



Прикладные аспекты применения статистических методов в среде R

Н.с. лаборатории биохимии
Жиров А.М.

Ставрополь 2022



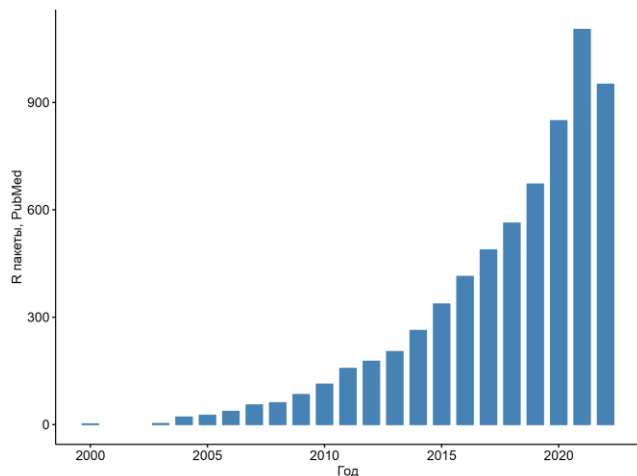
Что такое R?

- Интерпретируемый язык программирования
- R доступен для операционных систем Windows, Mac и Linux
- R предоставляет инструментарий и компоненты для создания собственных статистических программ
- Сообщество разработчиков CRAN, Comprehensive R archive network, Комплексная сеть архивов R





Почему R?



- Сделано статистиками для статистиков
- Открытость R и RStudio
- Поддержка и сообщество:
 - CRAN **18703** пакета
 - Bioconductor **2140** пакетов
- Полная поддержка предыдущих версий





Почему R?

Цена

Excel 70\$ / год

Statistica 13 800\$ / год

SPSS 3200\$

R 0\$

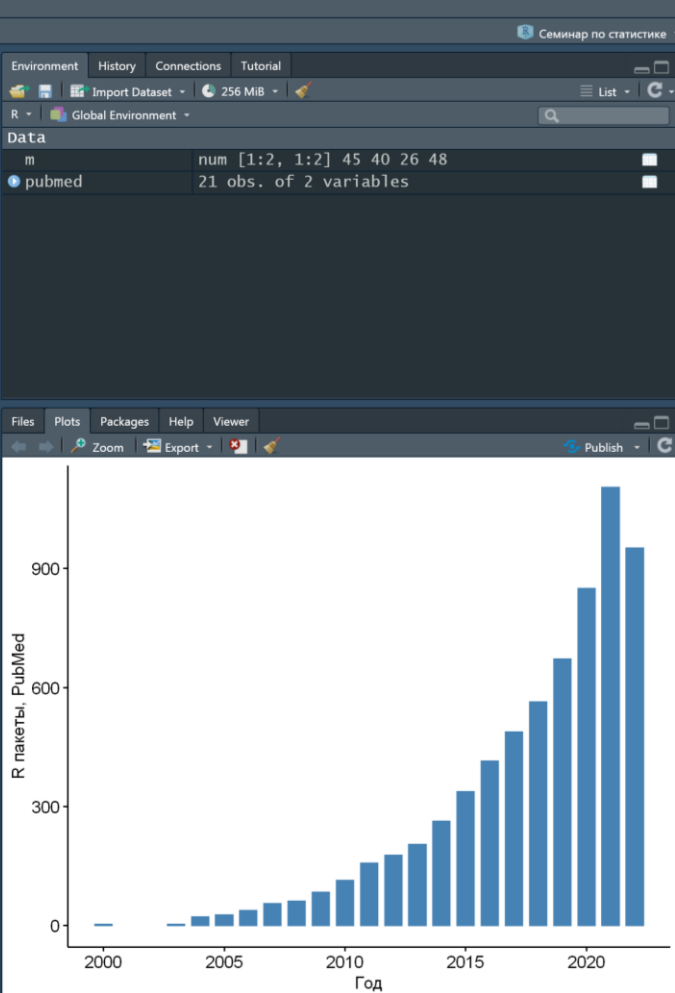
Удобство

- В виде модулей под R реализованы практически все виды графиков и статистических задач общего и частного характера
- Однажды написав скрипт R под свои рутинные задачи, вы тратите лишь на его запуск, **повторно** используя уже готовый программный код
- Большое количество скриптов уже было написано

```
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function Addins
script.R scr.R
Source on Save Run
1 library(dplyr)
2 library(ggpubr)
3 library(ggstatsplot)
4
5 pubmed <- read.csv("pubmed.csv", skip = 1)
6
7 ggbarplot(pubmed, x = "Year", y = "count",
8           ylab = "R пакеты, PubMed",
9           fill = "steelblue", col = "steelblue",
10          xlab = "Год",
11          legend.title = "MPG Group")
12 )
13 ggsave("pm.png", device = "png", width = 8, height = 6, dpi = 300)
14 dev.off()
15 options(device = "RStudioGD")
16 plot(mtcars)
17 getOption("device")
18
19 t.test(x, y, paired = FALSE)
20
21 t.test(x, mu = )
22
23 m <- matrix(c(45, 26, 40, 48), nrow = 2, byrow = T)
24 m
25 chisq.test(m)
26 fisher.test(m)
27
28 ?var.test()
29 ?kruskal.test
30 chisq.test()
31 fisher.test()
32
26:1 (Top Level) R Script
Console Jobs
R 4.2.1 H:/Семинар по статистике/
+ xlab = "Год",
+ legend.title = "MPG Group"
+ )
> m <- matrix(c(45, 26, 40, 48), nrow = 2, byrow = T)
> chisq.test(m)

Pearson's Chi-squared test with Yates' continuity correction

data: m
X-squared = 4.3802, df = 1, p-value = 0.03636
> )
```





t-критерий Стьюдента

t.test()



- Одновыборочный `t.test(x, mu)`
- Двухвыборочный для независимых выборок `t.test(x, y, paired = FALSE)`
- Двухвыборочный для зависимых выборок `t.test(x, y, paired = TRUE)`

Условия применения:

- Нормальность распределений: Тест Шапиро-Уилка `shapiro.test()`
- Равенство дисперсий: Критерий Фишера `var.test()`

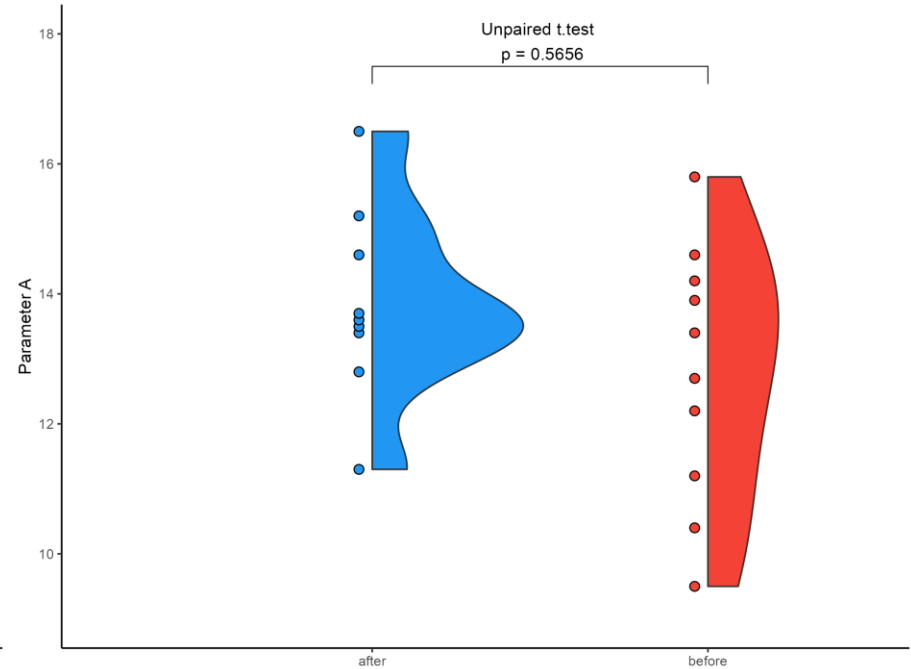
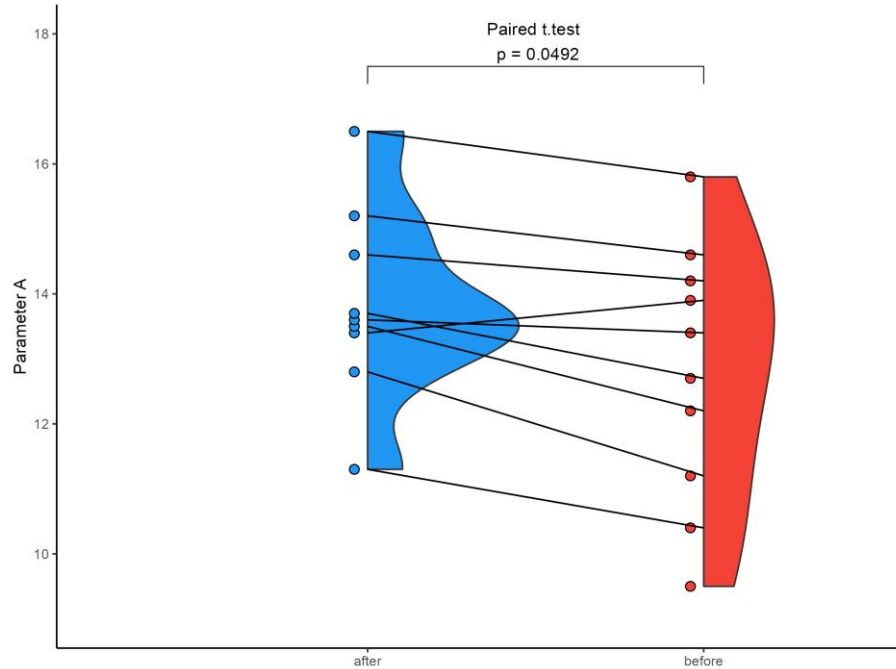
Непараметрический аналог:

Критерий Уилкоксона-Манна-Уитни `wilcox.test()`



t-критерий Стьюдента

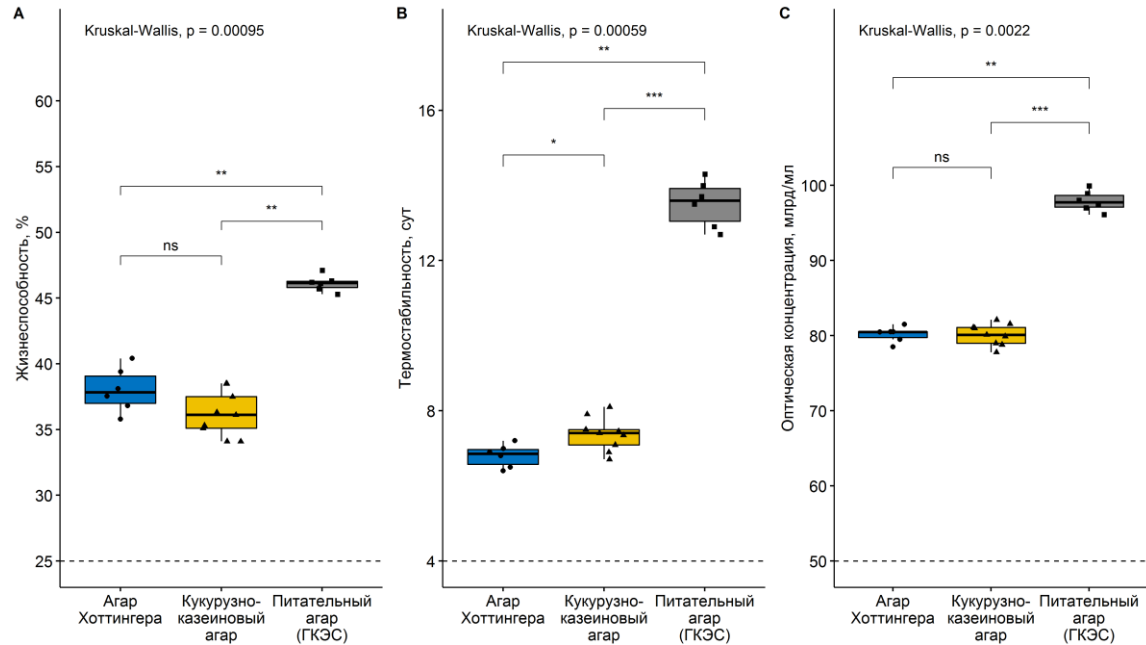
t.test()





Критерий Краскела-Уоллиса `kruskal.test()`

Параметрический аналог:
Дисперсионный анализ/
`ANOVA aov()`

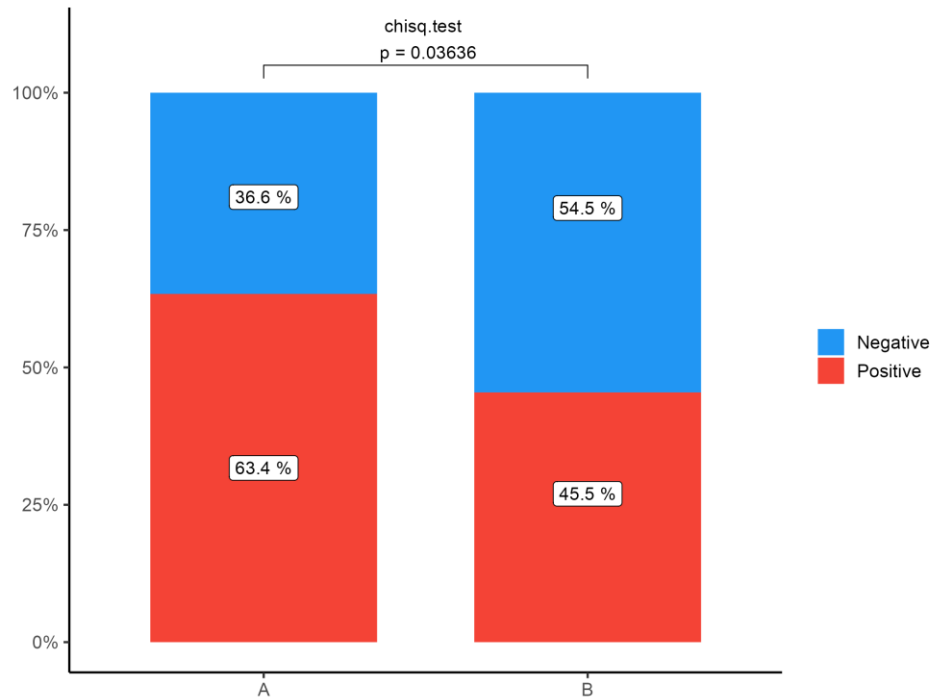




Хи-квадрат Пирсона `chisq.test()`

	Выжило	Умерло
Препарат А	45	26
Препарат В	40	48

Количество наблюдений в группе > 5

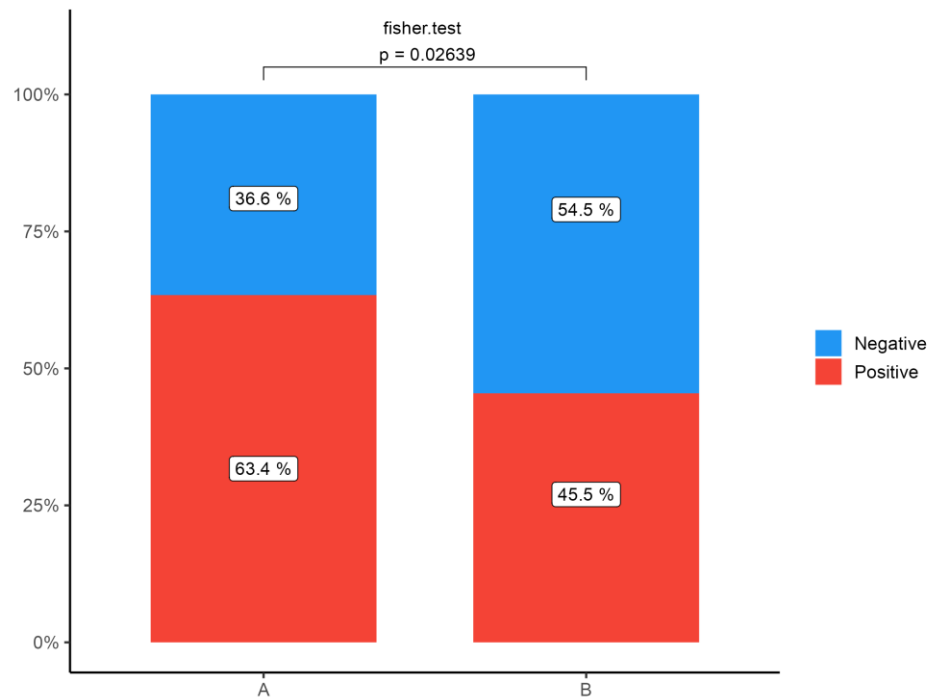




Точный тест Фишера `fisher.test()`

	Выжило	Умерло
Препарат А	45	26
Препарат В	40	48

Могут быть сложности при
большом количестве наблюдений





Анализ выживаемости

График Каплана-Мейера

Пакет “**survival**” и “**survminer**”

`survfit()` – расчет значений

`ggsurvplot()` - визуализация

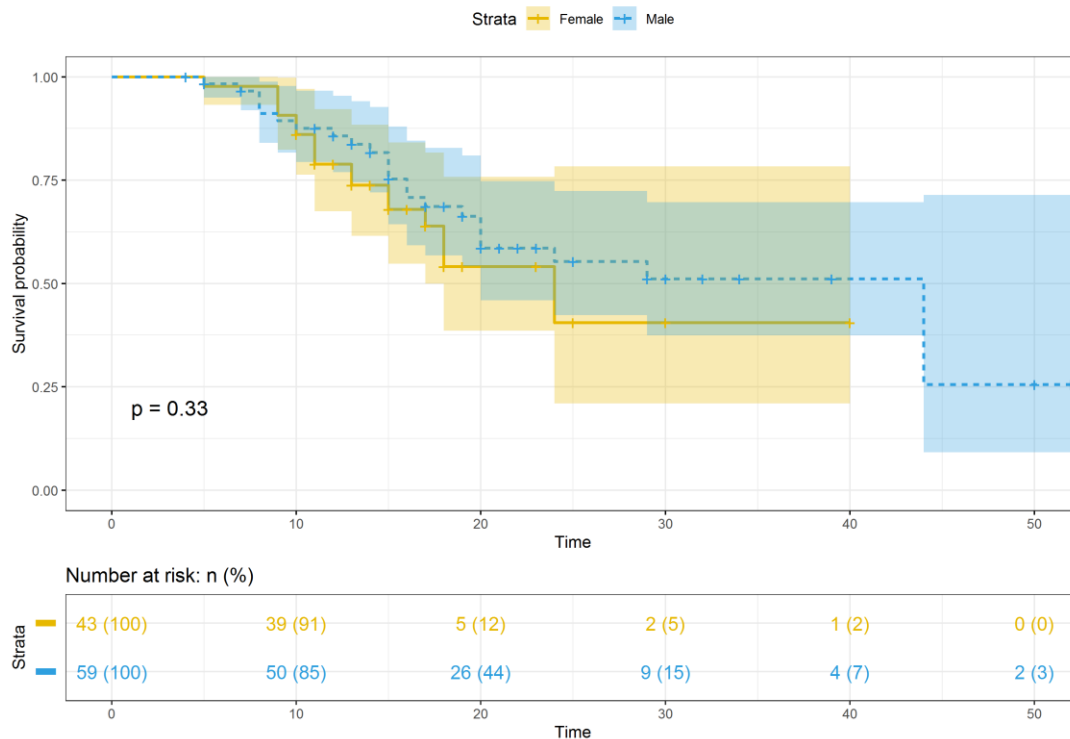
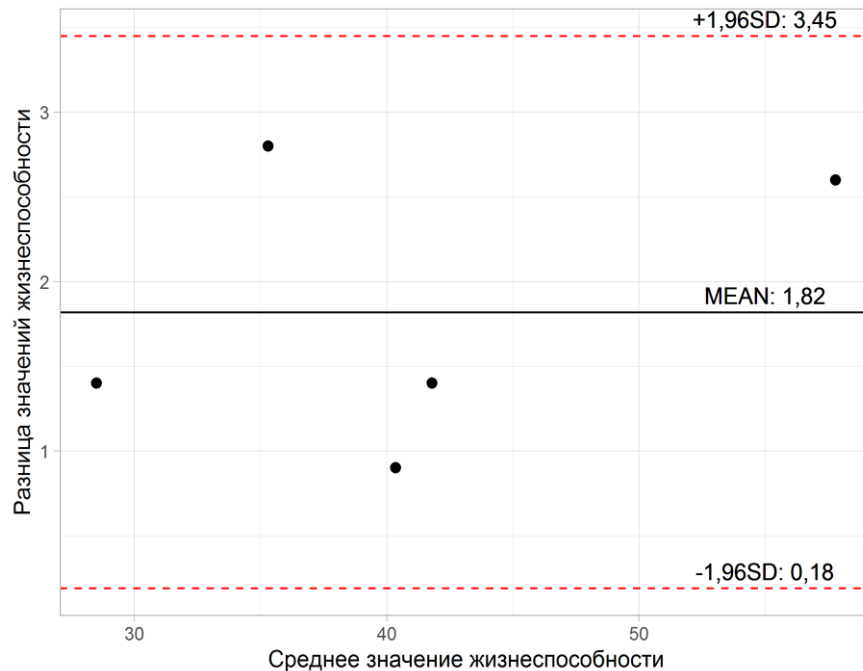




Диаграмма рассеяния разности и средних значений (Блэнда-Альтмана)

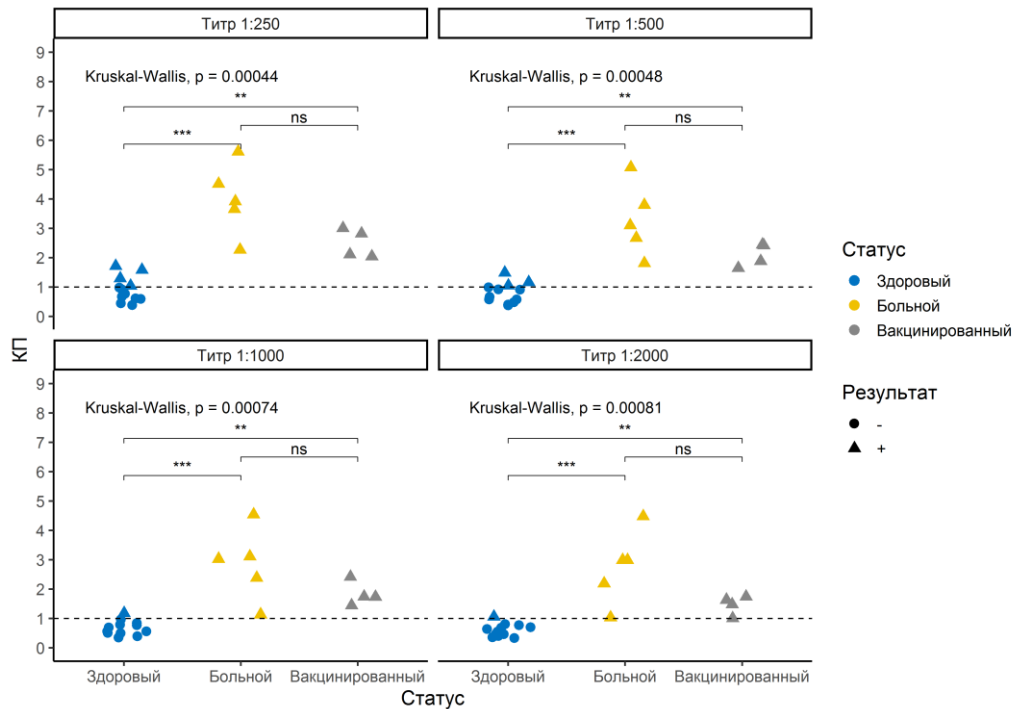
Сравнение результатов измерений, выполненных двумя методами





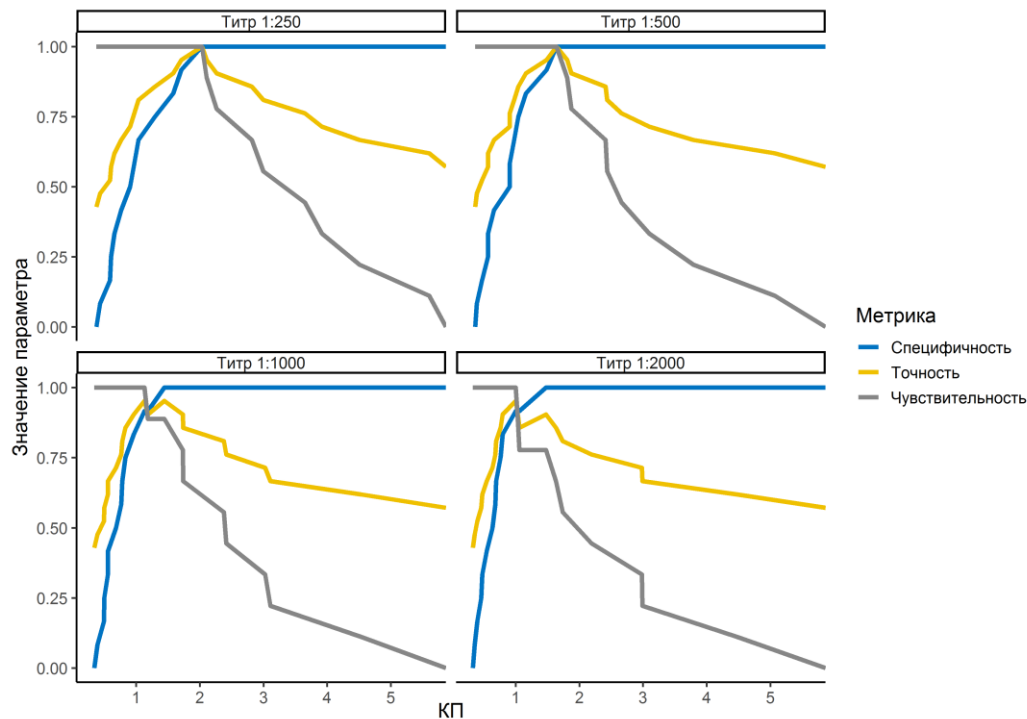
Бинарная классификация

Пакет “ROCR”, performance()





Бинарная классификация Пакет “ROCR”, performance()



$$\text{Специфичность} = \frac{\text{ИП}}{\text{ИП} + \text{ЛО}}$$

Способность выявлять здоровых

$$\text{Чувствительность} = \frac{\text{ИО}}{\text{ИО} + \text{ЛП}}$$

Способность выявлять больных

$$\text{Точность} = \frac{\text{ИП} + \text{ИО}}{\text{ИП} + \text{ИО} + \text{ЛП} + \text{ЛО}}$$

Доля правильных предсказаний

Спасибо за внимание