

# Эксперименты и математическая статистика

# Статистические критерии

8	3.3. Критерии проверки экспоненциальности распределения . . . . .	273
	3.3.1. Критерий Шапиро-Уилка (279). 3.3.2. Критерий типа Колмогорова-Смирнова (282). 3.3.3. Критерий типа Смирнова-Крамера-фон Мизеса для цензурированных данных (286). 3.3.4. Критерий Фроцини (288). 3.3.5. Корреляционный критерий экспоненциальности (288). 3.3.6. Регрессионный критерий Брейна-Шапиро (290). 3.3.7. Критерий Кимбера-Мичела (292).	
	3.3.8. Критерий Фишера (293). 3.3.9. Критерий Бартлетта-Морана (294).	
	3.3.10. Критерий Климко-Антла-Радемакера-Рокетта (294). 3.3.11. Критерий Холлендера-Прощана (295). 3.3.12. Критерий Кочара (298). 3.3.13. Критерий Эпса-Палли-Черго-Уэлча (299). 3.3.14. Критерий Бергмана (301).	
	3.3.15. Критерий Шермана (303). 3.3.16. Критерий наибольшего интервала (304). 3.3.17. Критерий Хартли (305). 3.3.18. Критерий показательных методов (305). 3.3.19. Ранговый критерий независимости интервалов (306).	
	3.3.20. Критерии, основанные на трансформации экспоненциального распределения в равномерное (308). 3.3.20.1. Критерий $\hat{U}$ (308). 3.3.20.2. Критерий $\hat{U}$ (309). 3.3.20.3. Критерий Гринвуда (309). 3.3.21. Критерий Манн-Фертига-Шуера для распределения Вейбулла (311). 3.3.22. Критерий Дешпанде (316). 3.3.23. Критерий Лоулесса (317).	
319	3.4. Критерии согласия для равномерного распределения . . . . .	319
	3.4.1. Критерий Шермана (319). 3.4.2. Критерий Морана (320). 3.4.3. Критерий Ченга-Смирнига (322). 3.4.4. Критерий Саркади-Косика (323). 3.4.5. Энтропийный критерий Дудевича-ван дер Мюлена (324). 3.4.6. Критерий Хегзи-Грина (326). 3.4.7. Критерий Янга (328). 3.4.8. Критерии типа Колмогорова-Смирнова (330). 3.4.9. Критерий Фроцини (331). 3.4.10. Критерий Гринвуда-Кэнберри-Миллера (332). 3.4.11. „Сглаженный“ критерий Неймана-Барттона (333).	
336	3.5. Критерии симметрии . . . . .	336
	3.5.1. „Быстрый“ критерий Кенуя (336). 3.5.2. Критерий симметрии Смирнова (337). 3.5.3. Знаковый критерий симметрии (337). 3.5.4. Одновыборочный критерий Вилкоксона (339). 3.5.5. Критерий Антилла-Керстинга-Цуккини (340). 3.5.6. Критерий Бхатачарьи-Гаствирта-Райта (модифицированный критерий Вилкоксона) (342). 3.5.7. Критерий Финча (344). 3.5.8. Критерий Босса (345). 3.5.9. Критерий Гупты (348). 3.5.10. Критерий Фрезера (350).	
352	3.6. Подбор кривых распределения вероятностей по экспериментальным данным . . . . .	352
	3.6.1. Кривые распределения Джонсона (352). 3.6.1.1. Семейство распределений $S_L$ Джонсона (353). 3.6.1.2. Семейство распределений $S_B$ Джонсона (355). 3.6.1.3. Семейство распределений $S_U$ Джонсона (357). 3.6.2. Кривые распределений Пирсона (368). 3.6.2.1. Кривые Пирсона типа I (369). 3.6.2.2. Кривые Пирсона типа II (375). 3.6.2.3. Кривые Пирсона типа III (377). 3.6.2.4. Кривые Пирсона типа IV (378). 3.6.2.5. Кривые Пирсона типа V (380). 3.6.2.6. Кривые Пирсона типа VI (381). 3.6.2.7. Кривые Пирсона типа VII (382). 3.6.3. Разложение теоретических распределений (384). 3.6.4. Метод вкладов (385).	
388	Глава 4. Проверка гипотез о значениях параметров распределений . . . . .	388
	4.1. Сравнение параметров распределений . . . . .	389
	4.1.1. Сравнение параметров нормальных распределений (389). 4.1.1.1. Сравнение двух средних значений (389). 4.1.1.1.1. Сравнение при известных дисперсиях $\sigma_1^2$ и $\sigma_2^2$ (389). 4.1.1.1.2. Сравнение при неизвестных равных дисперсиях (390). 4.1.1.1.3. Сравнение при неизвестных неравных дисперсиях (391). 4.1.1.1.3.1. Критерий Кохрана-Кокса (391). 4.1.1.1.3.2. Критерий Сатервайта (391). 4.1.1.1.3.3. Критерий Уэлча (392). 4.1.1.1.4. Модифицированный критерий Стьюдента (392). 4.1.1.1.5. Парный $t$ -критерий сравнения средних (393). 4.1.1.1.6. Критерий Уолда (393).	

451	4.1.1.1.7. Двухступенчатый двухвыборочный медианный критерий Волфа (395). 4.1.1.1.8. $F$ -критерий для сравнения двух средних с одинаковыми дисперсиями (396). 4.1.1.2. Сравнение нескольких ( $k > 2$ ) средних (397).	
	4.1.1.2.1. Модифицированный критерий Стьюдента (397). 4.1.1.2.2. Критерий „стъюдентизированного“ размаха (399). 4.1.1.2.3. Дисперсионный критерий (399). 4.1.1.2.4. Критерий Польсона (402). 4.1.1.2.5. Метод прямого сравнения (критерий Тьюки) (403). 4.1.1.2.6. Критерий „стъюдентизированного“ максимума (обобщенный критерий Тьюки) (405). 4.1.1.2.7. Критерий Шеффе (406). 4.1.1.2.8. Критерий Стьюдента-Ньюмена-Кейлса (407).	
	4.1.1.2.9. Критерий Дункана (408). 4.1.1.2.10. Критерий Линика-Уоллеса (408). 4.1.1.3. Сравнение двух дисперсий (412). 4.1.1.3.1. Критерий Фишера (412). 4.1.1.3.2. Критерий Романовского (413). 4.1.1.3.3. Критерий отношения размахов (414). 4.1.1.3.4. Критерий „стъюдентизированного“ размаха (415). 4.1.1.3.5. Критерий Аризено-Охты (415). 4.1.1.4. Сравнение нескольких ( $k > 2$ ) дисперсий (416). 4.1.1.4.1. Критерий Бартлетта (417). 4.1.1.4.2. Критерий Кохрана (418). 4.1.1.4.3. Критерий Неймана-Пирсона (критерий отношения правдоподобия) (419). 4.1.1.4.4. Критерий Бл исса-Кохрана-Тьюки (421).	
	4.1.1.4.5. Критерий Хартли (421). 4.1.1.4.6. Критерий Кэдуэлла-Лесли-Брауна (422). 4.1.1.4.7. Критерий Самиуддина (423). 4.1.2. Сравнение параметров экспоненциальных распределений (424). 4.1.2.1. Сравнение двух параметров (424). 4.1.2.1.1. Критерий Фишера (424). 4.1.2.1.2. Критерий Фишера при сравнении интенсивностей отказов ( $\lambda$ ) (425). 4.1.2.1.3. Двухвыборочный пуассоновский критерий (426). 4.1.2.1.4. Сравнение значения параметра с заданным (426). 4.1.2.2. Сравнение нескольких ( $k \geq 2$ ) параметров (429). 4.1.2.2.1. Критерий Дэвида (429). 4.1.2.2.2. Критерий максимального правдоподобия (430). 4.1.2.2.3. Критерий отношения правдоподобия (критерий Нагаренкера) (431). 4.1.2.2.4. Критерий Чена для двухпараметрических экспоненциальных распределений (432). 4.1.2.2.5. Комбинированный критерий Сингха (433). 4.1.3. Сравнение параметров биномиальных распределений (435). 4.1.3.1. Сравнение двух параметров (435). 4.1.3.2. Сравнение значения параметра с заданным (436). 4.1.3.3. Сравнение нескольких параметров ( $k \geq 2$ ) (437). 4.1.4. Последовательные методы проверки гипотез о значениях параметров распределений (последовательный анализ Вальда) (438). 4.1.4.1. Проверка гипотез о параметрах нормального распределения (439). 4.1.4.1.1. Проверка гипотезы о значении среднего (439). 4.1.4.1.2. Проверка гипотезы о значении дисперсии (446). 4.1.4.2. Проверка гипотезы о параметре экспоненциального распределения (447). 4.1.4.3. Проверка гипотезы о параметре биномиального распределения (449).	
451	4.2. Непараметрические (свободные от распределения) критерии однородности статистических данных . . . . .	451
	4.2.1. Непараметрические критерии сдвига (452). 4.2.1.1. Сравнение параметров сдвига двух совокупностей (452). 4.2.1.1.1. Быстрый (грубый) критерий Кенуя (452). 4.2.1.1.2. Ранговые критерии сдвига (453). 4.2.1.1.2.1. Быстрый (грубый) ранговый критерий (453). 4.2.1.1.2.2. Критерий Манна-Уитни-Вилкоксона (454). 4.2.1.1.2.3. Критерий Фишера-Йэтса-Терри-Гёфдинга (459). 4.2.1.1.2.4. Критерий Ван дер Вардена (460). 4.2.1.1.2.5. Медианный критерий (462). 4.2.1.1.2.6. Критерий Мостеллера (464). 4.2.1.1.2.7. Критерий Розенбаума (464). 4.2.1.1.2.8. Критерий Хаги (464). 4.2.1.1.2.9. Е-критерий (465). 4.2.1.2. Сравнение параметров сдвига нескольких ( $k > 2$ ) совокупностей (466). 4.2.1.2.1. Критерий Крускала-Уоллеса (466). 4.2.1.2.2. Критерий Гемени (469). 4.2.1.2.3. Критерий Вилкоксона-Вилкокса (471). 4.2.1.2.4. „Быстроый“ критерий Кенуя (473). 4.2.1.2.5. Критерий Фишера-Терри-Йэтса-Гёфдинга (473). 4.2.1.2.6. Критерий Ван дер Вардена (475). 4.2.1.2.7. Медианный критерий (475). 4.2.1.2.8. Критерий Хеттманспергера (476). 4.2.1.2.10. Критерий Мостеллера (477).	

# Параметрика

Берем в учет выборочный  
параметр

- Критерий Стьюдента
- Колмогоров-Смирнов
- ANOVA
- BOOTSTRAP

И тп

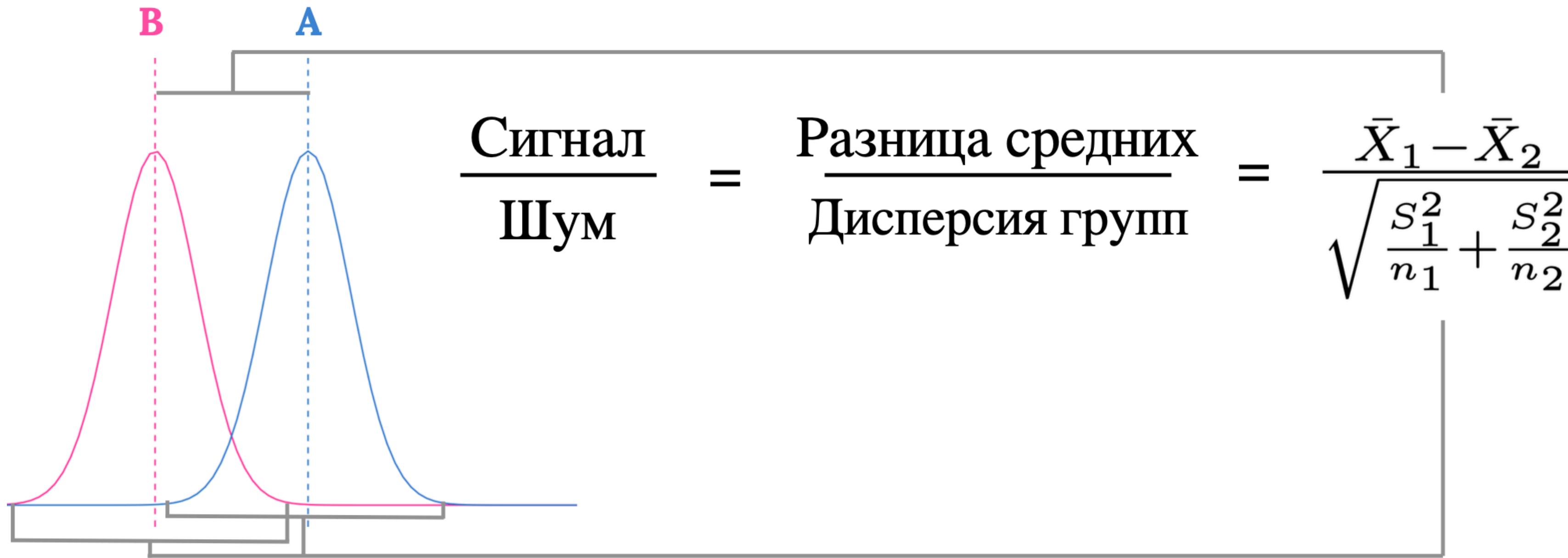
# Непараметрика

Не берет в учет  
выборочный параметр

- Манн-Уитни
- Q критерий Кохрена
- Хи-квадрат
- BOOTSTRAP

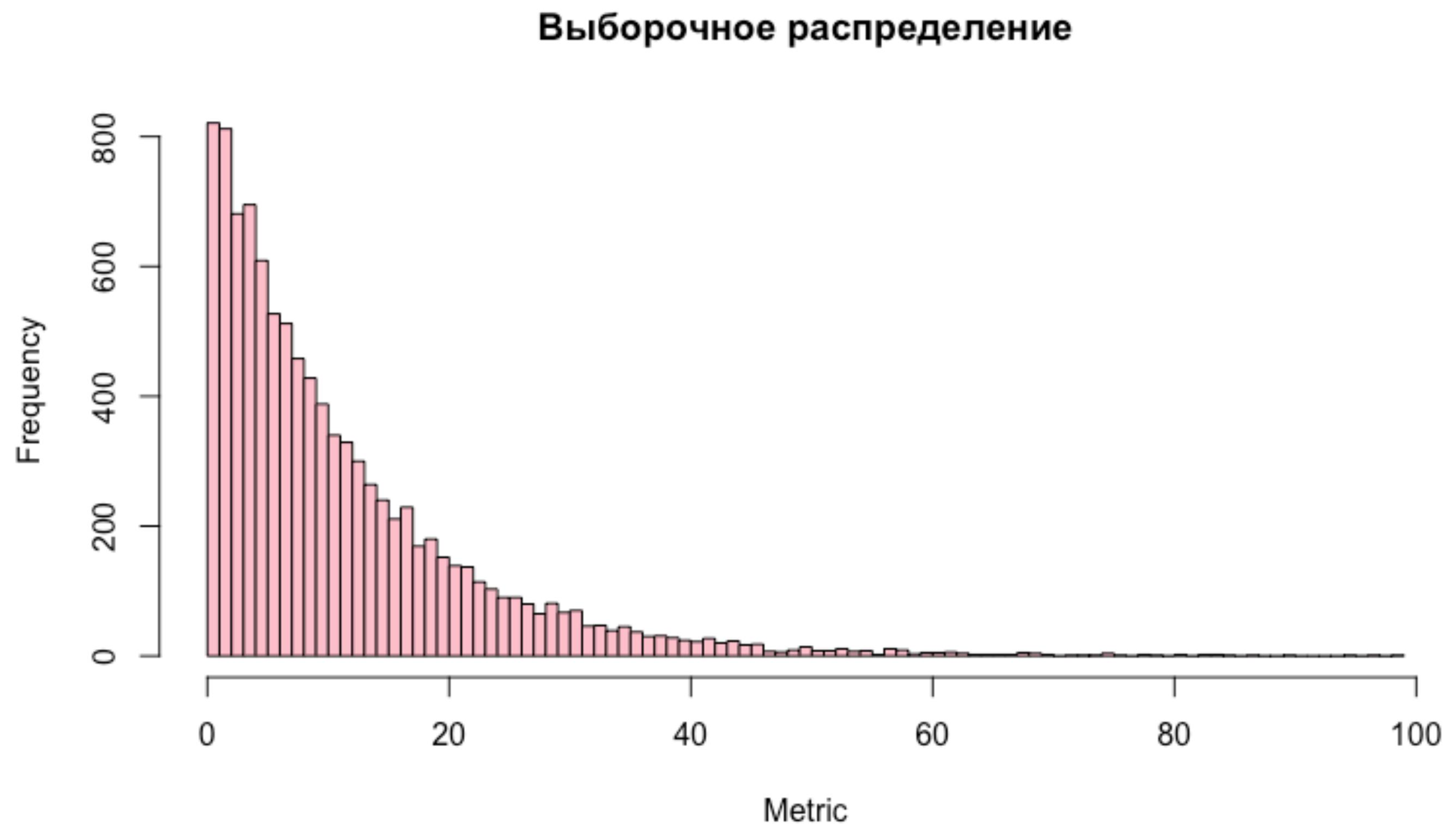
И тп

# Параметрика



$$\nu = \frac{\left( \frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} \right)^2}{\frac{S_1^4}{n_1^2(n_1-1)} + \frac{S_2^4}{n_2^2(n_2-1)}}$$

# Параметрика



- Чувствительна к большой дисперсии
- Чувствительна к дисбалансу

1. Мы «склеиваем» таблицы для двух групп и отсортируем их по возрастанию

Группа	A	A	B	B	A	A	A	B	A	A	B	B	B	A	B	A	A	B	B	
Чек	0	0	0	0	100	200	300	400	400	500	500	500	700	700	1000	1100	1200	1200	1400	3300



2. Присвоим каждому значению свой ранг. Сначала, пусть это будет порядковый номер

Группа	A	A	B	B	A	A	A	B	A	A	B	B	B	A	B	A	A	B	B	
Чек в рублях	0	0	0	0	100	200	300	400	400	500	500	500	700	700	1000	1100	1200	1200	1400	3300
Ранг	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

2. Присвоим каждому значению свой ранг. Сначала, пусть это будет порядковый номер

Группа	A	A	B	B	B	A	A	A	B	A	A	B	B	B	A	B	A	A	B	B
Чек в дублях	0	0	0	0	100	200	300	400	400	500	500	500	700	700	1000	1100	1200	1200	1400	3300
Ранг	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20



3. Но можно заметить, что некоторые значения повторяются, имея при этом разные ранги. Очень важно это поправить, взяв среднее между рангами для повторяющихся значений.

Группа	A	A	B	B	B	A	A	A	B	A	A	B	B	B	A	B	A	A	B	B
Чек в дублях	0	0	0	0	100	200	300	400	400	500	500	500	700	700	1000	1100	1200	1200	1400	3300
Ранг	2.5	2.5	2.5	2.5	5	6	7	8.5	8.5	11	11	11	13.5	13.5	15	16	17.5	17.5	19	20

Группа	A	A	B	B	A	A	A	B	A	A	B	B	B	A	B	A	A	B	B	
Чек в рублях	0	0	0	0	100	200	300	400	400	500	500	500	700	700	1000	1100	1200	1200	1400	3300
Ранг	2.5	2.5	2.5	2.5	5	6	7	8.5	8.5	11	11	11	13.5	13.5	15	16	17.5	17.5	19	20

4. А теперь просуммируем ранги погруппно, получаем  $R_1 = 98.5$  и  $R_2 = 111.5$ , где  $R_1$  это вариант A, и  $R_2 = B$ , соответственно. Подставляем эти значения в формулу

$$U_1 = n_1 \times n_2 + n_1 \times (n_1 + 1) \div 2 - R_1$$

$$U_2 = n_1 \times n_2 + n_2 \times (n_2 + 1) \div 2 - R_2$$

В итоге получаем  $U_1 = 56.5$  и  $U_2 = 43.5$ , берем меньший, т.е. 43.5

5. Ищем в таблице критических значений (как мы делали до этого с z и t критериями) соответствующее нашему и видим  $U_{\text{критическое}} = 23$ .

$U_{\text{критическое}} < U$ , отвергаем нулевую гипотезу о равенстве двух распределений

# BOOTSTRAP

# BOOTSTRAP

Параметрика

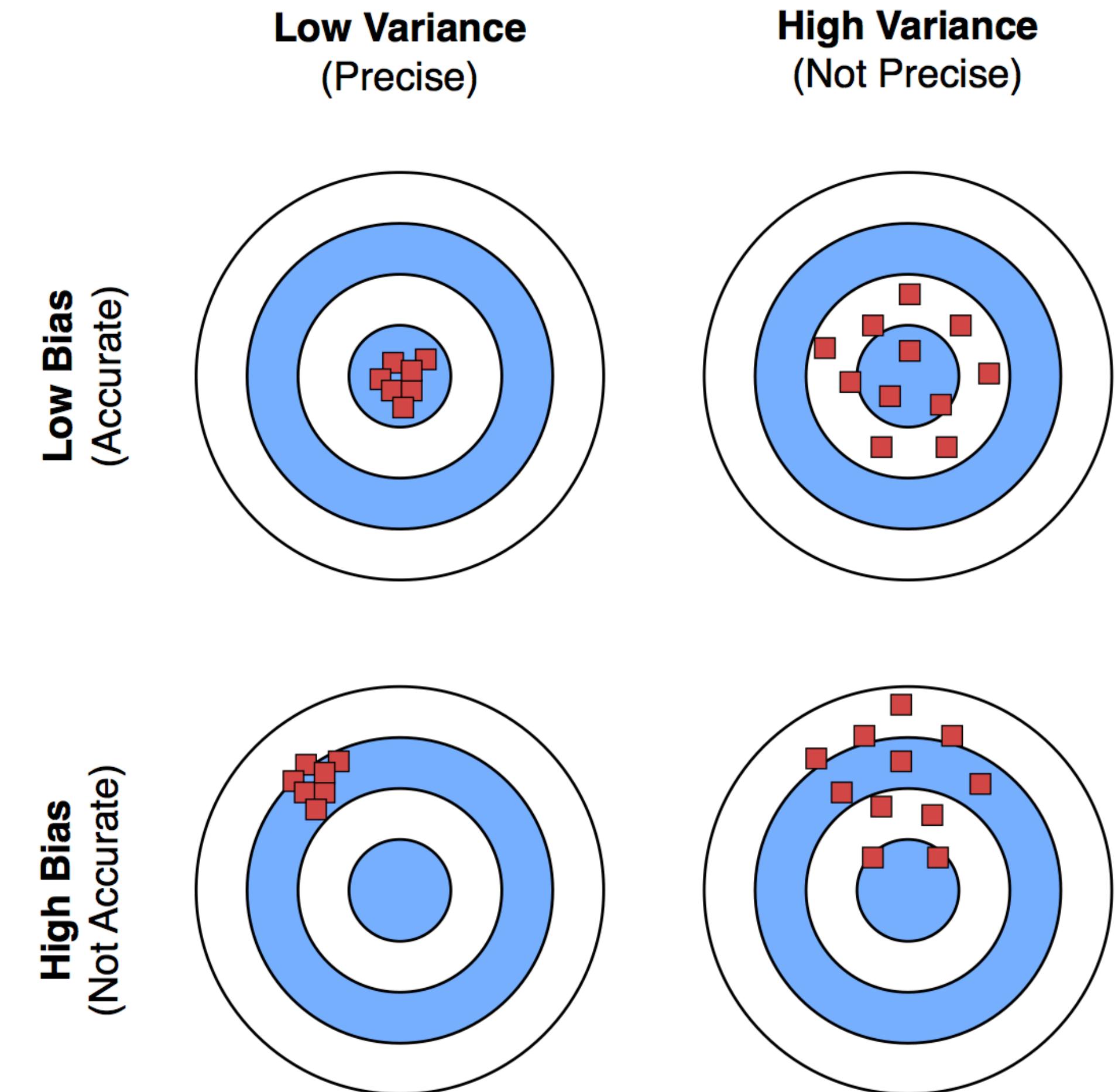
- $<\text{bias}$
- $>\text{D}$

Учитывает статистику популяции

Непараметрика

- $>\text{bias}$
- $<\text{D}$

Не учитывает статистику популяции



## Доверительные интервалы

1.  $[\bar{x} - \delta_{.1}^*, \bar{x} - \delta_{.9}^*]$

Интервал будет уже

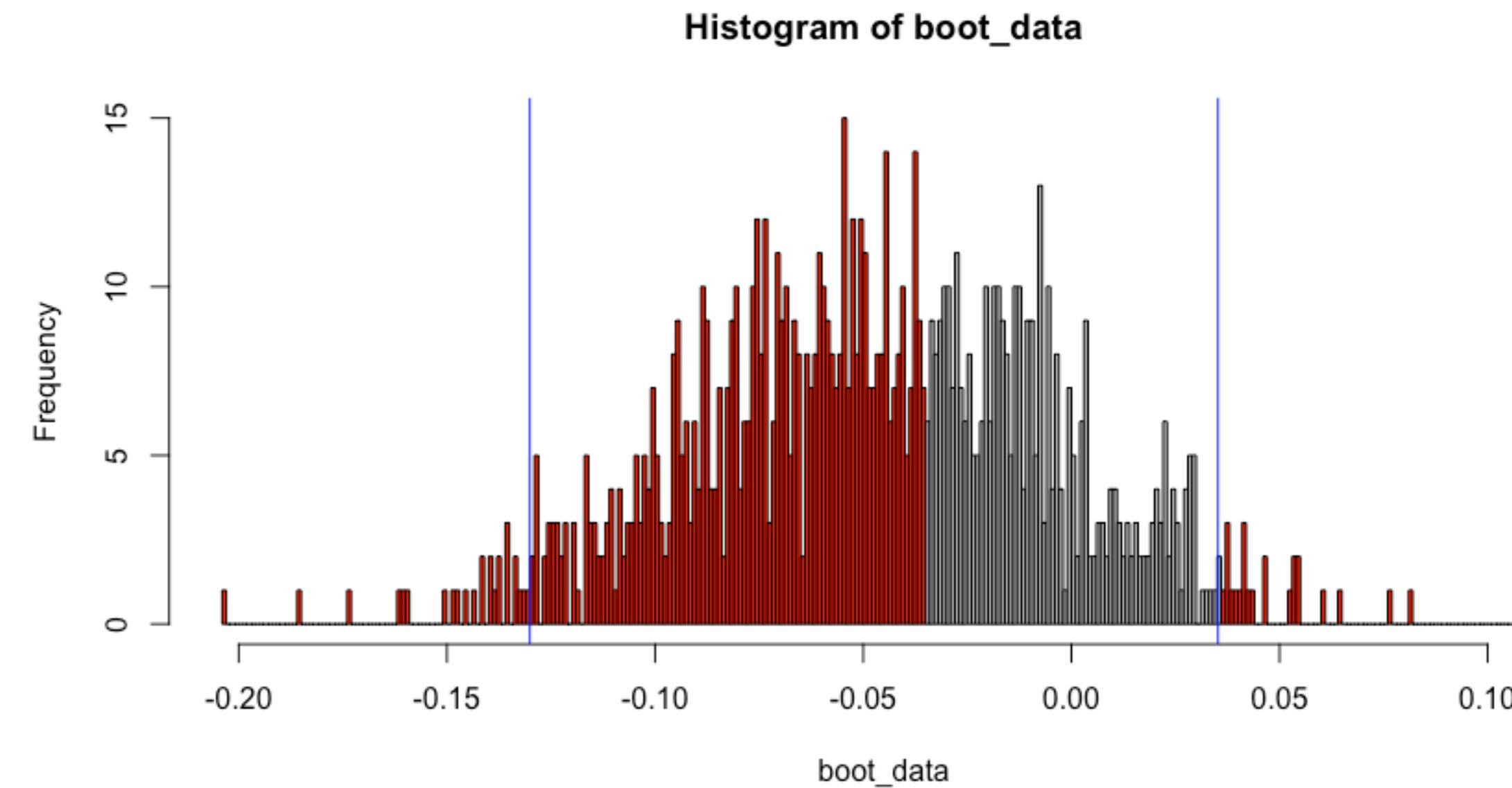
2.  $\delta_{.1}^* \quad \bar{x} \quad \delta_{.9}^*$

Интервал будет шире

## pValue для bootstrap

```
1.pval_1 = min(sum(S0>= 0,S0<=0 ))*2/N
```

```
2.pval_2 = min( (1+sum(S0>= 0,S0<=0 ))) *2/(N+1)
```



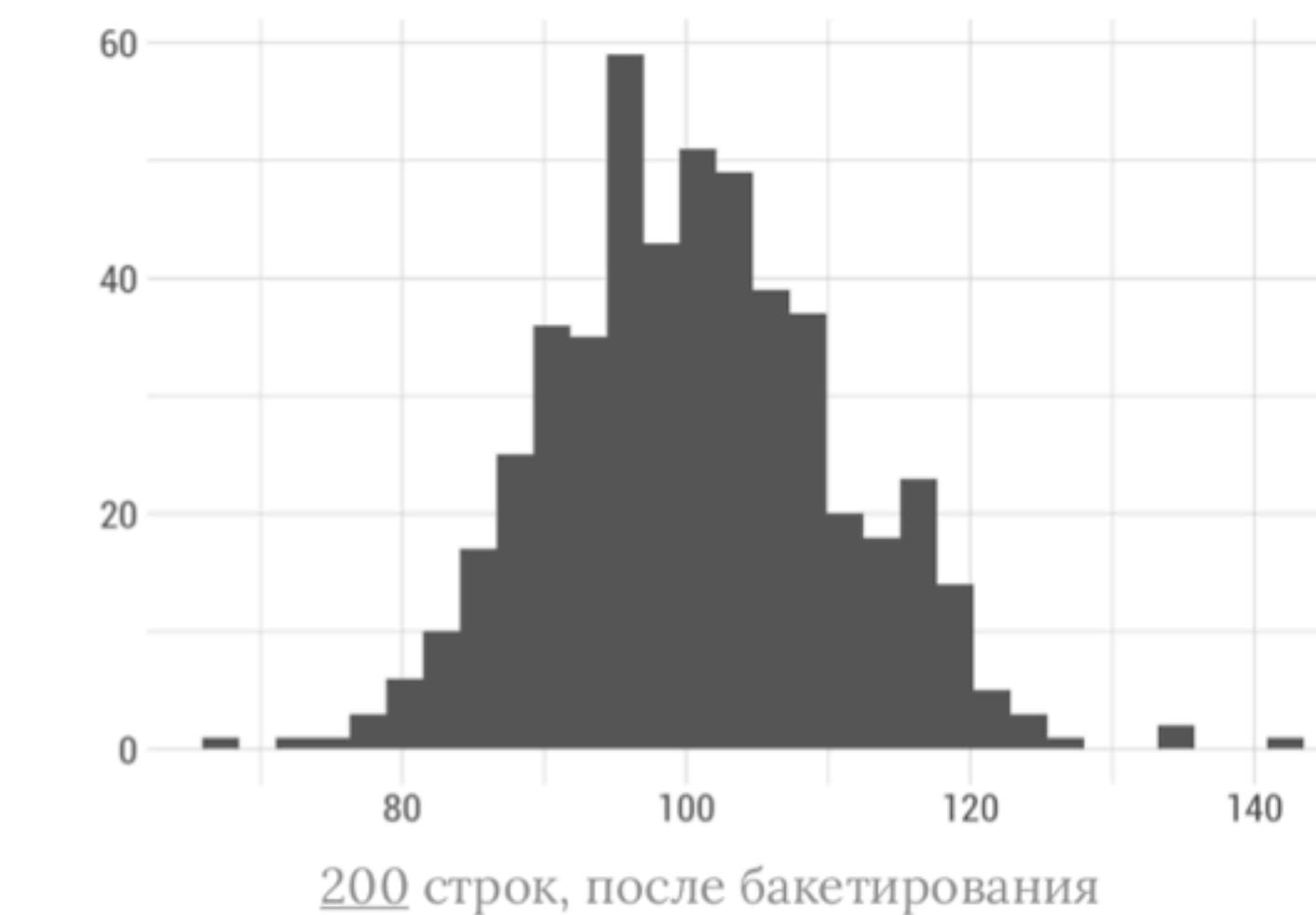
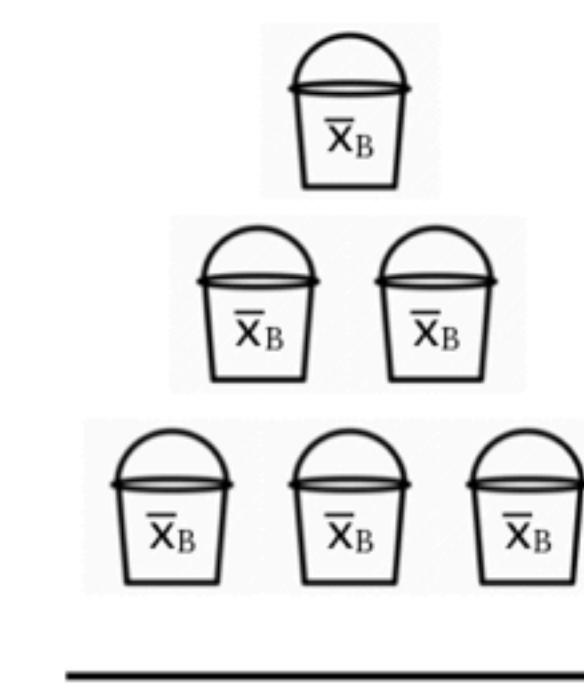
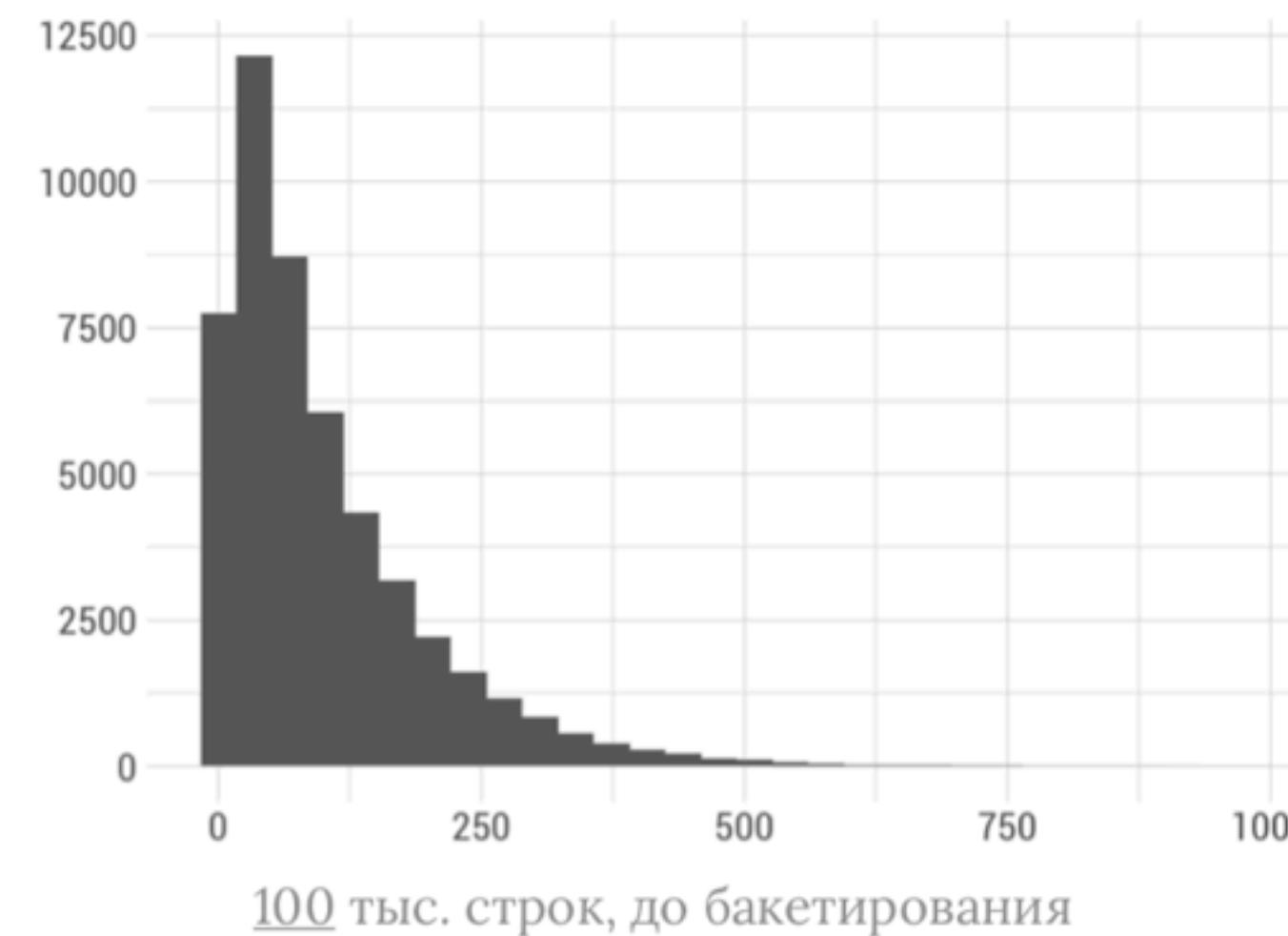
pValue можно выразить как площадь перекрытия под кривой распределения статистики

## Что стоит запомнить

- Bootstrap **НЕ** универсален
- Bootstrap **НЕ** сокращает дисперсию
- Bootstrap **НЕ** нормализует выборки
- Bootstrap sample **НЕЛЬЗЯ** использовать в качестве выборки для критерия
- Bootstrap **может быть быстрым** даже на больших данных в условиях оптимизации

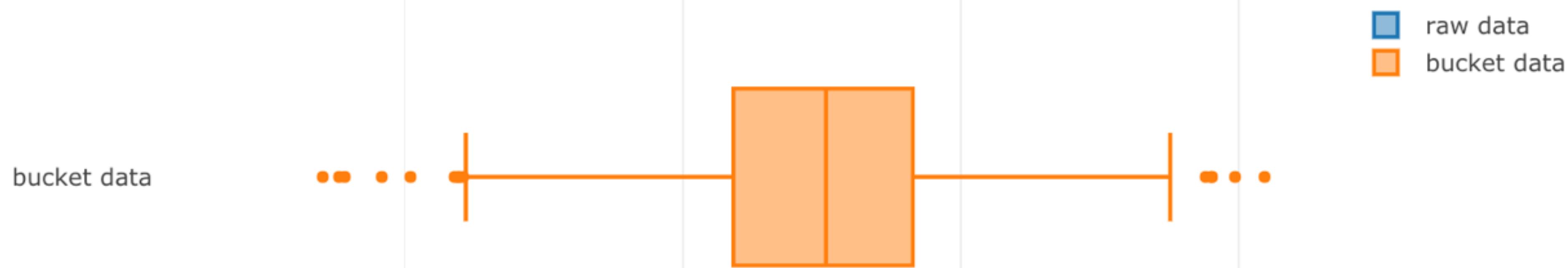
# ЦПТ + bootstrap

1. Рандомно присваиваем номер группы от  $B_1$  до  $B_n$  (где  $B_n$  = оптимальное число групп, которые мы усредним, напр., 500)
2. Усредняем значения в каждой группе
3. Из усредненных значений получаем распределение близкое к нормальному

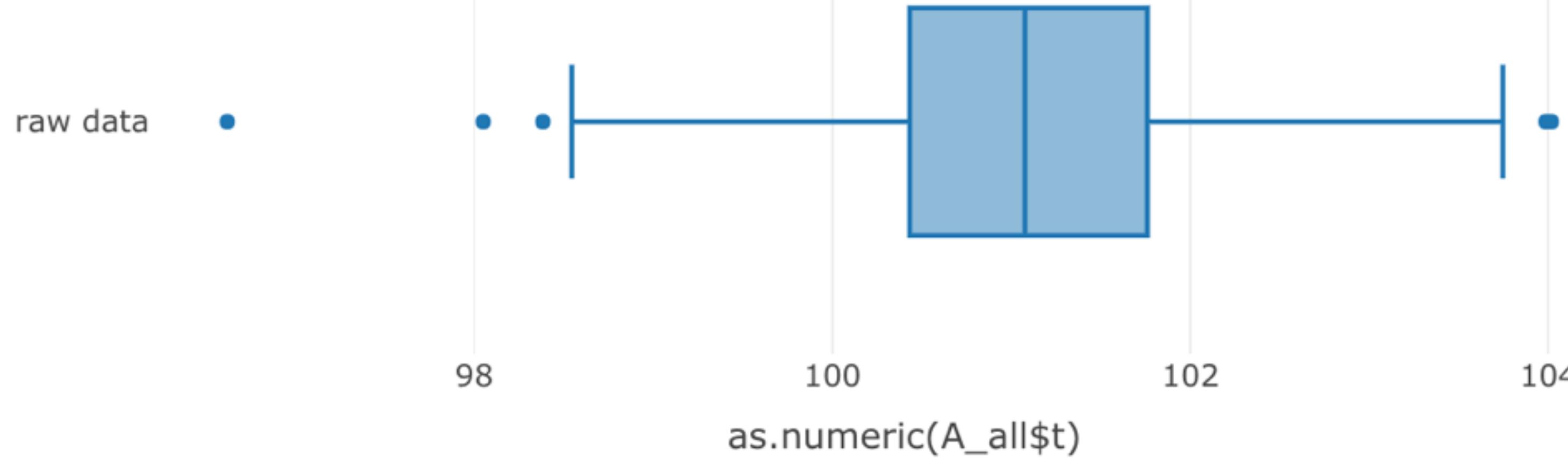


Быстрее без потери качества

**100 бакетов**



**10000 пользователей**



Но требует небольшой коррекции

# Дизайн эксперимента

## **Чувствительность**

Способность увидеть значимые различия в метрике там, где они на самом деле должны быть называется **чувствительностью**

Высокая чувствительность метрики позволяет:

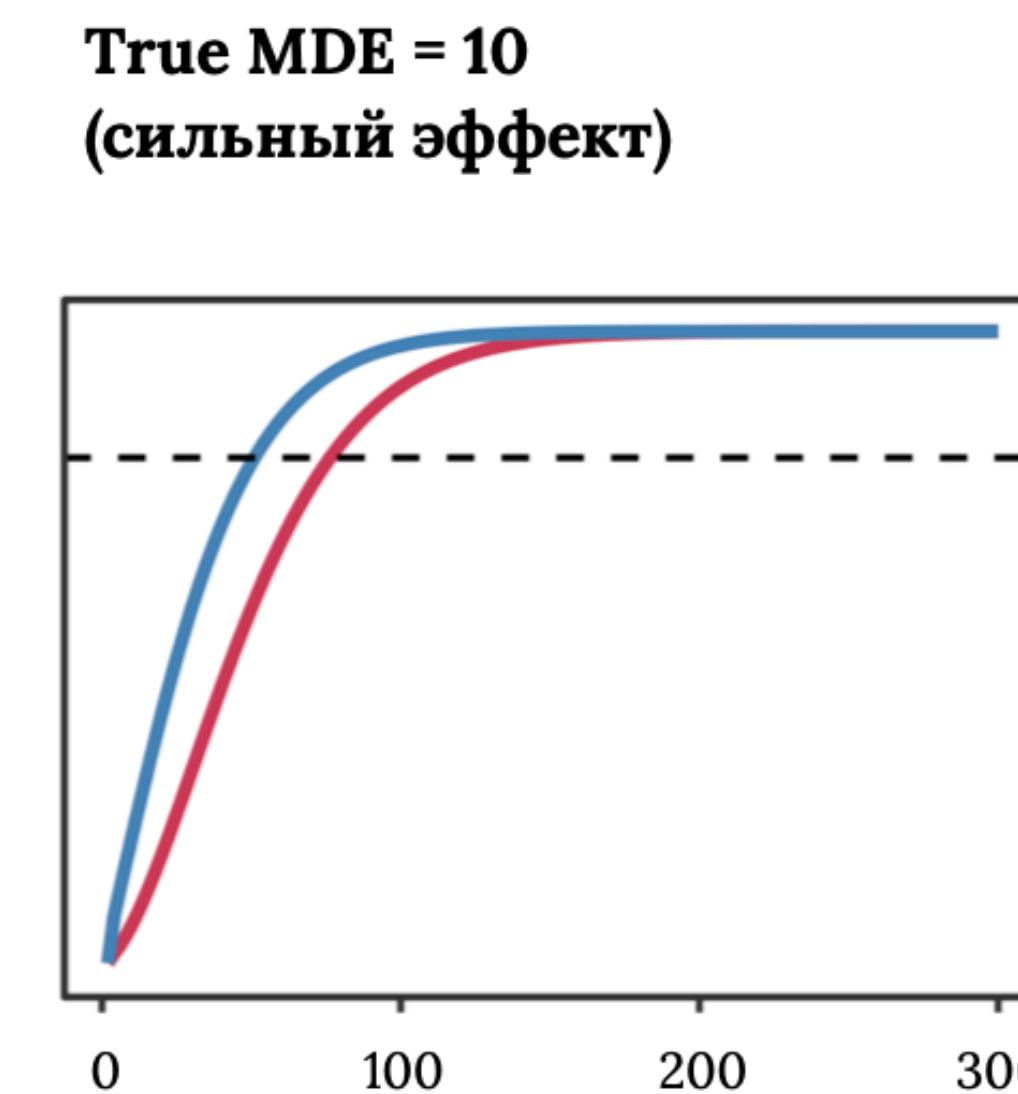
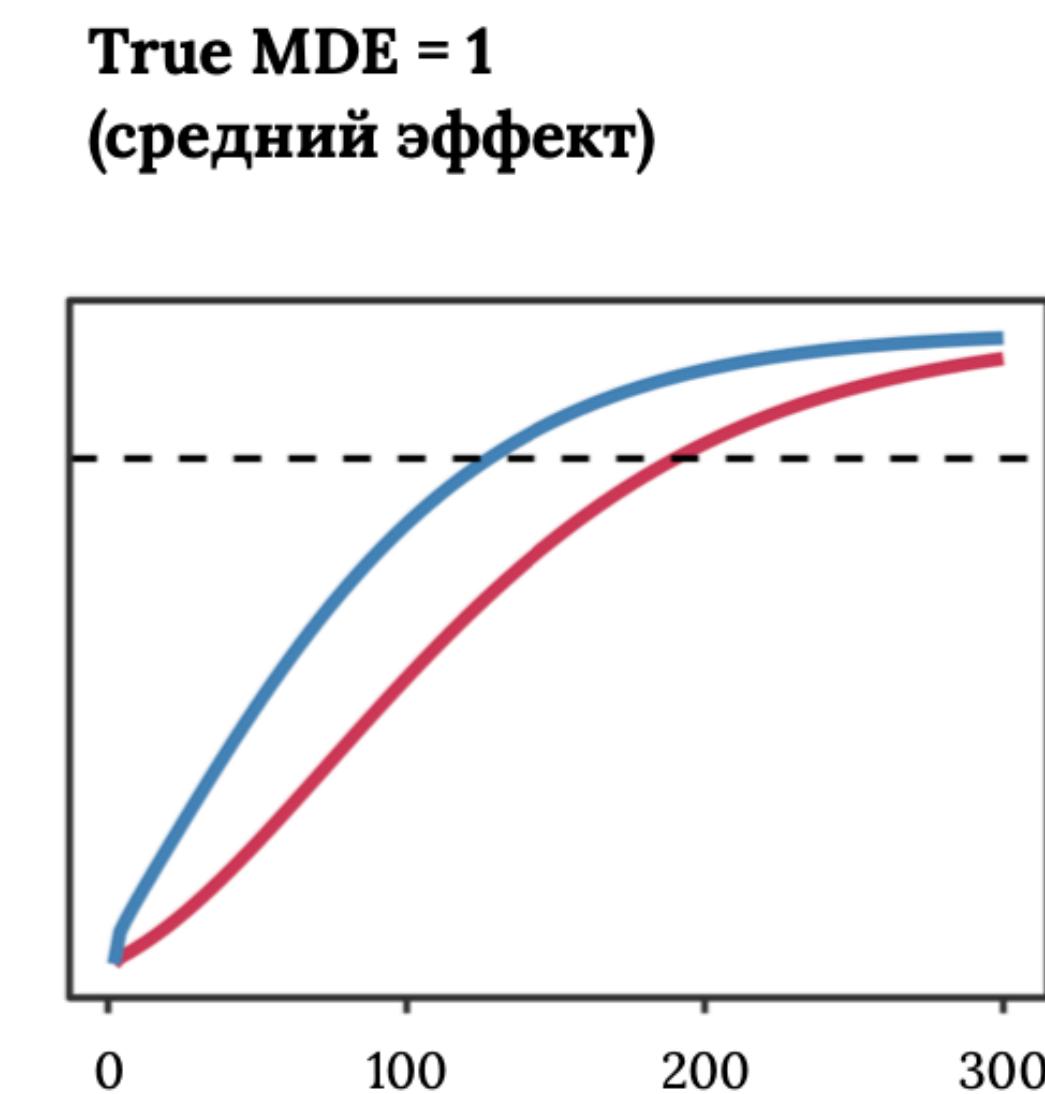
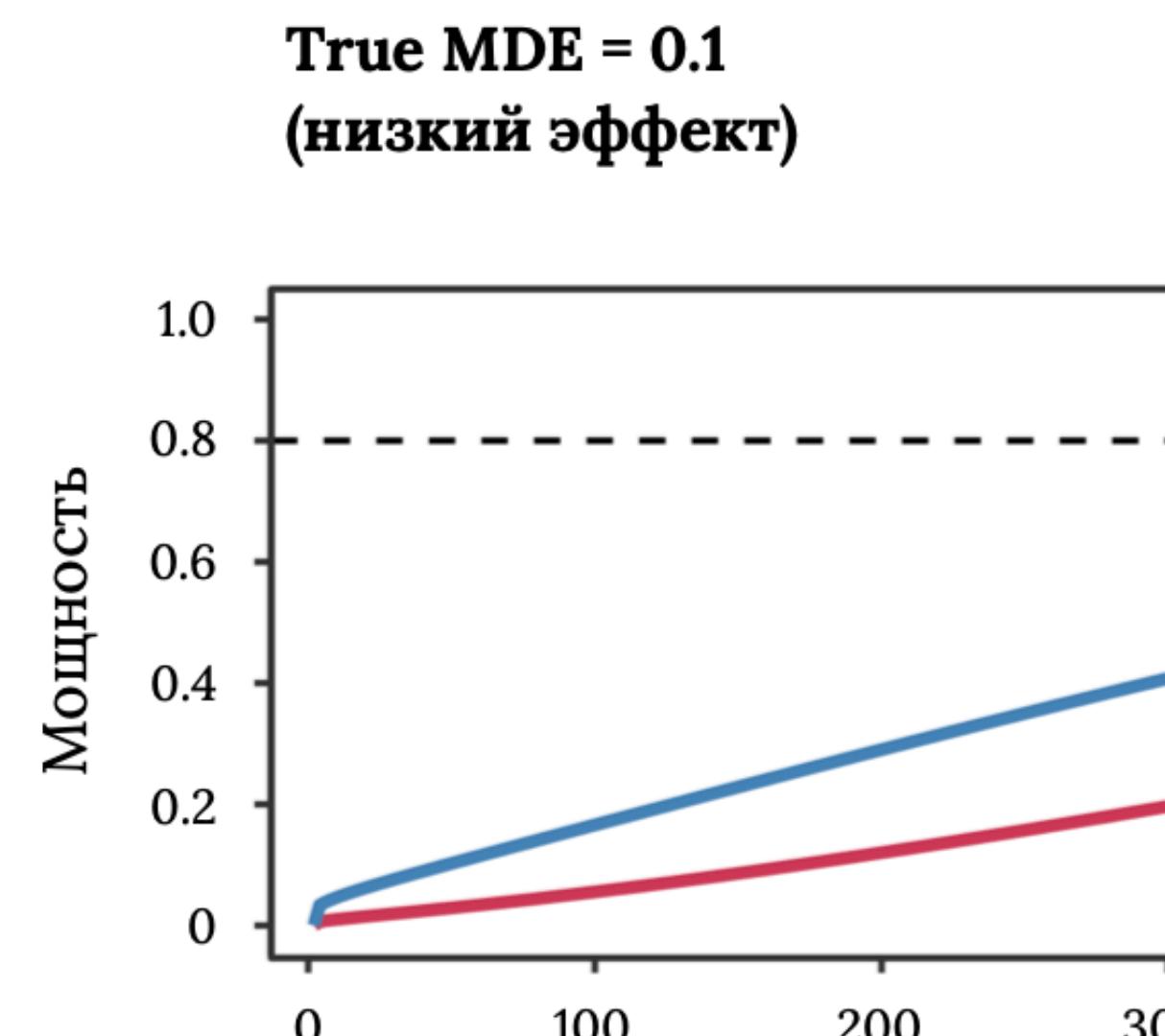
- видеть достаточно маленькие изменения
- или использовать меньшее количество пользователей

## **Чувствительность зависит от мощности**

$$\text{Мощность} = \frac{1 - \text{вероятность совершения}}{\text{статистической ошибки второго типа}}$$

Мощность теста – важнейший параметр при расчете объема выборки. Мощность можно также понимать как вероятность обнаружения существующей закономерности

Мощность теста увеличивается по мере увеличения объема выборки



—  $\alpha = 0.01$   
—  $\alpha = 0.05$

## **Сколько пользователей нужно для эксперимента? 1/2**

Fixed horizon

1. Оценка по MDE
2. Оценка по выборке
3. Оценка по выборке и MDE

## Симуляция A/B

CR A: 0,25

CR B: 0,26

MDE: 0,15

Power: 0,80

Alpha = 0,05

3. p-value до сих пор блуждает.

Истинный MDE точно ниже установленного при расчетах дизайна эксперимента, либо его вовсе нет (как в нашем случае)



## Итого

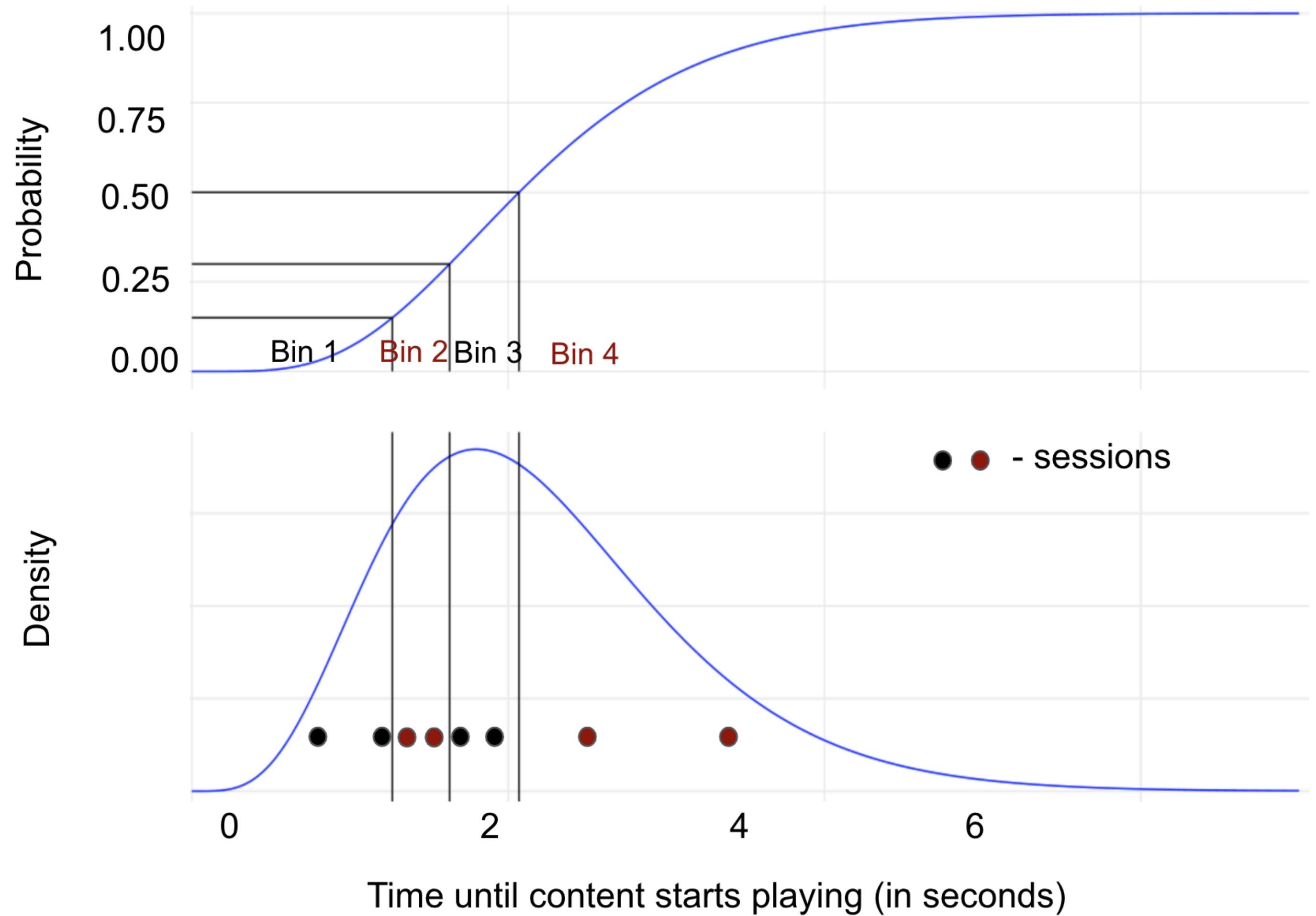
- Мощность считается определяющим параметром при расчете длительности эксперимента
- Fixed-horizon является классическим подходом к расчету необходимого времени на эксперимент, в то же время не самым эффективным для А/В-тестов
- Соблюдение процедур Sequential testing является более подходящем решением для мониторинга и отслеживания результатов эксперимента

# Управление чувствительностью метрик

Чувствительность метрики - это свойство метрики «прокрашиваться» под влиянием какого то изменения.

Что отвечает за чувствительность метрики?

# Изменение размерности без сокращения данных



# Про придется помнить про поправки

## Бонферони

$1 - (1 - \alpha)^m$ ; где :  $\alpha$  – альфа,  $m$  – сравнения

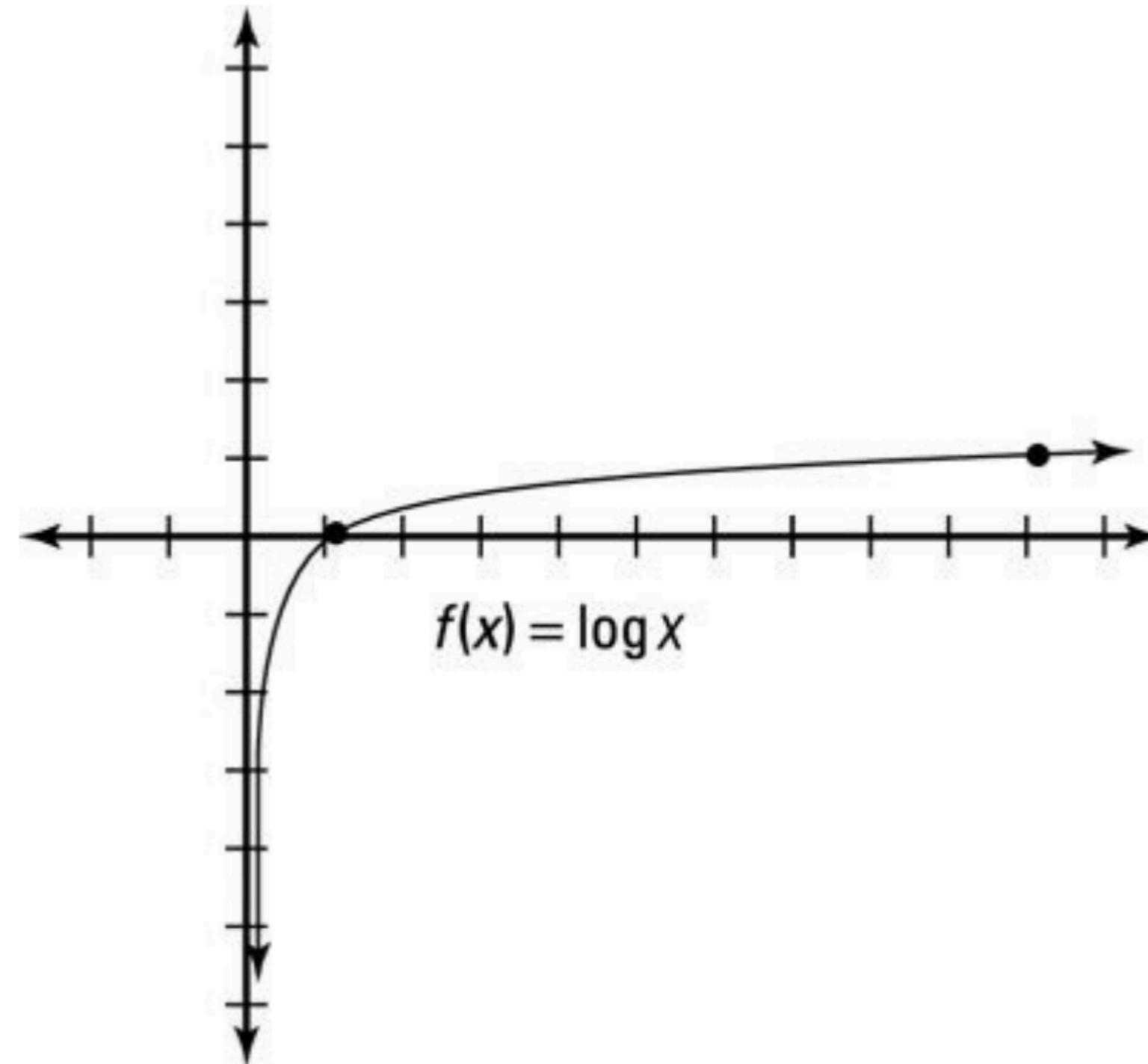
## Бенджамини-Хохберг

$$FDR = \mathbf{E} \left( \frac{V}{R} \right)$$

$$p_{(4)} = 0.0095 \leq (4/15)0.05 = 0.013$$

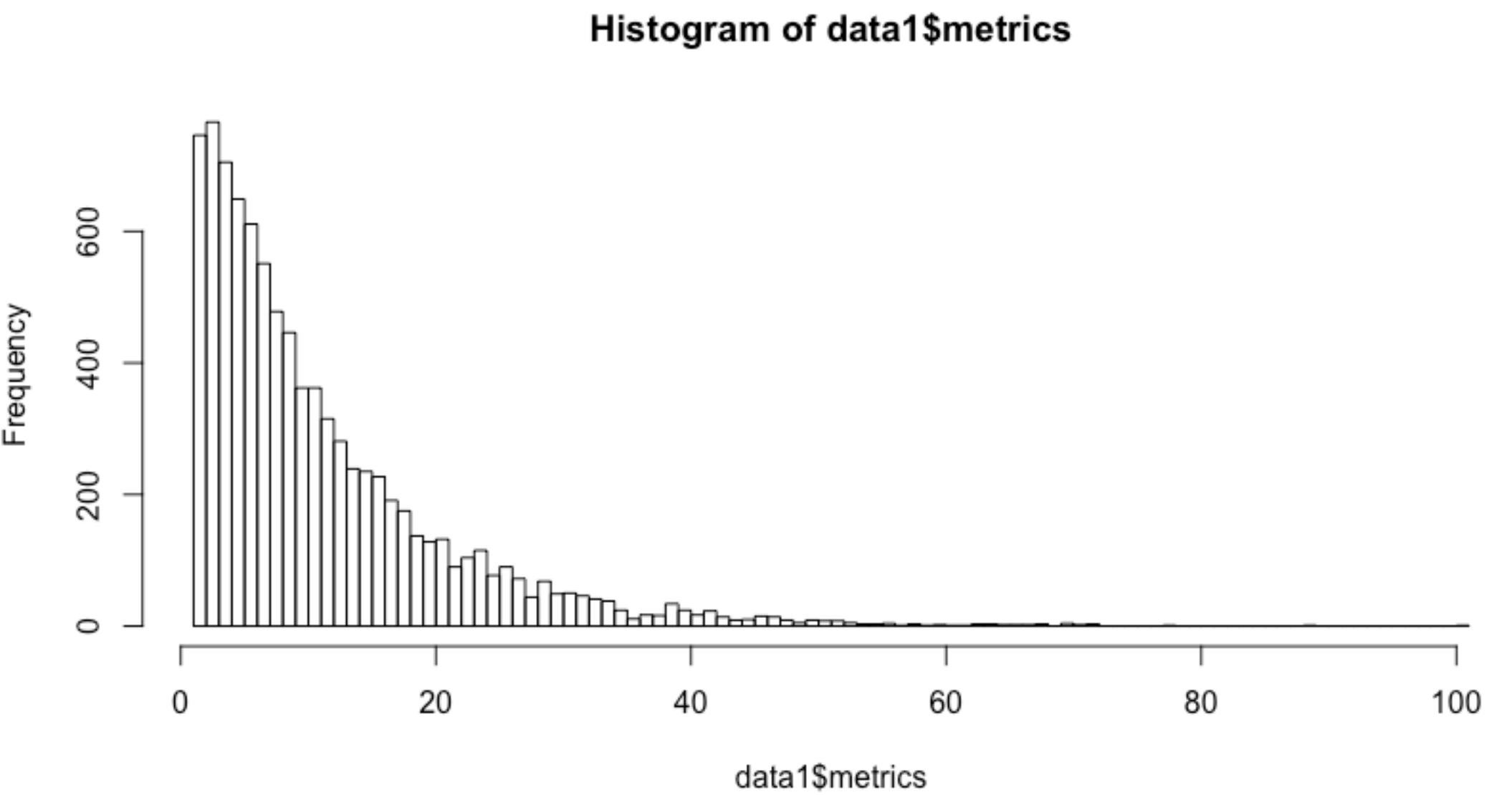
Множественная проверка требуется корректировки альфы, чтобы сократить вероятность ошибок.

# Трансформация - частный случай Бокса-Кокса

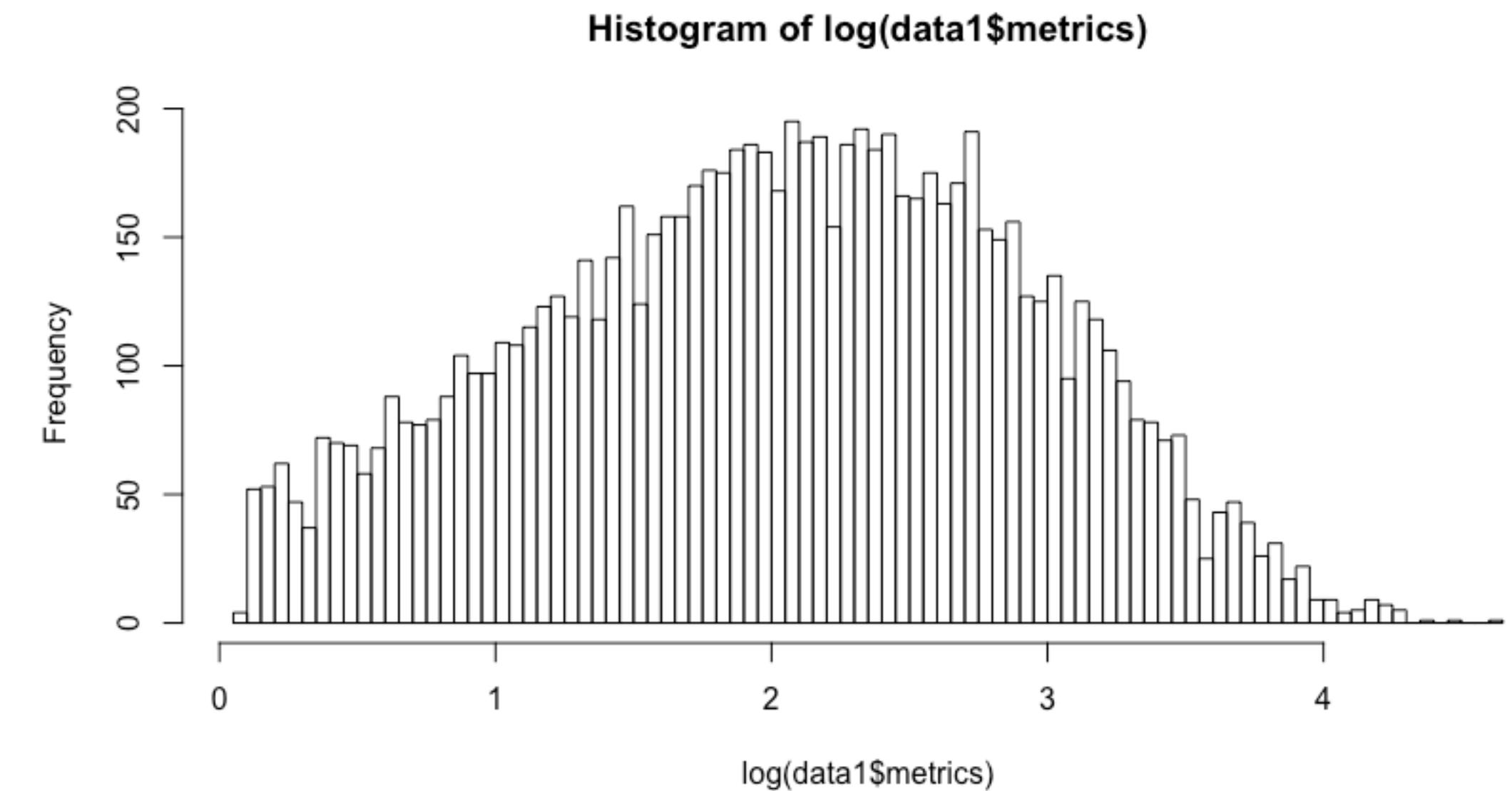


X	$\log(x)$
1	0
2	0.69
3	1.10
100	4.61
1000	6.91

# Трансформация - частный случай Бокса-Кокса

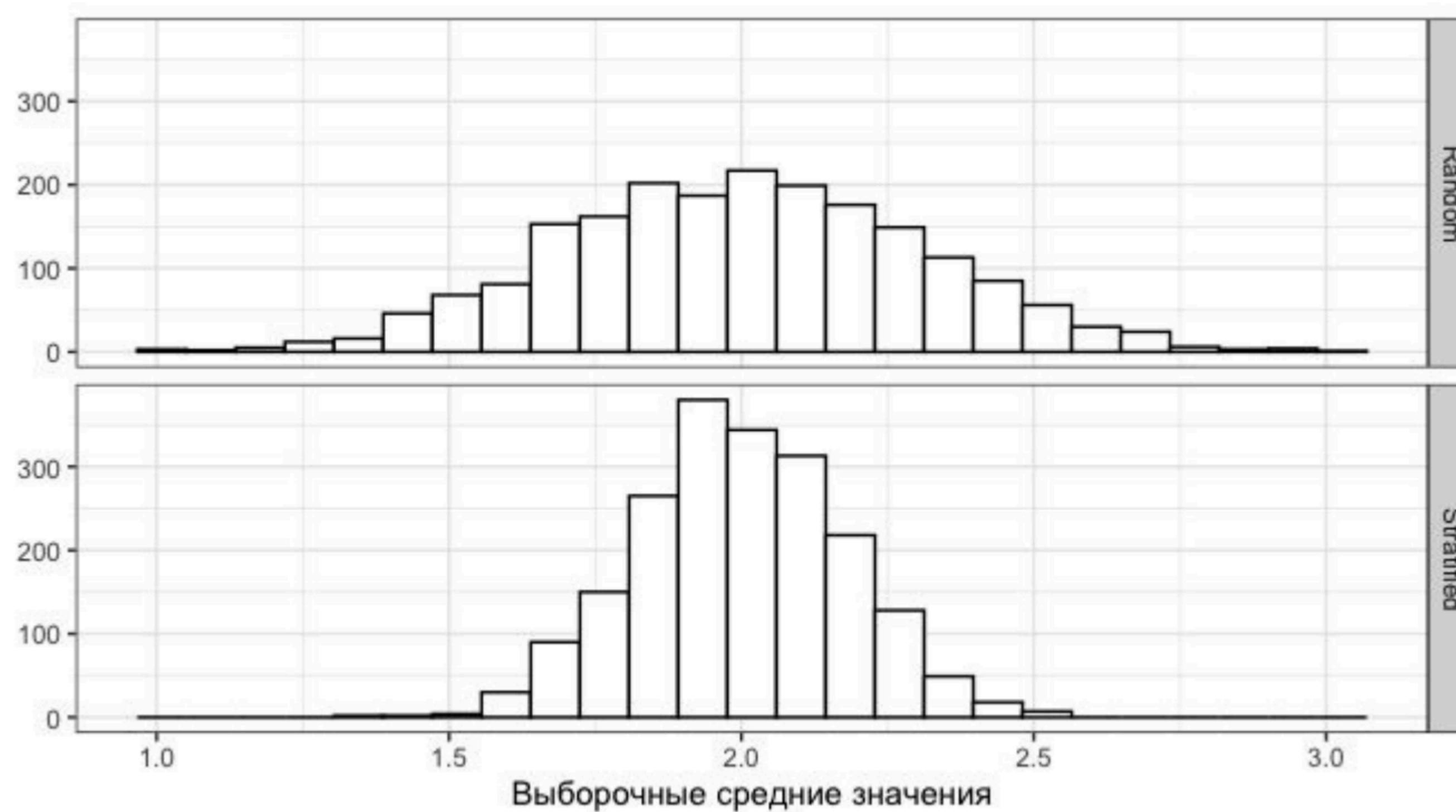


**X=11.05078**  
**D=97.73333**  
**SD=9.886017**  
**fsize=1.06060**



**X=2.035973**  
**D=0.7829451**  
**SD=0.8848418**  
**fsize=1.06585**

# Стратификация - пред и пост



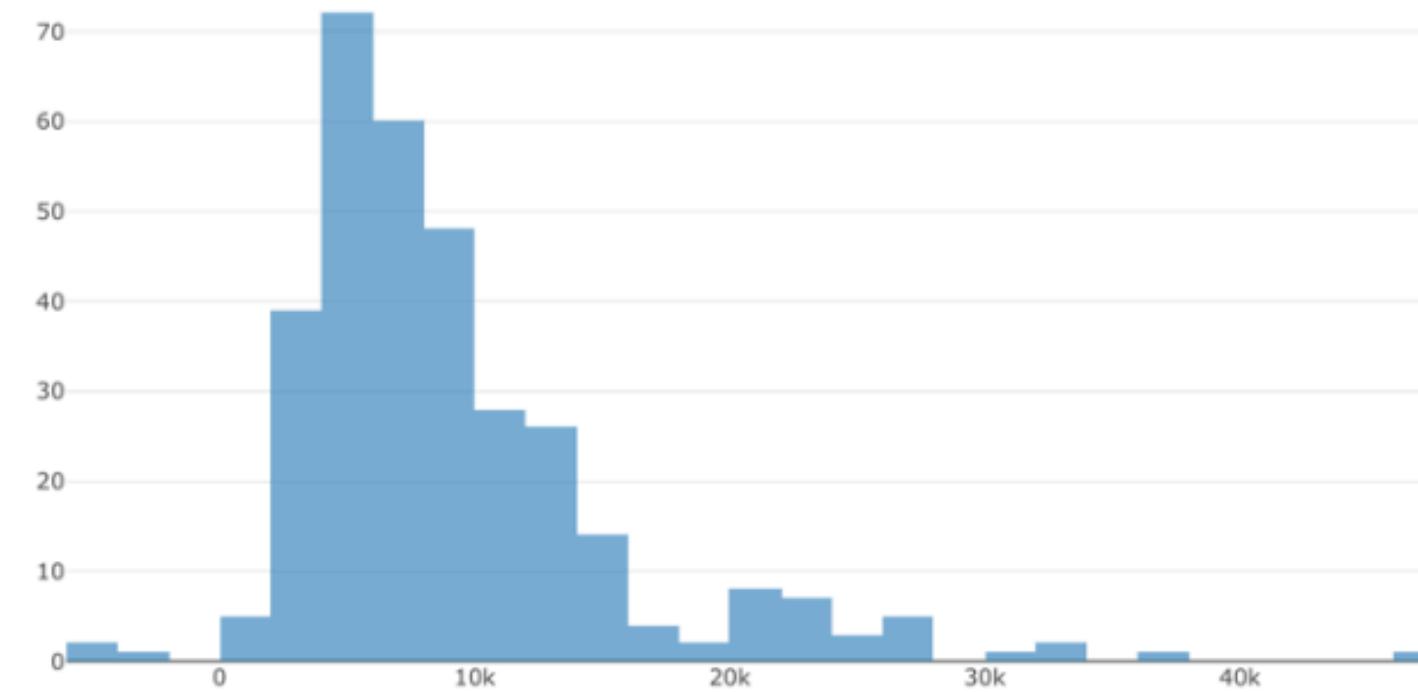
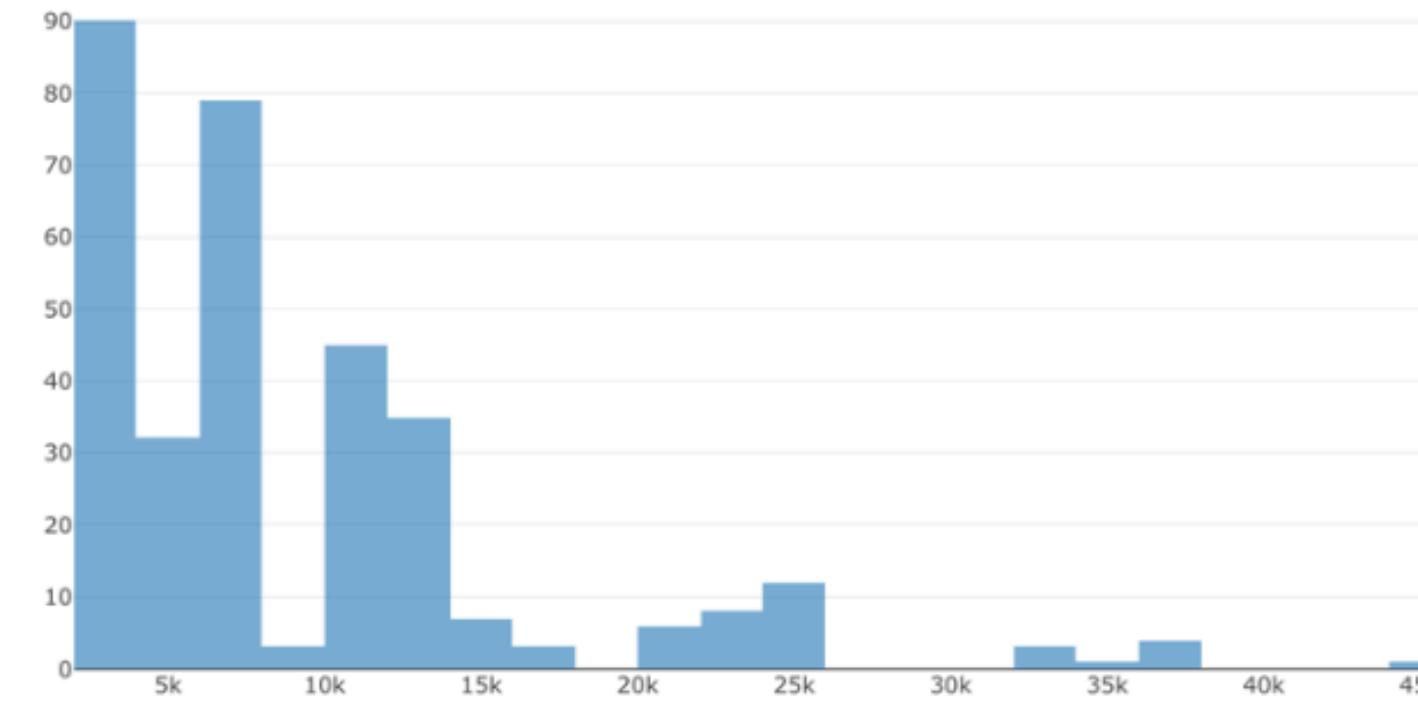
# Вычитание предсказания и базовые предикты

<b>Method</b>	<b>Effect</b>	<b>Standard error</b>	<b>P-value</b>
Before and after (SF)	2.9%	0.0377	0.4531
Simple regression	2.6%	0.0125	0.0802
Multivariate regression	3.0%	0.0067	0.0003

$$CUPED = metric - (covariate - \text{mean}(covariate)) * \theta$$

- **covariate** — метрика до эксперимента
- **metric** — метрика после эксперимента
- **theta** вычисляется как

$$\frac{\text{covariance}(metric, covariate)}{\text{variance}(covariate)}$$

**CUPED****До CUPED****Вариант**

как есть  
CUPED

**Среднее**

9 229  
9 229

**Дисперсия**

52633186  
42323576

**Стандартное отклонение**

7 254  
6 505