

# Статистическая проверка гипотез



Логистическая регрессия. Сигмоида. Функция ошибки в модели логистической регрессии. Кросс-валидация. Квантиль и квартиль. Дизайн эксперимента и статистические гипотезы о данных. Ошибки 1-го и 2-го рода. Статистическая значимость p-value. A/B тестирование. Тесты на нормальность. Корреляционные тесты. Проверка гипотезы t-критерия Стьюдента. Одновыборочный t-критерий. Двухвыборочный t-критерий для независимых выборок. Множественный тест (ANOVA).

**Даниил Корбут**

Специалист по Анализу Данных



НЕТОЛОГИЯ



**Даниил Корбут**  
DL Researcher  
Insilico Medicine, Inc

Окончил бакалавриат ФИВТ  
МФТИ (Анализ данных) в 2018г  
Учусь на 2-м курсе  
магистратуры ФИВТ МФТИ  
Работал в Statsbot и Яндекс.  
Алиса.  
Сейчас в Insilico Medicine, Inc,  
занимаюсь генерацией  
активных молекул и  
исследованиями старения с  
помощью DL.

# Логистическая регрессия

Задача логистической регрессии – определить вероятность принадлежности к классу.

Построена на основе линейной функции.

$$h(x) = \theta^T x$$

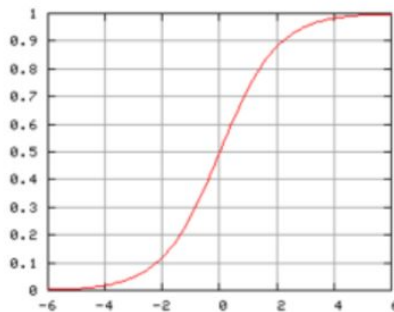
К линейной функции применяется функция активации:

$$h(x) = \sigma(\theta^T x)$$

Функция активации:

$$\sigma(t) = \frac{1}{1 + e^{-t}}$$

# Сигмоида



$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

Производная сигмоиды:

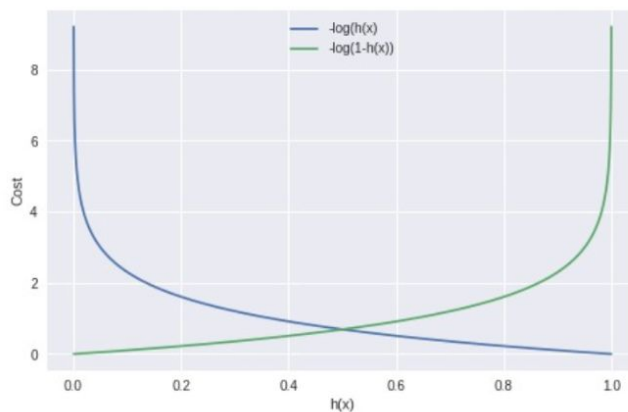
$$\sigma'(x) = \sigma(x) \cdot (1 - \sigma(x))$$

# Функция ошибки логистической регрессии

Модель ищет параметры, которые минимизируют функцию ошибки:

$$cost = \begin{cases} -\log(h(x)), & \text{if } y = 1 \\ -\log(1 - h(x)), & \text{if } y = 0 \end{cases}$$

Чем выше вероятность определения класса 1 при верном классе 0, тем выше стоимость ошибки.



# Функция ошибки логистической регрессии

Общий вид функции ошибки для модели:

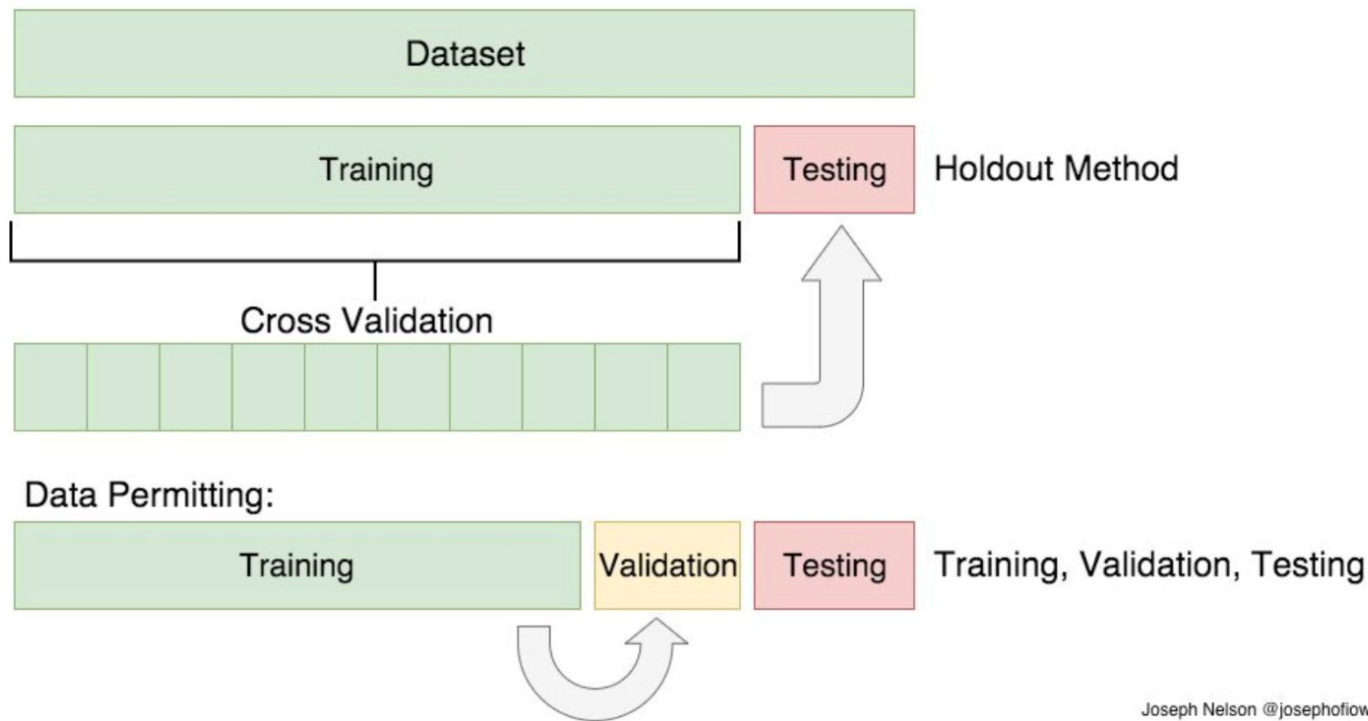
$$\text{cost}(h(x), y) = -y \cdot \log(h(x)) - (1 - y)\log(1 - h(x))$$

Ошибка для всех данных датасета:

$$J(\theta) = -\frac{1}{m} \sum_{i=1}^m [y^i \log(h(x^i)) + (1 - y^i) \log(1 - h(x^i))]$$

Где  $m$  – количество элементов.

# Кросс-валидация

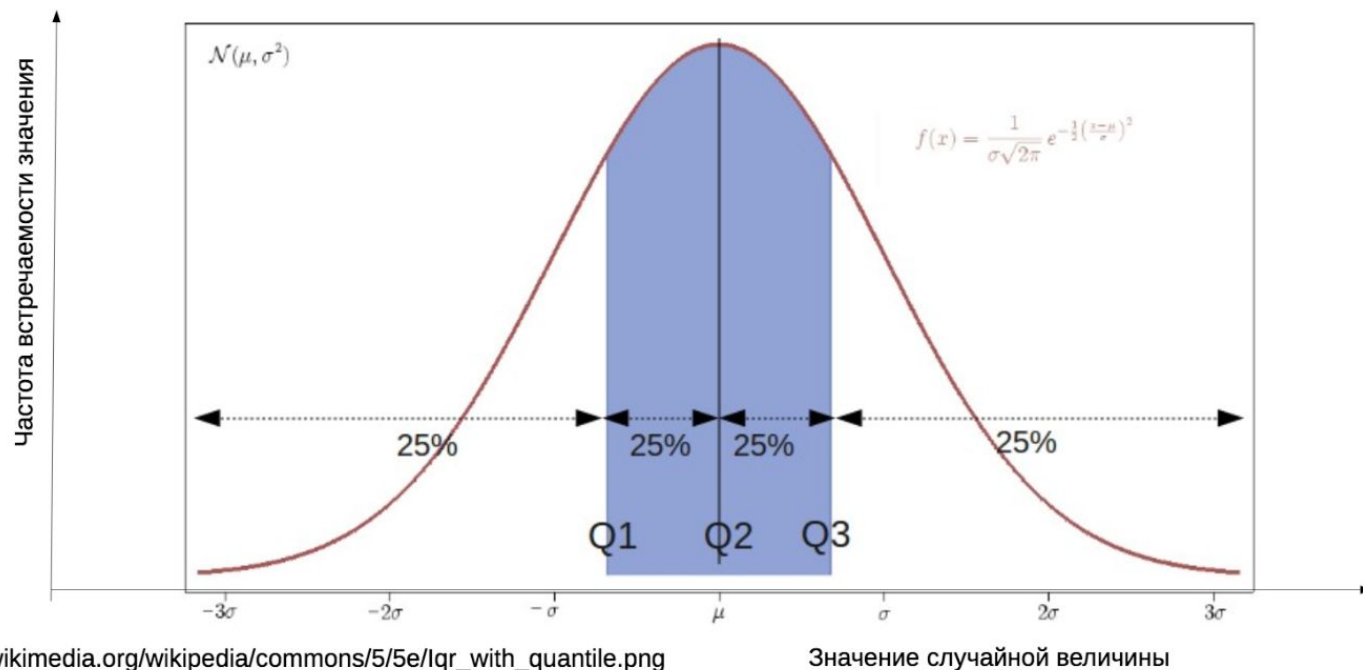


Joseph Nelson @josephofiowa

<https://discuss.pytorch.org>

# Квантиль

значение, которое заданная случайная величина не превышает с фиксированной вероятностью

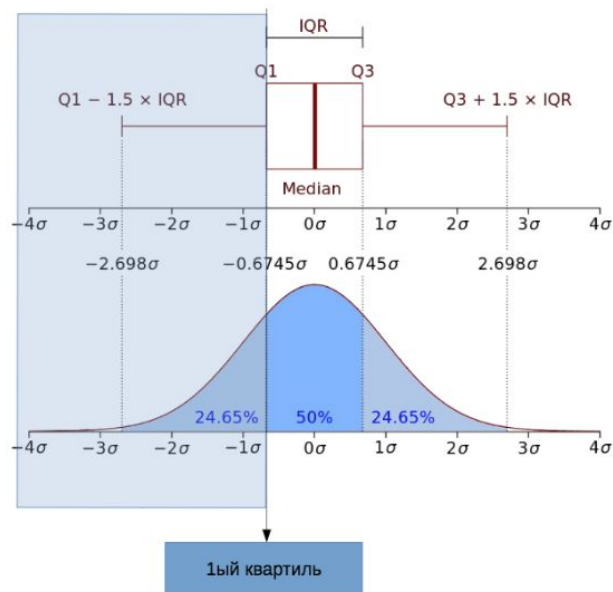


[https://upload.wikimedia.org/wikipedia/commons/5/5e/lqr\\_with\\_quantile.png](https://upload.wikimedia.org/wikipedia/commons/5/5e/lqr_with_quantile.png)

Значение случайной величины



# Квартиль



Предоставляют важную информацию о структуре **вариационного** (колонок таблицы) ряда признака. Вместе с медианой они делят вариационный ряд на 4 равные части. Квартилей две, их обозначают символами  $Q$ , верхняя и нижняя квартиль. 25% значений меньше, чем нижняя квартиль, 75% значений меньше, чем верхняя квартиль.

# Статистические гипотезы о данных

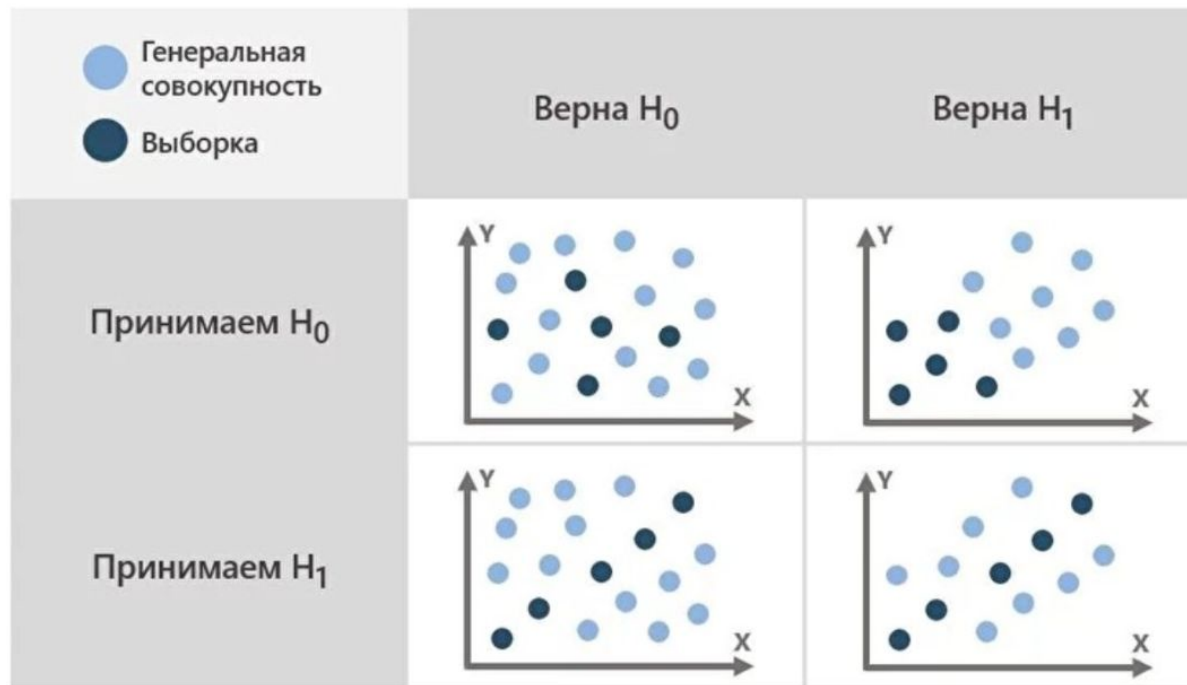
**Выборочная совокупность** — множество всех объектов, отобранных случайно из генеральной совокупности для изучения.



**Нулевая гипотеза ( $H_0$ )**— гипотеза о сходстве

**Альтернативная гипотеза, конкурирующая, ( $H_1$ )**— гипотеза о различиях

# Нулевая и альтернативная гипотезы



# Примеры основной и альтернативной гипотез

Основная гипотеза:

$$H_0 : a = 368$$

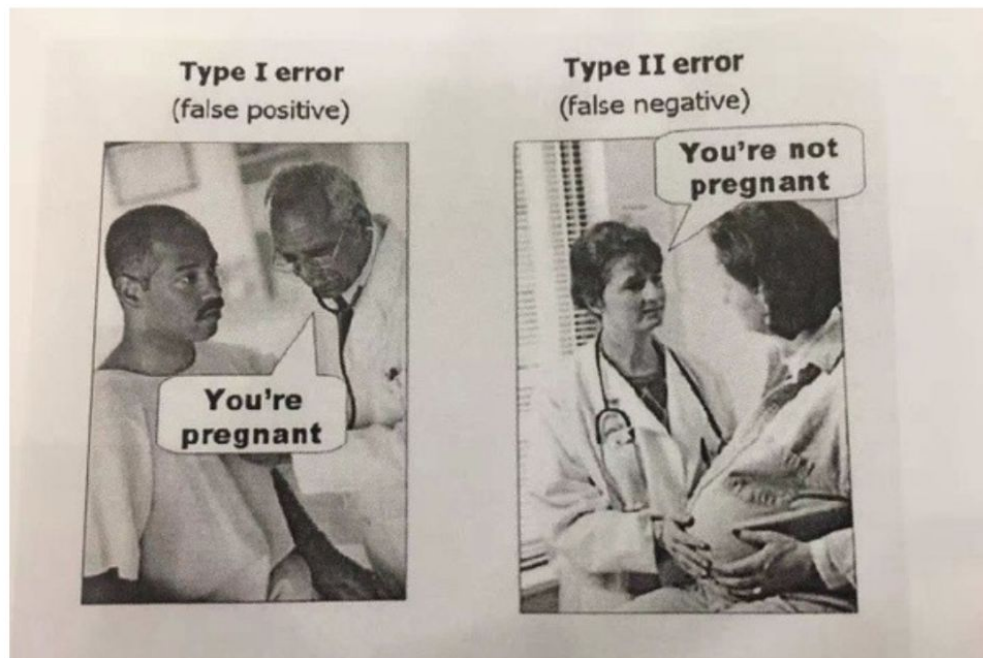
Средний вес выпускаемых коробок равен 368 г,  
конвейер работает нормально

Альтернативная гипотеза:

$$H_1 : a \neq 368$$



Средний вес выпускаемых коробок отличен от 368 г,  
конвейер требует наладки

## Пример: тест на беременность



<https://blog.mathquant.com/2019/01/26/why-do-you-have-to-learn-to-lose-money-in-the-futures-market.html>

# Статистические гипотезы о данных

	Disease present	Disease absent
Positive	<b>a</b> True positive	<b>b</b> False positive 
Negative	<b>c</b> False negative 	<b>d</b> True negative

**Ошибка 1 рода:**  
Вероятность отвергнуть гипотезу,  
но в действительности она верна

Критически значимый уровень  
**alpha** = 0.05

**Ошибка 2 рода:**  
Вероятность принять гипотезу,  
но в действительности она неверна  
**beta** — вероятность ошибки.  
Мощность исследования = 1-beta.

[https://www.youtube.com/watch?v=4eyEp\\_NTxAU](https://www.youtube.com/watch?v=4eyEp_NTxAU)

<https://www.nejm.org/doi/full/10.1056/NEJM199908193410823>

# Статистическая значимость

СТАТИСТИЧЕСКАЯ ЗНАЧИМОСТЬ (ЗНАЧЕНИЕ P)  
– РАСЧЕТНАЯ ВЕРОЯТНОСТЬ ОШИБКИ  
ПЕРВОГО РОДА, КОТОРАЯ РАССЧИТЫВАЕТСЯ С  
ПОМОЩЬЮ РАЗЛИЧНЫХ СТАТИСТИЧЕСКИХ  
КРИТЕРИЕВ



$$P < 0,05$$

# Виды статистических критериев

## Критерии согласия -

исследуемая случайная величина подчиняется предполагаемому закону.

```
1 If Data Is Gaussian:  
2   Use Parametric Statistical Methods  
3 Else:  
4   Use Nonparametric Statistical Methods
```

## Параметрические критерии -

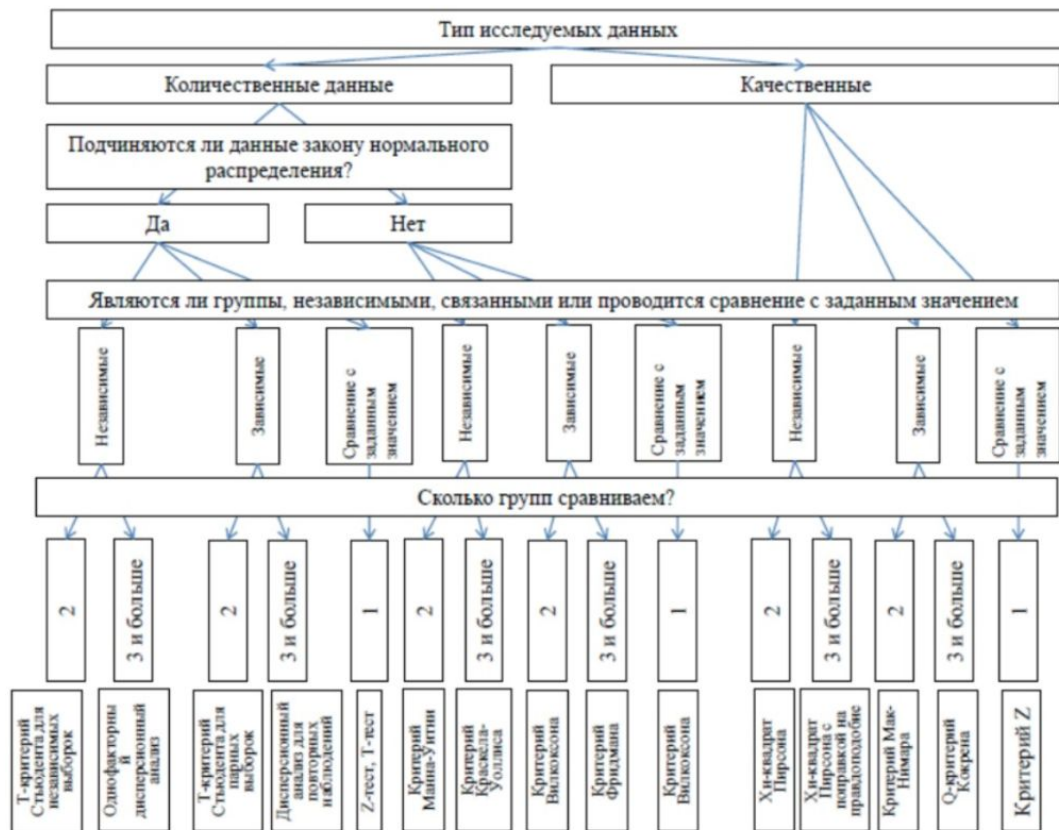
включают в расчет параметры вероятностного распределения признака (средние и дисперсии).

## Непараметрические критерии -

которые не включают в расчёт параметры вероятностного распределения и основаны на оперировании частотами или рангами.



# Схема применения критериев



# Параметрические критерии

Параметрические критерии - группа статистических критериев, которые включают в расчет параметры вероятностного распределения признака (средние и дисперсии).

t-критерий Стьюдента

Критерий Фишера

Критерий отношения правдоподобия

Критерий Романовского

## t-критерий Стьюдента

Случайная величина  $t$  имеет распределение Стьюдента с  $n-1$  степенями свободы, где  $n$  — размер выборки.

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

Критические значения  $t$  для заданного уровня доверия можно взять из таблицы:

<https://www.kontrolnaya-rabota.ru/s/teoriya-veroyatnosti/tablica-studenta/>

## t-критерий Стьюдента

Если t-критерий НЕ превышает пороговое  $t$ , при  $p\text{-level} = 0.05$  — это значит, что у нас **нет оснований для отклонения нулевой гипотезы**.

Если фактическое  **$t$**  превышает критическое табличное значение при  **$p=0.05$** , мы отклоняем нулевую гипотезу, — это означает, что мы обнаружили значимую закономерность!

# Непараметрические критерии

Непараметрические критерии

Группа статистических критериев, которые не включают в расчёт параметры вероятностного распределения и основаны на оперировании частотами или рангами.

Q-критерий Розенбаума

U-критерий Манна — Уитни

Критерий Уилкоксона

Критерий Пирсона

Критерий Колмогорова — Смирнова

# A/B тесты

A/B тестирование — это мощный маркетинговый инструмент для повышения эффективности работы вашего интернет-ресурса.

Ниже на картинках приведены примеры распределения значений показателя в сегментах.



# Пример А/В теста

Компания WallMonkeys решила оптимизировать веб-сайт на клики и конверсию.



# Пример А/В теста

1 тест: 27% кликов.



2 тест: 550% кликов





# Пример А/В теста

## CONTROL

OBAMA BIDEN

OBAMA CLASSIC 2012

You could be there

Michael Jordan, Patrick Ewing, Magic Johnson, Charles Oakley, and many more from basketball's greatest years are all coming up to support President Obama. You and a group will meet the President — and share dinner and more of the most legendary sports stars.

Watch Obama in action!

**Contribute**

First name:

Last name:

Address:

City:  State:  Zip:

Email:  Phone number:

**Select amount**

☐ \$10 ☐ \$25 ☐ \$50 ☐ \$100 ☐ \$250 ☐ \$500 ☐ \$1,000

Other:

**Credit card**

Card number:

Expiration: Month  Year

☐ Make this a recurring donation to support this campaign in the time you have left — Election Day is closer than you think.

**Employment**

Employer:

Occupation:

**DONATE NOW**

## “SEQUENTIAL”

OBAMA BIDEN

OBAMA CLASSIC 2012

You could be there

Michael Jordan, Patrick Ewing, Magic Johnson, Charles Oakley, and many more from basketball's greatest years are all coming up to support President Obama. You and a group will meet the President — and share dinner and more of the most legendary sports stars.

Watch Obama in action!

**How much would you like to donate today?**

**Select amount**

☐ \$15 ☐ \$25 ☐ \$50

☐ \$100 ☐ \$250 ☐ \$500

☐ \$1,000 ☐ Other amount:

**CONTINUE**

↑ +5%

<http://ly.lerank.net>

**Спасибо за внимание!**