Vilniaus Universitetas

Cenzuruotų imčių analizė

Laboratorinis darbas

Darbą atliko:
Matas Gaulia, Matas Kamarauskas
Duomenų Mokslas
4 kursas 1 gr.

Vilnius, 2022

Naudoti metodai

Darbas atliktas naudojant R.

Naudoti R paketai:
tidyverse
dplyr
survival
ggfortify
ggplot2
muhaz

Duomenys ir jų šaltiniai

Duomenys apie žmones sergančius COVID-19 ir jų išgyvenamumą.
Duomenų šaltinis - Kaggle. Prieiga per internetą: Lung-Survival Dataset | Kaggle

„time“ – laikas iki mirties dienomis.
„sex“ – lytis, 1 – vyras, 2 - moteris
„status“ – statusas, 1 – cenzūruotas, 2 - mirtis
„age“ - amžius
„ph.ecog“ – Neįgalumo statusas
0 – pilnos judėjimo galimybės
1 – galimi tik fiziškai lengvi darbai
2 – gali pasirūpinti savimi bet neįgalus dirbti
3 – ne pilnai gali savimi pasirūpinti
4 – pilnas neįgalumas

Tikslas ir uždaviniai

Tikslas: Ištirti COVID-19 išgyvenamumą ir patikrinti homogeniškumo hipotezes.

Uždaviniai:

Atlikti pirminę duomenų analizę

Ištriti gyvenamumą pagal lytį ir neįgalumo statusą

Patikrinti homogeniškumo hipotezes

Atliktos analizės aprašymas

Duomenų aibę sudaro duomenys 227 stebėjimai
Lyčių pasiskirstymas: 137 vyrai ir 90 moterys.
Cenzūravimo pasiskirstymas: 63 cenzūruoti, 164 – ne.
Neįgalumo pasiskirstymas:
0 – 63 stebėjimai
1 – 113 stebėjimai
2 – 50 stebėjimai
3 – 1 stebėjimai
4 – 0 stebėjimų

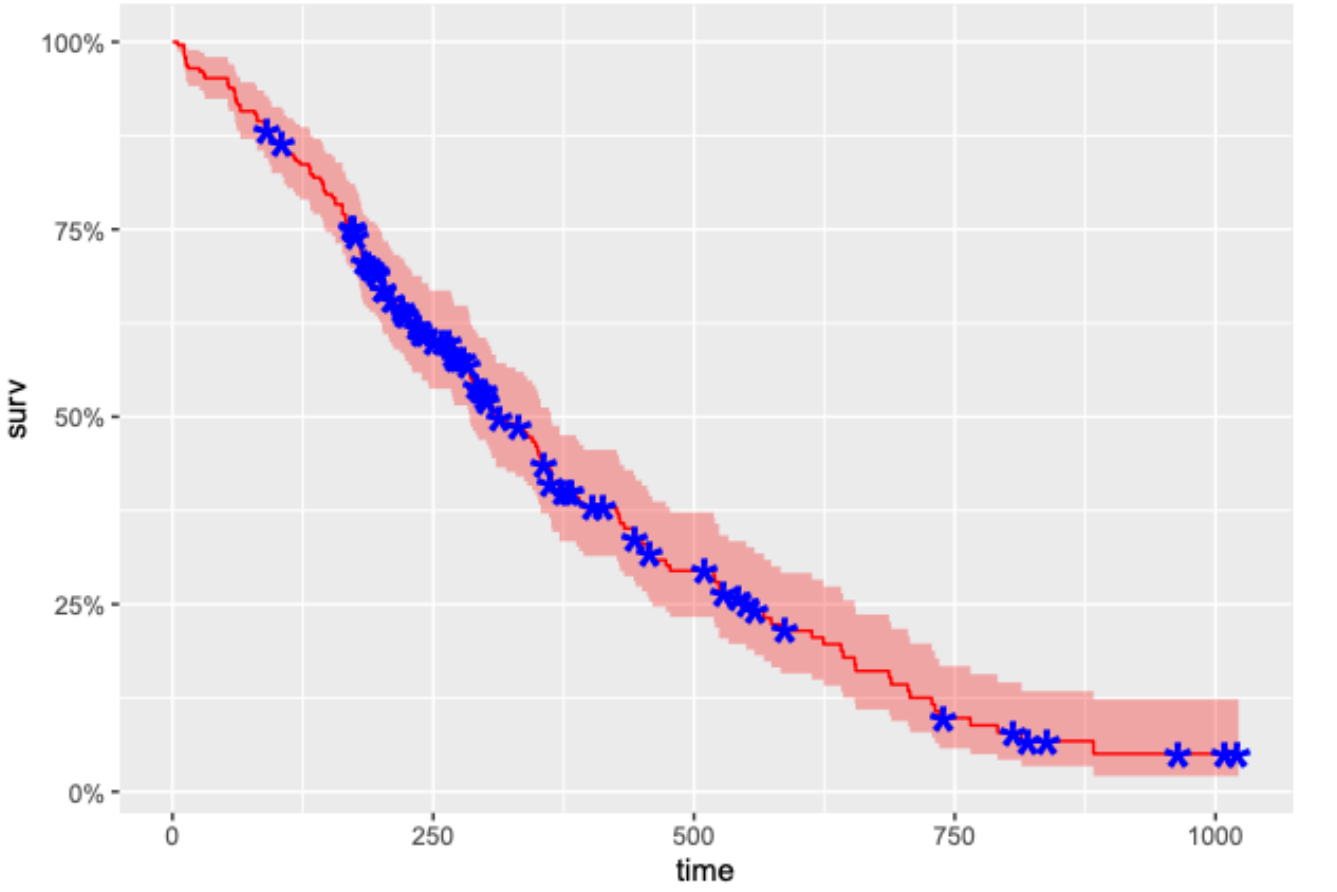
```
d <- read_csv("SurvivalCovid.csv") %>%
      select("time", "sex", "ph.ecog", "status") %>%
      drop_na()
d$time <- as.integer(d$time)
d$ph.ecog <- as.factor(d$ph.ecog)
d$sex <- as.factor(d$sex)
d$status <- as.integer(d$status)
#perkoduojame stulpeli (buvo 1-cenzuruota, dabar 0-cenzuruota)
d$status <- d$status - 1

NROW(d)
```
227

```
summary(as.factor(d$status))
```
0 63
1 164

```
summary(as.factor(d$ph.ecog))
```
0 63
1 113
2 50
3 1

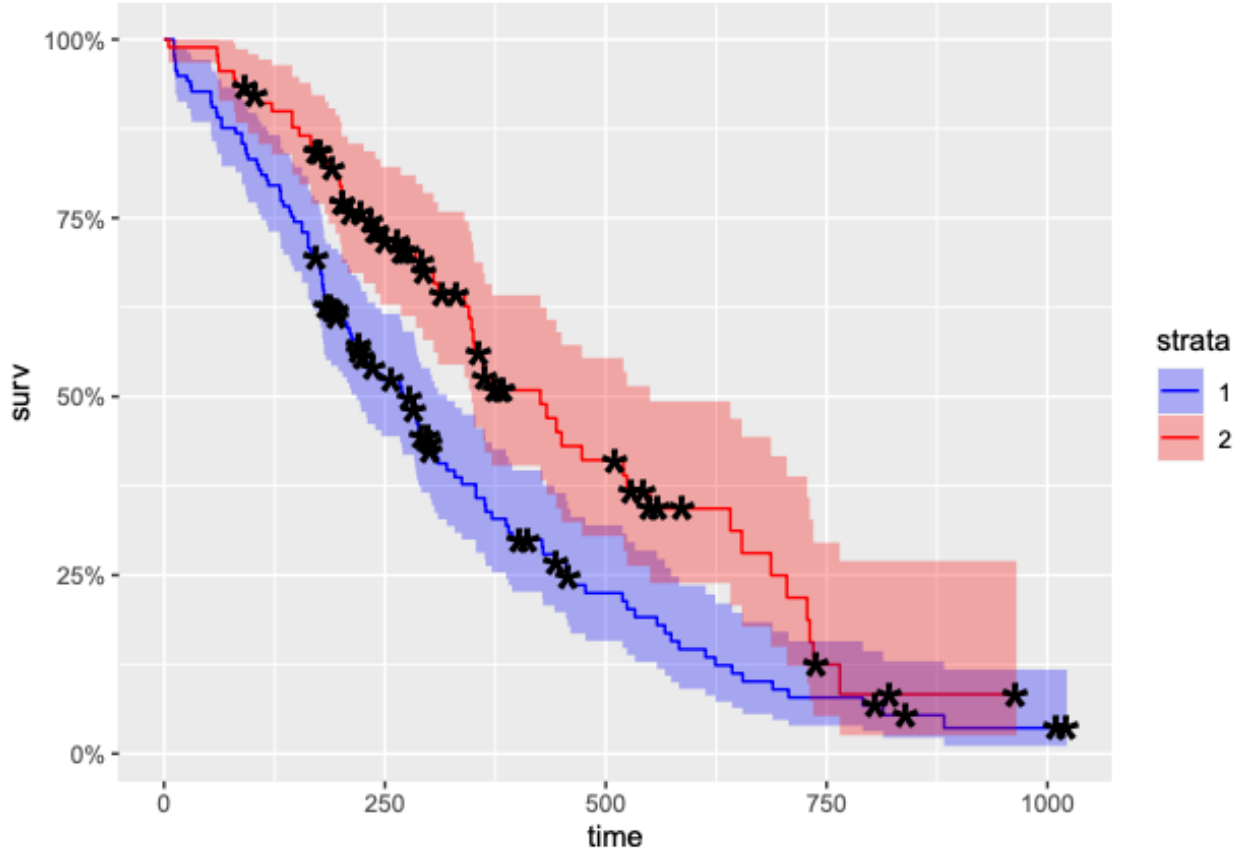
Išgyvenamumas

```
s <- Surv(d$time, d$status)
autoplot( survfit(s ~ 1),
      censor.shape = '*',
      censor.size = 10,
      surv.colour = 'red',
      censor.colour = 'blue')
```



Išgyvenamumas pagal lytį

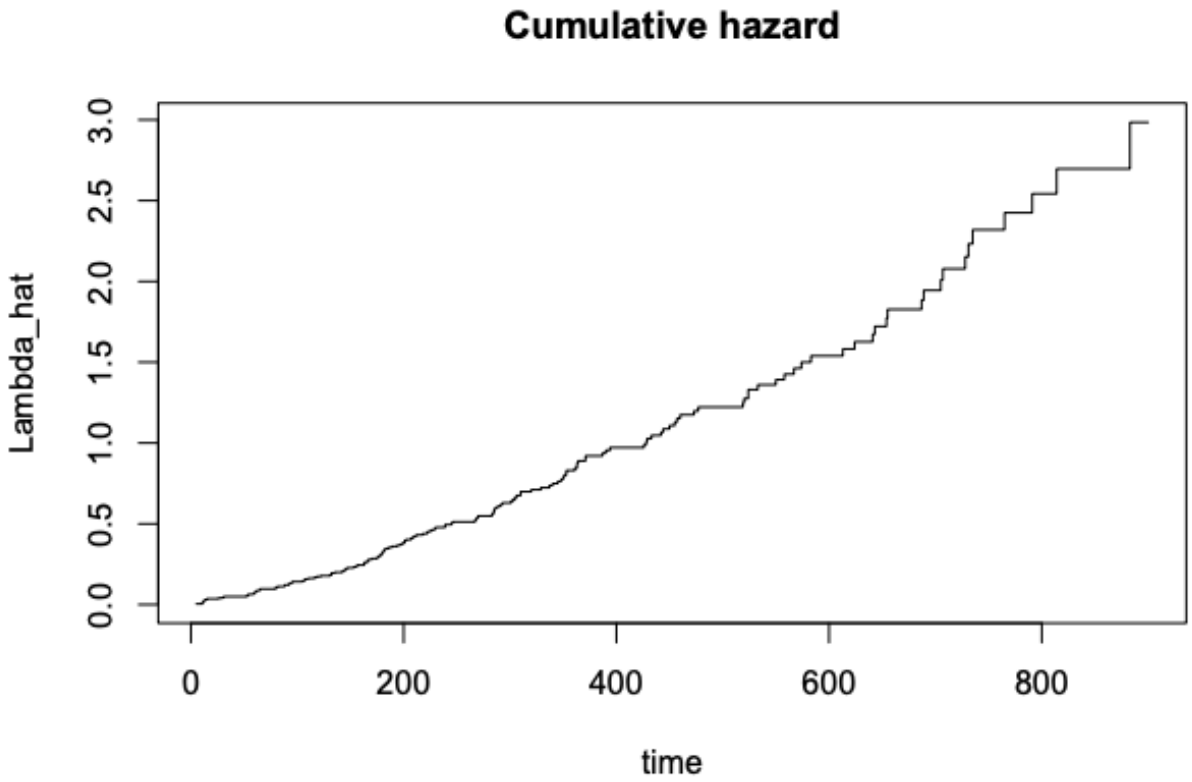
```
autoplot(survfit(s ~ d$sex),
      censor.shape = '*',
      censor.size = 10) +
  scale_color_manual(values = c("blue","red")) +
  scale_fill_manual(values = c("blue","red"))
```



Matome kad vyrų išgyvenamumas mažesnis nei moterų bet kuriuo laiko momentu.

Sukauptos rizikos įvertis

```
Iverciai <- summary(survfit(s ~ 1, type="kaplan-meier"))
Lambda_hat <- (-1)*log(Iverciai$surv)
Lambda_hat <- c(Lambda_hat, tail(Lambda_hat, 1))
plot(c(Iverciai$time, 900), Lambda_hat,
      xlab="time",
      ylab="Lambda_hat",
      main="Cumulative hazard",
      ylim=range(Lambda_hat),
      type="s")
```

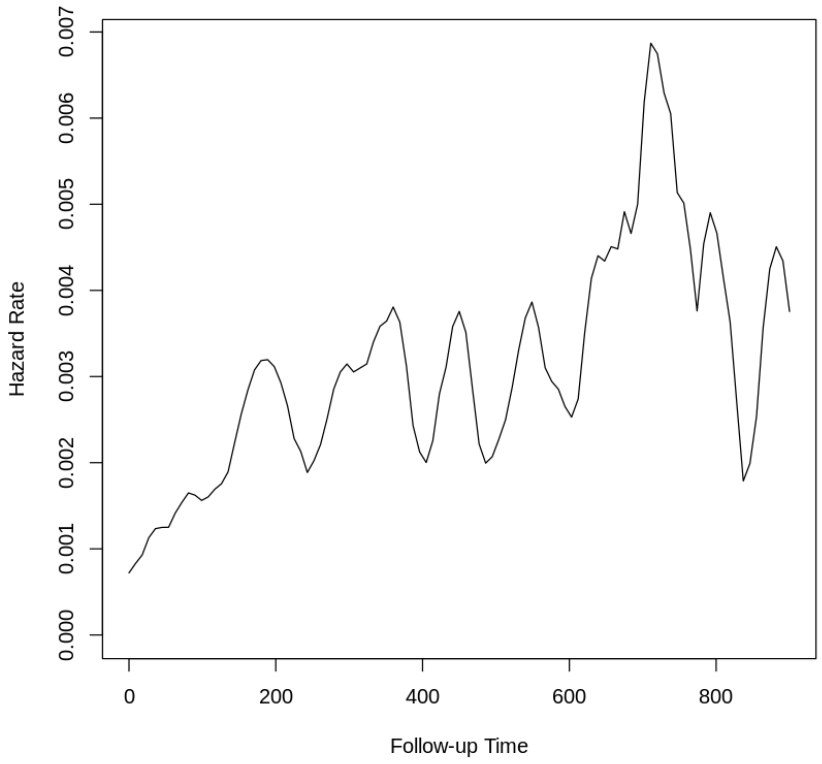


Matome kad rizika gana pastoviai auga.

```
print(survfit(s ~ 1), print.rmean=TRUE)

Call: survfit(formula = s ~ 1)

       n events rmean* se(rmean) median 0.95LCL 0.95UCL
[1,] 227    164    378      19.7    310     285     363
    * restricted mean with upper limit =  1022
```

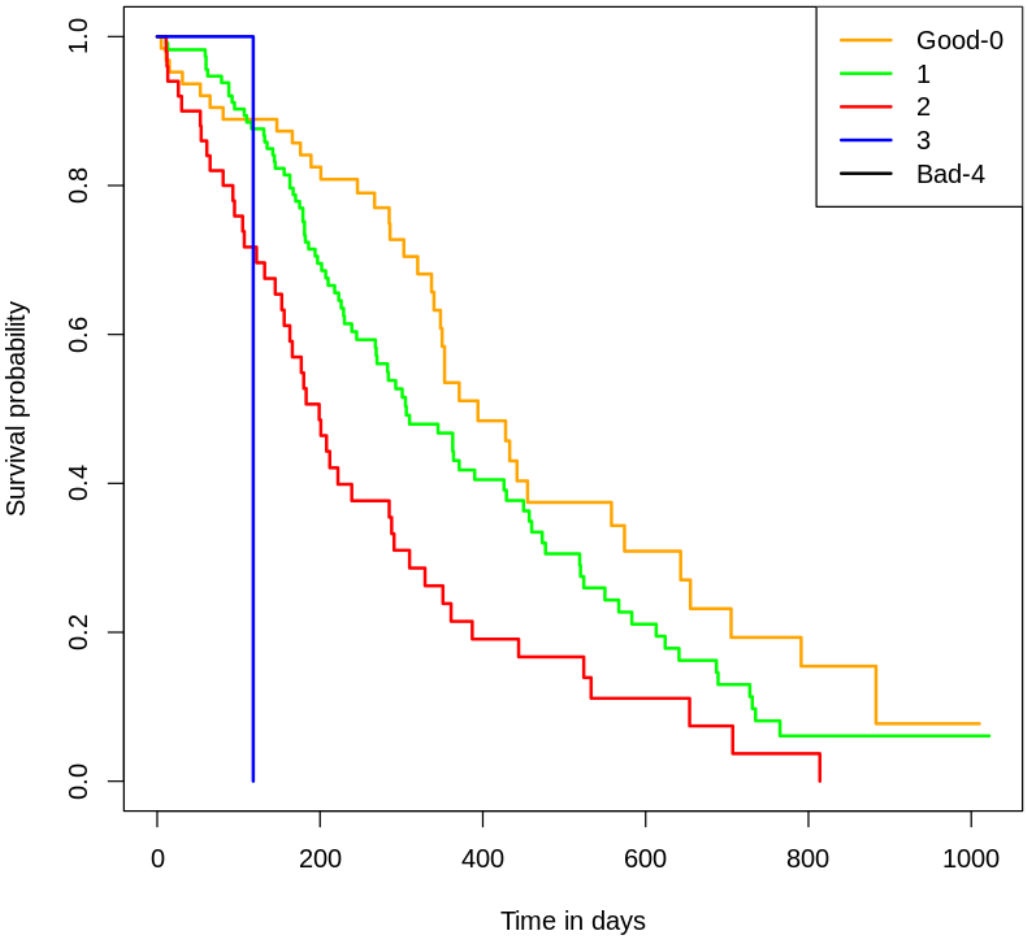
Branduolinis rizikos įvertis

```
result.simple <- muhaz(d$time,
      d$status,
      max.time=900,
      bw.method="global",
      b.cor="none")
plot(result.simple)
```



Išgyvenamumas pagal neigalumą

```
ecog <- s~d$ph.ecog
colors <- c("orange", "green", "red", "blue", "black")
plot(survfit(ecog),
  xlab="Time in days",
  ylab="Survival probability",
  col=colors,
  lwd =2)

legend("topright",
  legend=c("Good-0", "1", "2", "3", "Bad-4"),
  col=colors, lwd =2)
```



Matome kad gerokai skiriasi išgyvenamumas pagal neįgalumo grupes, blogiausiai grupei (pilnas neįgalumas) nepriklausė nei vienas pacientas, o trečiąjai grupei priklausė tik vienas pacientas.

Homogeniškumo hipotezių tikrinimas

Lograginis kriterijus:

```
survdiff(ecog, rho=0)
```

```
               N Observed Expected (O-E)^2/E (O-E)^2/V
d$ph.ecog=0   63       37   54.153    5.4331    8.2119
d$ph.ecog=1  113       82   83.528    0.0279    0.0573
d$ph.ecog=2   50       44   26.147   12.1893   14.6491
d$ph.ecog=3    1        1    0.172    3.9733    4.0040

 Chisq= 22  on 3 degrees of freedom, p= 7e-05
```

Matome, kad p-reikšmė yra mažiau nei 0.05, tad galime teigti kad egzistuoja statistiškai skirtingas skirtumas tarp grupių išgyvenamumo.

Gehan-Wilcoxon kriterijaus Peto ir Peto modifikacija:

```
survdiff(ecog, rho=1)
```

```
               N Observed Expected (O-E)^2/E (O-E)^2/V
 d$ph.ecog=0   63   19.877   31.190     4.103     8.605
 d$ph.ecog=1  113   47.814   50.346     0.127     0.367
 d$ph.ecog=2   50   30.002   16.843    10.279    16.964
 d$ph.ecog=3    1    0.846    0.159     2.967     3.236

  Chisq= 23.4  on 3 degrees of freedom, p= 3e-05
```

Matome, kad p-reikšmė yra mažiau nei 0.05, tad galime teigti kad egzistuoja statistiškai skirtingas skirtumas tarp grupių išgyvenamumo.

Rezultatai

Remiantis atliktu tyrimu galime teigti kad lytis ir neįgalumo statusas turi įtakos žmogaus išgyvenamume nuo COVID-19 ligos. Gauti rezultatai kad vyrų mirtingumas didesnis nei moterų, taip pat didesnis neįgalumas daro neigiamą įtaką neįgalumui.