6, 13, 20 МАЯ, 19:00 МСК
ОНЛАЙН

РАЗРУШИТЕЛИ СТАТИСТИЧЕСКИХ МИФОВ
2 СЕЗОН

МИФ №4:
КАТЕГОРИЧЕСКИ КАТЕГОРИЧНО, ИЛИ «ПРОСТО РАЗБЕЙ НА ГРУППЫ»

МИФ №5:
РАНДОМИЗАЦИЯ – СМЕШАТЬ И НЕ ВЗБАЛТЫВАТЬ

МИФ №6:
ВИЗУАЛИЗАЦИЯ – ЭТО ПРОСТО КРАСИВЫЕ ГРАФИКИ

6 МАЯ 19:00 МСК
ОНЛАЙН

РАЗРУШИТЕЛИ СТАТИСТИЧЕСКИХ МИФОВ
2 СЕЗОН

АЛЕКСЕЙ ГЛАЗКОВ | МИФ №4:
КАТЕГОРИЧЕСКИ КАТЕГОРИЧНО, ИЛИ «ПРОСТО РАЗБЕЙ НА ГРУППЫ»

HTTPS://T.ME/CHAT_BIOSTAT_R

О себе

Глазков Алексей

- Выпускник ФФМ МГУ, к. м. н., старший научный сотрудник
- 12 лет опыта
- 200+ статистических расчётов
- 95 рецензий на диссертационные работы
- 283 рецензии на научные статьи
- Преподаю в Институте биоинформатики:

  «Специфика медицинских данных»,

  «Количественное планирование исследований»,

  «Корреляционный анализ»

Дисклеймер

- Мы сами много и увлеченно ошибались, чему, наверняка, в Интернете можно найти свидетельства :)
- В чем-то мы ошибаемся и сейчас (но пока этого не поняли).
- Какие-то ошибки нас ещё ждут.

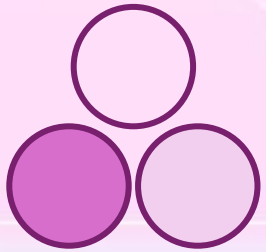
Цель данного мероприятия – поделиться своим опытом и опытом коллег с сообществом, чтобы снизить количество ошибок в статистическом анализе данных.

Синхронизируемся

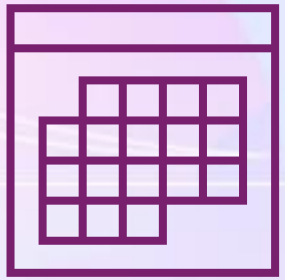
Типы переменных

Количественные
(шкалы, scale, quantitative)

Порядковые
(ordinal)



Качественные
(номинальные, nominal)



Дата/время

Что такое категоризация?

Преобразование непрерывной (количественной) переменной в категориальную путём разделения её значений на группы или интервалы.

Количественные
(шкалы, scale, quantitative)

Порядковые
(ordinal)

Качественные
(номинальные, nominal)

Пример

ИМТ, кг / м^2 – индекс массы тела

27.2 кг/м^2

< 18 кг/м^2 – **недостаточный** ИМТ

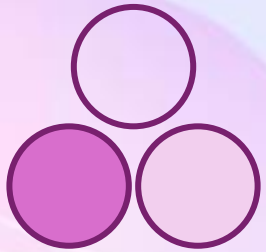
≥ 18 и < 25 кг/м^2 – **нормальный** ИМТ

≥ 25 и < 30 кг/м^2 – **избыточный** ИМТ

≥ 30 и < 35 кг/м^2 – **ожирение I** степени

≥ 35 и < 40 кг/м^2 – **ожирение II** степени

≥ 40 кг/м^2 – **ожирение III** степени



< 30 кг/м^2 | **≥ 30 кг/м^2**

Нет ожирения | **Есть ожирение**

Процедуры категоризации

1. Категоризация по клиническому опыту или эмпирическим допущениям

Пример: деление на группы «низкий/средний/высокий» по шкале тревожности, без нормативов.

2. Клинически обоснованные или нормативные пороги

Пример: категории ИМТ (нормальный, избыточный, ожирение), уровни АД по ESC/ESH.

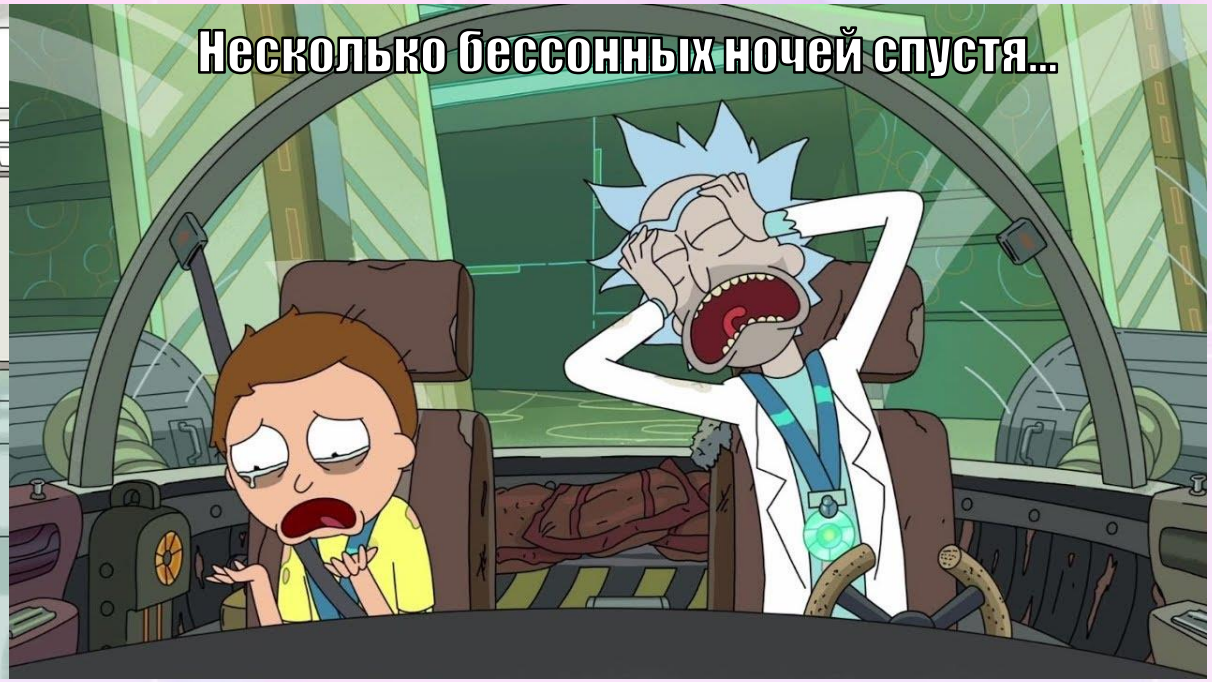
3. Статистическое разбиение по выборке (квартили, тертили, медиана)

Пример: деление на 4 группы по квартилям уровня ферритина.

4. «Оптимальные» категории (data-driven пороги)

Пример: определение порога по максимальному индексу Юдена на ROC-кривой.

Проблемы категоризации



Проблемы категоризации

Несоответствие биологической реальности

— Категории вводят искусственные границы там, где процесс имеет непрерывный характер.

— Например, риск сердечно-сосудистых заболеваний обычно нарастает плавно с ростом давления или холестерина, а не скачком в одной точке.

Потеря информации

— При группировке значений стираются различия внутри категорий.

— Особенно при дихотомизации (деление на две группы).

Увеличение частоты ошибок I рода (ложноположительных)

— Множественные сравнения между группами (например, между квартилями) увеличивают вероятность случайного нахождения значимого результата.

Увеличение ошибок II рода (ложноотрицательных)

— Уменьшение чувствительности к истинным слабым связям из-за упрощённой структуры данных.

— Эффект может «размазаться» внутри широкой категории и стать незаметным.

Искажение оценки эффекта

— Границы между категориями создают иллюзию резких изменений, приводя к завышенным или заниженным оценкам эффекта.

Проблемы воспроизводимости результатов

— Категории, основанные на характеристиках конкретной выборки (например, медиана), не воспроизводимы в других исследованиях.

Проблемы категоризации

Сложности в мета-анализе

— Разные критерии категоризации в исследованиях делают невозможным объединение результатов.

Риск «подгонки» (p-hacking)

— Исследователь может неосознанно (или осознанно) выбрать ту границу категорий, которая даёт статистически значимый результат.

— Это увеличивает риск ложноположительных выводов.

Неполная корректировка смешивающих факторов

— При категоризации ковариат (например, возраста) может сохраняться остаточное смешение, особенно если внутри категорий остаётся большой разброс.

Скрытие нелинейных зависимостей

— Категоризация «затирает» U-образные, S-образные и другие формы реальных связей.

— Модель видит ступеньки вместо плавных изменений.

Ограничения при использовании гибких моделей

— Современные методы (регрессии с нелинейными членами, сплайны, машинное обучение) лучше работают с непрерывными переменными.

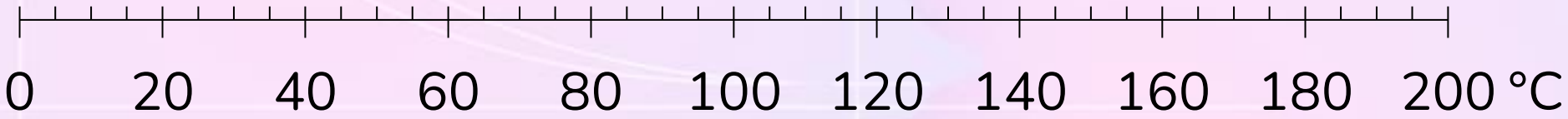
Интерпретационные искажения

— Врачи или читатели могут переоценивать значение искусственного порога, придавая ему клинический смысл, которого нет.

Часть 1. В которой мы рассуждаем о «биологической» реальности

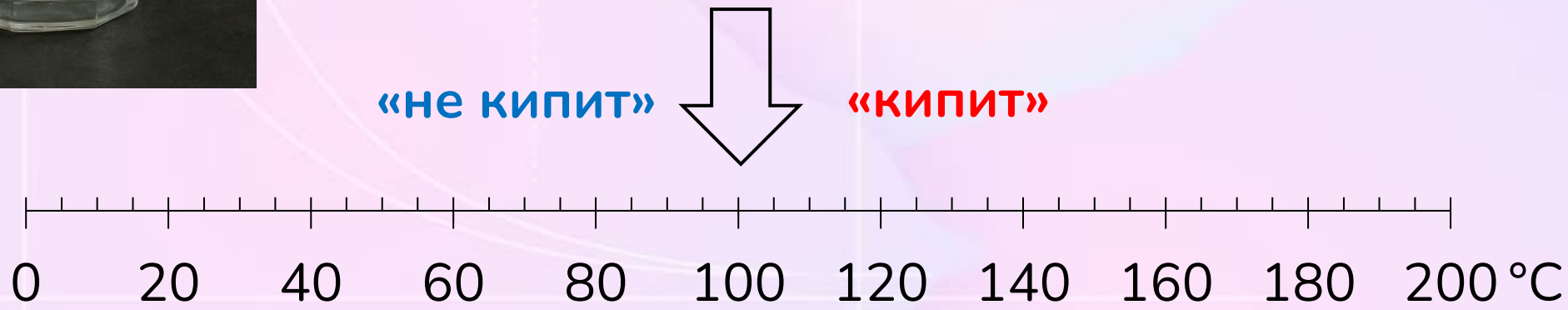
Пример «порога» в физическом процессе

При каком значении температуры вода в стакане переходит из состояния «не кипит» в состояние «кипит»?

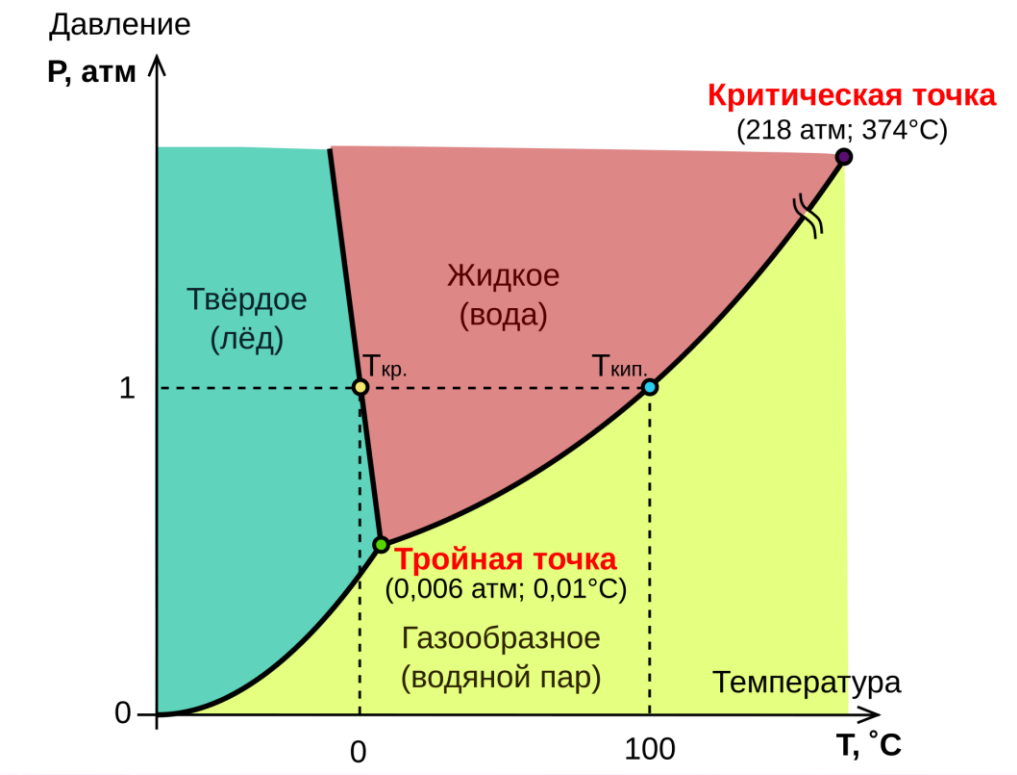


Пример «порога» в физическом процессе

При каком значении температуры вода в стакане переходит из состояния «не кипит» в состояние «кипит»?



Но всё не так просто...

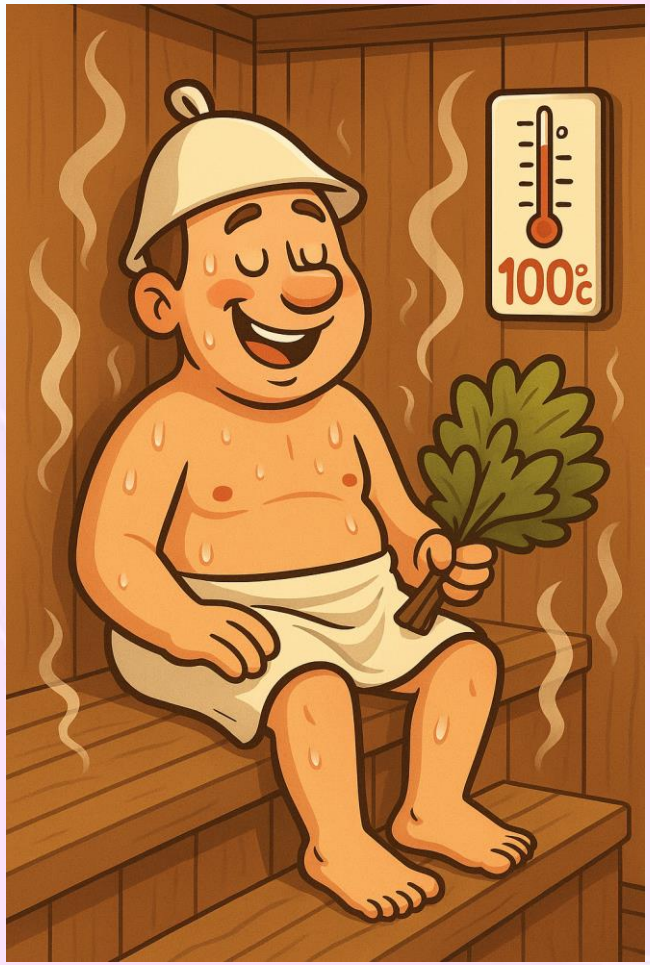


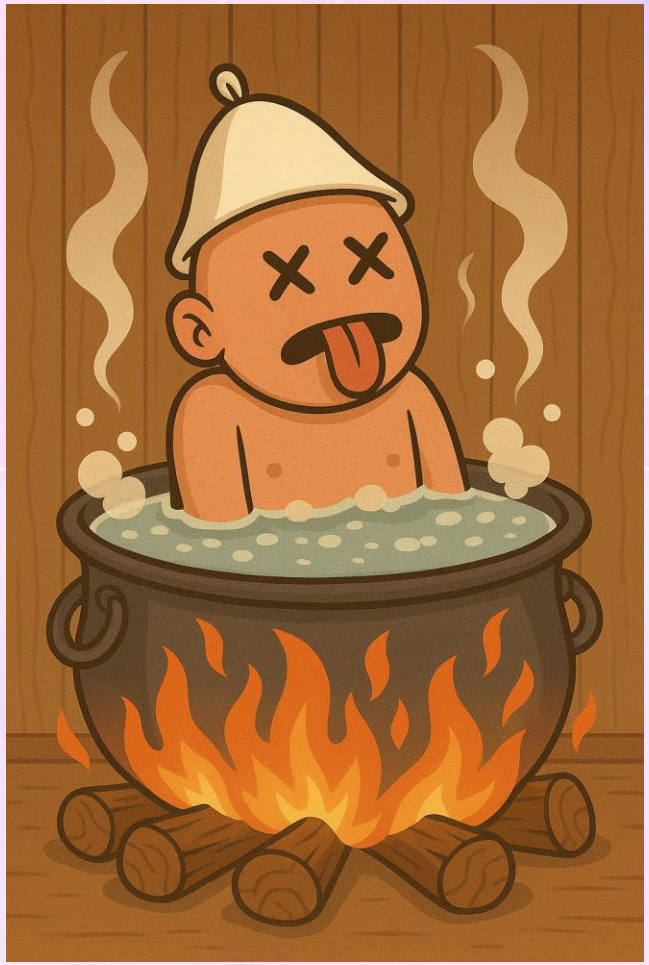
Температуры кипения водных растворов неорганических веществ - солей, кислот, оснований в зависимости от концентрации при атмосферном давлении 101,3 кПа.

Вещество	Температура кипения (°C) при массовой доле вещества в растворе (кг/кг)														
	0,05	0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,50	0,55	0,60	0,65	0,70	0,75
$Al_2(SO_4)_3$ -сульфат алюминия	100,2	100,4	100,7	101,1	101,6	102,2	-	-	-	-	-	-	-	-	-
$BaCl_2$ - хлорид бария, хлористый барий	100,3	100,7	101,1	101,6	102,2	103,0	103,9	-	-	-	-	-	-	-	-
$Ba(NO_3)_2$ - нитрат бария, азотнокислый барий, бариевая селитра	100,2	100,4	100,7	101,0	101,3	-	-	-	-	-	-	-	-	-	-
$CaCl_2$ - хлорид кальция	100,9	101,9	103,2	105,0	107,4	110,5	114,4	119,0	124,1	129,7	135,9	143,0	152,0	162,6	175,7
$Ca(NO_3)_2$ - кальция нитрат,кальциевая селитра, азотнокислый кальций	100,5	101,1	101,8	102,5	103,4	104,3	105,4	106,7	108,3	110,5	113,7	118,1	123,4	130,0	137,9
$CuSO_4$ - сульфат меди(II), медный купорос	100,1	100,2	100,4	100,6	100,9	101,3	101,8	102,8	104,1	-	-	-	-	-	-
$FeSO_4$ - сульфат железа(II), железо (II), железный купорос	100,1	100,3	100,5	100,7	101,0	101,6	-	-	-	-	-	-	-	-	-
K_2CO_3 - карбонат калия, углекислый калий, поташ	100,4	101,0	101,6	102,3	103,2	104,4	105,9	108,0	110,6	114,2	118,8	124,4	-	-	-
KCl - хлорид калия	100,5	101,3	102,1	103,2	104,6	106,1	107,9	-	-	-	-	-	-	-	-
KH_2PO_4 - дигидрофосфат калия, монофосфат калия	100,2	100,4	100,7	101,0	101,4	101,9	102,4	102,9	103,4	104,0	104,6	-	-	-	-
KNO_3 - нитрат калия, азотнокислый каоий, калиевая селитра, калийная селитра, индийская селитра	100,4	100,7	101,2	101,6	102,1	102,7	103,4	104,1	105,0	106,0	107,1	108,4	109,9	111,7	113,8
KOH - гидроксид калия, kalium hydroxidum, potassium hydroxide, калиевый щёлок, едкое кали, каустический поташ	101,1	102,4	104,1	106,4	109,5	113,3	118,2	124,6	133,4	145,0	160,2	178,4	200,2	226,6	255,5

На самом деле даже такой «простой» процесс зависит от множества факторов и этот «порог» не жёстко зафиксирован.

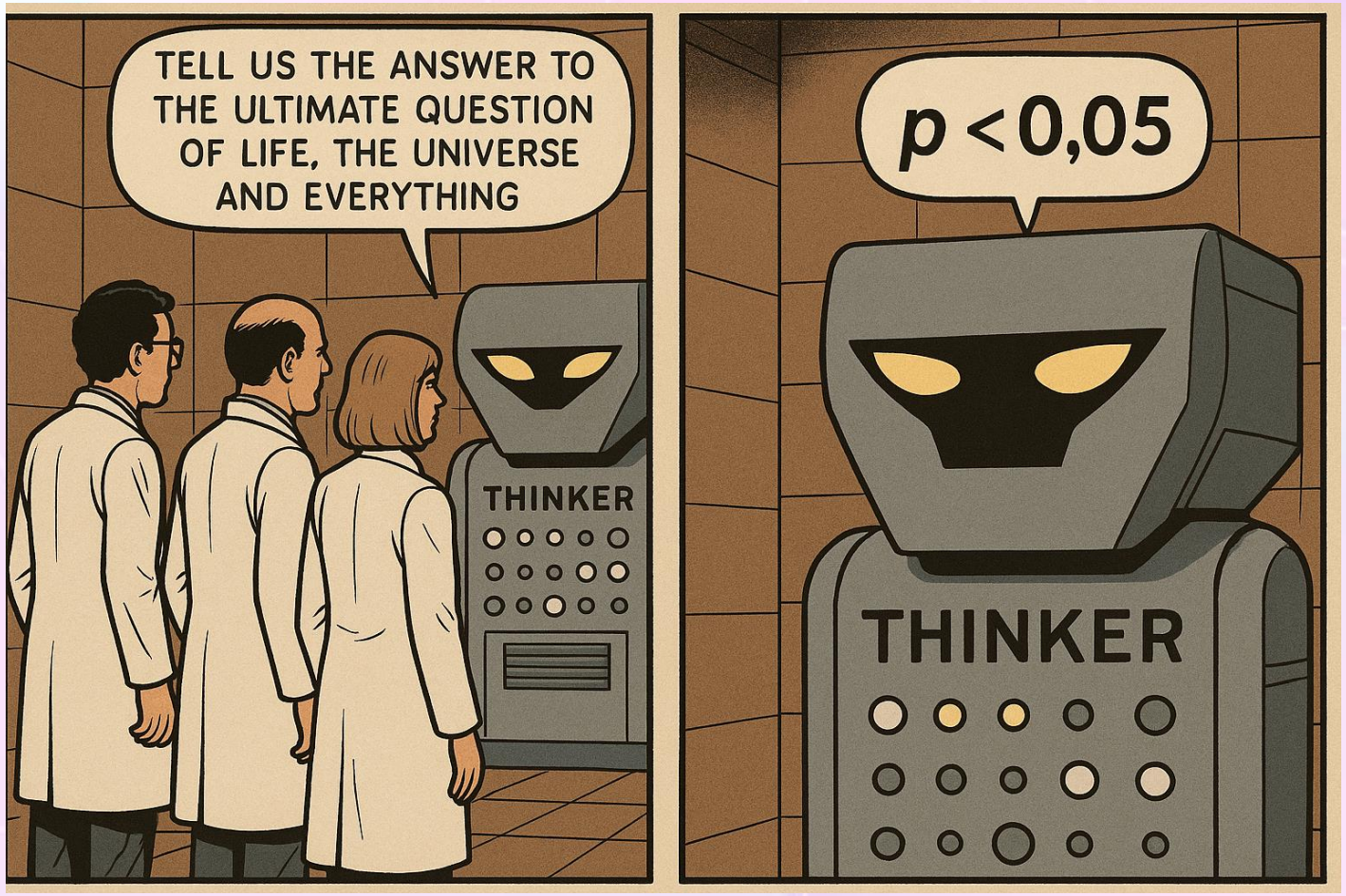
«Пороговые значения» очень редко встречаются в биологической реальности





И тут, и тут температура окружающей среды равна 100°C, но есть нюанс…

На какой вопрос мы чаще всего хотим ответить, проводя статистическое тестирование гипотез?



NB! Это субъективное мнение

Чаще всего мы хотим установить наличие ассоциации между какой-либо «независимой» и «зависимой» переменной

«Независимая»  «Зависимая»

p < 0.05 – «связь есть»

«Независимая»  «Зависимая»

p > 0.05 – не можем сделать вывод о наличии связи

NB! Это субъективное мнение

«Независимая» ? «Зависимая»

Количественные
(шкалы, scale, quantitative)

Порядковые
(ordinal)

Качественные
(номинальные, nominal)

Количественные
(шкалы, scale, quantitative)

Порядковые
(ordinal)

Качественные
(номинальные, nominal)

Дальше мы рассмотрим эффекты, которые возникают в разных ситуациях при категоризации как «независимых», так и «зависимых» переменных.

Часть 2. В которой мы совершаем ошибки I рода

Увеличение частоты ошибок I рода

Смоделируем исследование

Изучение ассоциации между уровнем С-реактивного белка (мг/л) и длительностью нетрудоспособности у пациентов с ОРВИ.

Набрали выборку пациентов

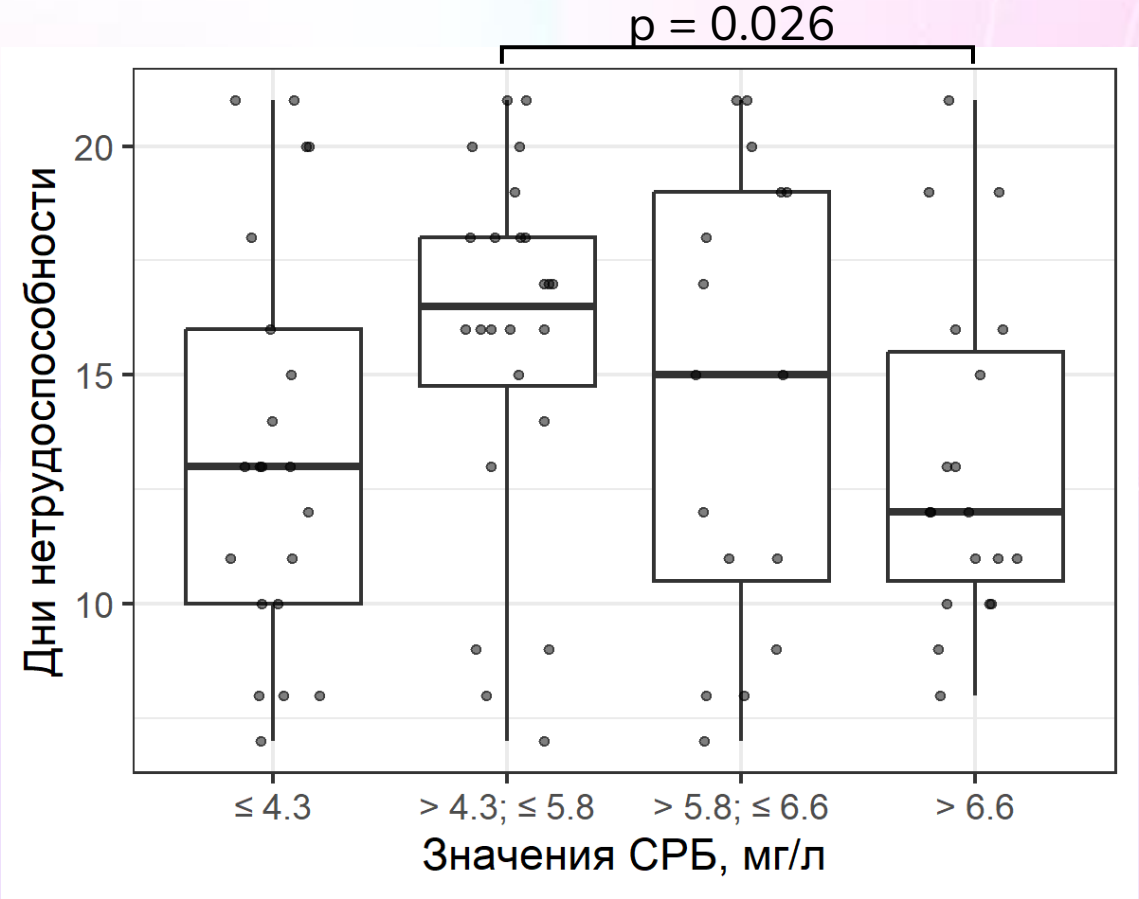
В исследование включили 80 пациентов,
у них зарегистрировали уровень СРБ в день открытия листка нетрудоспособности и длительность больничного.

4 уровня СРБ:

- **низкий**
- **средний**
- **высокий**
- **очень высокий**

	A	B
1	Disability_duration	CRP
2	8	3,2
3	13	3,5
4	7	6,6
5	12	5,9
6	16	6,9
7	20	5,8
8	9	6
9	21	7,6
10	18	4,6
11	21	3,6
12	9	5,2
13	21	6,1
14	16	0,8
15	8	6,4
16	17	5,8
17	11	3,6
18	10	3,6
19	12	7,8
20	16	5,7
21	15	6,6
22	14	5,4
23	11	7,7
24	16	4,9
25	10	7,7
26	10	7,7
27	20	3,9

Уровень СРБ ассоциирован с длительностью нетрудоспособности

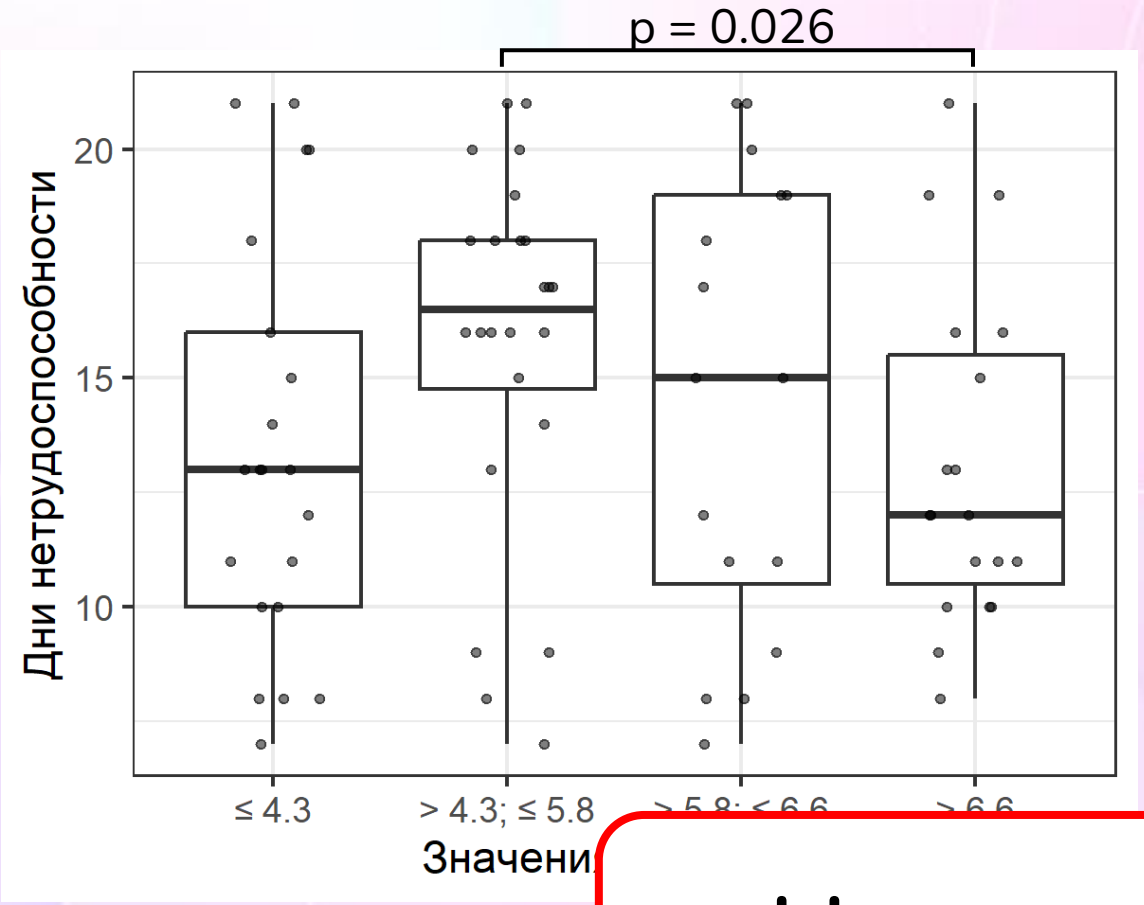


Пациенты были разбиты на 4 группы в зависимости от уровня С-реактивного белка.

Сравнение групп было проведено с помощью критерия Манна-Уитни.

Обнаружены статистически значимые различия в длительности нетрудоспособности между группами 2 и 4 ($p = 0.026$).
Между группами 1 и 2 имелась тенденция к значимости ($p < 0.1$).

Уровень СРБ ассоциирован с длительностью нетрудоспособности



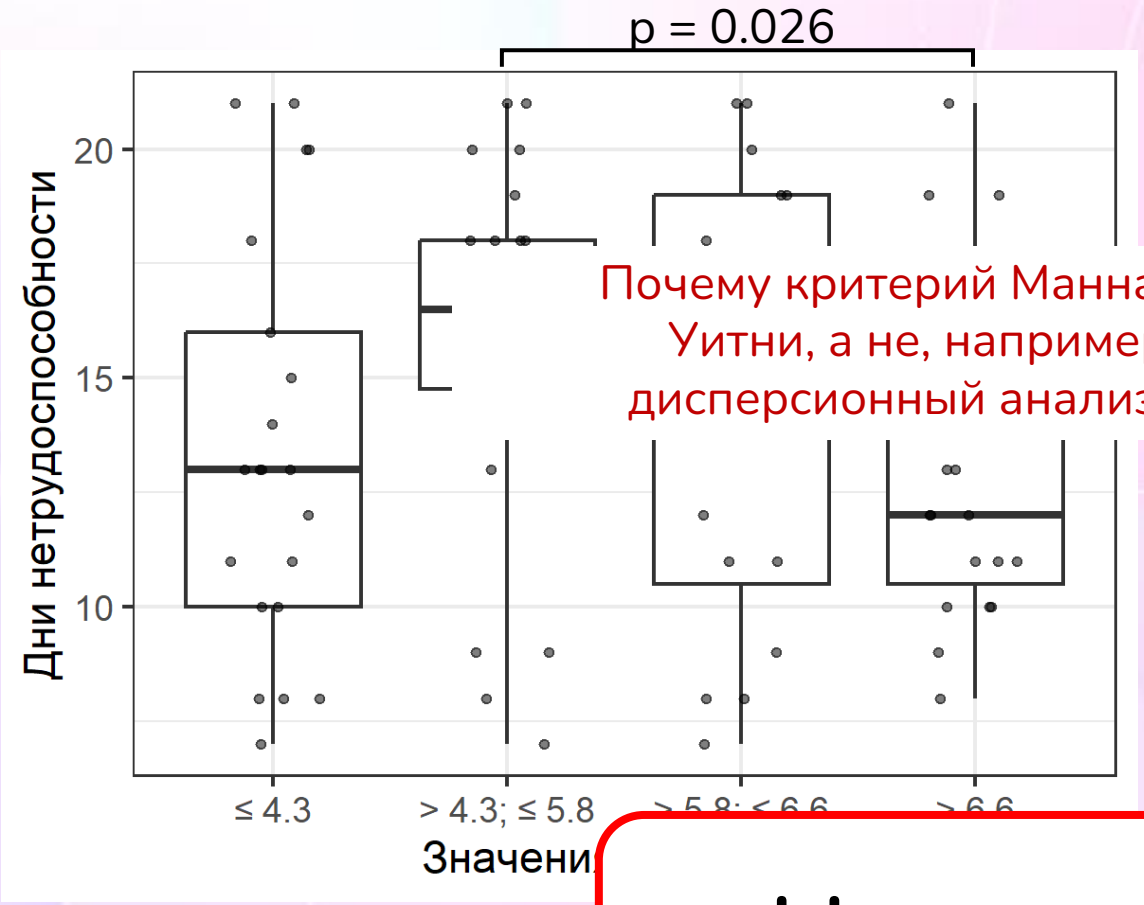
Пациенты были разбиты на 4 группы в зависимости от уровня С-реактивного белка.

Сравнение групп было проведено с помощью критерия Манна-Уитни.

Обнаружены статистически значимые различия в длительности нетрудоспособности между группами 2 и 4 ($p = 0.026$).
Между группами 1 и 2 имелась тенденция к значимости ($p < 0.1$).

Что тут не так?

Уровень СРБ ассоциирован с длительностью нетрудоспособности



Почему критерий Манна-Уитни, а не, например, дисперсионный анализ?

Пациенты были разбиты на 4 группы в зависимости от уровня С-реактивного белка.

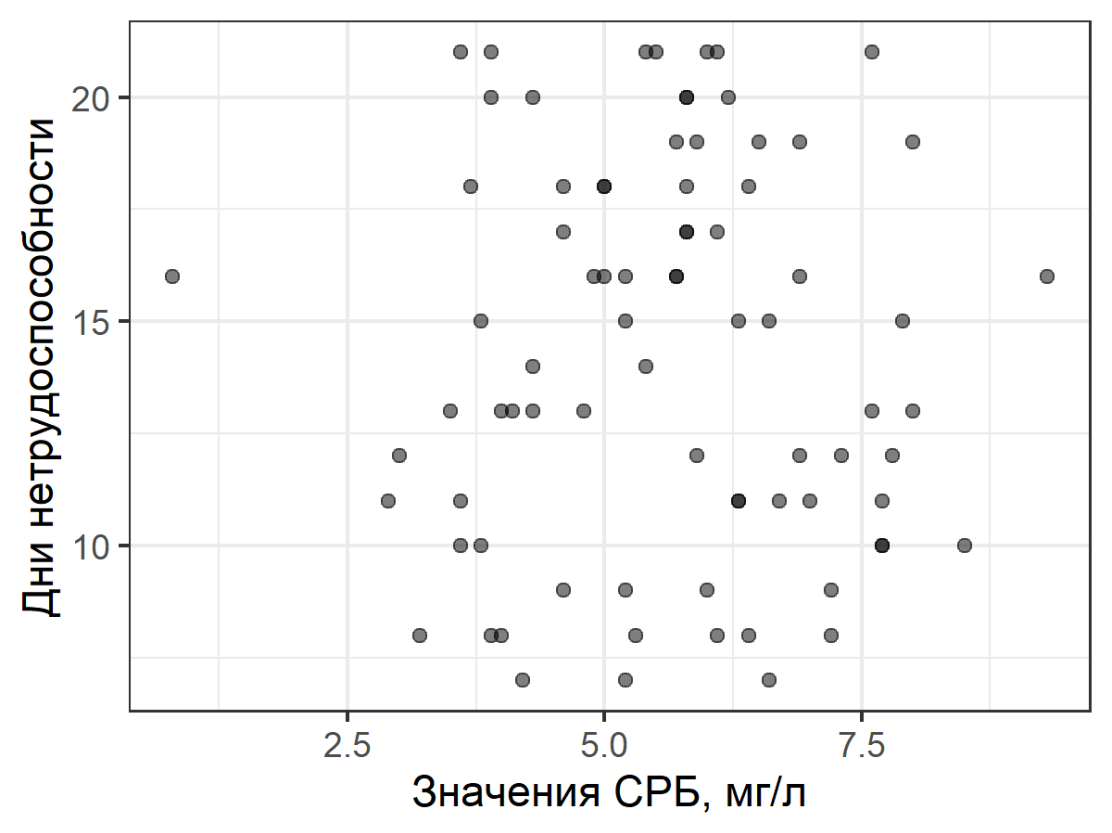
Сравнение групп было проведено с помощью критерия Манна-Уитни.

поправка на множественные сравнения?

Обнаружены статистически значимые различия в длительности нетрудоспособности между группами 2 и 4 ($p = 0.026$).
Между группами 1 и 2 **имелась тенденция к значимости** ($p < 0.1$).

Что тут не так?

Построим диаграмму рассеяния и проведём корреляционный анализ




```
##  Spearman's rank correlation rho
##
## data:  Disability duration and CRP
## S = 87458, p-value = 0.8254
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##         rho
## -0.02505709
```

Статистически значимой корреляции между уровнем СРБ и длительностью периода нетрудоспособности
не выявлено.

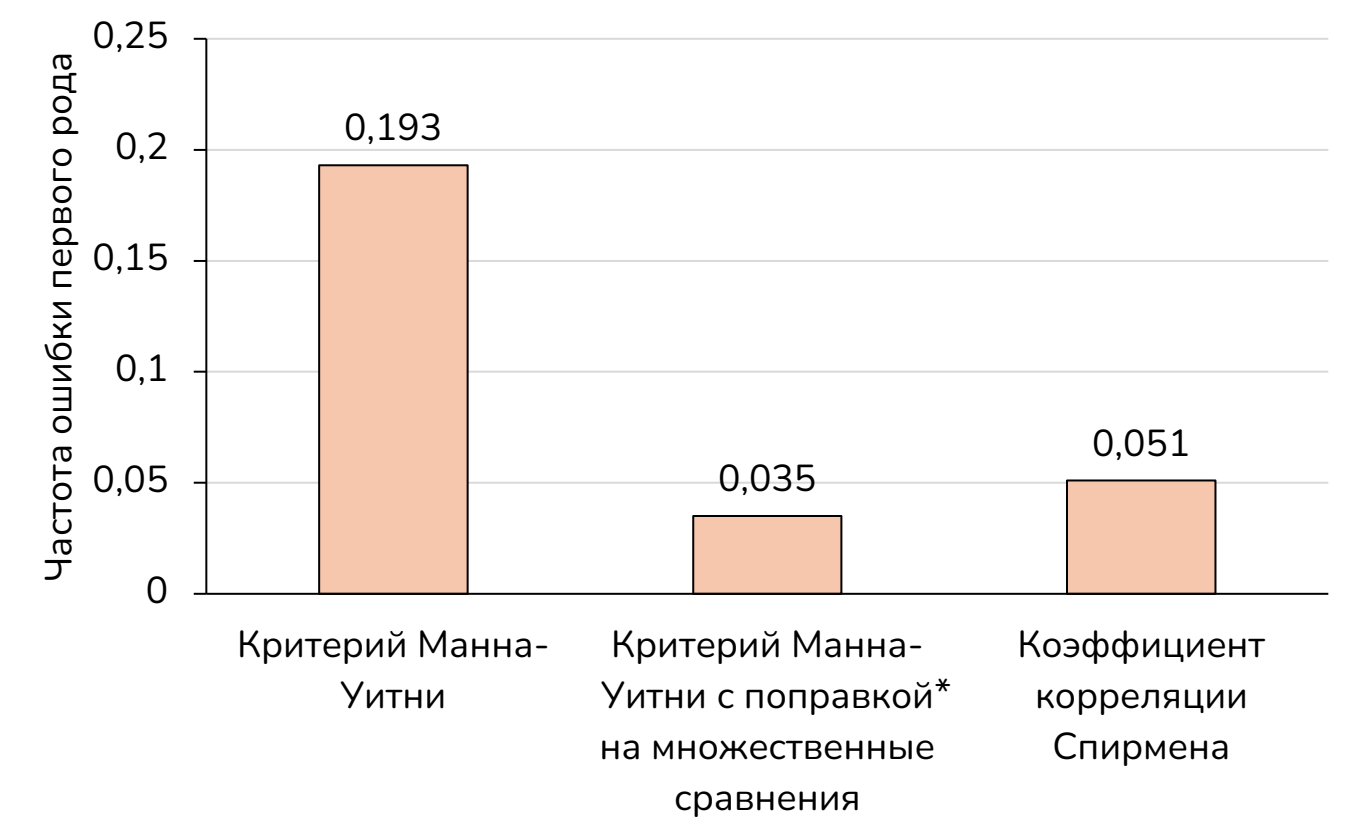
Выборка была сгенерирована

```
N <- 80

tibble_1 <- tibble(
    Disability_duration = runif(N, 7, 22) %>% floor(),  # Длительность нетрудоспособности и уровень СРБ сгенерированы независимыми
    CRP = rnorm(N, 5.5, 1.5) %>% round(1)
)

head(tibble_1)

#> # A tibble: 6 × 2
#>   Disability_duration   CRP
#>                 <dbl> <dbl>
#> 1                  12   5.4
#> 2                  11   3.7
#> 3                  11   7.1
#> 4                  13   4.5
#> 5                  12   6.4
#> 6                  11   5.7
```

Этот эксперимент моделирует «верные» H_0 для тестов Манна-Уитни и Спирмена – ассоциация между переменными отсутствует.

Что произошло?

Мы можем повторить эксперимент 1000 раз...



Из-за отсутствия поправки на множественные сравнения, ошибка первого рода была совершена в 193 исследованиях из 1000. Сильно выше 0.05.

При применении поправки, тест Манна-Уитни контролирует ошибку первого рода на уровне «меньше 0.05», но становится слишком консервативным.

*поправка Хольма-Бонферрони

Часть 3. В которой нам не хватает мощности

Падение мощности исследования

Смоделируем исследование № 2

Изучение ассоциации между уровнем С-реактивного белка (мг/л) и длительностью гипертермии у пациентов с ОРВИ.

Набрали выборку пациентов

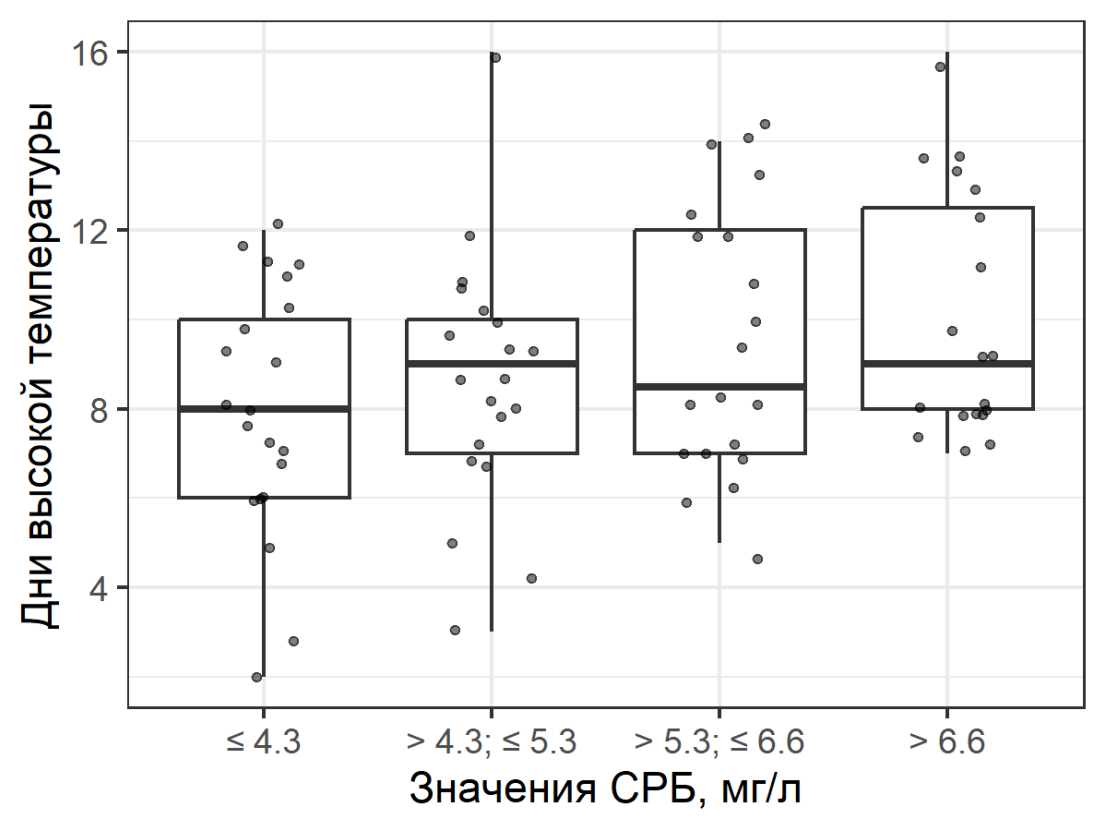
В исследование включили 80 пациентов,
у них зарегистрировали уровень СРБ в день открытия листка нетрудоспособности и длительность больничного, а также записали количество дней, в которые пациенты отмечали температуру выше 37°C

4 уровня СРБ:

- **низкий**
- **средний**
- **высокий**
- **очень высокий**

	A	B	C
1	Disability_duration	CRP	Temp_duration
2	8	8,2	8
3	12	6,6	12
4	16	3,2	6
5	11	5	11
6	8	7,5	8
7	11	4,4	10
8	21	3,2	2
9	21	6,9	14
10	8	9	8
11	10	4,3	10
12	11	7,2	8
13	8	5,4	8
14	7	7,7	7
15	8	5,8	8
16	21	6,3	14
17	9	4,8	9
18	14	4,1	9
19	18	4	12
20	12	5,1	11
21	15	7	8
22	21	8,4	13
23	18	6,6	9
24	19	7,8	16
25	19	3,8	7
26	9	5,9	6
27	17	6,6	7

Не выявлено ассоциаций между уровнем С-РБ и длительностью гипертермии

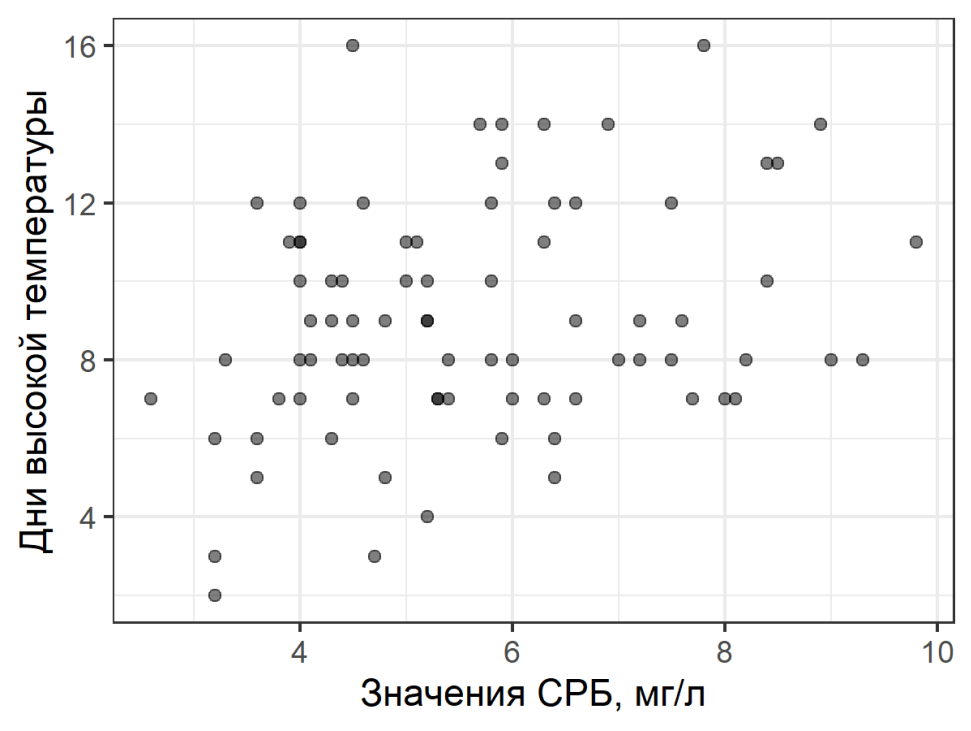


Пациенты были разбиты на 4 группы в зависимости от уровня С-реактивного белка.

Сравнение групп было проведено с помощью дисперсионного анализа.

Статистически значимых различий длительности высокой температуры в зависимости от группы пациентов выявлено не было ($p = 0.134$).

Построим диаграмму рассеяния и проведём корреляционный анализ




```
##  Spearman's rank correlation rho
##
## data:  Temp_duration and CRP
## S = 65457, p-value = 0.03769
## alternative hypothesis: true rho is not
equal to 0
## sample estimates:
##       rho
## 0.2328091
```

Выявлена статистически значимая ассоциация между длительностью гипертермии и значениями СРБ на старте болезни.

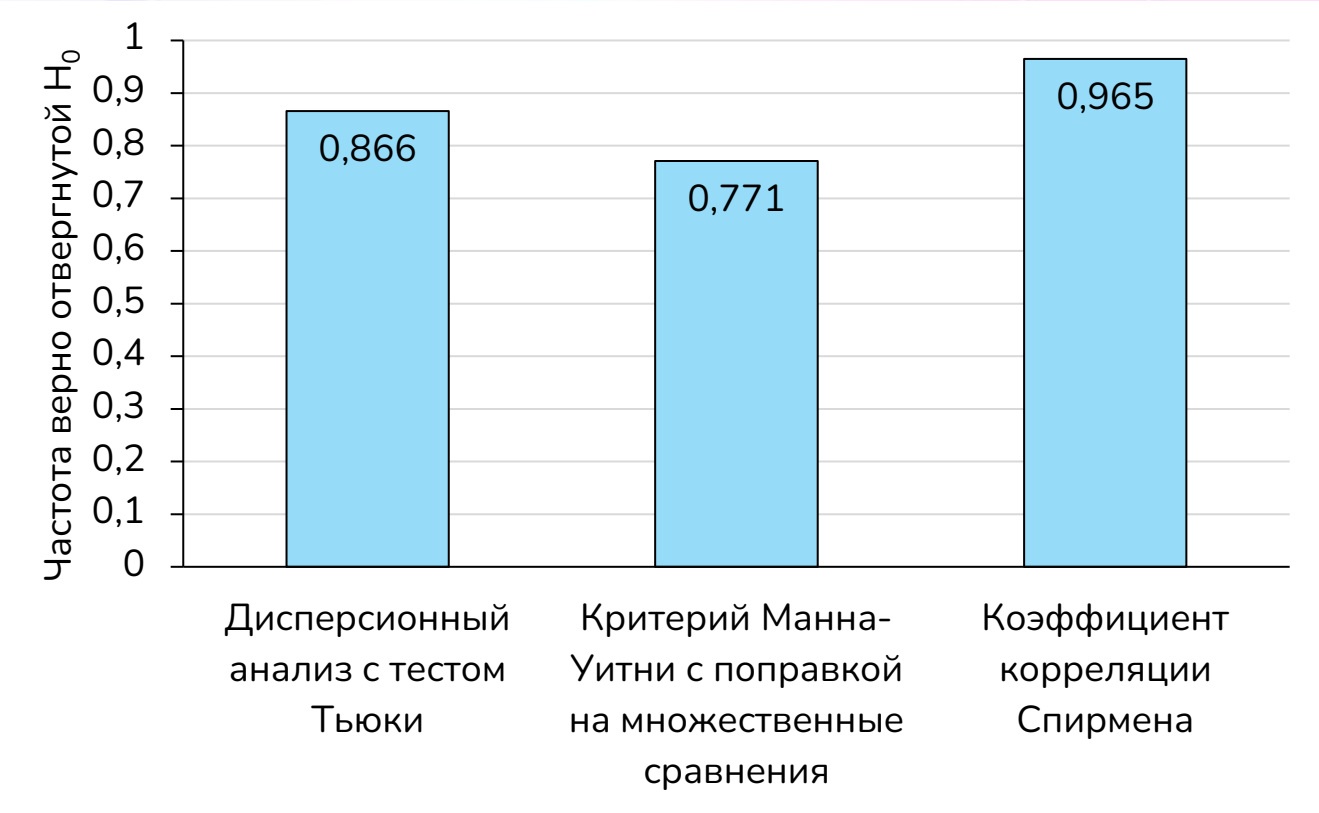
Выборка была сгенерирована

```
N <- 80

tibble_2 <- tibble(
    Disability_duration = runif(N, 7, 22) %>% floor(),
    CRP = rnorm(N, 5.5, 1.5) %>% round(1),
    Temp_duration = (1.2 * CRP + rnorm(N, 4, 3.2)) %>% floor())        # Моделируется линейная зависимость между CRP и Temp_duration
)

head(tibble_2)
#> # A tibble: 6 × 3
#>   Disability_duration   CRP Temp_duration
#>                 <dbl> <dbl>         <dbl>
#> 1                  20   3.1             6
#> 2                   9   3.9             6
#> 3                  15   7.2            10
#> 4                  15   8.1            12
#> 5                  12   8              12
#> 6                  14   5.7            13
```

Этот эксперимент моделирует ситуацию с наличием ассоциации между уровнем СРБ и длительностью гиперемии (верная Ha).

Что произошло?

Мы можем повторить эксперимент 1000 раз...



Мы отвергли H_0

- в 965 случаях из тысячи для коэффициента корреляции Спирмена,
- в 866 случаях из 1000 для дисперсионного анализа с тестом Тьюки
- в 771 случае из 1000 для критерия Манна-Уитни

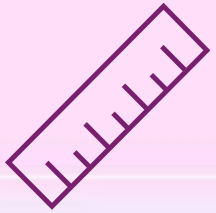
Что мы категоризировали?

«Независимая» «Зависимая»

С-РБ, мг / л

4 уровня СРБ:

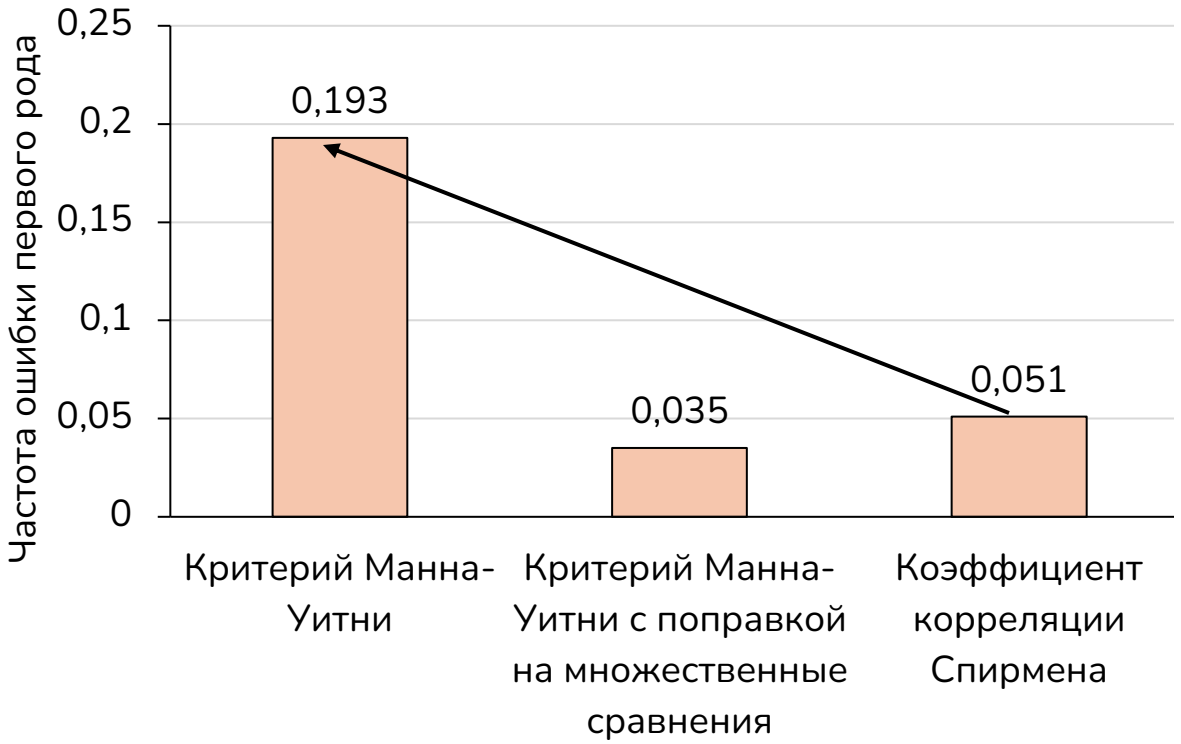
- **низкий**
- **средний**
- **высокий**
- **очень высокий**



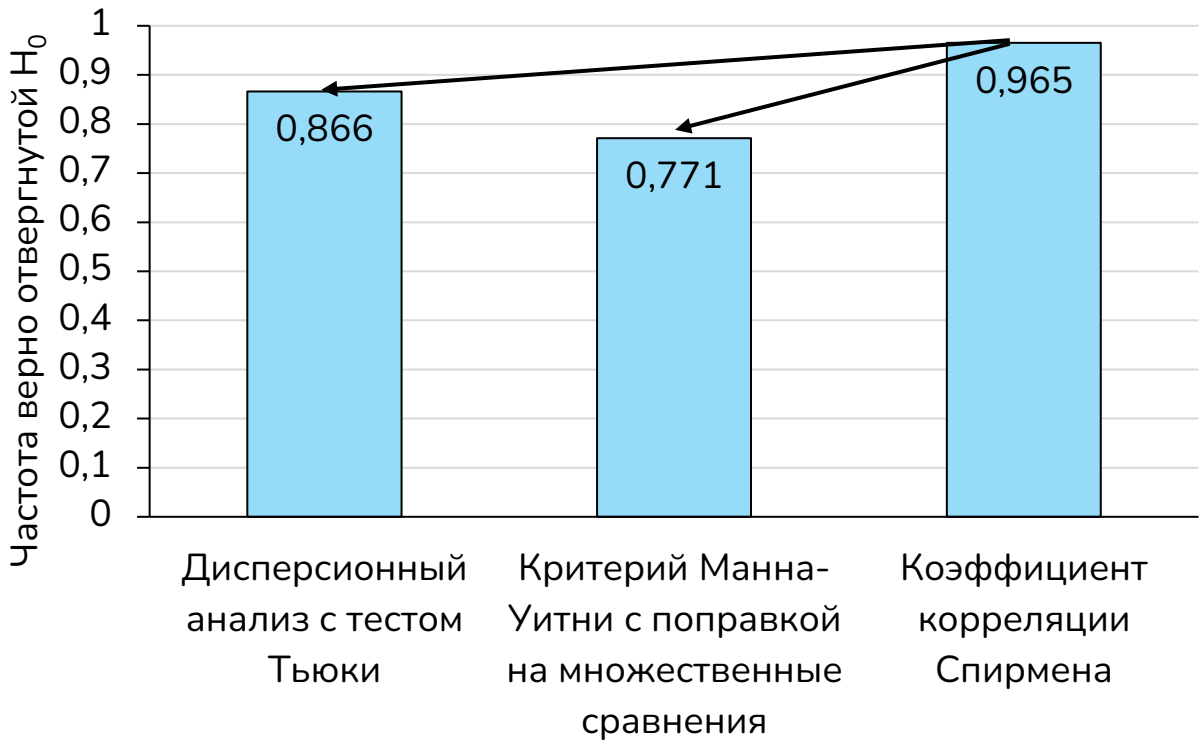
Количество дней нетрудоспособности, Длительность гипертермии

Что мы получили?

При верной H_0



При верной H_a



Возможность роста ошибки I рода

Рост ошибки II рода
(снижение мощности)

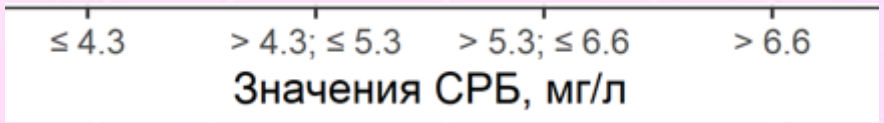
Часть 4. В которой страдает воспроизводимость

Проблемы с воспроизводимостью

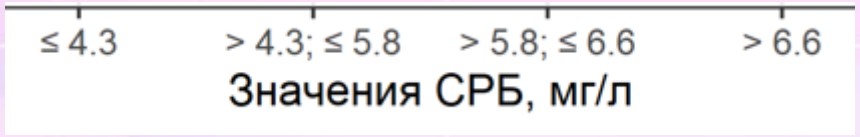
Мы делили апельсин...

4 уровня СРБ:

- **низкий**
- **средний**
- **высокий**
- **очень высокий**



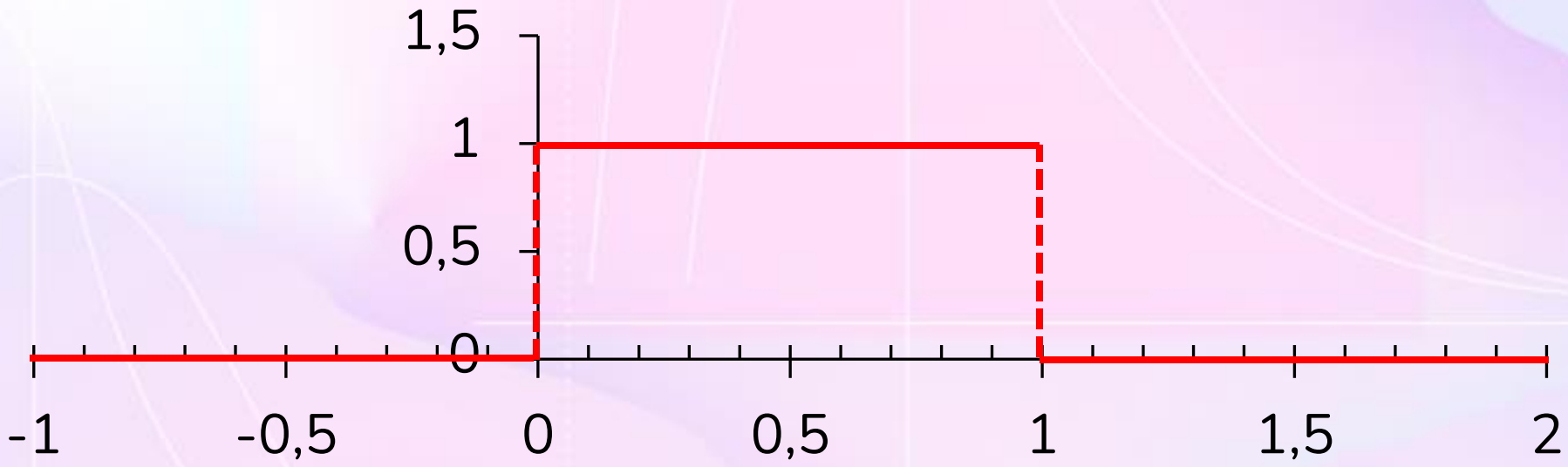
Первый пример



Второй пример

Медианы различаются на 0.5 мг/л

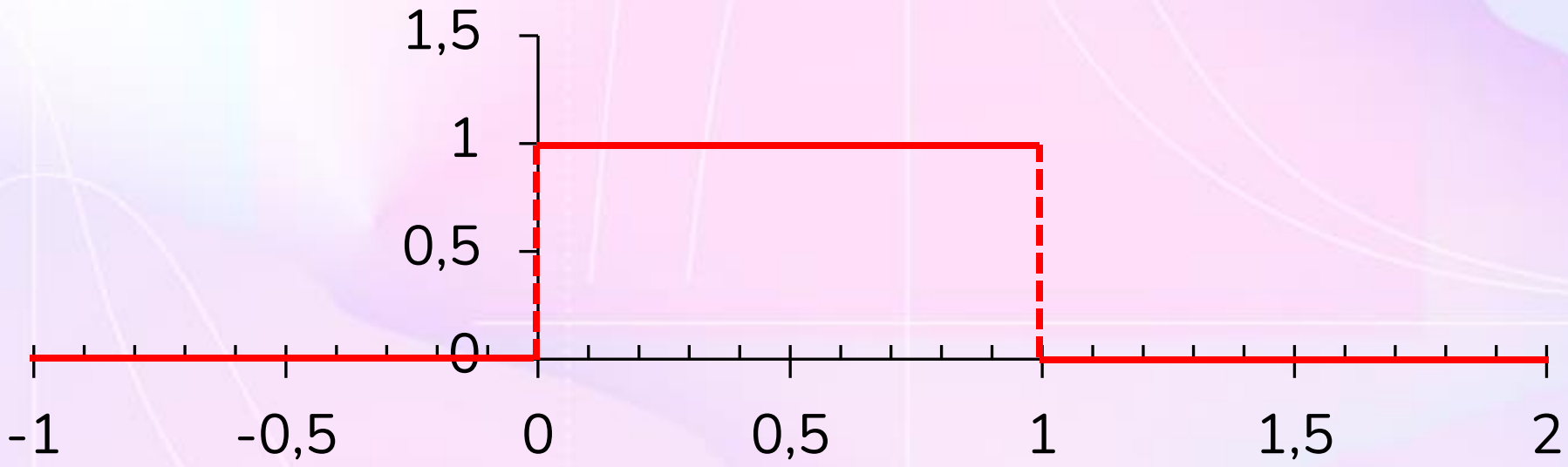
Непрерывное равномерное распределение



Какие квантили у этого распределения?

Min Q25 Me Q75 Max

Непрерывное равномерное распределение



Какие квантили у этого распределения?

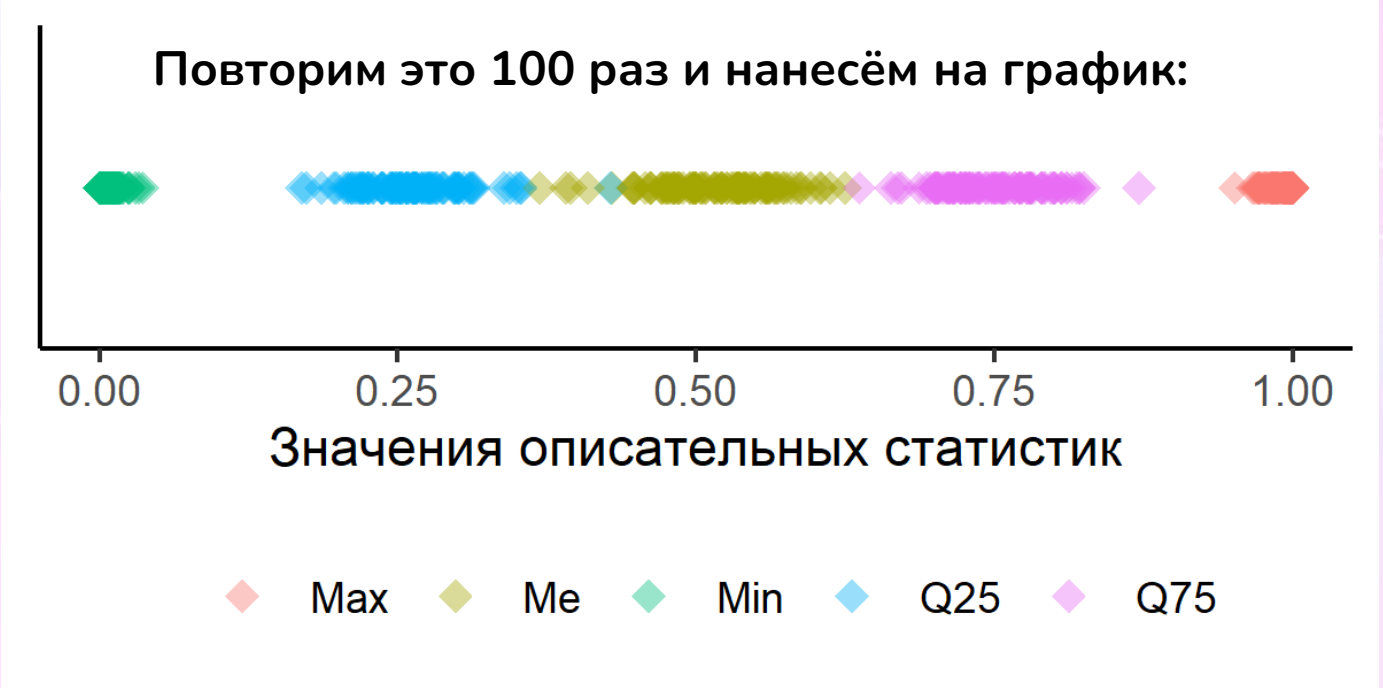
Min	Q25	Me	Q75	Max
0	0.25	0.5	0.75	1

Давайте посимулируем...

```
tibble(Quantiles = c("Min", "Q25", "Me", "Q75", "Max"),
       Values = quantile(runif(100)))
```

```
#> # A tibble: 5 × 2
#>   Quantiles  Values
#>   <chr>       <dbl>
#> 1 Min       0.00669
#> 2 Q25       0.302
#> 3 Me        0.571
#> 4 Q75       0.797
#> 5 Max       0.991
```

функция генерирует выборку из 100 наблюдений и считает на ней квартили



Если это было 100 разных исследований и по этому принципу происходило деление на группы, то как мы, например, сможем провести по ним мета-анализ?

Часть 5. В которой мы теряем информацию

А если мы категоризируем зависимую переменную?

У исследователя стоит задача сравнить значение ИМТ после лечения в зависимости от группы лечения (группа А и группа Б). Исследование было слепым и рандомизированным.

«Независимая» «Зависимая»

Группа

ИМТ, кг/м^2

Категоризируем ИМТ

27.2 кг/м²

< 18 кг/м² – **недостаточный** ИМТ

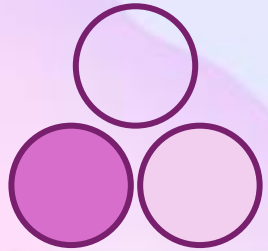
≥ 18 и < 25 кг/м² – **нормальный** ИМТ

≥ 25 и < 30 кг/м² – **избыточный** ИМТ

≥ 30 и < 35 кг/м² – **ожирение I** степени

≥ 35 и < 40 кг/м² – **ожирение II** степени

≥ 40 кг/м² – **ожирение III** степени



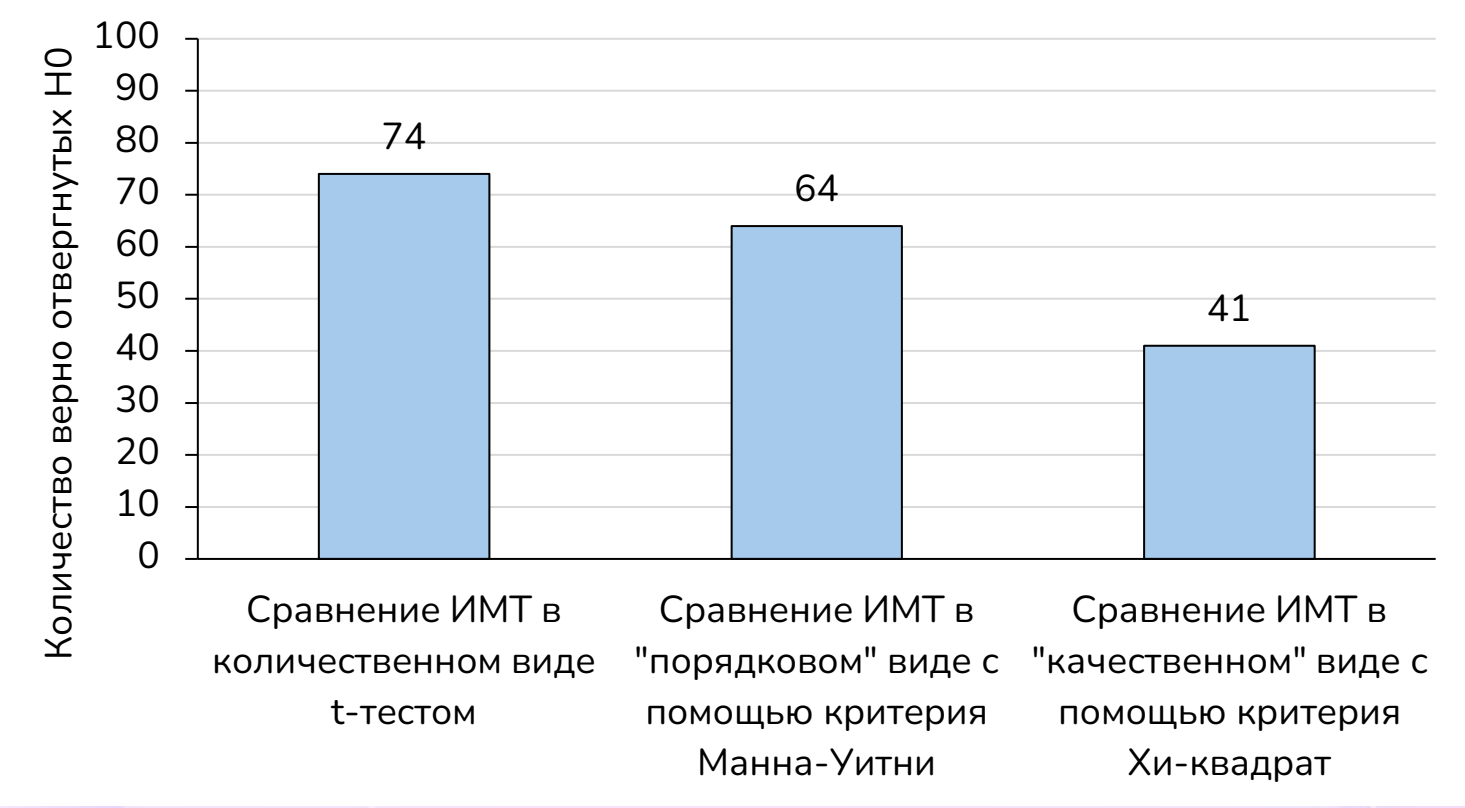
< 30 кг/м² | **≥ 30 кг/м²**

Нет ожирения | **Есть ожирение**

Перейдём к симуляциям

Моделируем при верной Ha

Мы задали $M_А = 29$, $M_Б = 31$, $SD_А = SD_Б = 4$, $n = 64$



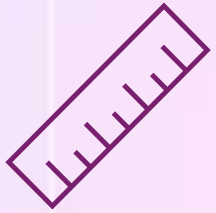
Мощность максимальна, когда мы анализируем данные в количественном виде, а не в порядковом или качественном!

Часть 6. В которой мы снова совершаем ошибку I рода

А теперь попробуем вот такое исследование…

Изучение ассоциации между уровнем С-реактивного белка (мг/л) и наличием бактериальных осложнений у пациентов с ОРВИ.

«Независимая» «Зависимая»



СРБ, мг / л

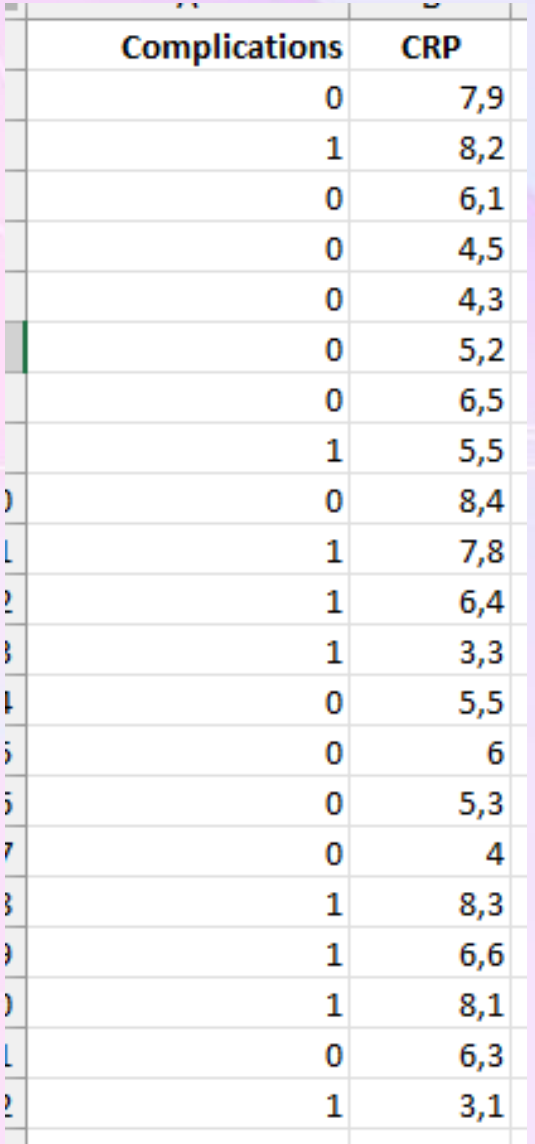
Наличие осложнений

Набрали выборку пациентов

В исследование включили 80 пациентов,
у них зарегистрировали уровень СРБ в день открытия листка нетрудоспособности, а также отметили развились (1) или нет (0) у них осложнения за период больничного.

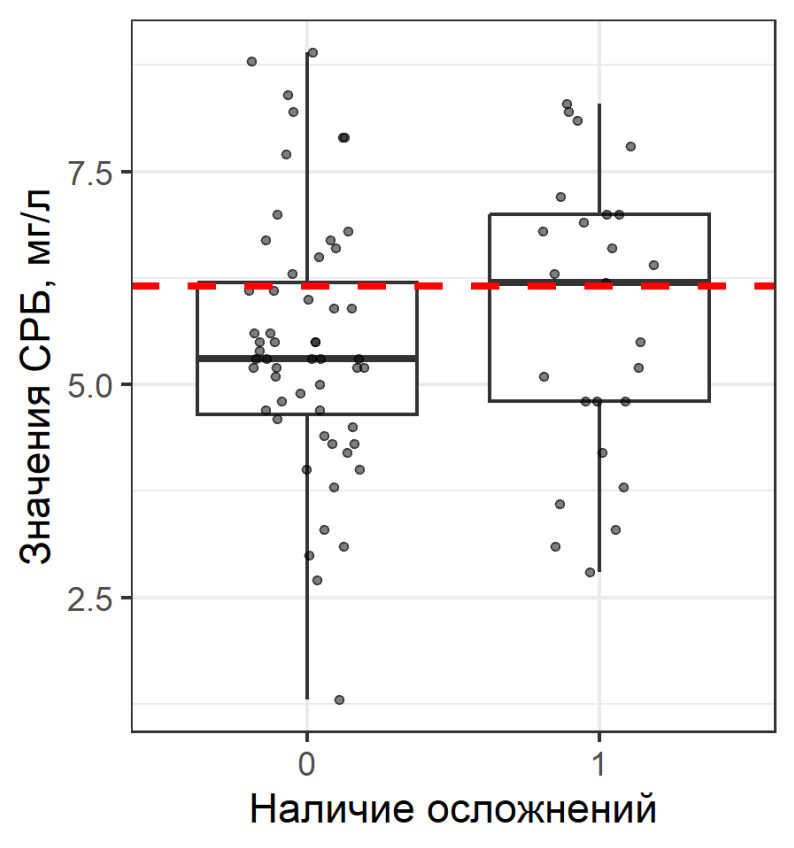
Исследователь хочет построить табличку:

Зависимая переменная	**Независимая переменная**	…	…
Есть осложнения	%	%	%
Нет осложнений	%	%	%



	Complications	CRP
	0	7,9
	1	8,2
	0	6,1
	0	4,5
	0	4,3
	0	5,2
	0	6,5
	1	5,5
)	0	8,4
L	1	7,8
2	1	6,4
3	1	3,3
4	0	5,5
5	0	6
5	0	5,3
7	0	4
3	1	8,3
9	1	6,6
)	1	8,1
L	0	6,3
2	1	3,1

Определили «оптимальную» точку отсечения



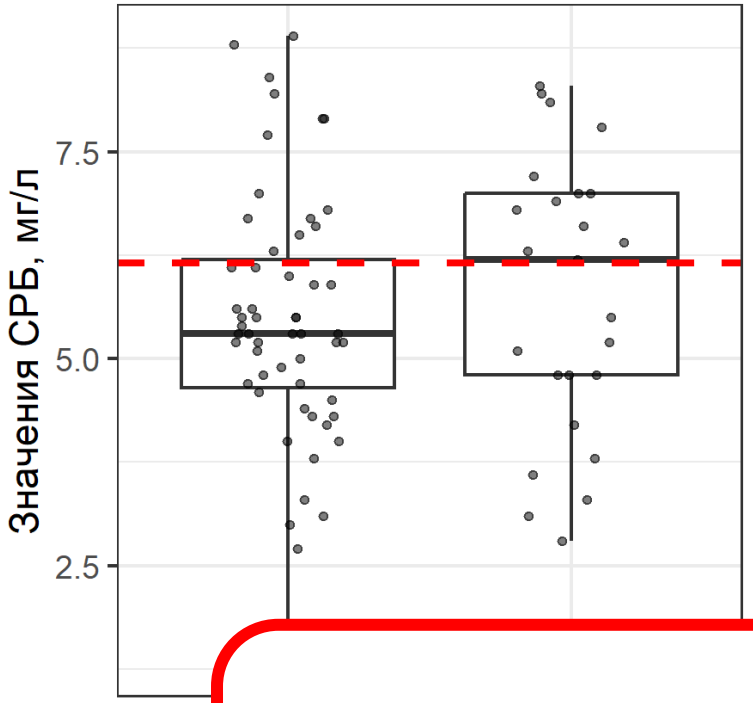
Был проведён ROC-анализ и выбрана «оптимальная» точка отсечения по максимальному индексу Юдена.

	СРБ > 6.15 мг / л	**СРБ < 6.15 мг / л**
Нет осложнений	14 (51,9%)	41 (77,4%)
Есть осложнения	13 (48,1%)	12 (22,6%)

Значение p, критерий Хи-квадрат = 0,02

Вывод: уровень СРБ выше 6.15 мг / л ассоциирован с развитием бактериальных осложнений на фоне ОРВИ

Определили «оптимальную» точку отсечения



	СРБ > 6.15 мг / л	СРБ < 6.15 мг / л
Нет осложнений	14 (51,9%)	41 (77,4%)
Есть осложнения	13 (48,1%)	12 (22,6%)

Значение p, критерий Хи-квадрат = 0,02

А если посчитать логистическую регрессию для СРБ в количественном виде, получим ОШ (95% ДИ), равное 1.12 (0.83; 1.54) и значение p = 0.460

ован с
е ОРВИ

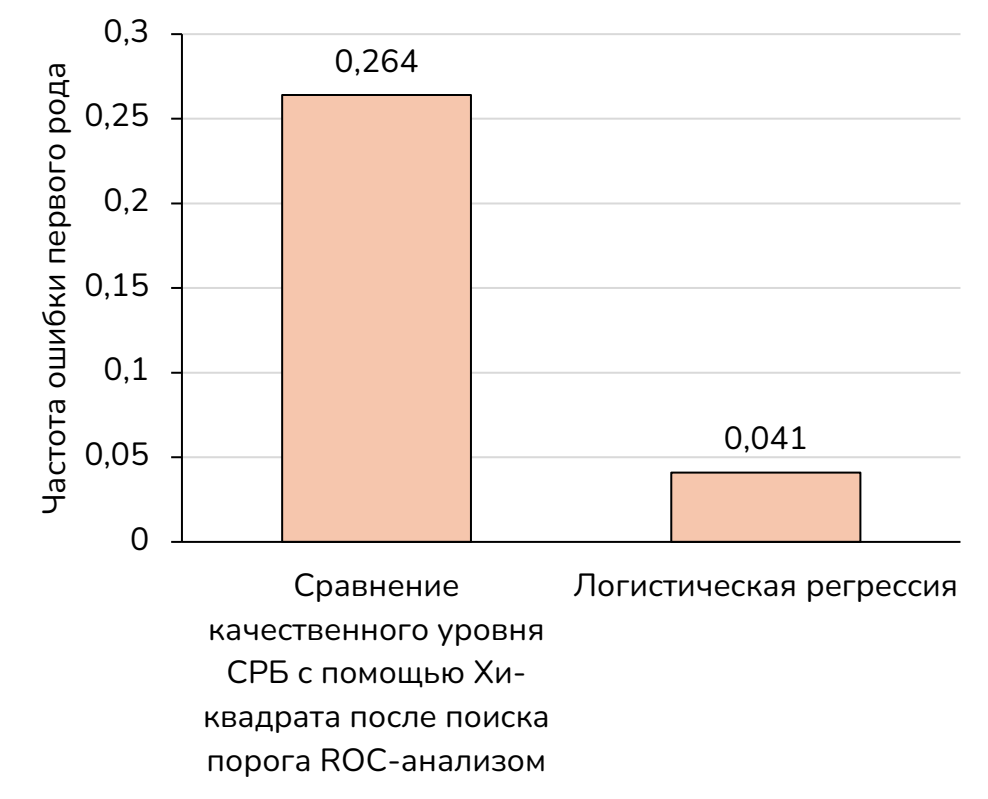
Выборка была сгенерирована

```
N <- 80

tibble_3 <- tibble(
    Complications = rbinom(N, 1, 0.3) %>% as.factor(),    # Осложнения не зависят от уровня СРБ
    CRP = rnorm(N, 5.5, 1.5) %>% round(1))

head(tibble_3)

#> # A tibble: 6 × 2
#>   Complications   CRP
#>   <fct>         <dbl>
#> 1 0               5.8
#> 2 1               5.5
#> 3 1               4.1
#> 4 1               7.3
#> 5 0               8.2
#> 6 0               8.2
```

Этот эксперимент моделирует ситуацию с отсутствием ассоциации между уровнем СРБ и осложнениями (верная H_0).

Что произошло?

Мы можем повторить эксперимент 1000 раз...



!

Из-за того, что порог определяется на той же выборке, мы по сути занимаемся p-hacking'ом, **сильно увеличивая вероятность ошибки первого рода**.

Часть 7. В которой мы оцениваем «риски»

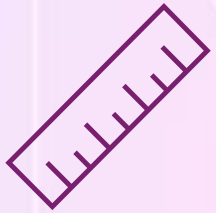
И ещё один пример…

Задача – оценить ассоциацию между уровнем PSA и наличием рака предстательной железы.
Предложить оптимальный вариант принятия дальнейших решений о тактике ведения пациента.

Наберём выборку

Проведём масштабное исследование, наберём 10000 пациентов группы риска (мужчины старше 60 лет). Распространённость рака ПЖ в этой категории пациентов ~ 5%.

«Независимая» «Зависимая»



PSA, нг / мл

Наличие рака ПЖ

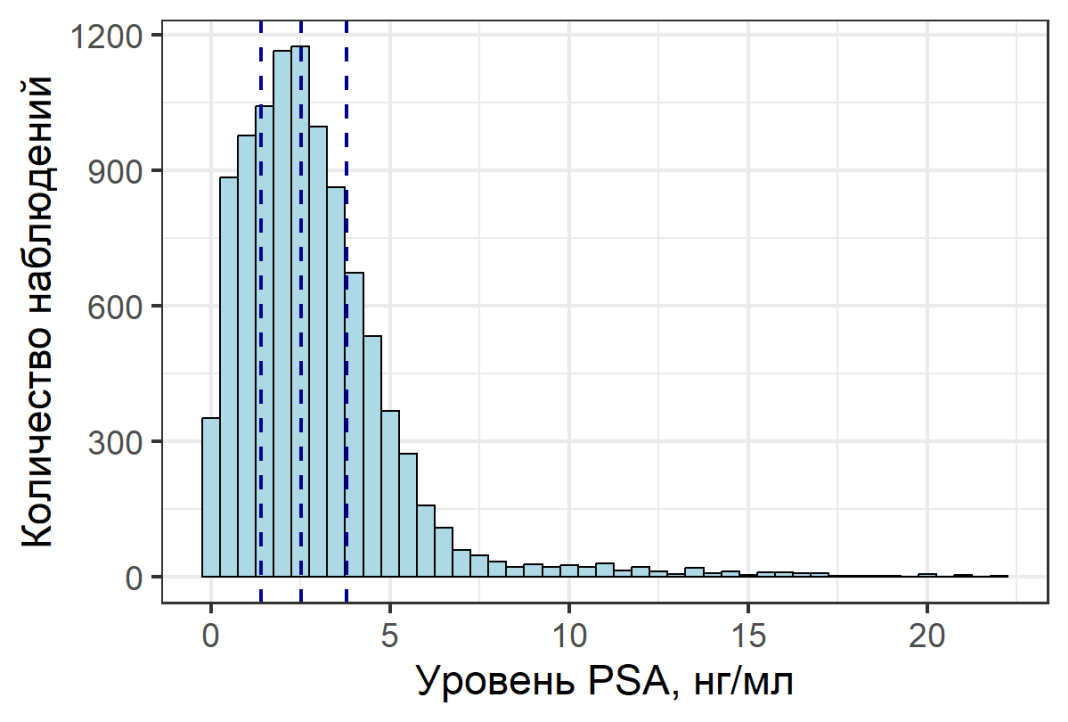
Категоризируем PSA по уровням

4 уровня PSA (нг / мл)

- **Q1 - низкий**
- **Q2 - средний**
- **Q3 - высокий**
- **Q4 - очень высокий**

Так как у нас выборка в 10000 наблюдений, будем считать наше деление по квартилям «золотым стандартом»

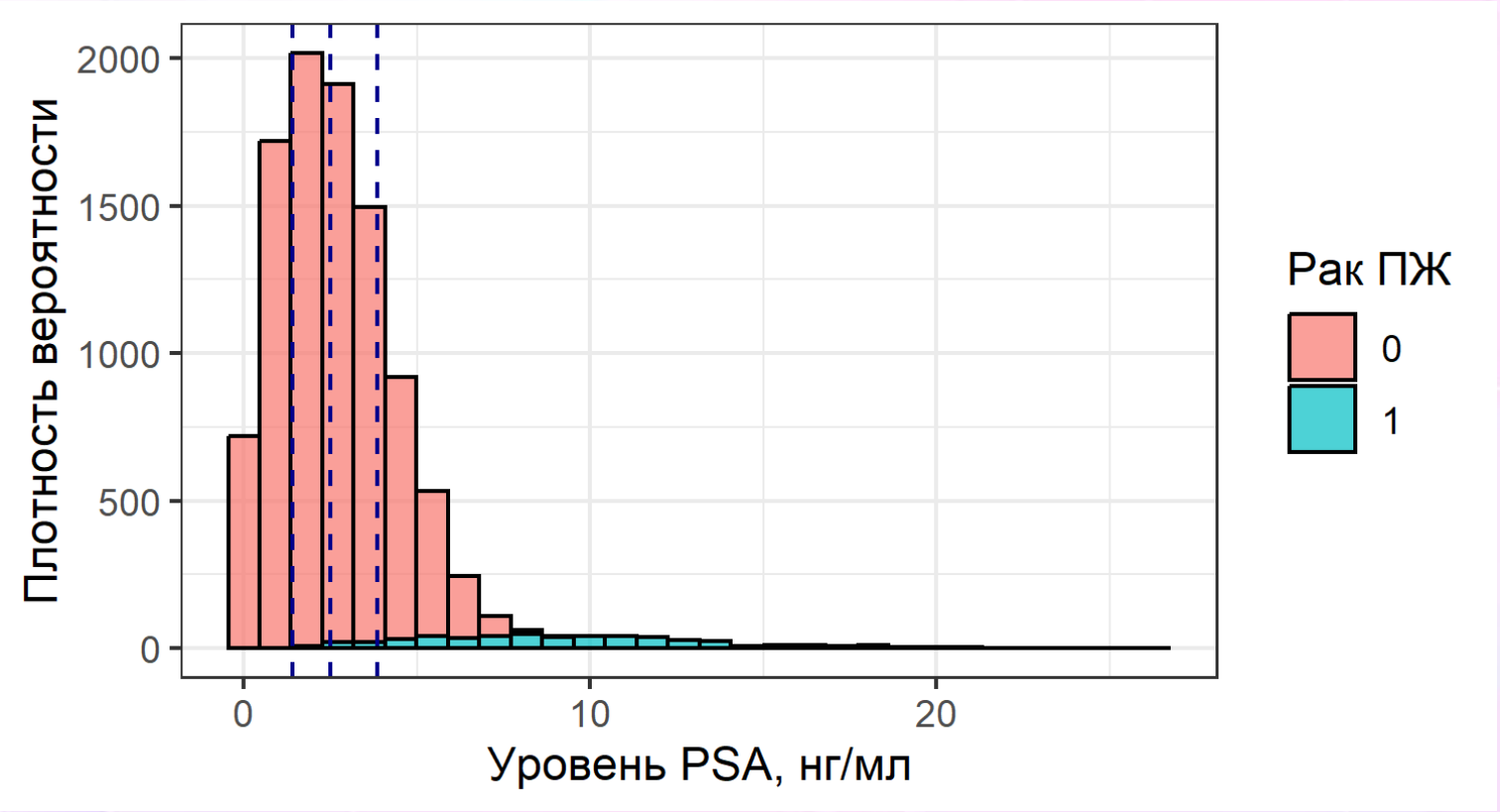
Как выглядит наше деление по квартилям на гистограмме распределения



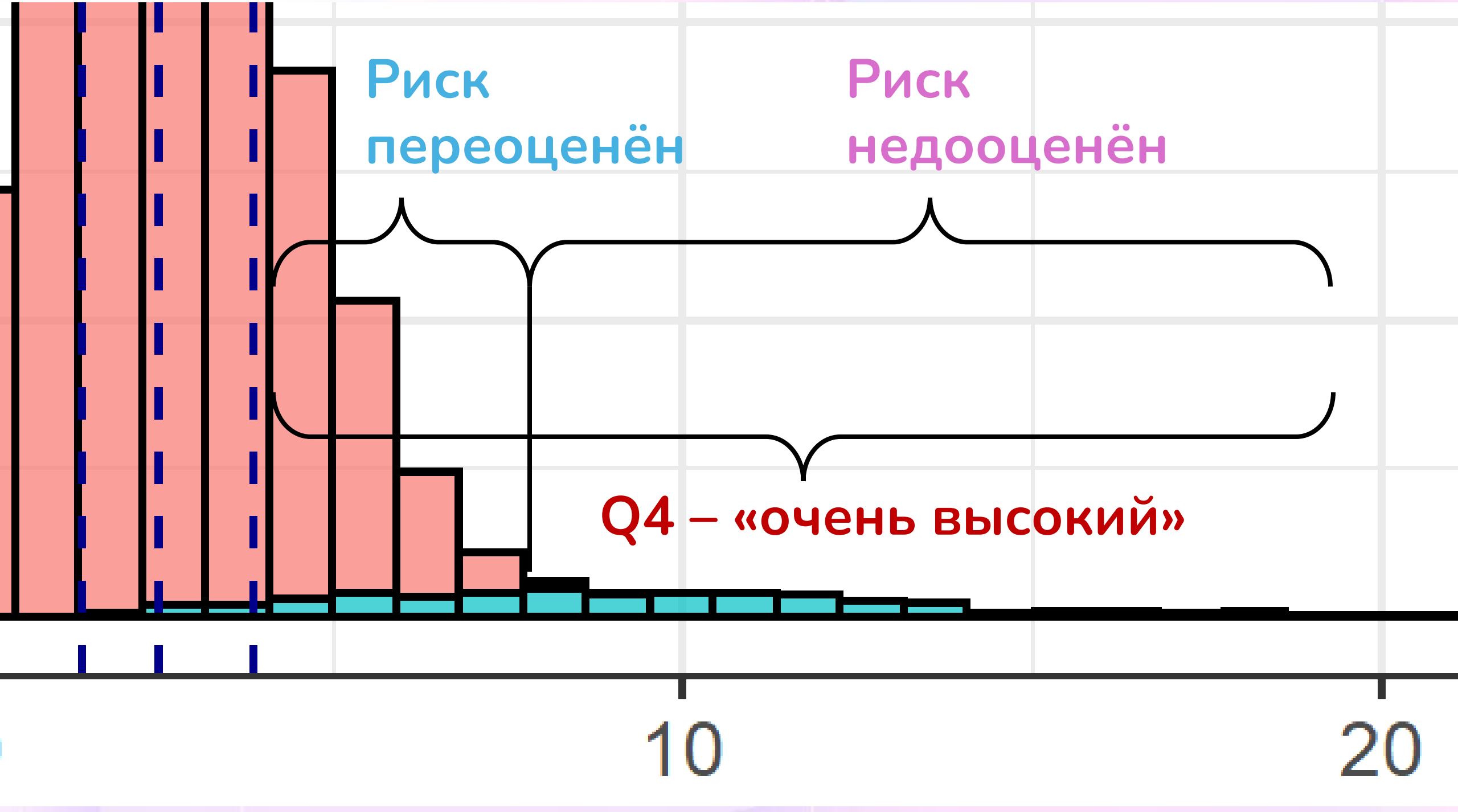
Параметр		Квартиль PSA			
		≤ 1.4	> 1.4; ≤ 2.5	> 2.5; ≤ 3.84	> 3.84
Cancer	0, n (%)	2531 (100%)	2450 (99.5%)	2488 (99%)	2031 (81.5%)
	1, n (%)	0 (0%)	13 (0.5%)	26 (1%)	461 (18.5%)

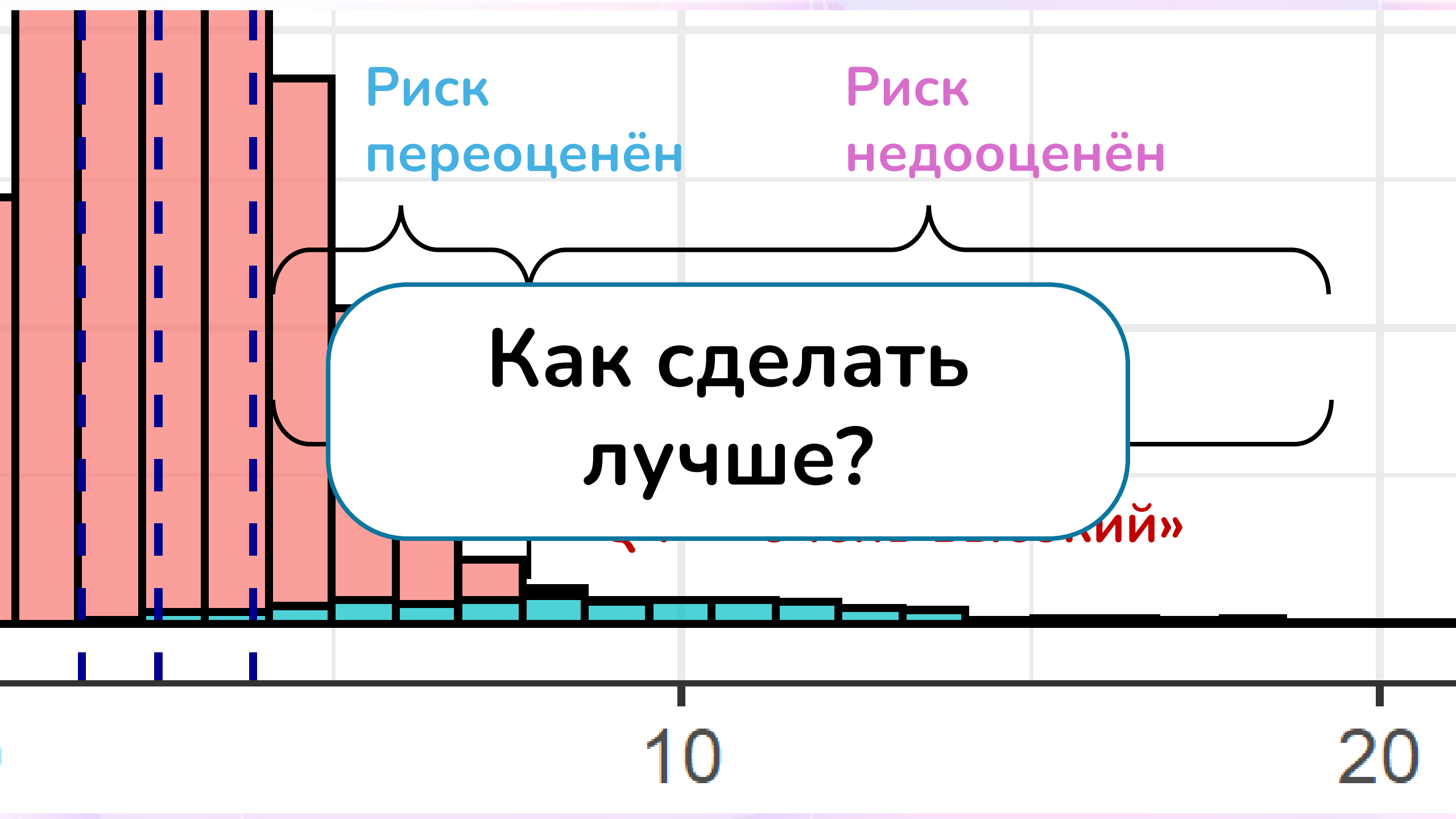
Вывод: Уровень PSA > 3.84 из 4 квартиля ассоциирован с высоким риском наличия рака предстательной железы!

Но эта оценка риска не идеальна...

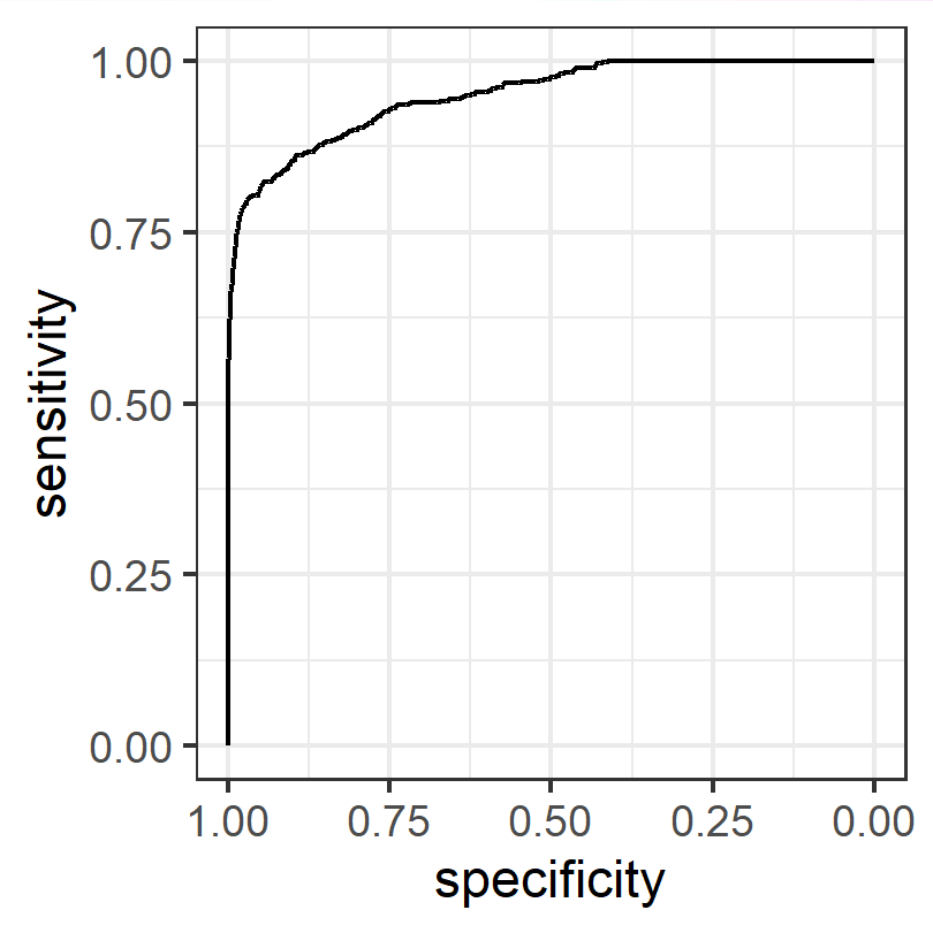


У некоторых пациентов из Q4 мы риск переоценили, а у некоторых мы риск сильно недооценили!


Риск переоценён
Риск недооценён
Q4 – «очень высокий»
10
20


Риск переоценён
Риск недооценён
Как сделать лучше?
10
20

ROC-анализ



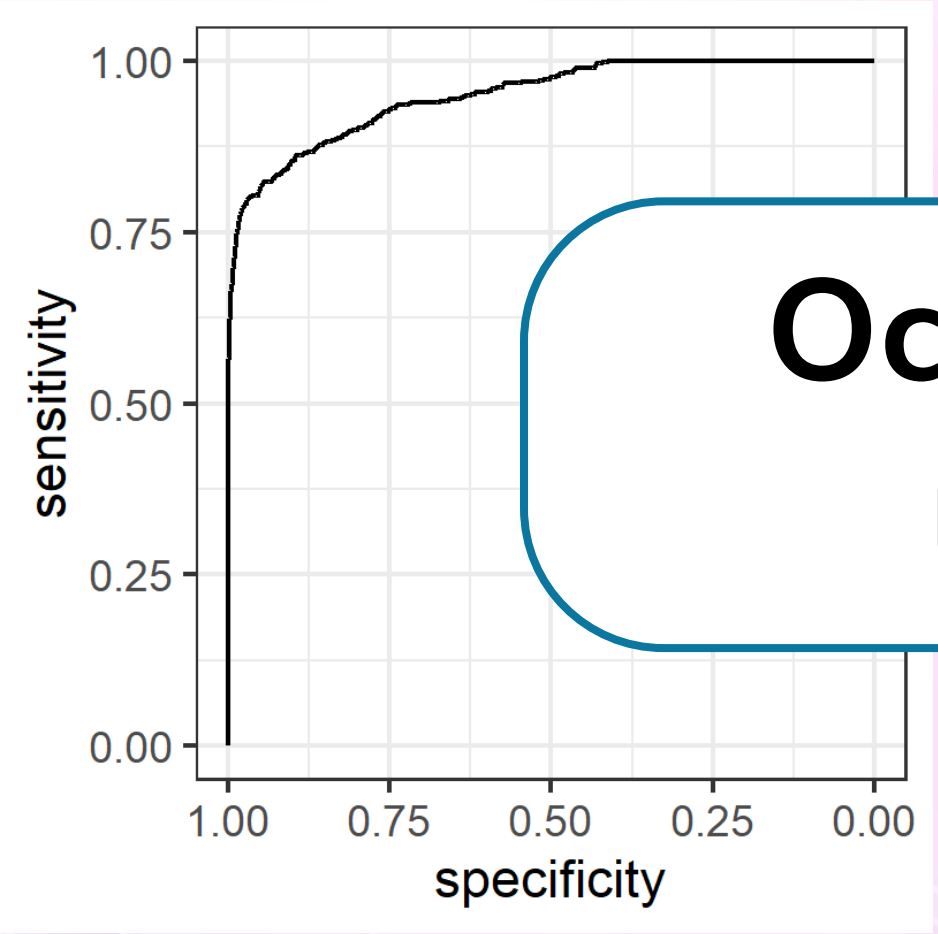
Площадь под ROC-кривой (95% ДИ):
0.951 (0.94; 0.961)

Оптимальный порог (индекс Юдена):
5.87

Чувствительность:
80%

Специфичность:
97%

ROC-анализ



Площадь под ROC-кривой (95% ДИ):
0.951 (0.94; 0.961)

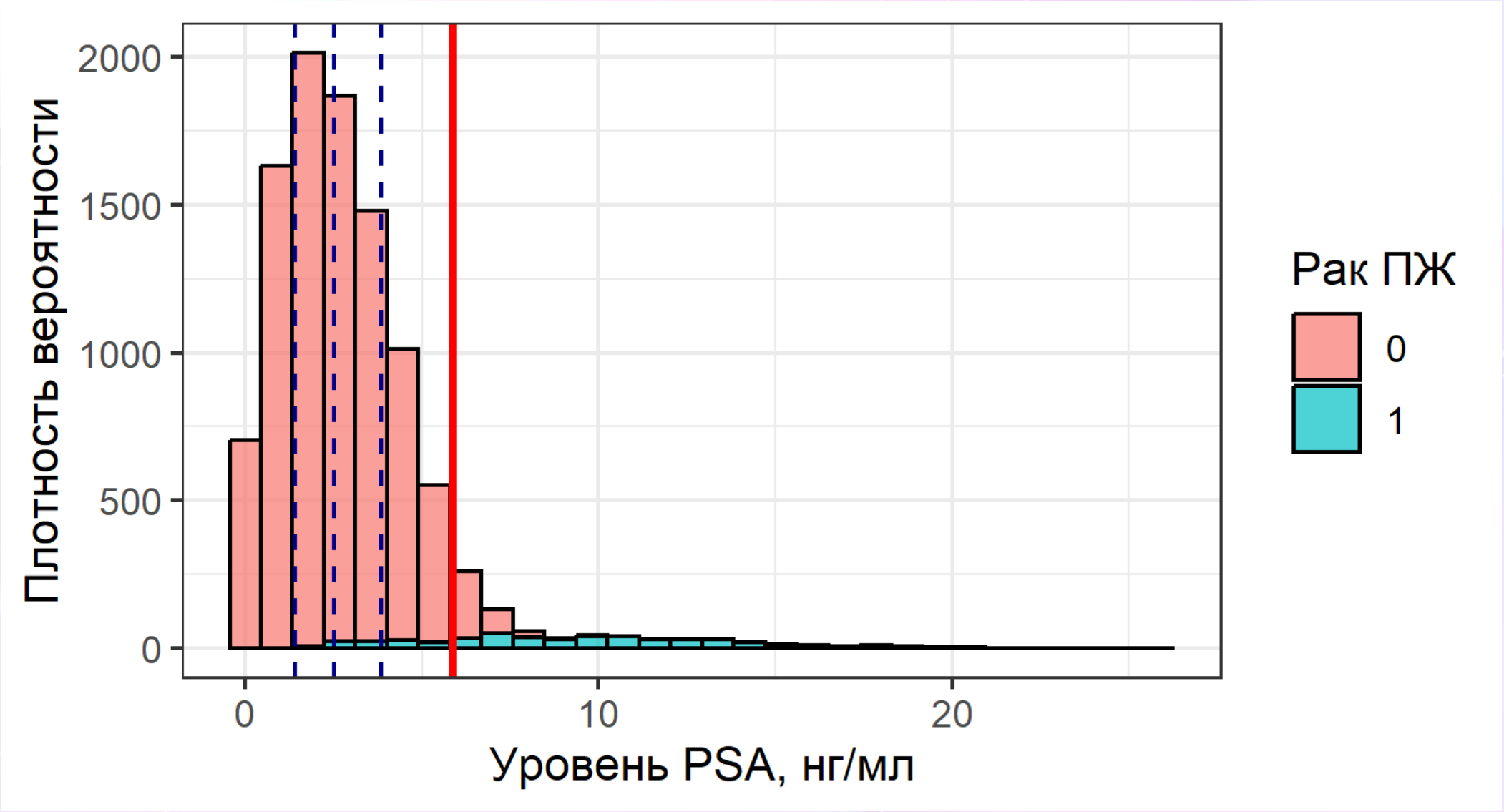
Остановимся на этом?

дена):

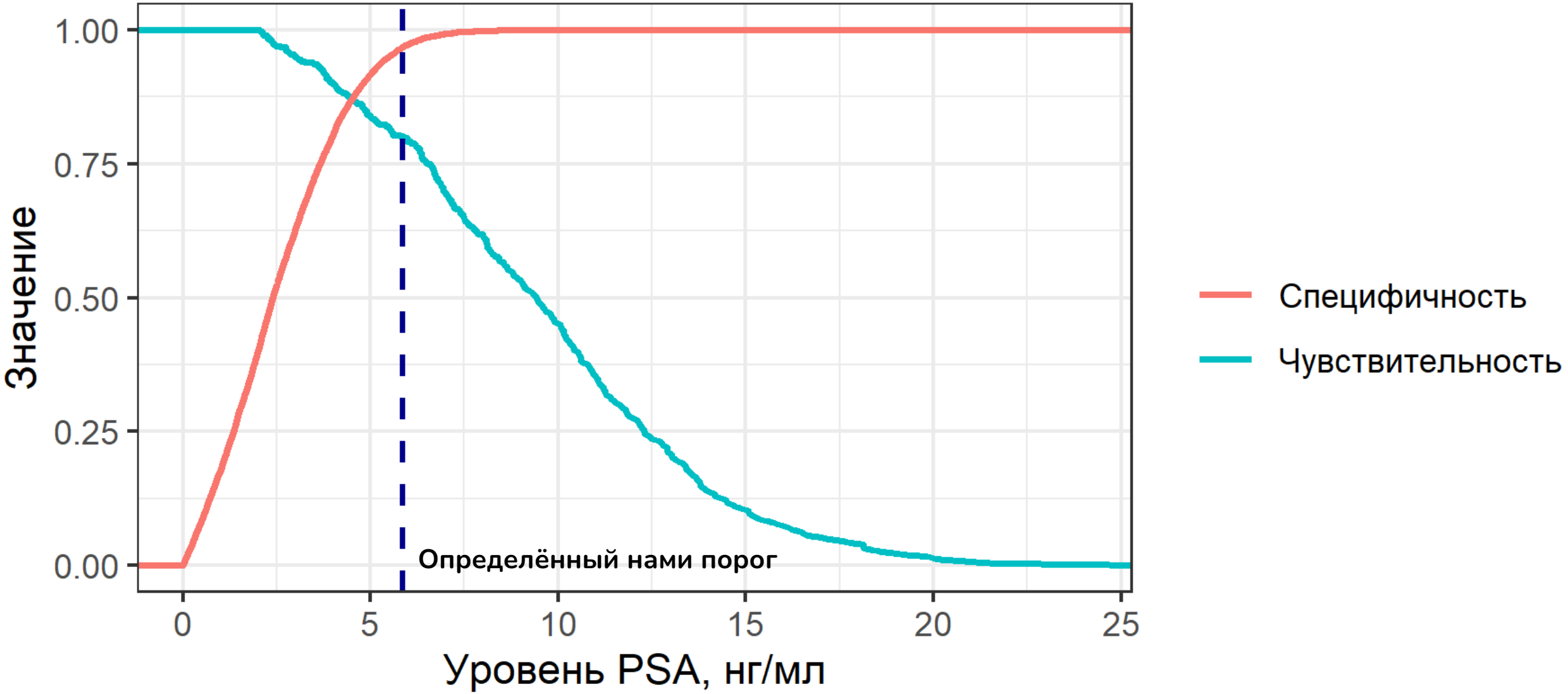
ность:
80%

Специфичность:
97%

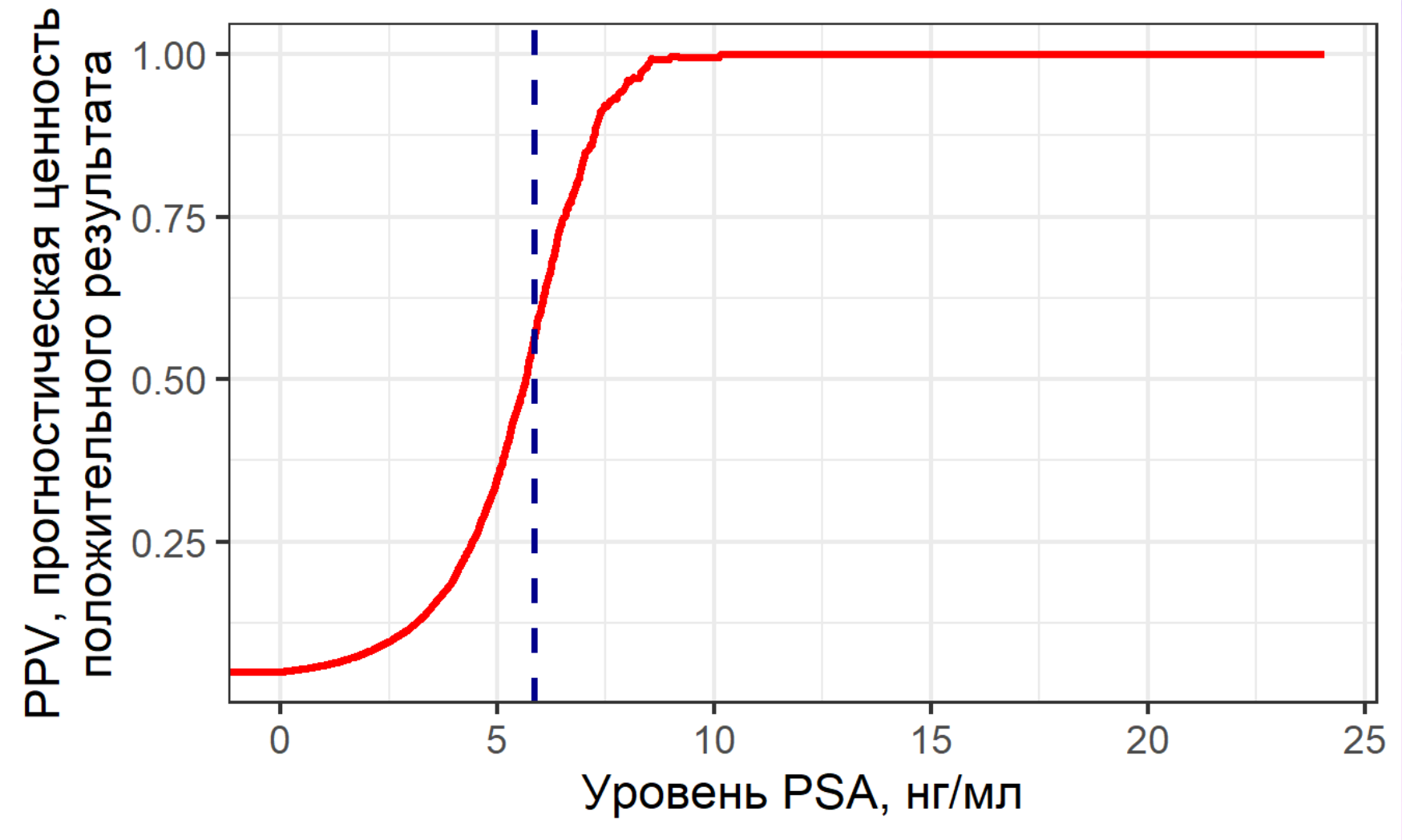
Нет!


Плотность вероятности
2000
1500
1000
500
0
0
10
20
Уровень PSA, нг/мл
Рак ПЖ
0
1

Альтернатива ROC-кривой



Прогностическая ценность



Часть 8. В которой мы выдыхаем…

Выводы

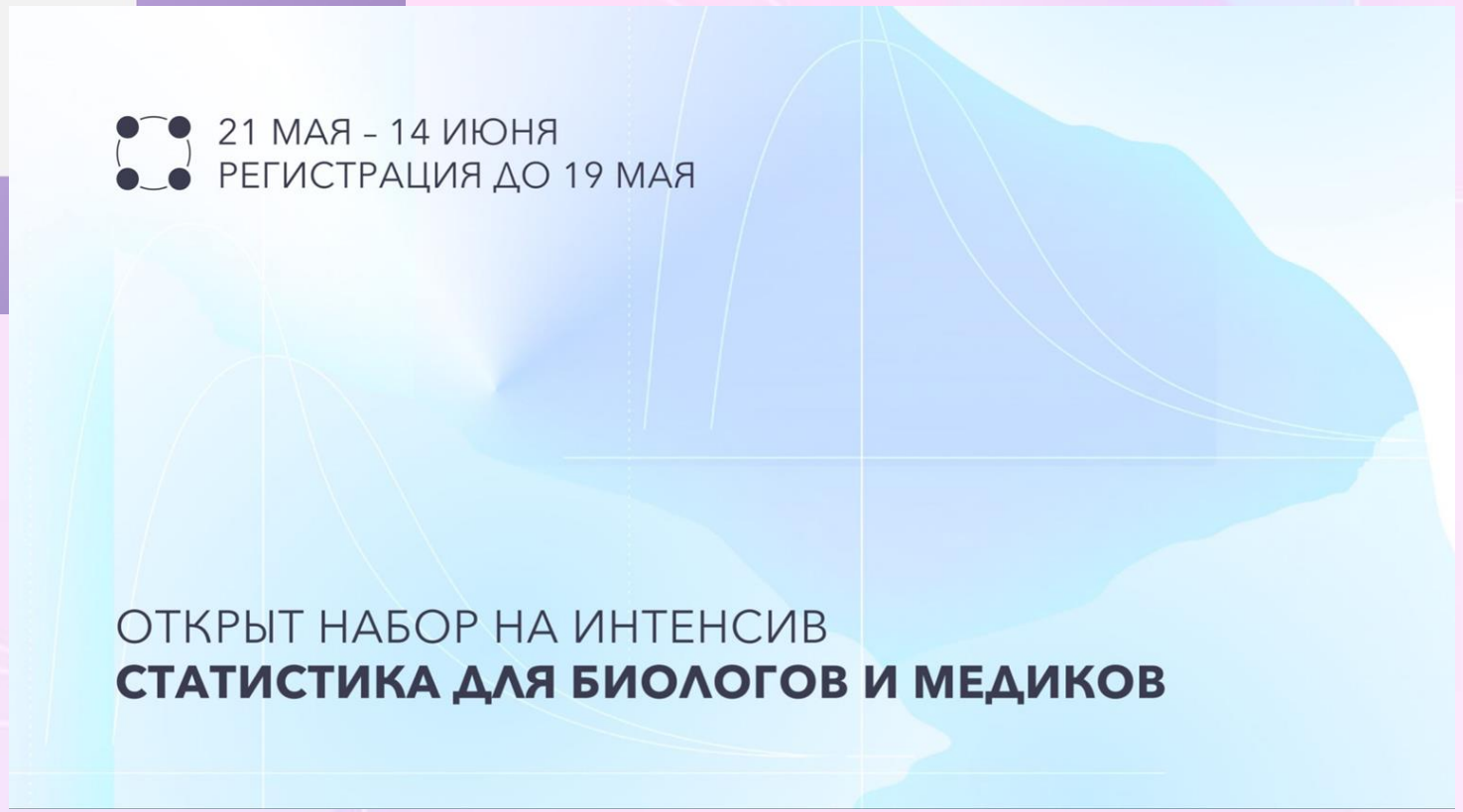
- **Категоризация непрерывных переменных — это упрощение, которое почти всегда сопровождается потерей информации.**
  Использование категорий может снижать статистическую мощность, искажать результаты и создавать иллюзию пороговых эффектов.
- **Разделение по медиане, квартилям и другим «выборкозависимым» критериям — технически удобно, но методологически уязвимо.**
  Такие пороги плохо воспроизводимы, не имеют клинического смысла и затрудняют сопоставимость исследований.
- **Категоризация может увеличивать риск ошибок.**
  Особенно опасна практика подбора «оптимального» cut-off, приводящая к росту ошибок первого рода.
- **Существуют обоснованные случаи для использования категорий — когда они основаны на клинических рекомендациях, диагностических порогах или заранее заданной структуре исследования.**
  Но даже в этих случаях желательно сохранять переменную в непрерывном виде на этапе моделирования и использовать категории только для интерпретации.

Источники

- Altman DG, Royston P. The cost of dichotomising continuous variables. BMJ. 2006 May 6;332(7549):1080. doi: 10.1136/bmj.332.7549.1080
- Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. Stat Med. 2006 Jan 15;25(1):127-41. doi: 10.1002/sim.2331.
- Bennette, C., Vickers, A. Against quantiles: categorization of continuous variables in epidemiologic research, and its discontents. *BMC Med Res Methodol* **12**, 21 (2012). https://doi.org/10.1186/1471-2288-12-21

bioinf.institute/hack2025

https://bioinf.me/education/workshops/stat



Институт биоинформатики в социальных сетях

Разрушители статистических мифов: bioinf.me/stat_myths

Чат по биостатистике и R: https://t.me/chat_biostat_R

По всем вопросам: biostat@bioinf.me

Сайт Института: bioinf.me

Институт в VK: vk.com/bioinf

Телеграм-канал Института: t.me/bioinforussia

Чат про образование и карьеру: t.me/bioinf_career

YouTube-канал: www.youtube.com/bioinforussia