Введение в математическую статистику II

Computer Science Club, 27 ноября 2021

Данные

Данные d — это реализация случайного элемента D, имеющего (неизвестное) распределение $\mathcal{P}_D$.

Примеры:

- 1 Тб фотографий с кошечками и собачками
- Цена биткоина за последний месяц
- Результат броска монетки
- Количество антител у участников испытания вакцины

Статистическая модель

Статистическая модель — это множество распределений $\mathfrak{P}$, которое, по нашему мнению, адекватно приближает $\mathcal{P}_D$.

Примеры:

- d — результат 10 подбрасываний монетки, $\mathfrak{P} = \left\{B(p)^{\otimes 10} \,\middle|\, p \in (0,1)\right\}$.
- d — 1 Тб фоточек, $\mathfrak{P}$ — все фоточки независимы, каждая фоточка имеет распределение, приближенно описываемое GAN'ом.

Статистическая модель

Статистические модели делят на

- параметрические, если $\mathfrak{P} = \{\mathcal{P}_\theta \big| \theta \in \Theta \subset \mathbb{R}^k\}$;

  Пример: $\mathfrak{P} = \{\mathcal{N}(\mu, \sigma^2) | \mu \in \mathbb{R}, \sigma^2 \geq 0\}$
- непараметрические, если $\mathfrak{P} = \{\mathcal{P}_\theta | \theta \in \Theta \subset V\}$, V не обязано быть конечномерным;

  Пример: $\mathfrak{P} = \left\{\mathcal{P}^{\otimes n} \middle| \int_{\mathfrak{X}} x\mathcal{P}(dx) = 0\right\}$
- семипараметрические, если $\mathfrak{P} = \{\mathcal{P}_\theta \big| \theta \in \Theta \subset \mathbb{R}^k \times V\}$;

  Пример: линейная регрессия $Y = X\beta + \varepsilon$, $\beta \in \mathbb{R}^k$, $\mathbb{E}\varepsilon = 0$, $\mathbb{D}\varepsilon = \sigma^2$

В байесовской статистике в модель также включается априорное распределение на Θ.

Пример: $\mathfrak{P} = \{\mathcal{N}(\mu, 1) \big| \mu \sim \mathcal{N}(\mu_0, \sigma_0^2)\}$.

PDMI COMPUTER SCIENCE CLUB

Выборка

Если $D = [X_1, \dots, X_n]$, X_i независимы и имеют одинаковое распределение, то D называется выборкой объема n и обозначается как $X_{[n]}$. Распределение $\mathcal{P}_X$, которому подчинены все X_i, называется генеральной совокупностью.

Модель принимает вид $\mathfrak{P} = \left\{\mathcal{P}^{\otimes n} \middle| \mathcal{P} \in \mathfrak{P}_X\right\}$, где $\mathfrak{P}_X$ — модель для $\mathcal{P}_X$.

Предположение о том, что некоторый набор данных является выборкой — это сильное предположение; всегда нужно подумать о том, адекватно ли оно!

Почему выборка? Количество → качество

Для выборки $x_{[n]}$ с одномерным $\mathcal{P}_X$ рассмотрим

– $\mathcal{P}_n^*\left(A \middle| X_{[n]} = x_{[n]}\right) = \frac{1}{n}\sum_{i=1}^n [x_i \in A]$ — эмпирическое распределение,

– $F_n^*\left(x \middle| X_{[n]} = x_{[n]}\right) = \frac{1}{n}\sum_{i=1}^n [x_i < x]$ — эмпирическую функцию распределения.

Теорема Гливенко—Кантелли

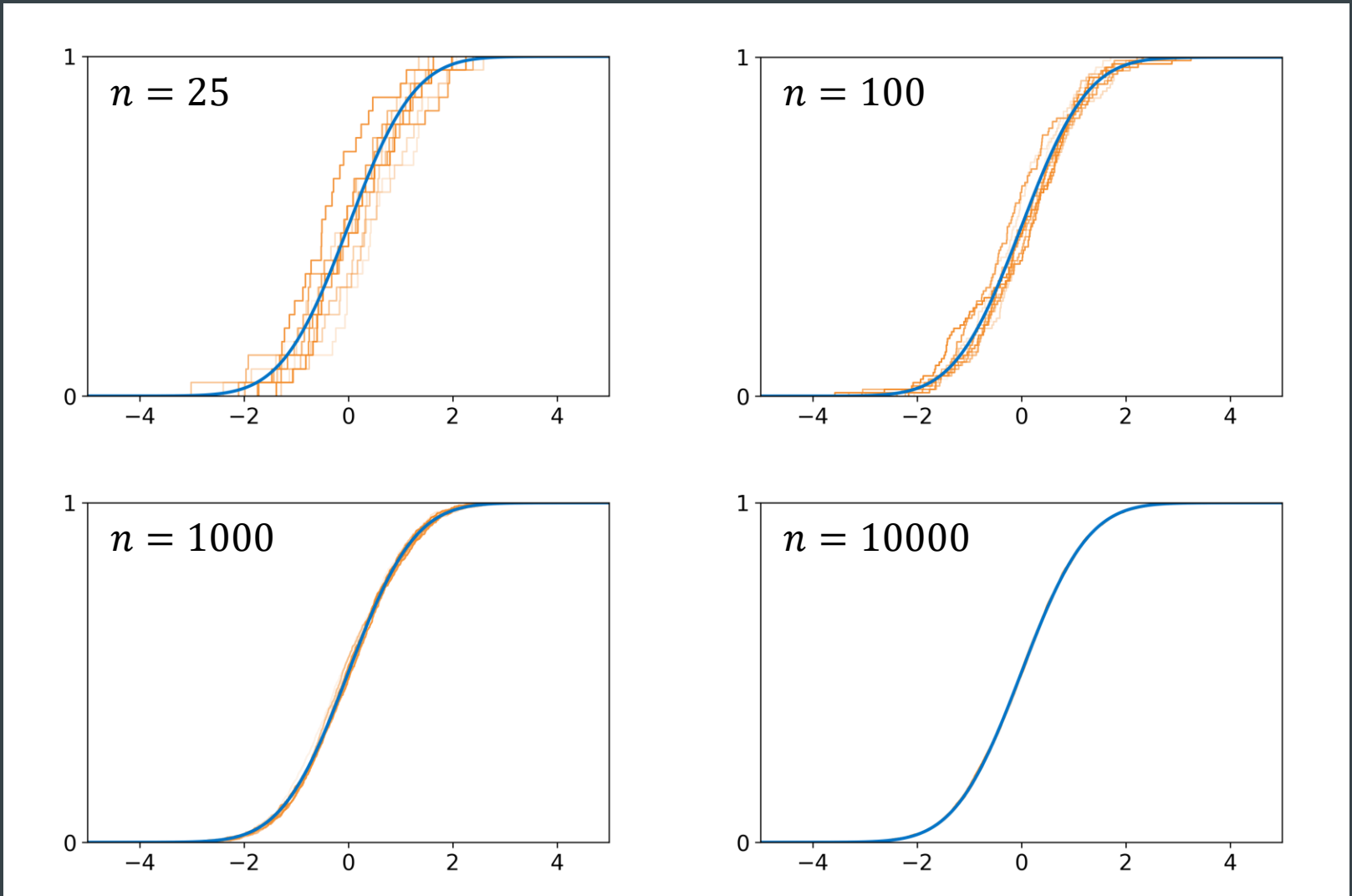
$$\sup_{x\in\mathbb{R}} |F_n^*(x) - F_X(x)| = \sup_{x\in\mathbb{R}} \left|\frac{1}{n}\sum_{i=1}^n [X_i < x] - F_X(x)\right| \xrightarrow{\text{п.н.}} 0.$$

Теорема Колмогорова

Если F_X непрерывна, то $\sqrt{n}\sup_{x\in\mathbb{R}} |F_n^*(x) - F_X(x)| \xrightarrow{d} K$ — распределение Колмогорова.

PDMI COMPUTER SCIENCE CLUB

Почему выборка? Количество → качество



PDMI
COMPUTER
SCIENCE
CLUB

Выборочные характеристики: моменты

- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ — выборочное среднее
- $\tilde{s}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$ — выборочная дисперсия, $s^2 = \frac{n}{n-1}\tilde{s}^2$ — исправленная дисперсия
- $\tilde{s}$ — выборочное стандартное отклонение, s — исправленное стандартное отклонение
- $\alpha_k^* = \frac{1}{n}\sum_{i=1}^{n} x_i^k$ — k-й выборочный момент, $\bar{x} = \alpha_1^*$
- $\mu_k^* = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^k$ — k-й центральный выборочный момент, $\tilde{s}^2 = \mu_2^*$

PDMI
COMPUTER
SCIENCE
CLUB

Вариационный и статистический ряды

Отсортируем выборку: $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$. Это вариационный ряд.

$x_{(k)}$ — k-я порядковая статистика.

Посчитаем уникальные значения: $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)} \to z_{(1)} < \cdots < z_{(k)}$.

	$z_{(1)}$	$z_{(2)}$	…	$z_{(k)}$
кратность	n_1	n_2	…	n_k
частота	n_1/n	n_2/n	…	n_k/n

Это статистический ряд.

PDMI COMPUTER SCIENCE CLUB

Квантили, квартили, медиана

Квантиль уровня α — число x_α (любое) такое, что $\mathbb{P}(X \le x_\alpha) \ge \alpha$ и $\mathbb{P}(X \ge x_\alpha) \ge 1 - \alpha$.

Выборочный квантиль x_α^*:

- если $\frac{k}{n} < \alpha < \frac{k+1}{n}$, то $x_\alpha^* = x_{(k+1)}$,
- если $\alpha = \frac{k}{n}$, то x_α^* — любое число в интервале $[x_{(k)}, x_{(k+1)}]$.

Выборочный квартиль:

$Q_1 = x_{0.25}^*$, $Q_2 = x_{0.5}^*$, $Q_3 = x_{0.75}^*$

Выборочная медиана med^*:

- если $n = 2k + 1$, то $med^* = x_{(k+1)}$
- если $n = 2k$, то $med^* = \frac{1}{2}\left(x_{(k)} + x_{(k+1)}\right)$

PDMI COMPUTER SCIENCE CLUB

Характеристики положения и разброса

Характеристики положения

- Выборочное среднее
- Выборочная медиана
- Выборочная мода
- Квантили, квартили…

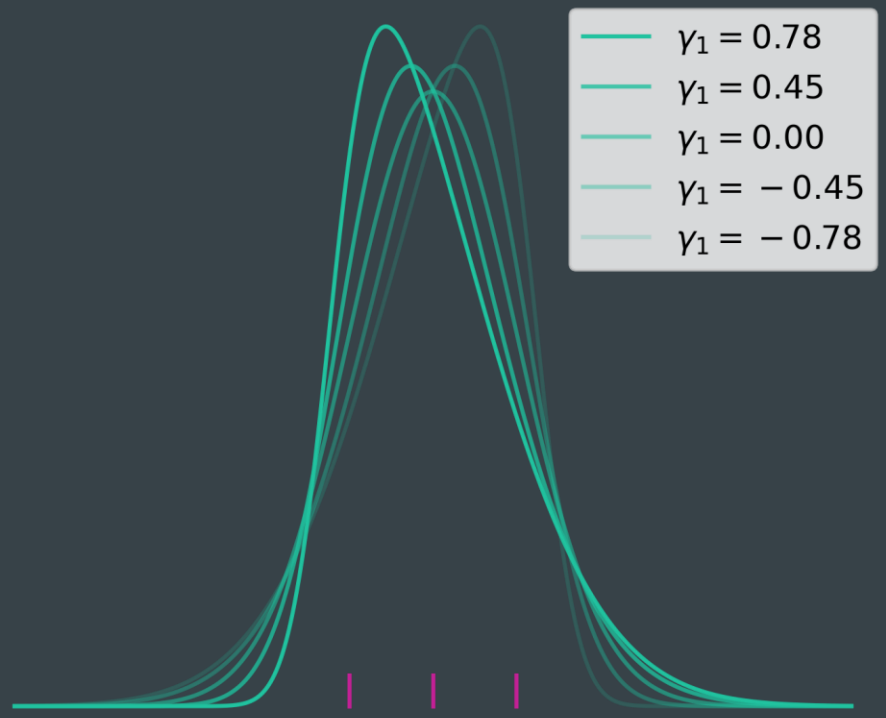
Характеристики разброса

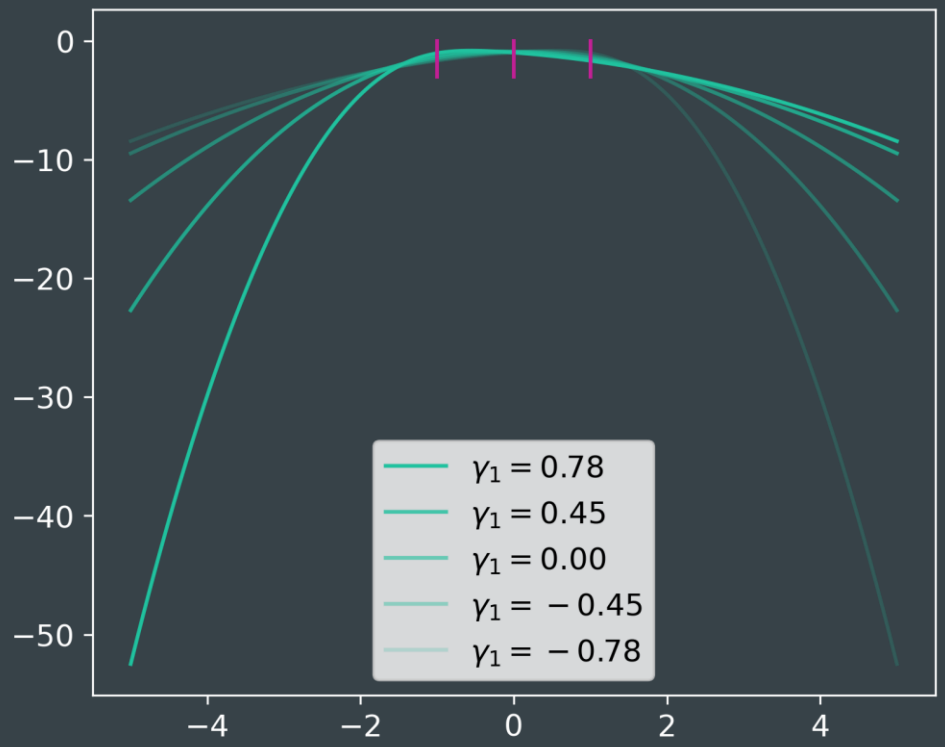
- Размах: $R = x_{(n)} - x_{(1)}$
- Межквартильный размах: $IQR = Q_3 - Q_1$
- Выборочное стандартное отклонение и исправленное стандартное отклонение
- Медианное абсолютное отклонение: $MAD = med(|x_{[n]} - med^*|)$

PDMI COMPUTER SCIENCE CLUB

Коэффициент асимметрии (skewness)

$$\gamma_1 = \frac{\mu_3^*}{(\mu_2^*)^{\frac{3}{2}}} = \frac{\mu_3^*}{\tilde{s}^3} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\tilde{s}}\right)^3$$

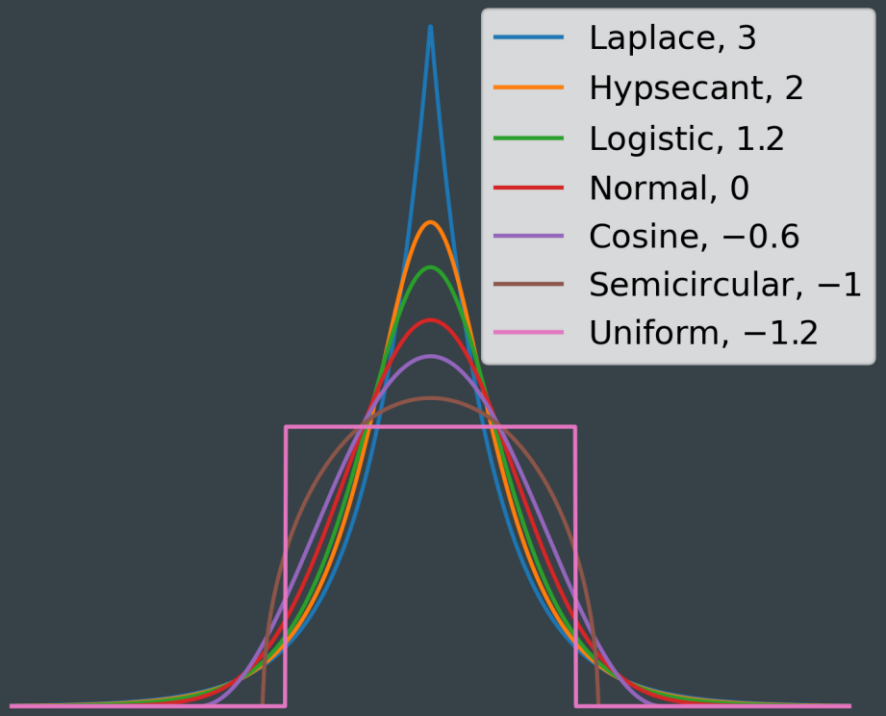


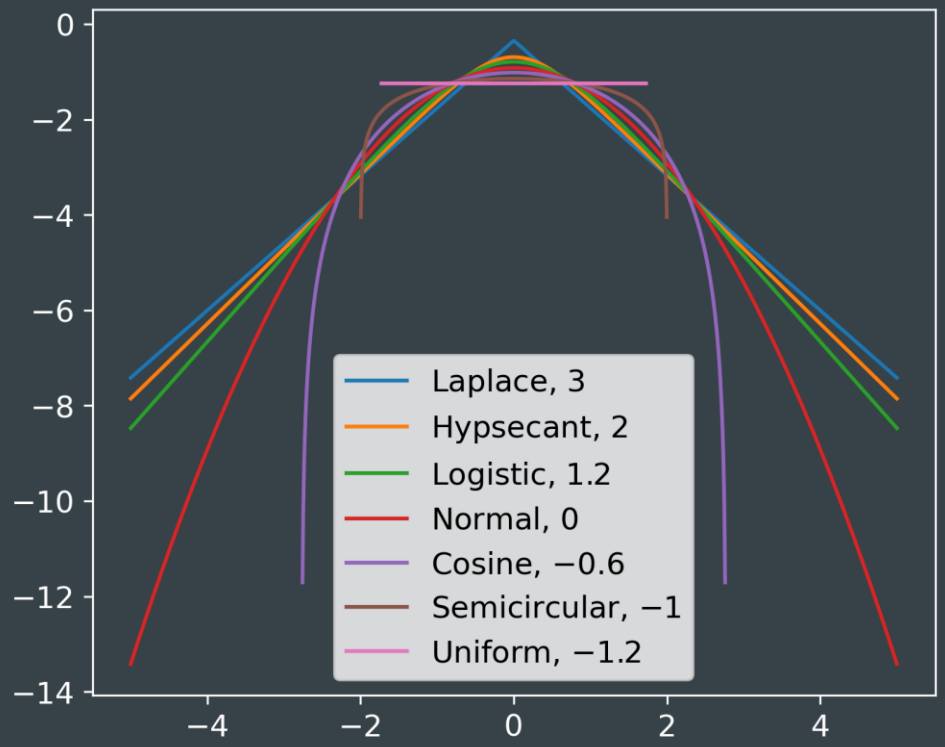


PDMI
COMPUTER
SCIENCE
CLUB

Коэффициент эксцесса (kurtosis)

$$\gamma_2 = \frac{\mu_4^*}{(\mu_2^*)^2} - 3 = \frac{\mu_4^*}{\tilde{s}^4} - 3 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\tilde{s}}\right)^4 - 3$$





PDMI
COMPUTER
SCIENCE
CLUB

Статистический вывод

Введение

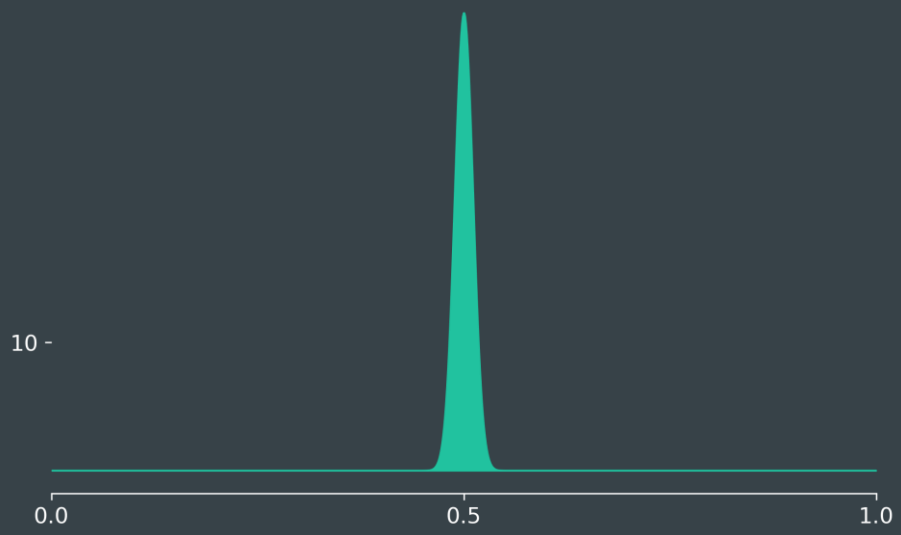
Задачи статистического вывода:

- Оценка параметров (и характеристик): точечные, интервальные
- Проверка гипотез
- Сравнение моделей
- ...

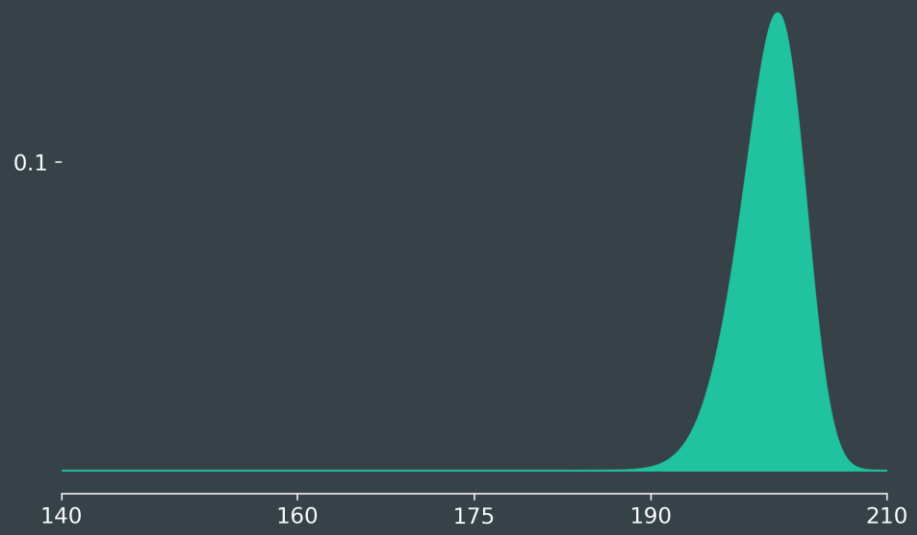
Байесовский подход

Основная идея — степень уверенности в значении величины численно выражаем вероятностью или плотностью.

Чем больше уверенность, тем больше вероятность.



вероятность выпадения орла
у монетки в вашем кармане



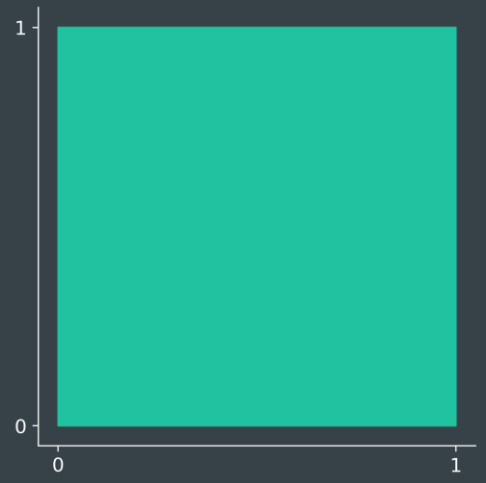
средний рост баскетболиста NBA

PDMI
COMPUTER
SCIENCE
CLUB

Априорное и апостериорное распределения

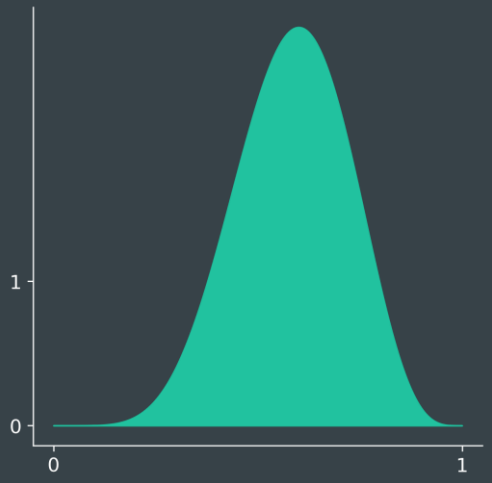
Наше ощущение относительно значения параметра до эксперимента называется **априорным распределением**.

В результате эксперимента мы получаем данные, которые некоторым образом меняют наше ощущение. Полученное ощущение называется **апостериорным распределением**.



априорное распределение

$$+ \quad d = [0,1,1,0,1,0,1,1,1,0] \quad =$$



апостериорное распределение

PDMI COMPUTER SCIENCE CLUB

Формула Байеса

p — вероятность
или плотность
d — данные
θ — параметры

правдоподобие
likelihood

априорное
распределение
prior

$$p(\theta|d) = \frac{p(d|\theta) \cdot p(\theta)}{p(d)}$$

posterior
апостериорное
распределение

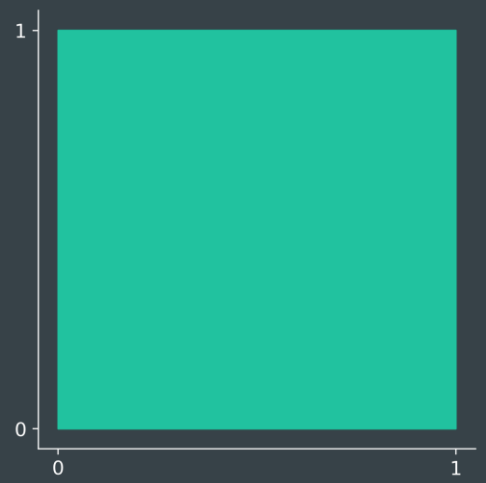
evidence
вероятность
данных

нормирующая
константа,
не зависит от θ!

PDMI
COMPUTER
SCIENCE
CLUB

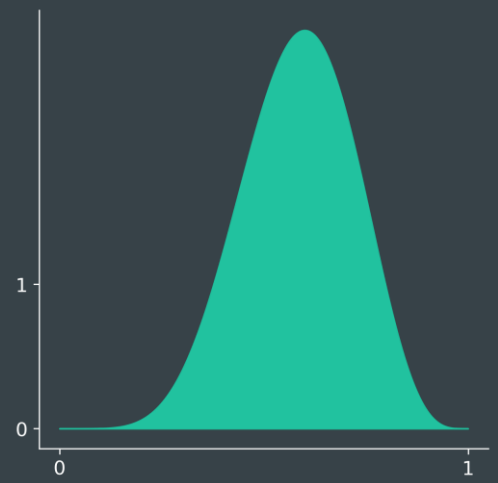
Априорное и апостериорное распределения

$$\Theta = [0,1], p(\theta) \equiv 1 \quad \times \quad p(d|\theta) = \theta^6(1-\theta)^4 \quad \propto \quad p(\theta|d) = \frac{\theta^6(1-\theta)^4}{\mathrm{B}(7,5)}$$



априорное распределение

$$+ \quad d = [0,1,1,0,1,0,1,1,1,0] \quad =$$



апостериорное распределение

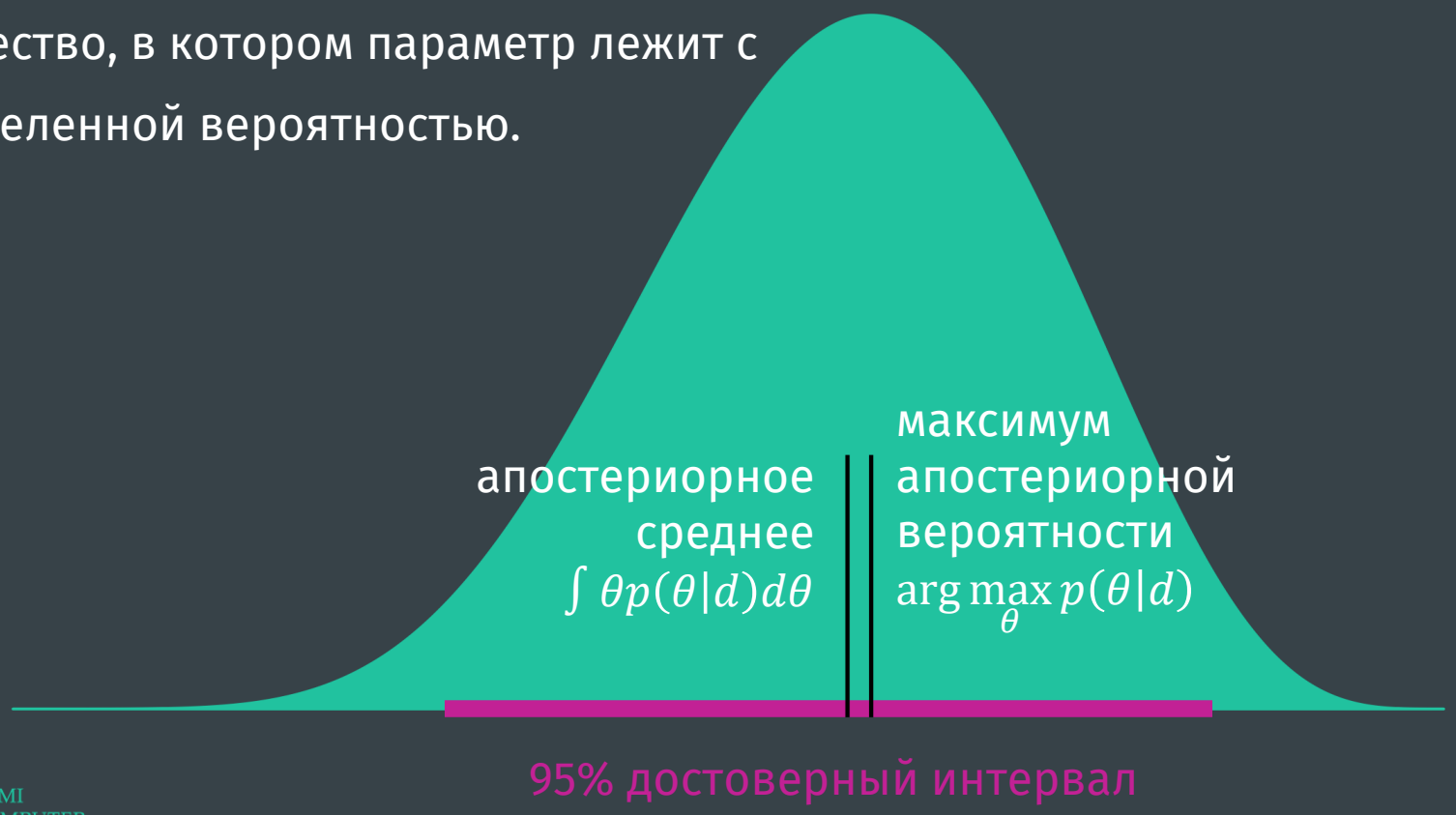
Распределение на $[0,1]$ с плотностью вида $p(\theta) = \theta^{\alpha-1}(1-\theta)^{\beta-1}/\mathrm{B}(\alpha,\beta)$, $\alpha, \beta > -1$, называется бета-распределением с параметрами α, β и обозначается как $Beta(\alpha, \beta)$.

PDMI COMPUTER SCIENCE CLUB

Оценка параметров

По апостериорному распределению можно оценивать параметры и погрешности.

Достоверный интервал (credible interval) — множество, в котором параметр лежит с определенной вероятностью.



PDMI COMPUTER SCIENCE CLUB

Проверка гипотез

Имеются гипотезы $\mathfrak{P}_1, \dots, \mathfrak{P}_k$ с априорными вероятностями $\mathbb{P}(\mathfrak{P}_1), \dots, \mathbb{P}(\mathfrak{P}_k)$, а также данные d.

априорные шансы	$\mathbb{P}(\mathfrak{P}_1) : \mathbb{P}(\mathfrak{P}_2) : \dots : \mathbb{P}(\mathfrak{P}_k)$
	$\times$
отношение правдоподобия	$\mathbb{P}(d \mid \mathfrak{P}_1) : \mathbb{P}(d \mid \mathfrak{P}_2) : \dots : \mathbb{P}(d \mid \mathfrak{P}_k)$
	$=$
апостериорные шансы	$\mathbb{P}(\mathfrak{P}_1 \mid d) : \mathbb{P}(\mathfrak{P}_2 \mid d) : \dots : \mathbb{P}(\mathfrak{P}_k \mid d)$

Если нас интересует, какая гипотеза скорее всего верна, то смотрим на апостериорные шансы.

Если нас интересует, какая гипотеза наиболее правдоподобна для наших данных, то смотрим на отношение правдоподобия.