



ТИНЬКОФФ

# Математическая статистика





**ТИНЬКОФФ**

# Введение

# Регламент



**Участвуем в интерактивах**



**Задаём вопросы устно в конце блоков**



***Вы не только слушатели, но и участники***

# Знакомство

## Учусь и работаю

- Тинькофф Страхование, Middle продуктовый аналитик
- Мехмат, МГУ, 6 курс
- Математический институт при РАН, стажёр-исследователь

## Из интересного

- Всё ещё занимаюсь математикой
- Разрабатываю общий регламент по А/В тестам
- Team Lead на 32 часа



# **Зачем аналитику нужна статистика?**

# Статистика в анализе данных



## ML

- Алгоритмы обучения
- Алгоритмы интерпретации
- Выбор лучшей модели



## A/B тесты

- Проверка наличия эффекта
- Проверка однородности
- Время на проведение теста
- Анализ теста



## Мониторинг метрик

- Обнаружение отклонений
- Выявление причин отклонений
- Частота нотификаций



## Риски

- Убыточность продуктов
- Откладывание резервов
- NPV и LTV

# Новый продукт

## Задача

- Продажа нового продукта
- 2 варианта продажи
- Нужно выбрать лучший

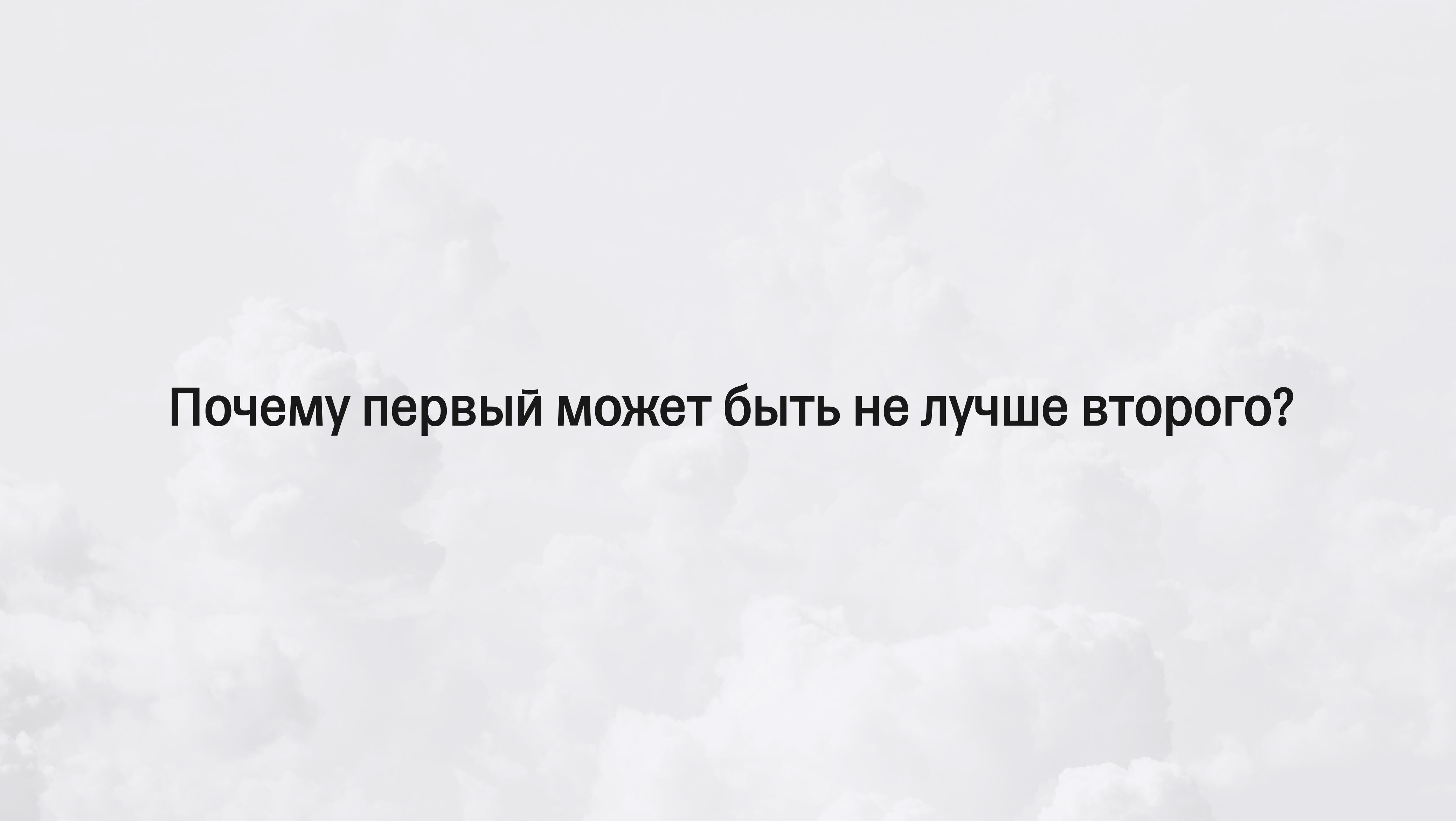
## Исходные данные

1. Предложили 100 клиентам, продали 30.
2. Предложили 1000 клиентам, продали 250.

## Решение

А какой лучше?

Первый, так как  $30\% > 25\%$ ?

The background of the slide features a soft, out-of-focus image of a cloudy sky, providing a calm and contemplative atmosphere.

**Почему первый может быть не лучше второго?**

# Сложность задачи

Мы даём прогноз на будущее,  
используя имеющиеся данные



**Клиенты были лучше?**

Плохой рандомизатор, фактически результаты получены на разных базах клиента.



**Просто повезло?**

Случайно в 1 клиенты покупали чаще, чем в 2.



**Другие факторы?**

Параллельно тестировали другую часть процесса, которая улучшила конверсию на 1, так как они позднее проходили заявочный процесс.



**Чтобы принимать решения на основе  
данных, нужна статистическая модель**



**ТИНЬКОФФ**

# Статистическая модель

# Статистическая модель

○

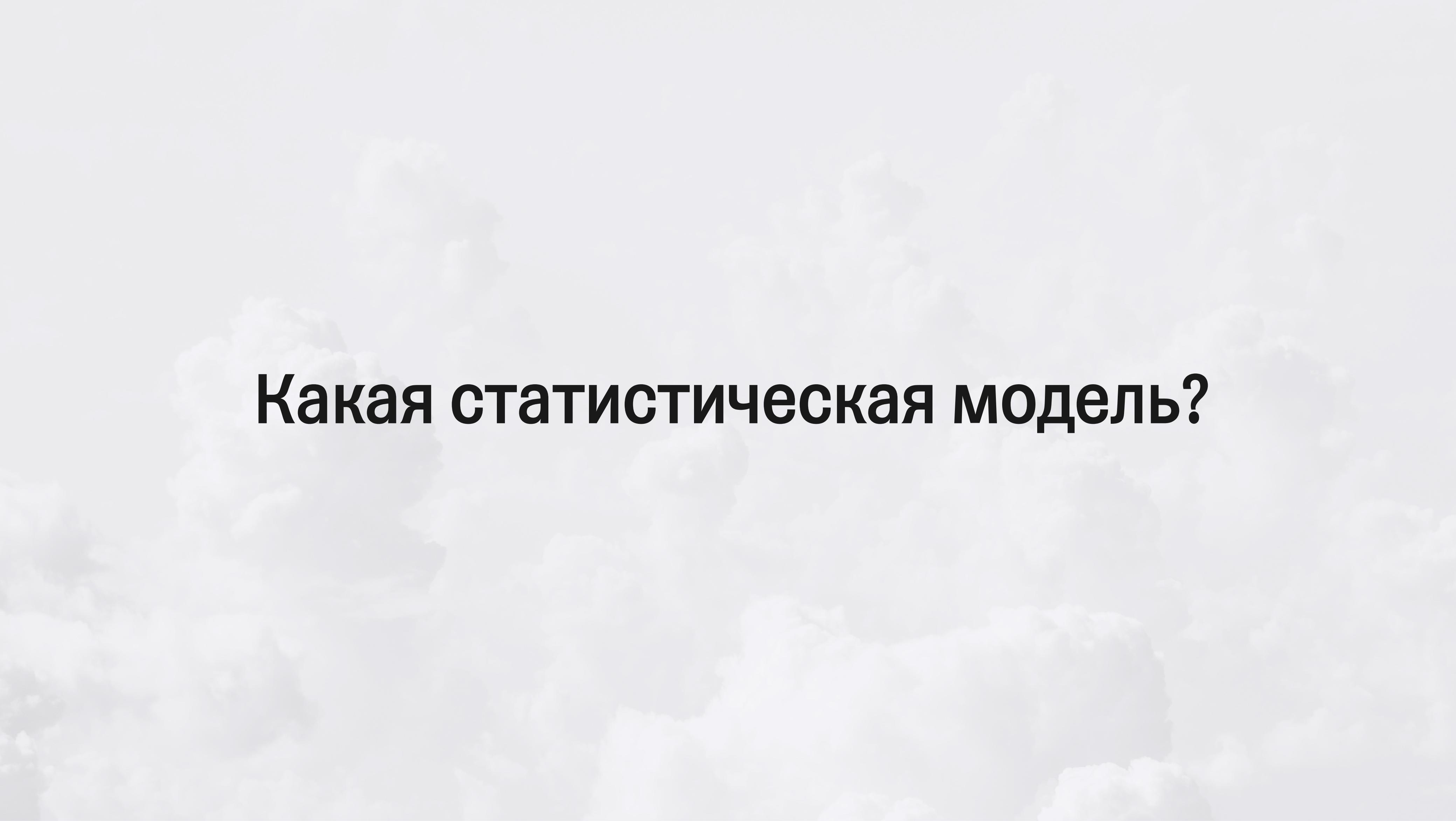
**Статистическая модель** – тройка  $(\mathcal{X}, \mathcal{F}, \mathcal{P})$ , где

- $\mathcal{X} \subset \mathbb{R}^n$  – выборочное пространство,
- $\mathcal{F} \subset 2^{\mathcal{X}}$  –  $\sigma$ -алгебра событий,
- $\mathcal{P} = \{\mathbb{P} \mid \mathbb{P}: \mathcal{F} \rightarrow [0, 1]\}$  – множество вероятностных мер.



Мы *знаем* выборку  $x \in \mathcal{X}$ , полученную с помощью  $\mathbb{P} \in \mathcal{P}$ .

Наша задача состоит в том, чтобы *имея конкретных что-то сказать* о  $\mathbb{P}$ .



# Какая статистическая модель?

# Статистические модели эксперимента

Первый вариант продажи

Множество вероятностных мер	Выборочное пространство	Вопрос
$\mathbb{P}_p^{n_1}, p \in [0, 1]$	$\{0, 1\}^{n_1}$	Просто повезло?
$\mathbb{P}_p^{n_{1,1}} \times \mathbb{P}_q^{n_{1,2}}$	$\{0, 1\}^{n_1}$	Проверка влияние акции
$\mathbb{P}^{n_1}, \mathbb{P} \in \mathcal{P}$	$\mathbb{N}^{n_1}$	Однородность возраста

Второй вариант продажи

Вопрос	Выборочное пространство	Множество вероятностных мер
Просто повезло?	$\{0, 1\}^{n_2}$	$\mathbb{P}_p^{n_2}, p \in [0, 1]$
Проверка влияние акции	$\{0, 1\}^{n_2}$	$\mathbb{P}_p^{n_{2,1}} \times \mathbb{P}_q^{n_{2,2}}$
Однородность возраста	$\mathbb{N}^{n_2}$	$\mathbb{P}^{n_2}, \mathbb{P} \in \mathcal{P}$

$n_1 = 100$  - размер первой выборки

$n_{1,1}$  - размер выборки до акции,  $n_{1,2}$  - размер выборки после акции

$\mathbb{P}_p^k$  - вероятностная мера из  $k$  независимых бернульиевских величин с параметром  $p$

$n_2 = 1000$  - размер второй выборки

$n_{2,1}$  - размер выборки до акции,  $n_{2,2}$  - размер выборки после акции

$\mathcal{P}$  – множество вероятностных мер со значениями в  $\mathbb{N}$

tf-data-analysis-lesson1-example / task1 / var0 / problem.md

Go

Latest commit 5ea3ebe 4 minutes ago

ributor

lines (24 sloc) 1.64 KB

Raw Blame

# Статистические модели в ДЗ

## Условие

Школа  $N$  имеет сильный состав для соревнования в прыжках в длину. В ней есть несколько сильных спортсменов, но на соревнования нужно отправить одного. Тренер Максим вычитал из книги, что длина прыжка имеет нормальное распределение, поэтому тренер решил выбрать лучшего школьника на основании оценки матожидания длины прыжка. Предполагая, что длины прыжков одного спортсмена независимы и имеют одинаковое для одного спортсмена распределение, помогите Максиму составить оценку этой величины для каждого студента.

## Входные данные

Одномерный массив `numpy.ndarray` длин прыжков (в сантиметрах) одного спортсмена.

## Возвращаемое значение

Оценка матожидания длины прыжка.

## Оценка

Максимальный балл: 4.

- +1 балл, если на выборках размера 1000 MSE оценки  $< 0.01$ .
- +1 балл, если на выборках размера 1000 MSE оценки  $< 0.005$ .
- +1 балл, если на выборках размера 100 MSE оценки  $< 0.015$ .
- +1 балл, если на выборках размера 10 MSE оценки  $< 0.09$ .



## Выборочное пространство

Пусть  $n$  – количество прыжков,  $m$  – количество спортсменов.

Какое выборочное пространство?

1.  $\mathbb{R}$
2.  $\mathbb{R}^n$
3.  $\mathbb{R}^m$
4.  $\mathbb{R}^{nm}$



## Множество вероятностных мер

Пусть  $\mathcal{P}$  – множество всех вероятностных мер на  $\mathbb{R}$ ,  $\mathcal{P}_{\mathcal{N}}$  – множество всех нормальных распределений на  $\mathbb{R}$ . Какое множество вероятностных мер?

1.  $\mathcal{P}$
2.  $\mathcal{P}_{\mathcal{N}}$
3.  $\mathbb{P}^n, \mathbb{P} \in \mathcal{P}$
4.  $\mathbb{P}^n, \mathbb{P} \in \mathcal{P}_{\mathcal{N}}$
5.  $\mathbb{P}_1 \times \dots \times \mathbb{P}_n, \mathbb{P}_i \in \mathcal{P}$
6.  $\mathbb{P}_1 \times \dots \times \mathbb{P}_n, \mathbb{P}_i \in \mathcal{P}_{\mathcal{N}}$

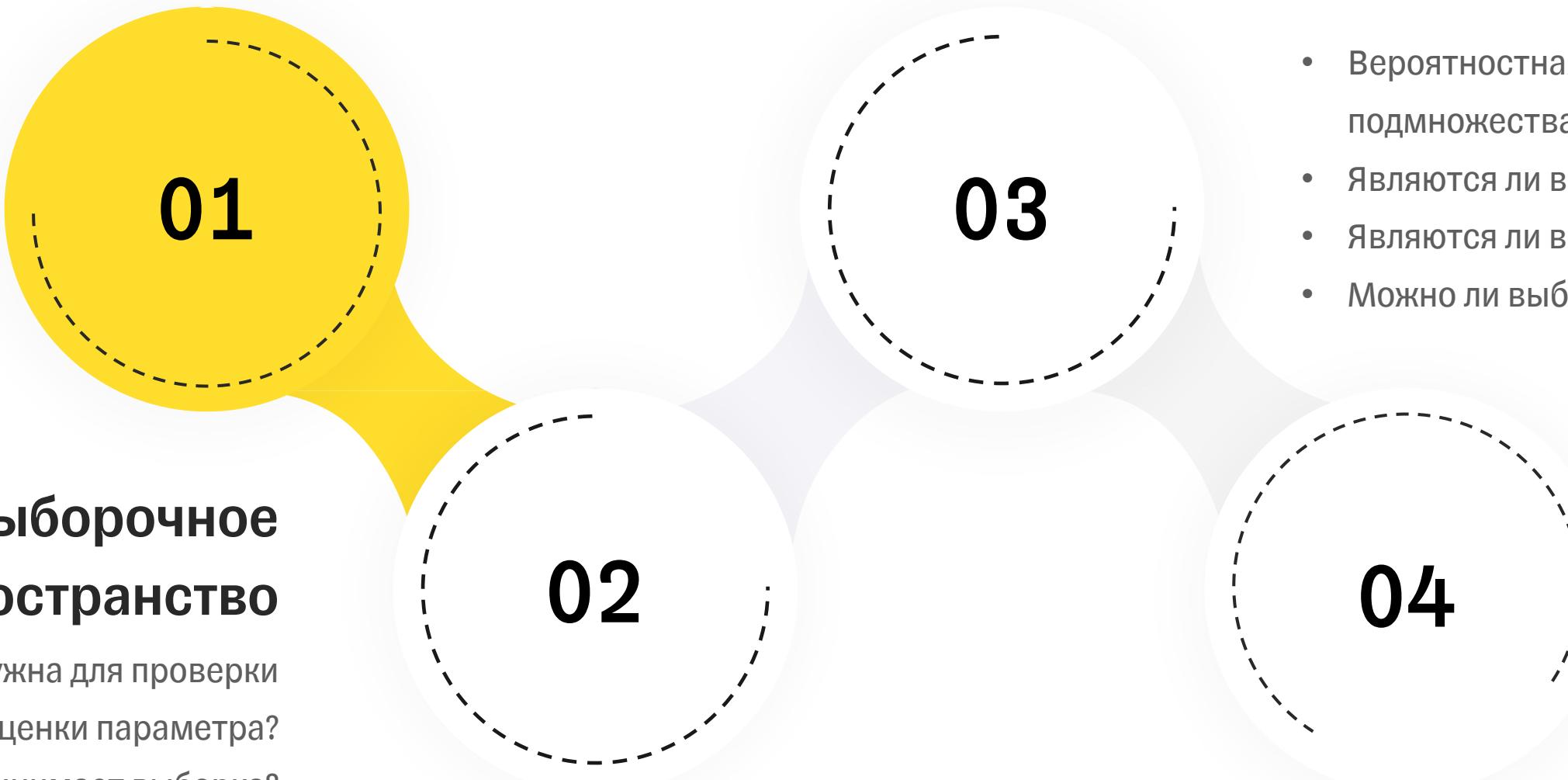
# Составление статистической модели

## Поставьте задачу

- В чём цель вашего эксперимента?
- Что вы хотите проверить или оценить?

## Составьте выборочное пространство

- Какая выборка мне нужна для проверки гипотезы или оценки параметра?
- Какие значения принимает выборка?

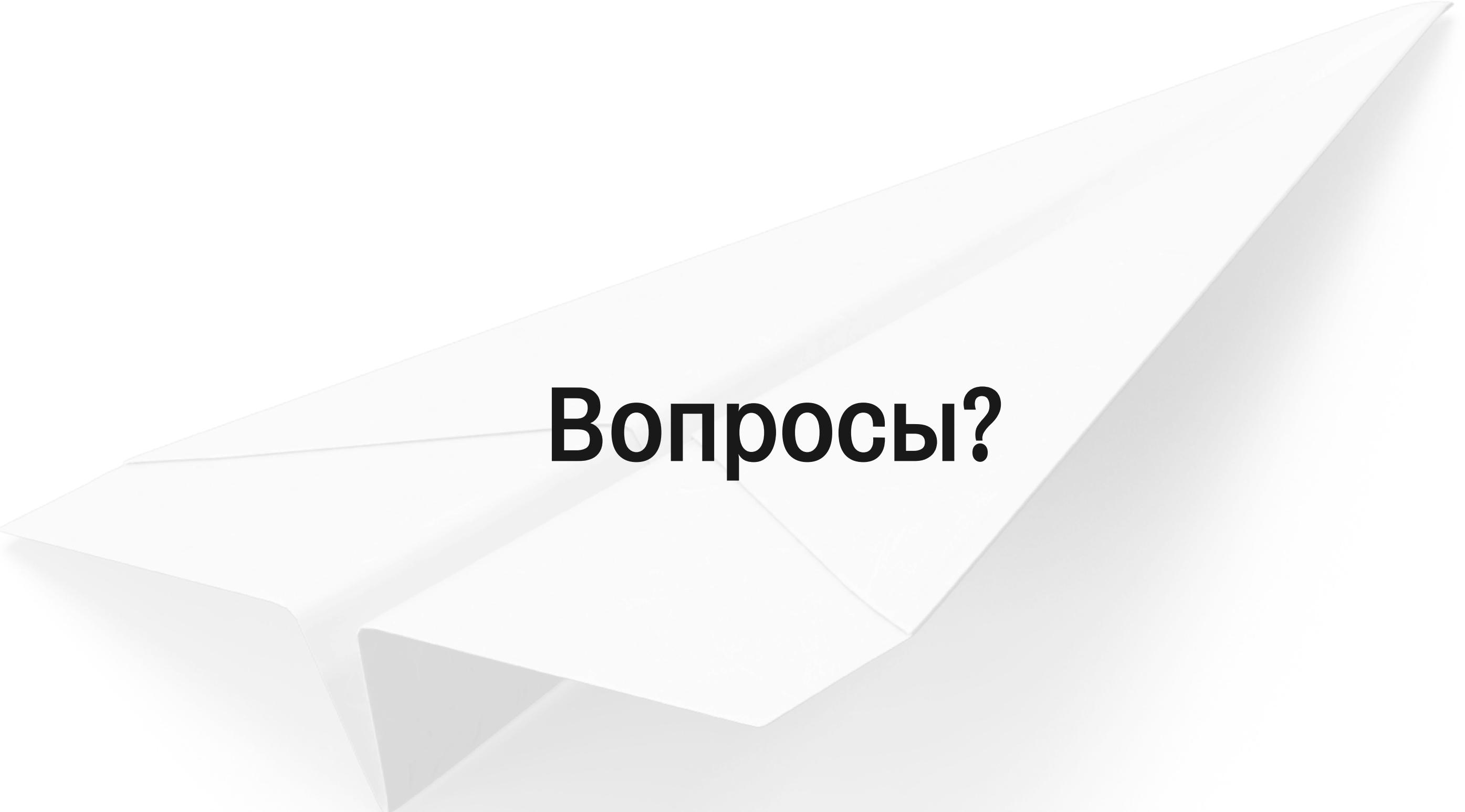


## Определите множество вероятностных мер

- Вероятностная мера определяет вероятность подмножества  $\mathcal{X}$ .
- Являются ли величины независимыми?
- Являются ли величины одинаково распределёнными?
- Можно ли выбрать параметрическое семейство мер?

## Ищите баланс

- Чем грубее модель, тем менее точной будет оценка.
- Не пренебрегайте конкретизацией модели и проверкой условий принадлежности параметрическому семейству.



**Вопросы?**

# Задачи статистики



## Точечное оценивание

Цель – составить оценку  $\varphi$  по выборке  $x$  параметра  $\theta$ , чтобы  $\varphi(x) \approx \theta$ .



## Интервальное оценивание

Цель – по выборке  $x$  определить такой интервал  $I = [\varphi(x), \psi(x)]$ , чтобы  $\mathbb{P}(\theta \in I) \geq p$ .



## Проверка гипотез

Цель – по выборке  $x$ , полученной с помощью неизвестной  $\mathbb{P}$ , понять, к какому из  $\mathcal{P}_i$  относится  $\mathbb{P}$ .



**ТИНЬКОФФ**

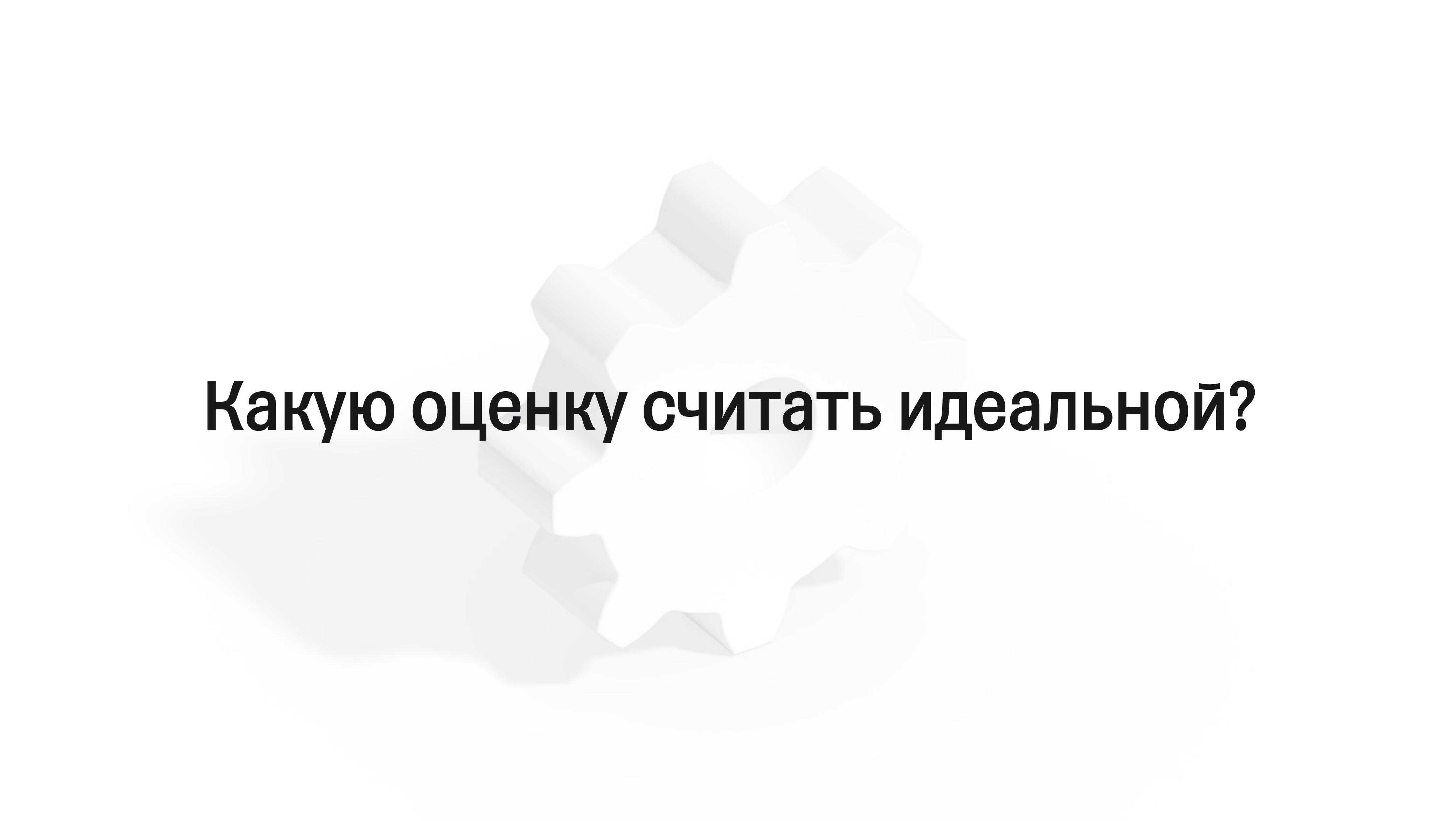
# Точечное оценивание

# Выборочные характеристики

Характеристика	Формула	Статистика
Математическое ожидание	$\bar{x} = \sum_{i=1}^n \frac{x_i}{n}$	Выборочное среднее
Дисперсия	$\overline{(x - \bar{x})^2}$	Выборочная дисперсия
$\alpha$ -квантиль распределения	$x_{([n \alpha])}$	Выборочный $\alpha$ -квантиль
Медиана	$\frac{x_{([n/2])} + x_{(n-[n/2])}}{2}$	Выборочная медиана

Пусть  $x = (x_1, \dots, x_n)$  – выборка.

Тогда  $x_{(i)}$  –  $i$ -е значение в отсортированном массиве из  $x$ .



# **Какую оценку считать идеальной?**

# Свойства точечных оценок

Свойство	Определение	Как доказывается
Несмешённость	$\mathbb{E}_\theta \varphi(X_1, \dots, X_n) = \theta$	Свойства распределения и матожидания
Состоятельность	$\varphi(X_1, \dots, X_n) \rightarrow \theta$	Закон больших чисел
Асимптотическая нормальность	$\sqrt{n} \frac{\varphi(X_1, \dots, X_n) - \theta}{\sigma(\theta)} \rightarrow \mathcal{N}(0, 1)$	Центральная предельная теорема

# Выборочное среднее – идеальная оценка?

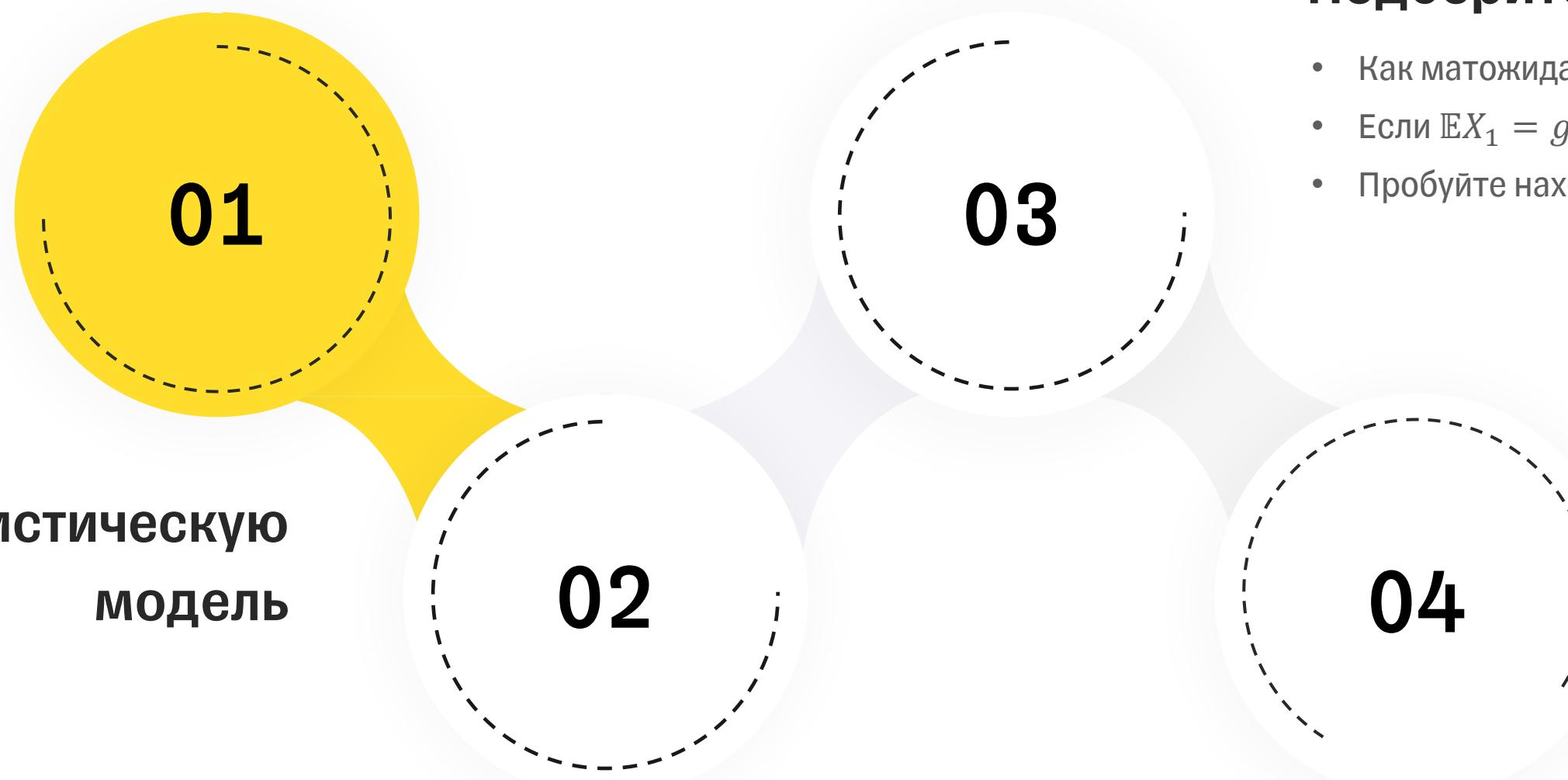
Свойство	Верно ли?	Если верно, то почему?
Несмешённость	Да (если $\mathbb{E} X_1  < \infty$ )	$\mathbb{E}\bar{X} = \sum_{i=1}^n \frac{\mathbb{E}X_i}{n} = \mathbb{E}X_1$
Состоятельность	Да (если $\mathbb{E} X_1  < \infty$ )	Закон больших чисел
Асимптотическая нормальность	Да (если $\mathbb{E}X_1^2 < \infty$ )	Центральная предельная теорема
Минимальность дисперсии	В общем случае <b>нет!</b> Но для $X_i \sim Bern(p)$ да	

Проверка свойств моделированием

# Составление точечной оценки

## Поставьте задачу

- В чём цель вашего эксперимента?
- Вам нужна именно точечная оценка?



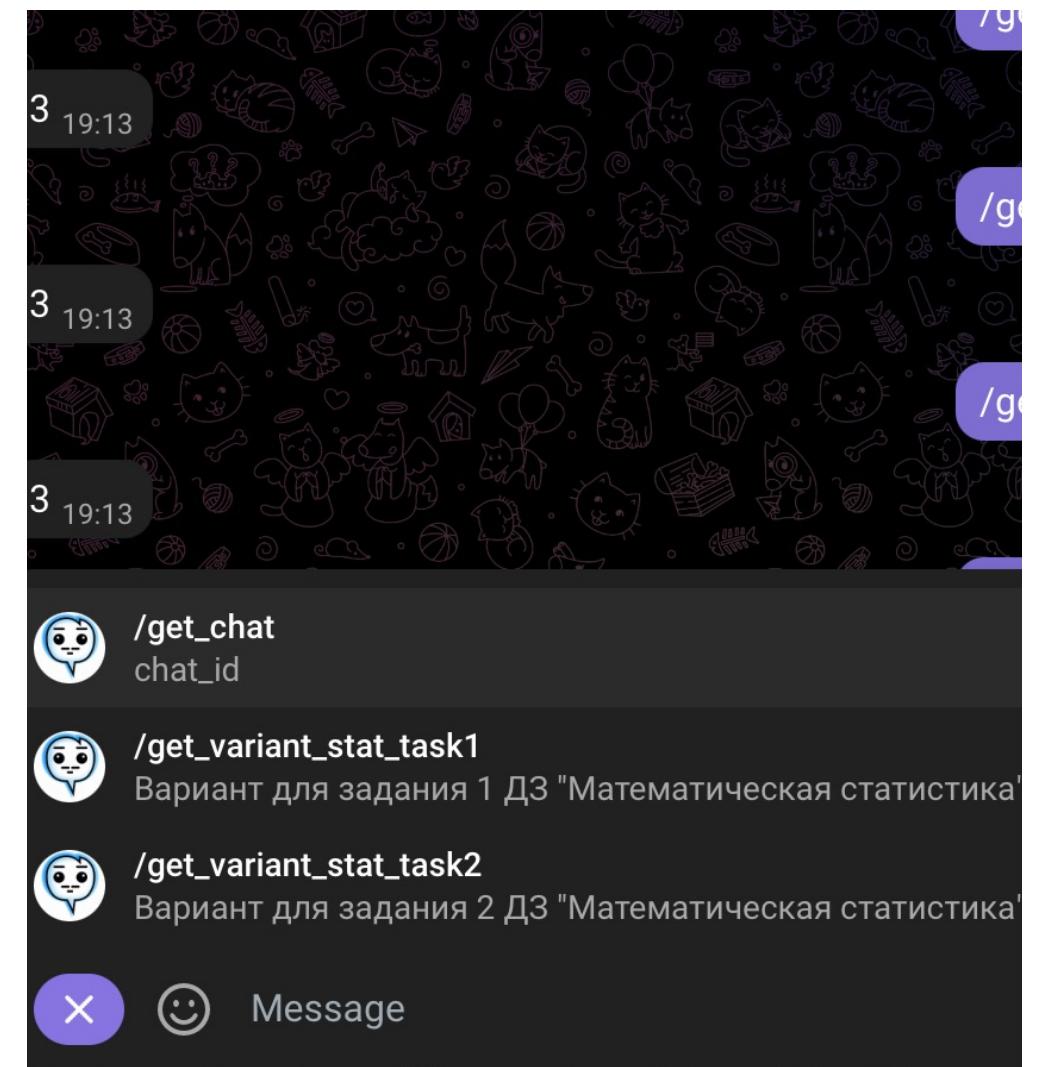
## Составьте статистическую модель

## Подберите оценку

- Как матожидание связано с оцениваемым параметром?
- Если  $\mathbb{E}X_1 = g(\theta)$ , то  $g^{-1}(\bar{X}) \rightarrow \theta$ .
- Пробуйте находить связь с параметром у  $\mathbb{E}X_1^k$ ,  $\mathbb{E}f(X_1)$ .

## Проверяйте оценку моделированием

- Несмешённость.
- Состоятельность.
- Асимптотическая нормальность.



## 1 Условие

Школа  $N$  имеет сильный состав для соревнования в прыжках в длину. В ней есть несколько сильных спортсменов, но на соревнования нужно отправить одного. Тренер Максим вычитал из книги, что длина прыжка имеет нормальное распределение, поэтому тренер решил выбрать лучшего школьника на основании оценки матожидания длины прыжка. Предполагая, что длины прыжков одного спортсмена независимы и имеют одинаковое для одного спортсмена распределение, помогите Максиму составить оценку этой величины для каждого студента.

## 2 Входные данные

Одномерный массив `piypru.ndarray` длин прыжков (в сантиметрах) одного спортсмена.

## 3 Возвращаемое значение

Оценка матожидания длины прыжка.

## 4 Оценка

Максимальный балл: 4.

- +1 балл, если на выборках размера 1000 MSE оценки  $\leq 0.0983$ .
- +1 балл, если на выборках размера 1000 MSE оценки  $\leq 0.00983$ .
- +1 балл, если на выборках размера 100 MSE оценки  $\leq 0.0296$ .
- +1 балл, если на выборках размера 10 MSE оценки  $\leq 0.112$ .

1

решено верно 1 из 2 • Баллы 4 из 10

Сегодня в 15:31

4 попытка

решено верно 1 из 2 • Баллы 4 из 10

Сегодня в 15:44

5 попытка

решено верно 1 из 2 • Баллы 4 из 10

Сегодня в 16:18

6 попытка

решено верно 1 из 2 • Баллы 4 из 10

Сегодня в 16:31

7 попытка

решено верно 1 из 2 • Баллы 9 из 10

Сегодня в 16:31

8 попытка

решено верно 1 из 2 • Баллы 9 из 10

Сегодня в 16:41

9 попытка

решено верно 0 из 2 • Баллы 5 из 10

Сегодня в 16:51

0 попытка

решено верно 1 из 2 • Баллы 9 из 10 • Засчитали эту попытку

Сегодня в 17:01

1 попытка

решено верно 0 из 2 • Баллы 0 из 10

Сегодня в 17:41

2 попытка

решено верно 1 из 2 • Баллы 8 из 10

Сегодня в 17:51

## Общаемся с ботом

[@TFDataAnalysisBot](#)

1. Находим [бота в Telegram](#)
2. Пишем боту `/start`
3. Получаем `chat_id`
4. Получаем условие задания 1

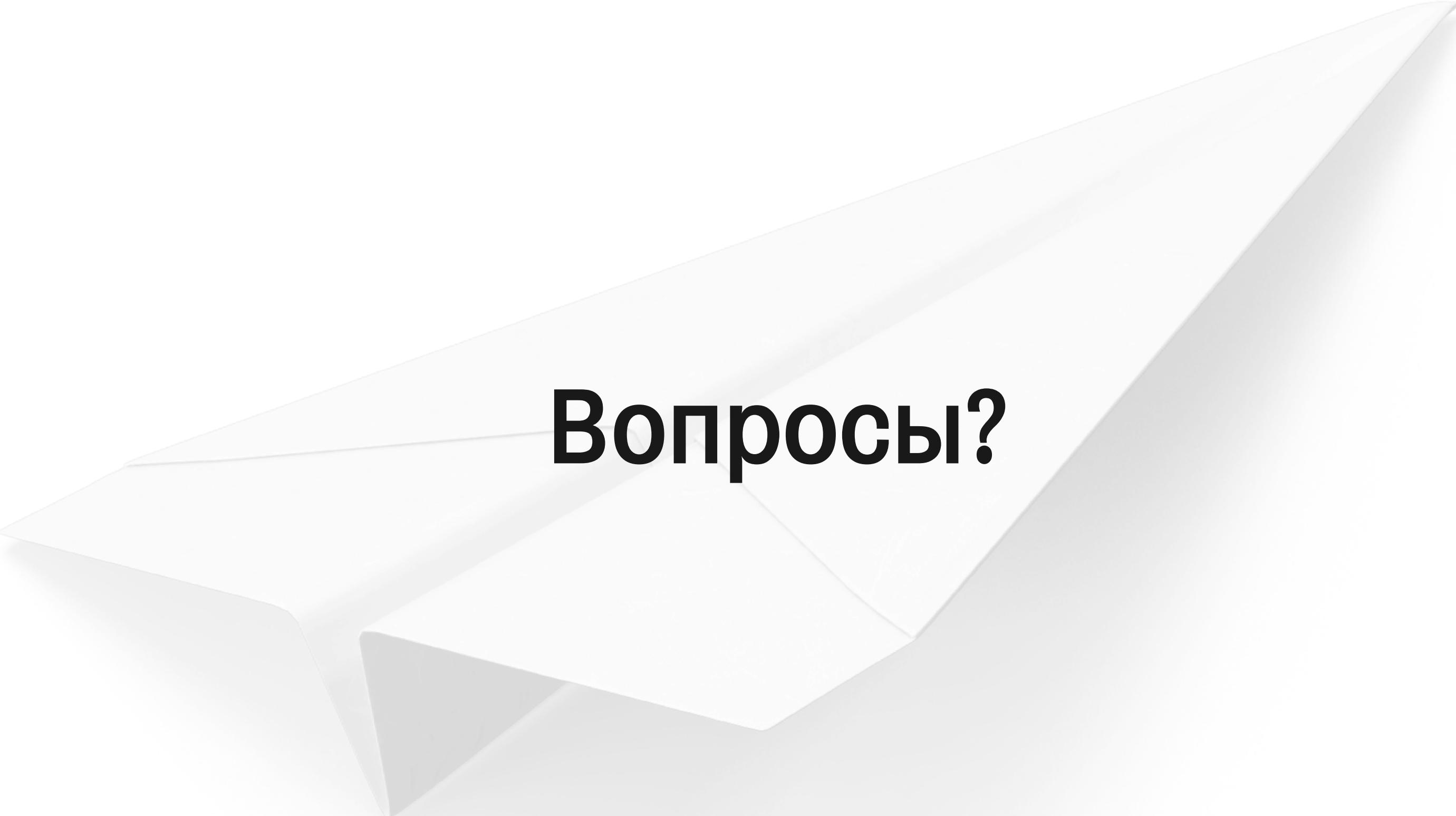
## Решаем задачу

1. Решаем задачу теоретически
2. Проверяем с помощью моделирования

1  
1

## Отправляем ответ

1. Регистрируемся на [GitHub](#)
2. Fork'аем [проект](#), открываем `solution.py`
3. Указываем свой `chat_id`
4. Пишем своё решение в `solution()`
5. Указываем ссылку на репозиторий на платформе `edu`



**Вопросы?**



**ТИНЬКОФФ**

# Интервальное оценивание

# Доверительное оценивание

Доверительный интервал – интервал  $I = [\varphi(x), \psi(x)]$ , где

$$\mathbb{P}_\theta(\theta \in I) = (\geq) p,$$

*p – уровень доверия.*

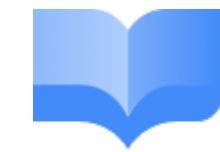
- Если  $\varphi \equiv -\infty$ , то интервал *левосторонний*.
- Если  $\psi \equiv \infty$ , то интервал *правосторонний*.



Противоречивость.

- Чем меньше длина  $I$ , тем точнее оценка.
- Чем больше длина  $I$ , тем больше  $p$ .

# Точный доверительный интервал



## Методика

Пусть  $g(\theta; X_1, \dots, X_n) \sim F$ , где  $F$  – распределение, не зависящее от  $\theta$ , а функция  $g$  возрастает по  $\theta$ . Тогда

$$\begin{aligned} & \mathbb{P}_\theta(z_\alpha \leq g(\theta; X) \leq z_{1-\beta}) = \\ & = \mathbb{P}_\theta(g^{-1}(z_\alpha; X) \leq \theta \leq g^{-1}(z_{1-\beta}; X)) = 1 - \alpha - \beta, \end{aligned}$$

где  $z_\gamma$  -  $\gamma$ -квантиль распределения  $F$ .



## Пример с равномерным распределением

Пусть  $X \sim R[0, a]$ ,  $a > 0$ . Тогда

$$X/a \sim R[0,1], \quad 1 - X/a \sim R[0,1].$$

Положим  $g(a; x) = 1 - x/a$ .

- Распределение  $g(a; X)$  не зависит от  $a$ .
- $z_\alpha = \alpha$ ,  $z_{1-\beta} = 1 - \beta$ .
- $g(a; x) = 1 - x/a = t$ ,  $a = g^{-1}(t; x) = x/(1 - t)$ .

В итоге:

$$a \in I = \left( \frac{X}{1 - \alpha}, \frac{X}{\beta} \right).$$

# Построение точного доверительного интервала



## Среднее нормального распределения

Пусть  $X = (X_1, \dots, X_n)$ ,  $X_i \sim \mathcal{N}(\mu, \sigma^2)$ , где  $\sigma$ - известна, а  $\mu$  нужно оценить. В силу [свойств нормального распределения](#)

$$X_1 + \dots + X_n \sim \mathcal{N}(\mu n, n\sigma^2).$$

Положим

$$g(\mu; x) = \frac{\mu n - (x_1 + \dots + x_n)}{\sigma \sqrt{n}} = \sqrt{n} \frac{\mu - \bar{x}}{\sigma}.$$

Тогда

$$g(\mu; X) \sim \mathcal{N}(0,1)$$

и  $z_\alpha, z_{1-\beta}$  - квантили  $\mathcal{N}(0,1)$ . Отметим, что

$$g(\mu; x) = t, \quad g^{-1}(t; x) = \bar{x} + t\sigma/\sqrt{n}.$$

В итоге:

$$\mu \in I = \left( \bar{X} + z_\alpha \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{1-\beta} \frac{\sigma}{\sqrt{n}} \right).$$



## Оценка интенсивности

Пусть  $X = (X_1, \dots, X_n)$ ,  $X_i \sim \exp(\lambda)$ , где  $\lambda$  нужно оценить. В силу [свойств экспоненциального распределения](#)

$$\min(X_1, \dots, X_n) \sim \exp(n\lambda).$$

Положим

$$g(\lambda; x) = n\lambda \min(x_1, \dots, x_n).$$

Тогда

$$g(\lambda; X) \sim \exp(1)$$

и  $z_\alpha, z_{1-\beta}$  - квантили  $\exp(1)$ . Отметим, что

$$g(\lambda; x) = t, \quad g^{-1}(t; x) = t/(n \min(x_1, \dots, x_n)).$$

В итоге:

$$\lambda \in I = \left( \frac{z_\alpha}{n \min(X_1, \dots, X_n)}, \frac{z_{1-\beta}}{n \min(X_1, \dots, X_n)} \right).$$

Проверка свойств моделированием

# Асимптотический доверительный интервал



## Методика

Пусть  $X_1, \dots, X_n$  - н.о.р.,  $\mathbb{E}X_1 = \theta$ ,  $\mathbb{D}X_1 = \sigma^2 < \infty$ . Тогда  
 $\sqrt{n}(\bar{X} - \theta)/\sigma \rightarrow \mathcal{N}(0,1)$ .

Отсюда

$$\mathbb{P}(\theta \in I) \approx 1 - \alpha - \beta, \quad I = \left( \bar{X} + z_\alpha \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{1-\beta} \frac{\sigma}{\sqrt{n}} \right).$$

Если  $\sigma$  неизвестна, то можно использовать оценку

$$S_X^2 = \overline{(X - \bar{X})^2} \rightarrow \sigma^2.$$



## Пример с равномерным распределением

Пусть  $X_1, \dots, X_n$  - н.о.р.,  $X_1 \sim R[0, a]$ ,  $a > 0$ . Тогда  
 $\mathbb{E}(2X_1) = a$ .

В итоге:

$$I = \left( 2\bar{X} + 2z_\alpha \frac{S_X}{\sqrt{n}}, 2\bar{X} + 2z_{1-\beta} \frac{S_X}{\sqrt{n}} \right).$$

Сравнение методов моделированием

# Приятие решений с помощью доверительного интервала

## Задача

**22%**

минимальное значение конверсии, при котором продукт окупаем. Хотим выбрать вариант продажи с наиболее вероятной окупаемостью.

## Данные

**30/100, 250/1000**

данные по тесту в формате продажи/лиды. В силу ЦПТ  
 $\mathbb{P}(p \geq \bar{X} - z_\alpha S_p / \sqrt{n}) \approx \alpha$ .  
В силу постановки задачи левая часть равна 0.22. Отсюда  
 $\alpha = \Phi(\sqrt{n}(\bar{X} - 0.22) / S_p)$ .

Расчёты приведены [здесь](#).

## Приятие решения

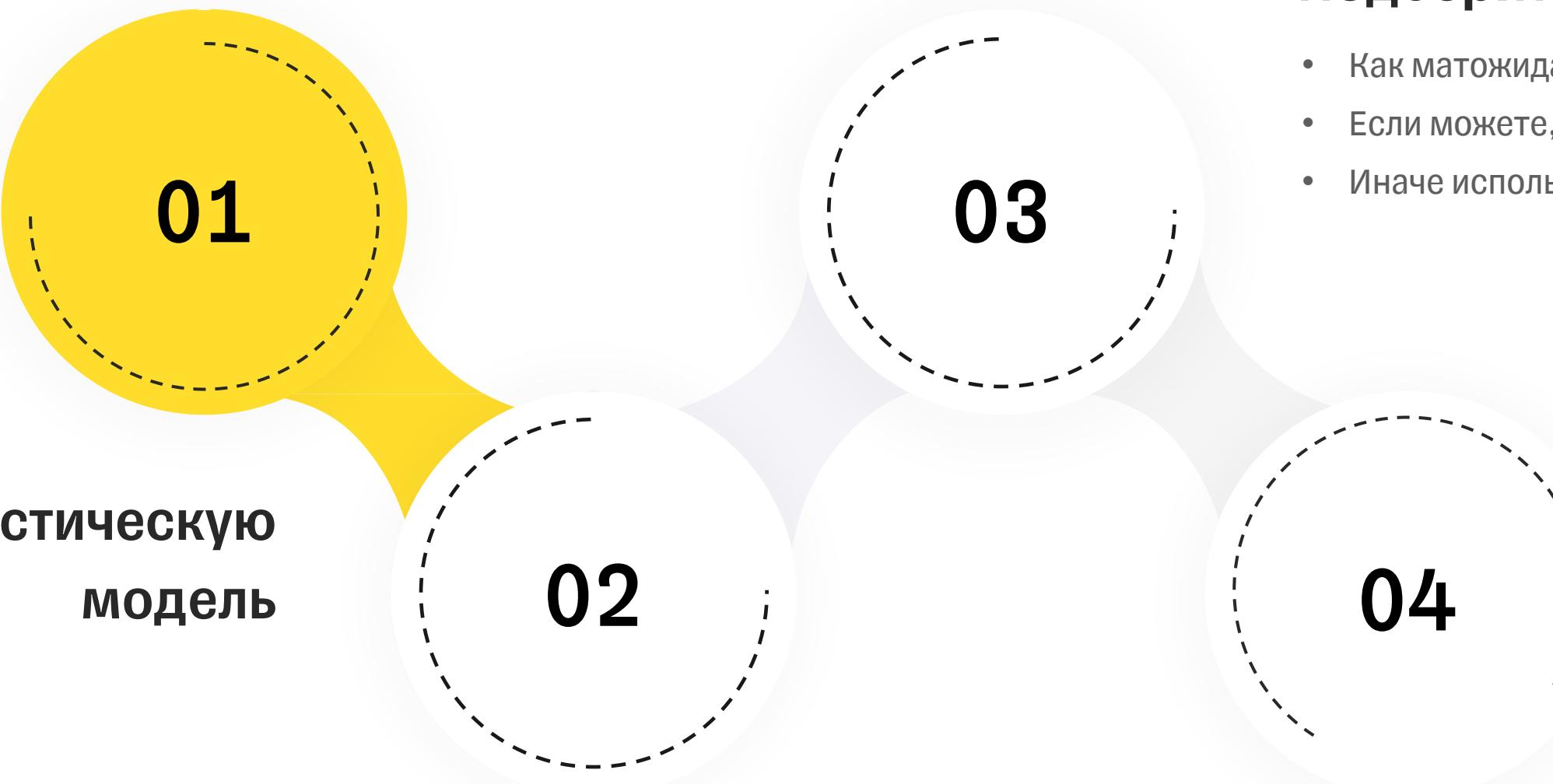
**96% vs 98.6%**

вероятность окупаемости продаж у первого и второго варианта соответственно. Второй лучше первого!

# Составление интервальной оценки

## Поставьте задачу

- В чём цель вашего эксперимента?
- Вам нужна именно интервальная оценка?



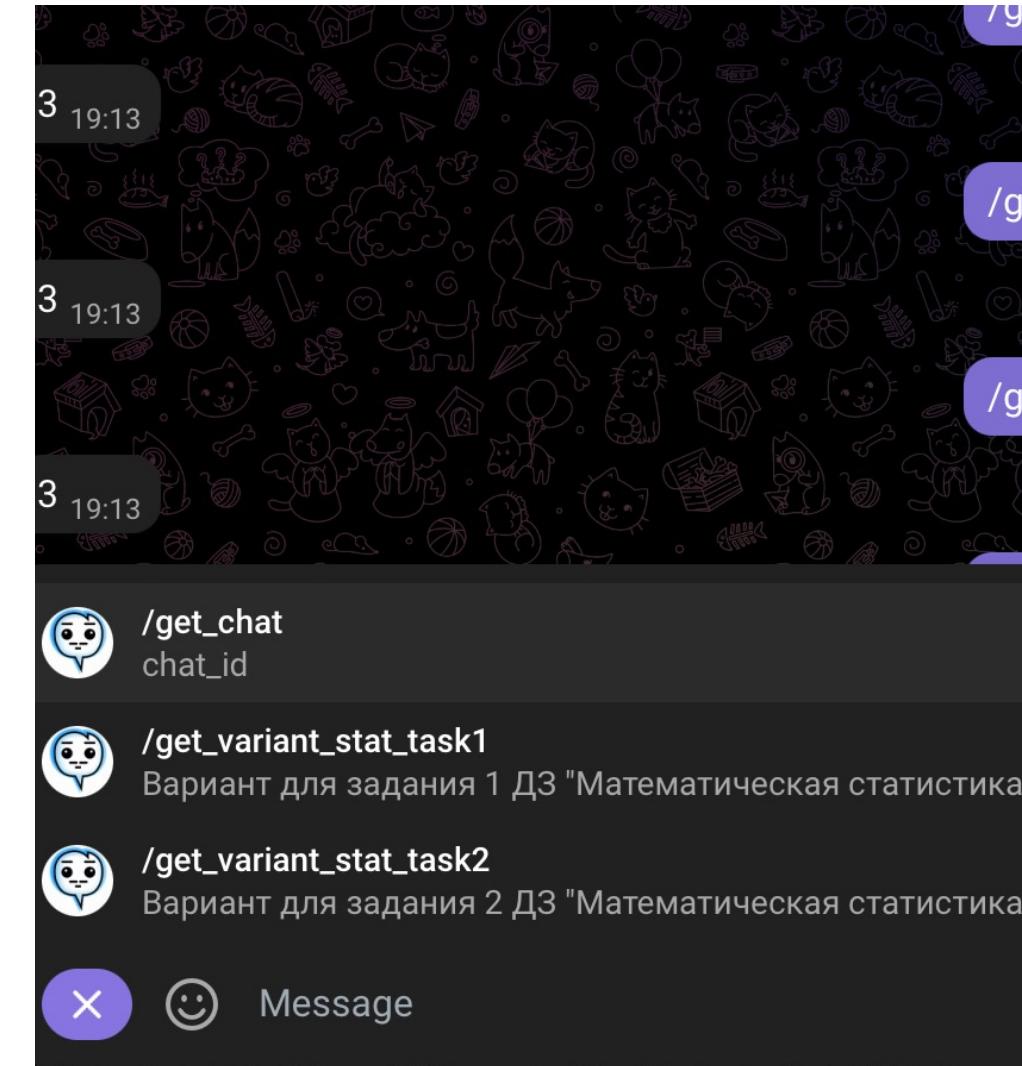
## Составьте статистическую модель

## Подберите оценку

- Как матожидание связано с оцениваемым параметром?
- Если можете, используйте распределение вашей оценки.
- Иначе используйте асимптотический интервал.

## Проверяйте оценку моделированием

- Частота попадания в интервал.
- Длина интервала.
- Изменение характеристик при росте выборок.



## Общаемся с ботом

**@TFDataAnalysisBot**

1. Находим [бота в Telegram](#)
2. Получаем `chat_id`
3. Получаем вариант задания 2

## Решаем задачу

1. Решаем задачу теоретически
2. Проверяем с помощью моделирования

### 1 Условие

Школа  $N$  имеет сильный состав для соревнования в прыжках в длину. В ней есть несколько сильных спортсменов, но на соревнования нужно отправить одного. Тренер Максим вычитал из книги, что длина прыжка имеет нормальное распределение с дисперсией 100, поэтому тренер решил выбрать лучшего школьника на основании оценки матожидания длины прыжка. Помогите Максиму составить симметричный доверительный интервал этой величины для каждого студента.

### 2 Входные данные

Два входных значения. Первое - уровень доверия, число от 0 до 1. Второе - одномерный массив `numPy.ndarray` длин прыжков (в сантиметрах) одного спортсмена.

### 3 Возвращаемое значение

Кортеж или список из двух значений, равных левой и правой границе доверительного интервала.

### 4 Оценка

Максимальный балл: 6.

Выборка	Доверие	Частота ошибок	Длина интервала
1000	0.99	0.13	32.6
1000	0.9	0.46	10.5
100	0.7	0.464	4.15
100	0.9	0.119	3.95
10	0.95	0.125	18.6
10	0.9	0.107	11.5

- Выборка - Размер выборки
- Доверие - Уровень доверия

1

решено верно 1 из 2 • Баллы 4 из 10

Сегодня в 15:31

4 попытка  
решено верно 1 из 2 • Баллы 4 из 10

Сегодня в 15:44

5 попытка  
решено верно 1 из 2 • Баллы 4 из 10

Сегодня в 16:18

6 попытка  
решено верно 1 из 2 • Баллы 4 из 10

Сегодня в 16:34

7 попытка  
решено верно 1 из 2 • Баллы 9 из 10

Сегодня в 16:37

8 попытка  
решено верно 1 из 2 • Баллы 9 из 10

Сегодня в 16:44

9 попытка  
решено верно 0 из 2 • Баллы 5 из 10

Сегодня в 16:51

0 попытка  
решено верно 1 из 2 • Баллы 9 из 10 • Засчитали эту попытку

Сегодня в 17:01

1 попытка  
решено верно 0 из 2 • Баллы 0 из 10

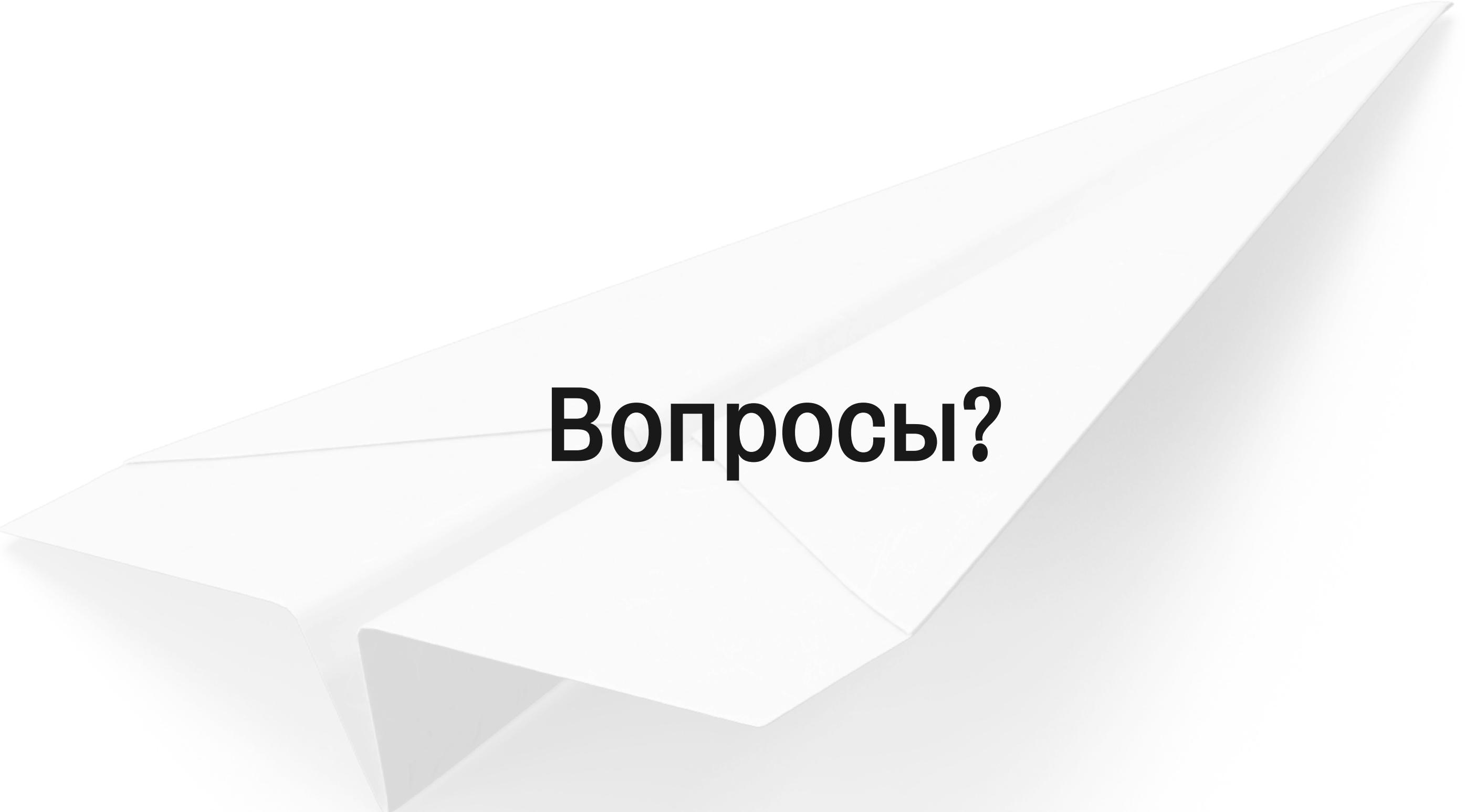
Сегодня в 17:41

2 попытка  
решено верно 1 из 2 • Баллы 8 из 10

Сегодня в 17:54

## Отправляем ответ

1. Fork'аем [проект](#), открываем `solution.py`
2. Указываем свой `chat_id`
3. Пишем своё решение в `solution()`
4. Указываем ссылку на репозиторий на платформе edu



**Вопросы?**

# Литература



Чернова Н.И.

[Теория вероятностей](#)



Коршунов Д.А., Фосс С.Г.

[Сборник задач и упражнений по теории вероятностей](#)



Чернова Н.И.

[Математическая статистика](#)



Лагутин М.Б.

[Наглядная математическая статистика](#)



Коршунов Д.А., Чернова Н.И.

[Сборник задач и упражнений по математической статистике](#)



ТИНЬКОФФ

Он такой один

