



# Статистика

## DS-поток

Лекция 12

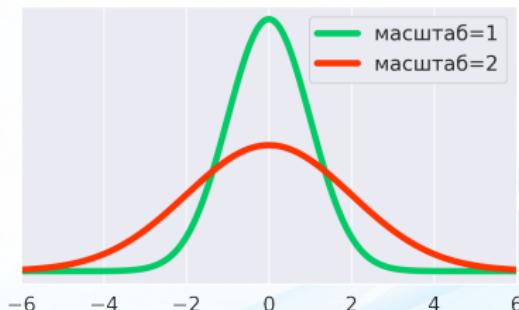
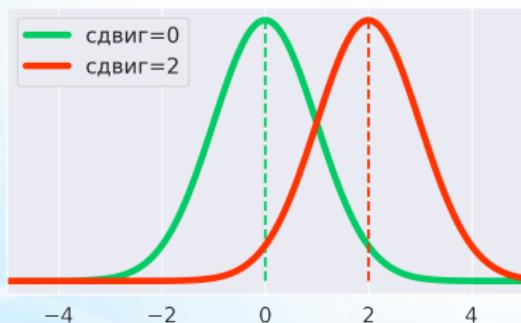
## 6.3 Обобщенная модель линейной регрессии

# Сдвиг и масштаб

$\mathcal{P} = \{P_\theta \mid \theta \in \Theta\}$  — семейство непрер. распр. с плотностью  $p_\theta(x)$ .

$\theta$  — параметр сдвига, если  $p_\theta(x) = p_0(x - \theta)$ .

$\theta$  — параметр масштаба, если  $p_\theta(x) = \frac{1}{\theta}p_1(x/\theta)$  и  $\Theta \subset (0, +\infty)$ .



Примеры:

$$\mathcal{P} = \{\mathcal{N}(a, \sigma^2) \mid a \in \mathbb{R}, \sigma > 0\}:$$

$a$  — параметр сдвига

$\sigma$  — параметр масштаба.

$$\mathcal{P} = \{U[0, \theta] \mid \theta > 0\}:$$

$\theta$  — параметр масштаба.

## Вспомним гауссовскую линейную модель

Данные  $Y = X\theta + \varepsilon \sim \mathcal{N}(X\theta, \sigma^2 I_n)$

Строим модель вида  $y(x) = x^T \theta$ .

### Что мы предсказываем?

Пусть  $x_0$  новый объект.

Тогда  $Y_0 = x_0^T \theta + \varepsilon_0 \sim \mathcal{N}(x_0^T \theta, \sigma^2)$ .

Т.е. в качестве предсказания оцениваем  $E Y_0$  — ожидаемый отклик.

### Итог

$y \in \mathbb{R}$  — значения наблюдаемого отклика

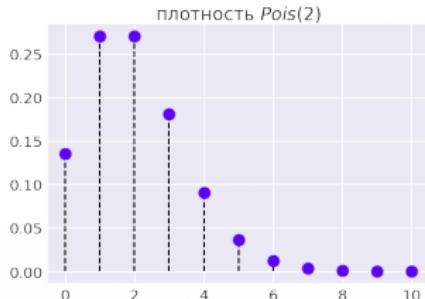
$E_x Y$  — ожидаемый отклик

$Y_i \sim \mathcal{N}(x_i^T \theta, \sigma^2)$  — наблюдаемый отклик

# Пуассоновское распределение

$$Pois(\lambda) : p(x) = \frac{\lambda^x}{x!} e^{-\lambda}, x \in \mathbb{Z}_+$$

**Смысл:** число событий, произошедших за единицу времени



## Условия:

1. события происходят с фиксированной интенсивностью  $\lambda$ .
2. независимо друг от друга.

**Утверждение:** время между двумя событиями имеет распр.  $Exp(\lambda)$  (см. пуассоновские случайные процессы)

## Примеры:

1. число клиентов в час
2. число запросов на сервер за минуту

# События разной интенсивности

Интенсивность может зависеть от каких-то факторов.

$X_1, \dots, X_n$  — факторы интервала времени

$Y_1, \dots, Y_n$  — число событий, произошедших за интервал времени

Тем самым имеется  $\lambda(x)$  — интенсивность событий для факторов  $x$ .

Получаем  $Y_i \sim Pois(\lambda(x_i))$

## Что предсказывать?

Нет смысла предсказывать сам  $Y_i$ ,

т.к. помимо  $\lambda(x_i)$  он содержит непрогнозируемый шум.

Тогда оценим  $EY_i = \lambda(x_i)$  — ожидаемый отклик.

Как параметризовать  $\lambda(x)$  для линейной модели?

## Определимся с требованиями

Пусть значению  $x_0^T \theta = 0$  соответствует интенсивность  $\lambda_0 = 1$ .

Значению  $x_1^T \theta$  сопоставим интенсивность  $\lambda_1 = 5$  событий в час.

Хотим чтобы значению  $-x_1^T \theta$  соответствовала

интенсивность  $1/\lambda_1 = 0.2$  событий в час.

## Линеаризация

Соответственно, нужно взять  $\lambda_\theta(x) = \exp(x^T \theta)$ .

Тогда  $\ln \lambda_\theta(x) = x^T \theta$ .

$g(z) = \ln z$  — **линеаризация** ожидаемого отклика.

## Итог

$y \in \mathbb{Z}_+$  — значения наблюдаемого отклика

$\lambda_\theta(x) = E_x Y$  — ожидаемый отклик

$Y_i \sim Pois(\lambda_\theta(x_i))$  — наблюдаемый отклик

Это **пуассоновская регрессия**.

# Случай бинарной классификации

$X_1, \dots, X_n$  — признаки объекта

$Y_1, \dots, Y_n$  — бинарный класс

Тем самым имеется  $\rho(x)$  — вероятность класса 1 для объекта  $x$ .

Получаем  $Y_i \sim Bern(\rho(x_i))$

**Что предсказывать?**

Оцениваем  $EY = \rho(x)$  — ожидаемый отклик.

Как параметризовать  $\rho(x)$  для линейной модели?

## Определимся с требованиями

Пусть значению  $x_0^T \theta = 0$  соответствует вероятность 0.5.

Значению  $x_1^T \theta$  сопоставим вероятность  $\rho_0 = 0.9$ .

Хотим чтобы значению  $-x_1^T \theta$  соответствовала вероятность 0.1.

Что такое в 2 раза более/менее вероятно?

Возможно, 0.95 и 0.05, но это не точно.

## Сведение к параметру масштаба

Заметим, что  $\frac{\rho}{1-\rho}$  — насколько чаще выпадает класс 1

по сравнению с классом 0. Тем самым это параметр масштаба.

## Линеаризация

Логит  $g(z) = \ln \frac{z}{1-z}$  — линеаризация ожидаемого отклика.

Т.е. параметризация  $\rho_\theta(x)$  должна быть такой, что  $\ln \frac{\rho_\theta(x)}{1-\rho_\theta(x)} = x^T \theta$ .

Тогда нужно взять  $\rho_\theta(x) = \frac{1}{1+\exp(-x^T \theta)}$ .

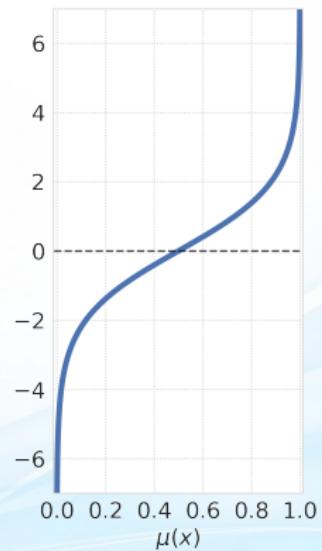
# Бинарный отклик

$y \in \{0, 1\}$  — значения наблюдаемого отклика

$\rho_\theta(x) = P_x(Y = 1)$  — ожидаемый отклик

$Y_i \sim Bern(\rho_\theta(x_i))$  — наблюдаемый отклик

Это **логистическая регрессия**.



# Обобщенная модель линейной регрессии

## Гауссовская линейная модель

Ожидаемый отклик:

$$y = \mu_\theta(x) = x^T \theta.$$

Наблюдаемый отклик:

$$Y_i = x_i^T \theta + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2).$$

$$\text{или } Y_i \sim \mathcal{N}(\mu_\theta(x_i), \sigma^2)$$

Оценка отклика:

$$\hat{y} = x^T \hat{\theta}.$$

## Generalized Linear Models (GLM)

Ожидаемый отклик:

$$y = \mu_\theta(x), \text{ причем } g(\mu_\theta(x)) = x^T \theta,$$

т.е.  $g$  — линеаризация ожид. отклика

Наблюдаемый отклик:

$$Y_i \sim P_{\mu_\theta(x_i)},$$

где  $\{P_\psi \mid \psi \in \Psi\}$  — семейство распр.

Оценка отклика:

$$\hat{y} = g^{-1}(x^T \hat{\theta}).$$

# Свойства GLM

В качестве  $\hat{\theta}$  берется ОМП (ищется численно)

$$L_X(\theta) = \prod_{i=1}^n p_{\mu_\theta(x_i)}(Y_i) = \prod_{i=1}^n p_{g^{-1}(x_i^T \theta)}(Y_i) \longrightarrow \max_{\theta}$$

Если  $\{P_\psi \mid \psi \in \Psi\}$  лежит в экспоненциальном классе, то  $\hat{\theta}$ :

1. существует и единственна;
2. состоятельна;
3. асимптотически нормальна:  $\sqrt{I(\theta)} (\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, I_d)$ ,

где  $I(\theta) = \left(-E \frac{\partial^2 \log L_X(\theta)}{\partial \theta_j \partial \theta_k}\right)_{jk}$  — информационная матрица Фишера.

Частные случаи:

1. Линейная (гауссовская):  $I(\theta) = \sigma^{-2} X^T X$ .
2. Логистическая:  $I(\theta) = X^T \cdot \text{diag} [\sigma(x_i^T \theta) (1 - \sigma(x_i^T \theta))] \cdot X$ .
3. Пуассоновская:  $I(\theta) = X^T \cdot \text{diag} [\exp(x_i^T \theta)] \cdot X$ .

Примечание. Свойства работают если верны предположения модели.

## Асимпт. доверительные интервалы в GLM

Для параметров ( $\implies$  критерий для гипотезы  $H_0: \theta_j = 0$ )

$$\theta_j \in \left( \hat{\theta}_j \pm z_{1-\alpha/2} \sqrt{\left( I^{-1}(\hat{\theta}) \right)_{jj}} \right)$$

Для преобразованного ожидаемого отклика

$$x_0^T \theta \in \left( x_0^T \hat{\theta} - \delta, x_0^T \hat{\theta} + \delta \right)$$

Для ожидаемого отклика

$$\mu(x_0) = g^{-1}(x_0^T \theta) \in \left[ g^{-1}\left(x_0^T \hat{\theta} - \delta\right), g^{-1}\left(x_0^T \hat{\theta} + \delta\right) \right],$$

$$\delta = z_{1-\alpha/2} \sqrt{x_0^T I^{-1}(\hat{\theta}) x_0}$$

Примечание. Для линейной регрессии вместо  $\sigma^2$  нужно взять ее несмешц. оценку.

# Проверка линейности логита в логистической регрессии

## Сглаженные диаграммы рассеяния

Пусть  $(x_1, Y_1), \dots, (x_n, Y_n)$  — обучающая выборка, где  $Y_i \in \{0, 1\}$ .

Выберем признак  $j$  и построим ядерную регрессию  $y \sim x_j$ :

$$\hat{y}(x_j) = \sum_{i=1}^n q\left(\frac{x_j - x_{ij}}{h}\right) Y_i \Bigg/ \sum_{i=1}^n q\left(\frac{x_j - x_{ij}}{h}\right),$$

где  $x_{ij}$  — признак  $j$  объекта  $i$ ,

$x_j$  — признак  $j$  нового объекта (по сетке),

$q$  — ядро,

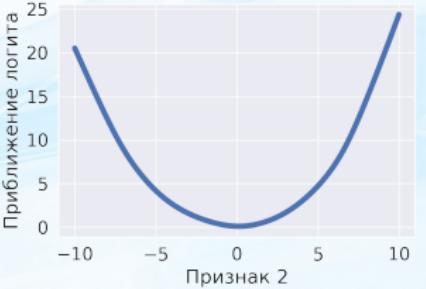
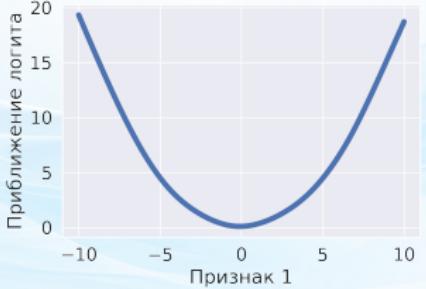
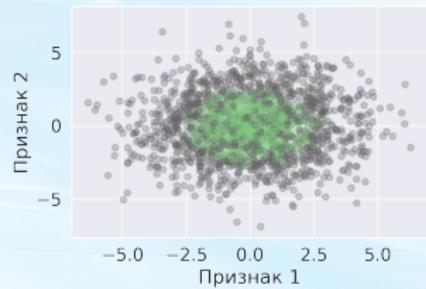
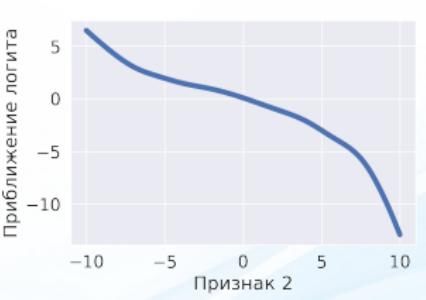
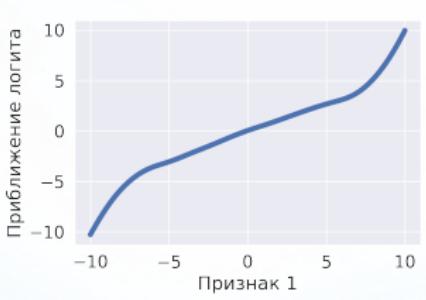
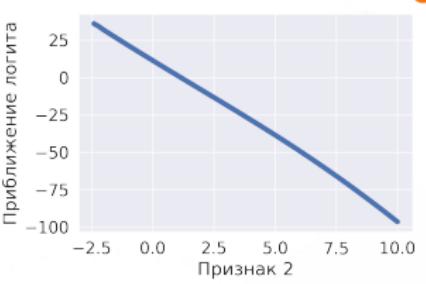
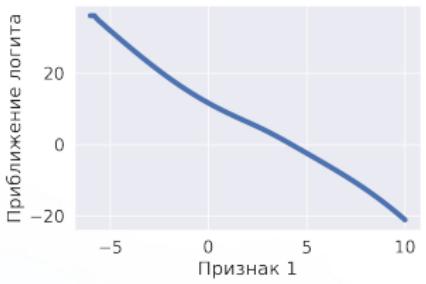
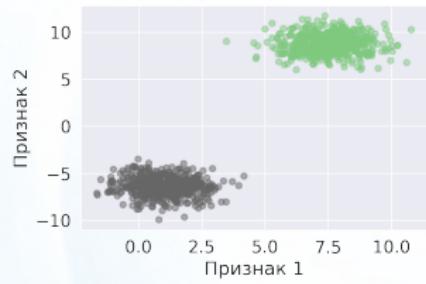
$h > 0$  — ширина ядра.

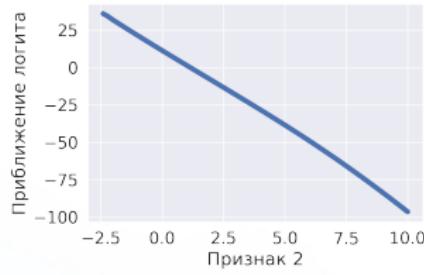
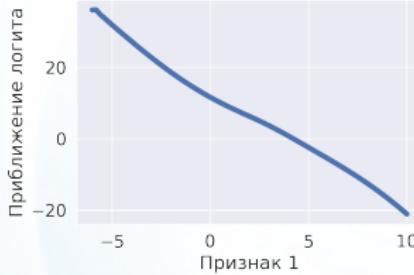
Эта регрессия — приближение вер-ти класса 1 в зависимости от  $x_j$ .

Отсюда делаем приближение логита:

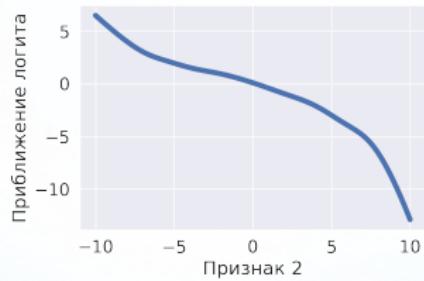
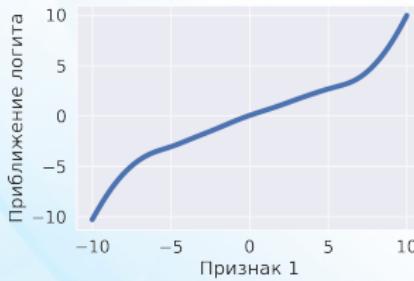
$$\text{logit}(x_j) = \log \frac{\hat{y}(x_j)}{1 - \hat{y}(x_j)}.$$

**Проверка:** график  $\text{logit}(x_j)$  похож на прямую.

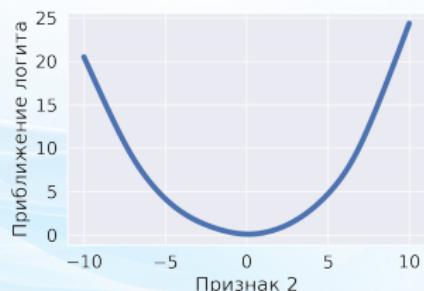
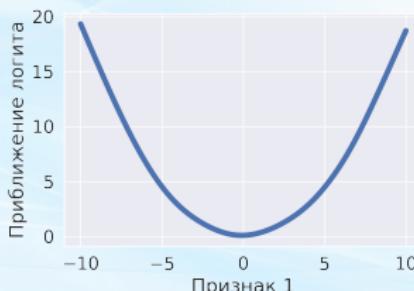




Логит линеен,  
все хорошо



Классы линейно  
разделимы, но зависи-  
мость нелинейна



Классы не являются  
линейно разделимыми

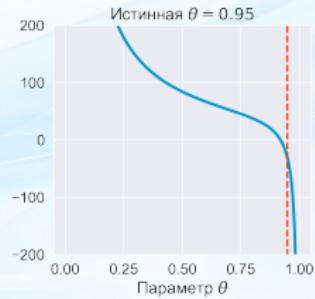
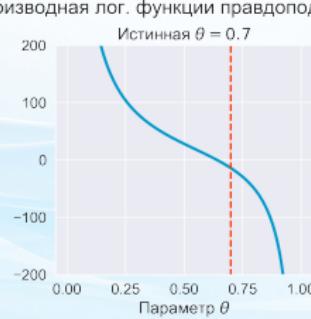
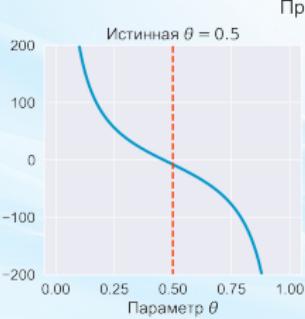
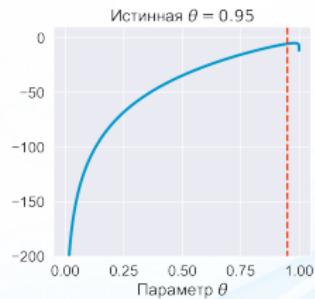
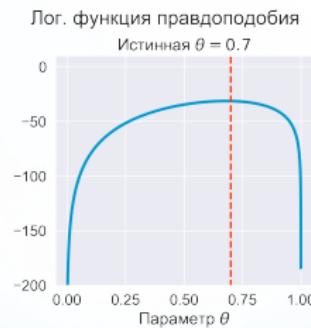
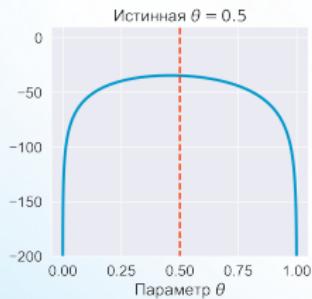
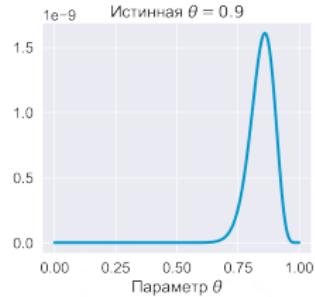
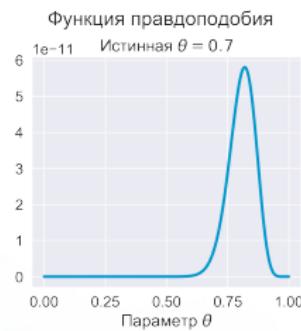
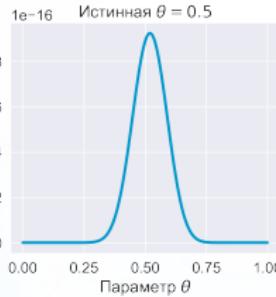
# 7. Теория наилучших оценок

## 7.1а Информация Фишера

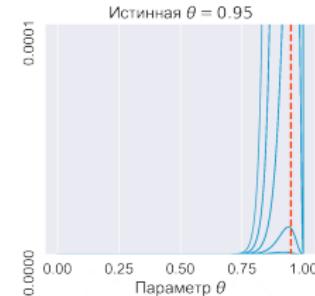
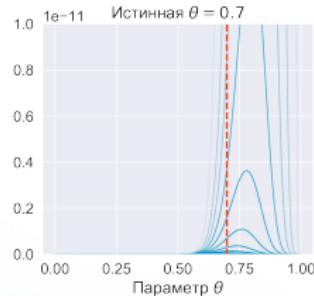
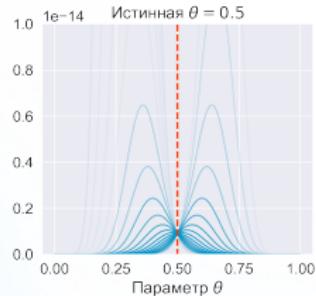
$$X = (X_1, \dots, X_n) \sim Bern(\theta)$$

Насколько легко оценить параметр в зависимости от того, чему равно его истинное значение?

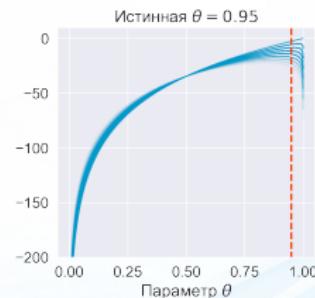
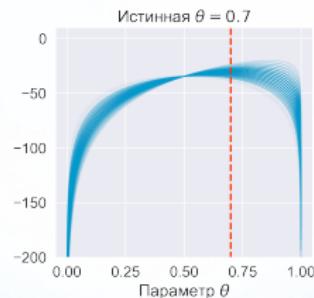
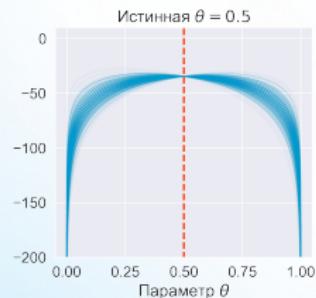
**Цель:** определить информацию о параметре  $\theta$ , которую несет выборка  $X$ .



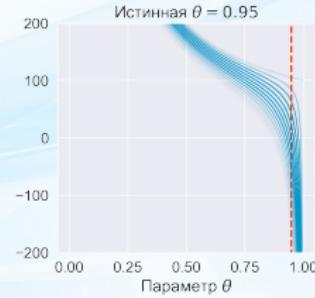
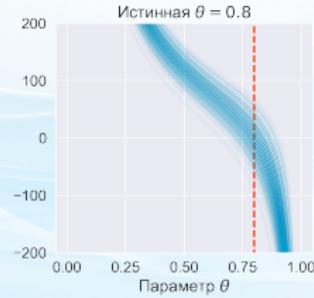
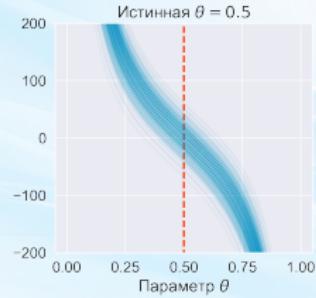
## Функция правдоподобия

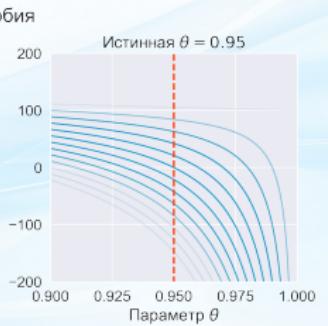
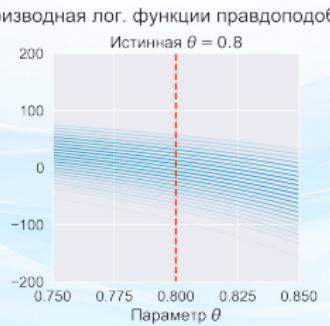
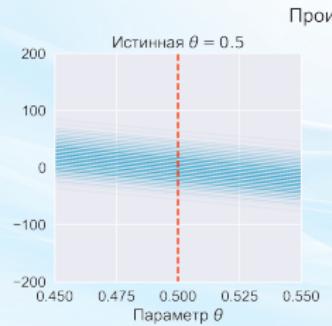
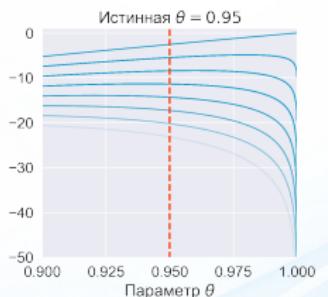
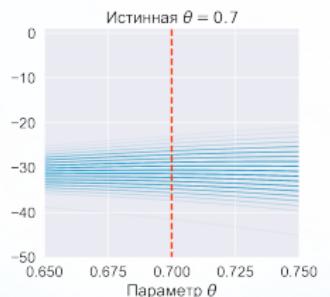
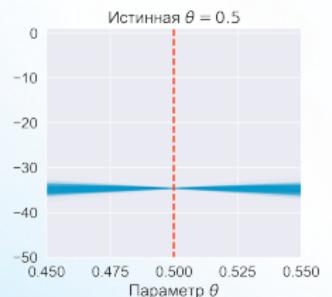
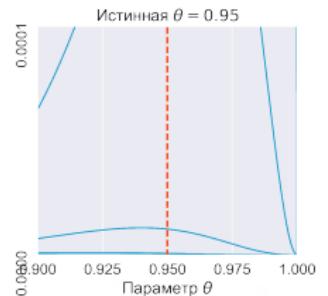
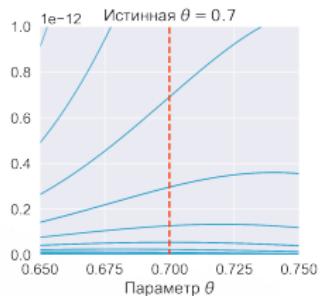
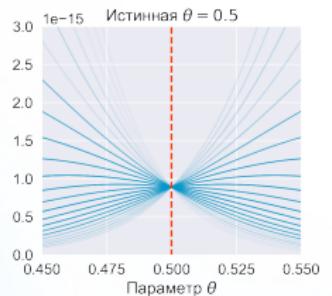


## Лог. функция правдоподобия



## Производная лог. функции правдоподобия

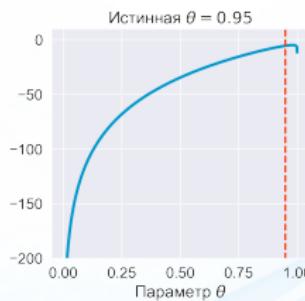
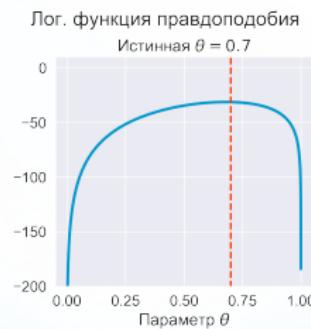
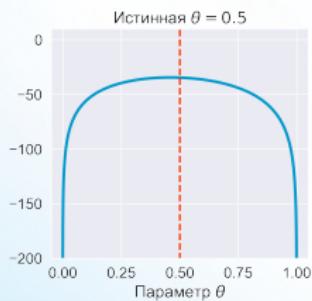
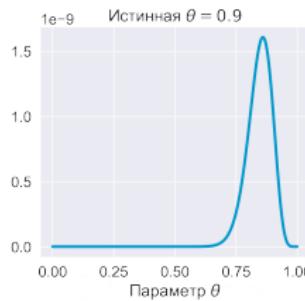
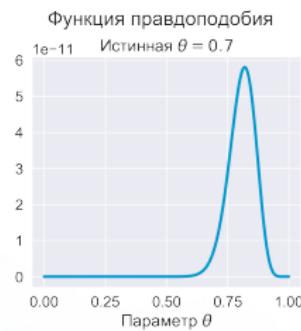
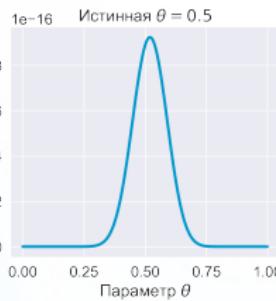




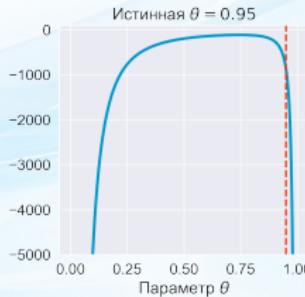
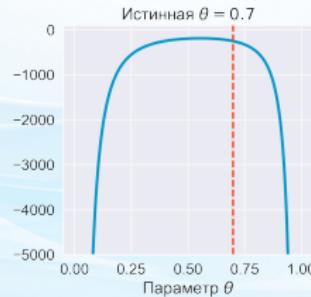
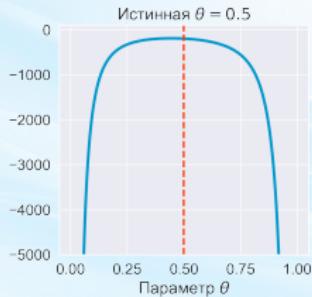
## Наблюдения:

1. Чем острее пик, тем больше  $L_X(\theta)$  реагирует на  $\theta$  в окрестности максимума
2. Чем острее пик, тем больше в нем дисперсия  $\frac{\partial \log L_X(\theta)}{\partial \theta}$
3. Информация Фишера выборки  $X$  о параметре  $\theta$ :

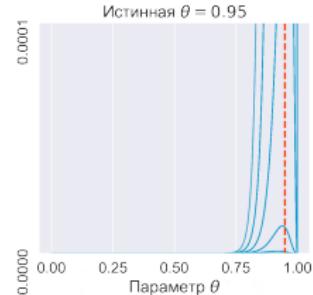
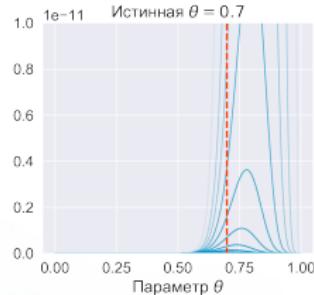
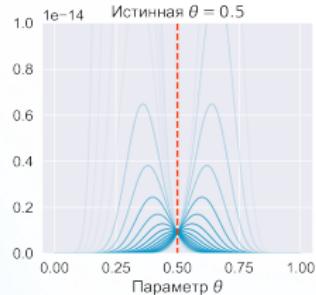
$$I_X(\theta) = D_\theta \frac{\partial \log L_X(\theta)}{\partial \theta}$$



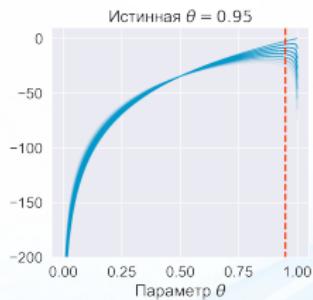
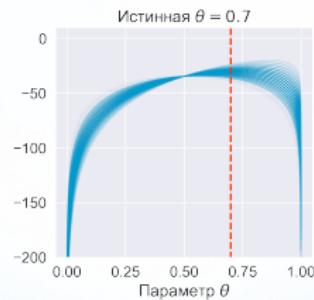
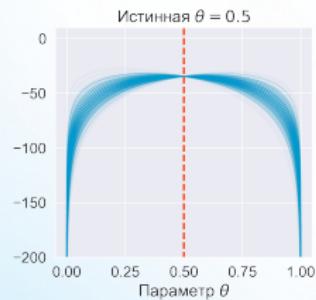
### 2-ая производная лог. функции правдоподобия



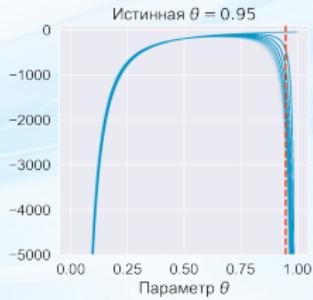
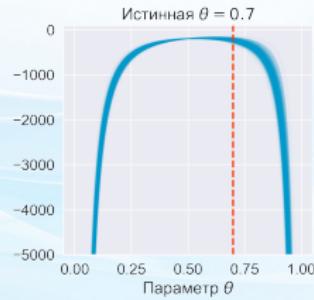
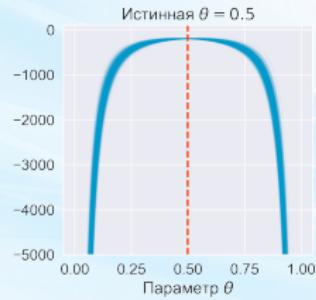
## Функция правдоподобия

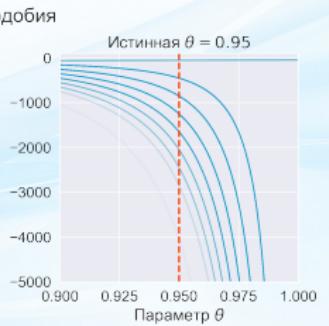
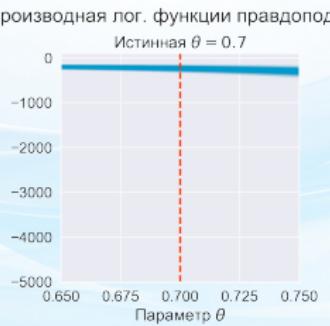
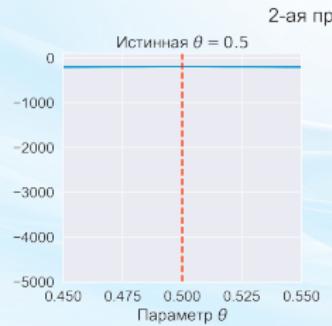
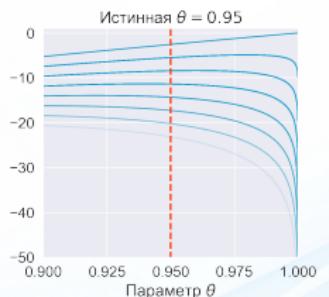
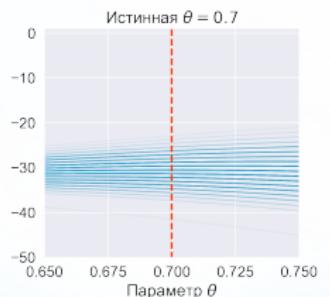
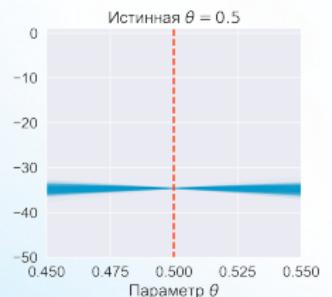
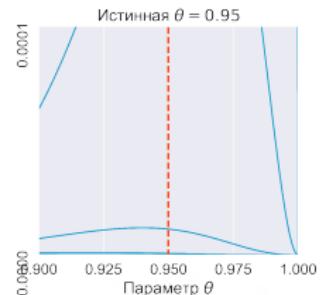
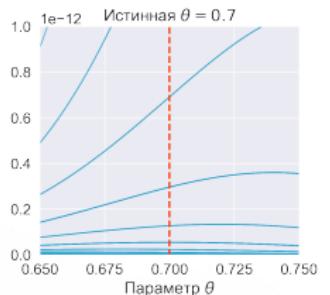
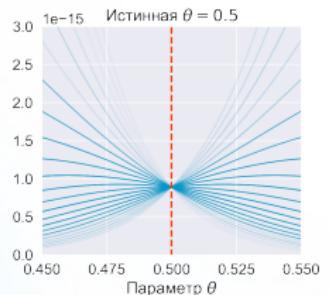


## Лог. функция правдоподобия



## 2-ая производная лог. функции правдоподобия



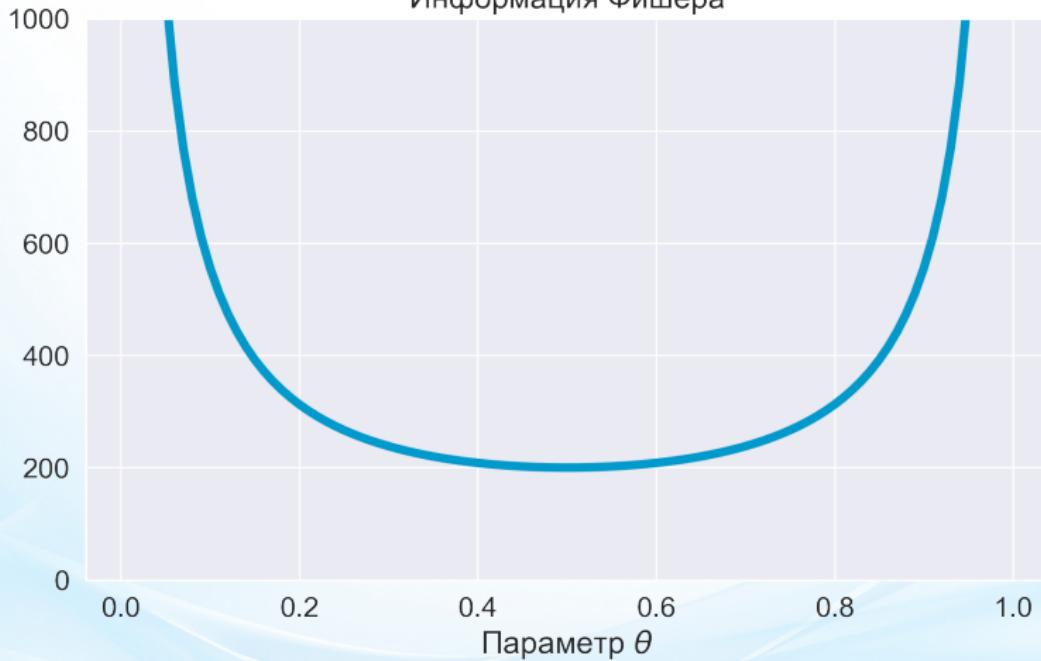


## Наблюдения:

1. Чем острее пик, тем больше  $L_X(\theta)$  реагирует на  $\theta$  в окрестности максимума  
     $\Rightarrow$  проще оценить  $\theta$   
     $\Rightarrow$  выборка содержит больше информации о  $\theta$
2. Чем острее пик, тем более вогнута  $L_X(\theta)$
3. Информация Фишера выборки  $X$  о параметре  $\theta$ :

$$I_X(\theta) = -E_\theta \frac{\partial^2 \log L_X(\theta)}{\partial \theta^2}$$

## Информация Фишера



## Информация Фишера



Theta



BCE !