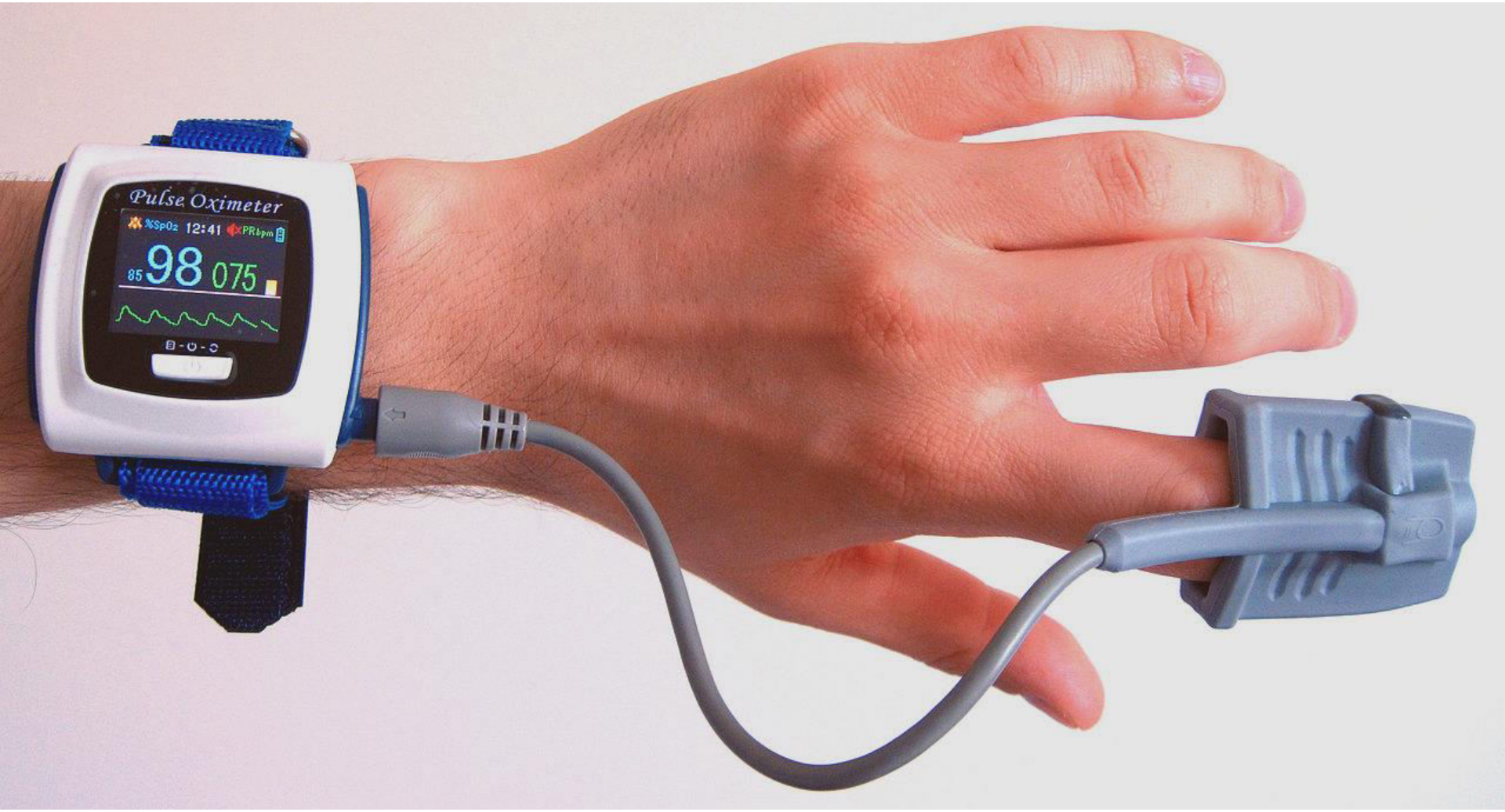
EKG i IR za merenje pulsa




```
> plot(vEKG - vIR, ylab = "vEKG - vIR [s]",
+      xlab = "redni broj pika")
> grid()
> abline(h = 0, col = "red")
> t.test(vEKG, vIR)

        Welch Two Sample t-test

data:  vEKG and vIR
t = -0.42519, df = 29.214, p-value = 0.6738
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.013451441  0.008819862
sample estimates:
mean of x mean of y
0.6048947 0.6072105
```

- Osmi korak: da li su podaci validni i da li su zaključci očekivani?
- Da li dobijeni rezultati odgovaraju očekivanim vrednostima (npr. ako je puls 400 bpm sigurno je greška u analizi signala).
- Dalja analiza ... ?

Zaključak?



By UusiAjaja - Own work, CC0, https://commons.wikimedia.org/w/index.php?curid=15840996

Dalji rad

- Analiza na većem skupu.
  - Duži vremenski intervali merenja.
  - Merenje na jednom ispitaniku tokom različitih protokola (u mirovanju, tokom i nakon fizičke aktivnosti i sl.)
  - Analiza na većem broju ispitanika.
- Analiza u slučaju patologije?

REZIME

Rezime

Take-home messages

- Osim t-testa postoji niz drugih statističkih testova koji se mogu koristiti u analizi podataka. **Potrebno je razumeti testove i metode analize pre nego što se primene na podacima i pre nego što se donesu zaključci. Dodatno, potrebno je poznavati i razumeti podatke.**
- Važno je primeniti odgovarajuće korake i provere u analizi signala kako bi se izbegle greške i kako bi se došlo do validnih zaključaka.
- Deo ove lekcije je inspirisan knjigom: R. A. Irizarry, M. I. Love, Data Analysis for the Life Sciences, Leanpub, 2016, https://leanpub.com/dataanalysisforthelifesciences.