

Введение в анализ данных

Лекция 5

Анализ данных, теория вероятностей и статистика

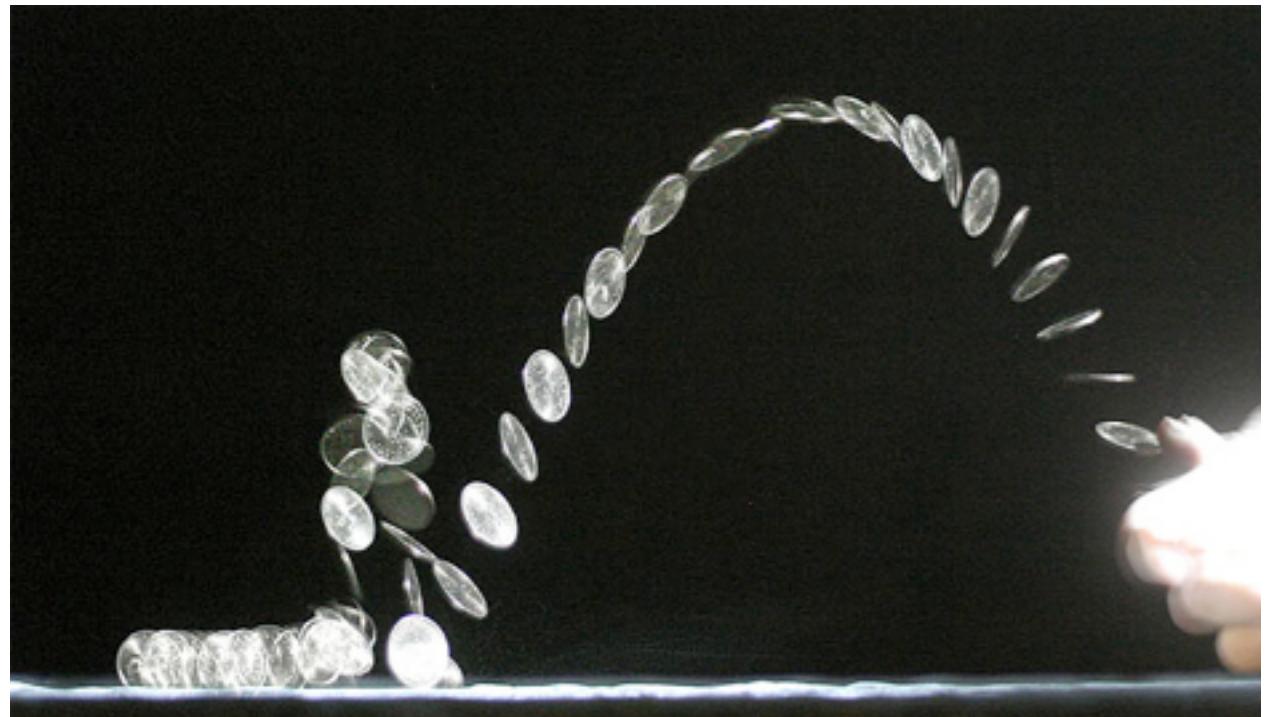
Евгений Соколов

sokolov.evg@gmail.com

НИУ ВШЭ, 2016

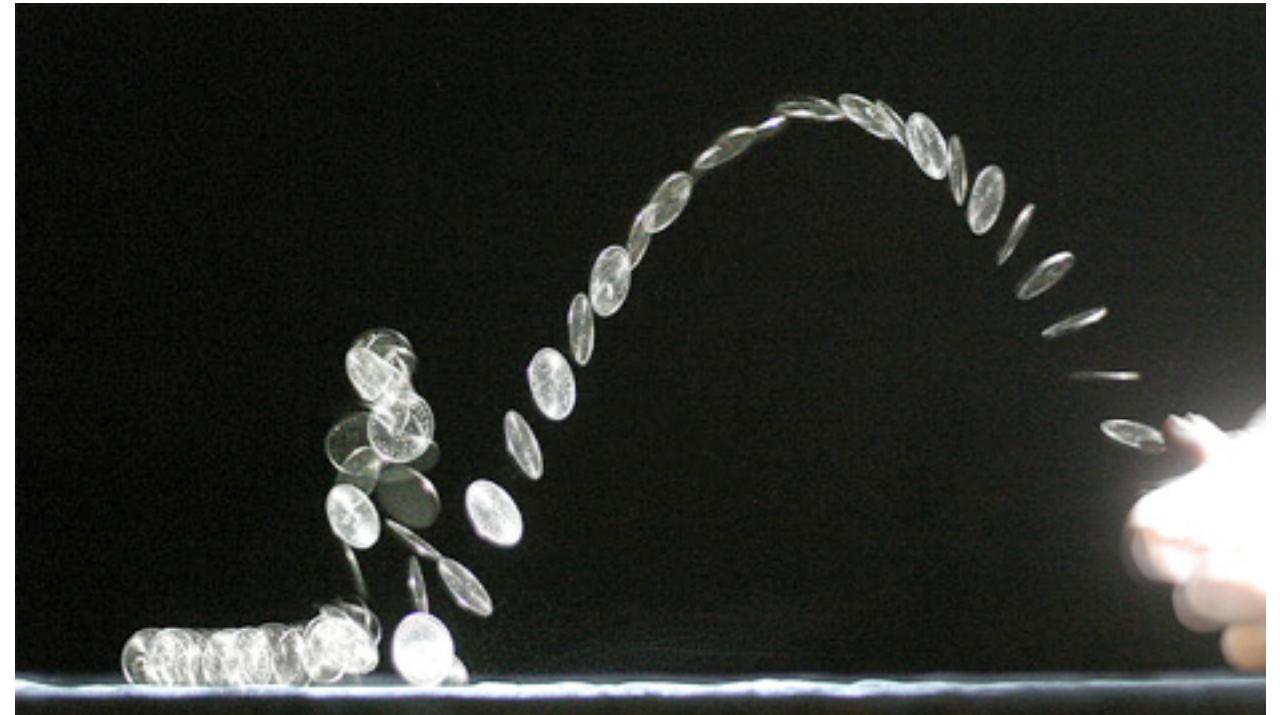
Подбрасывание монетки

- Случайный ли результат?



Подбрасывание монетки

- Случайный ли результат?
- Нет!
- Зависит от:
 - параметров броска
 - свойств монетки
 - свойств воздуха
 - ...
- Очень сложно описать с помощью формул



Случайность

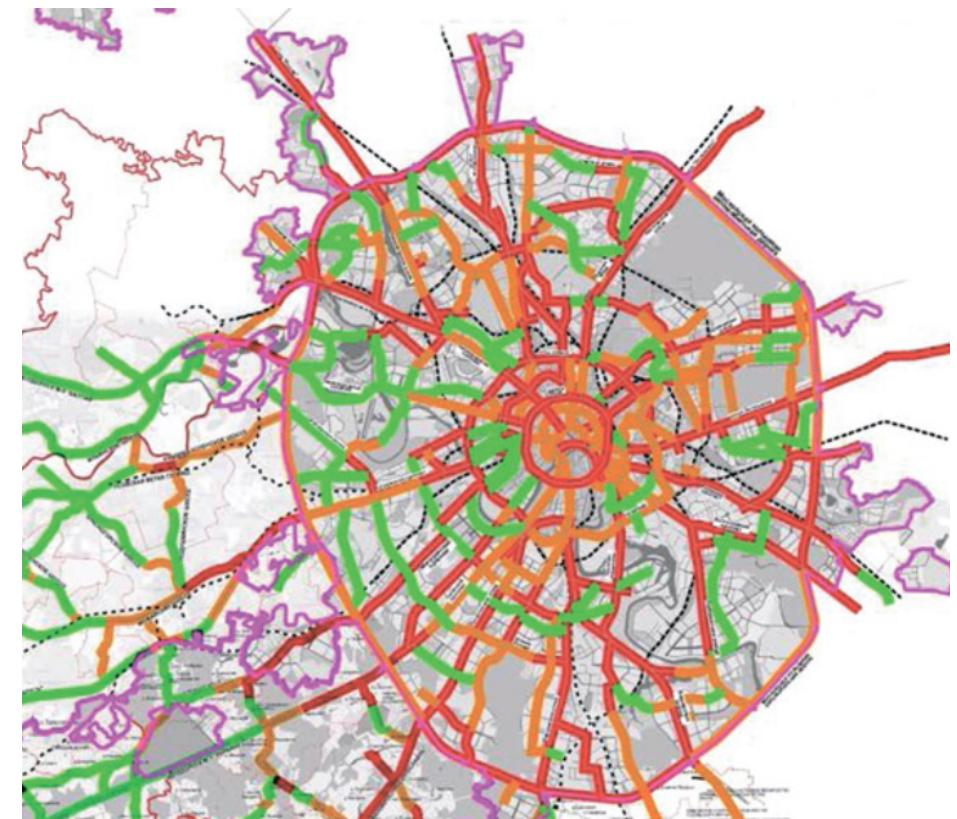
- Большинство процессов в мире можно описать уравнениями
- (кроме квантовых)
- Это может быть слишком сложно
- Проще считать исход случайным

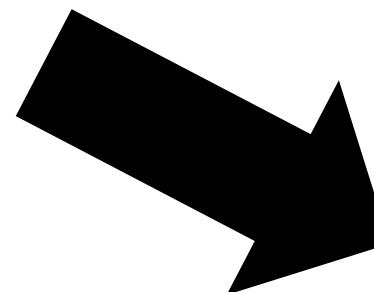
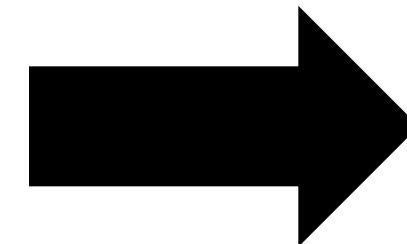
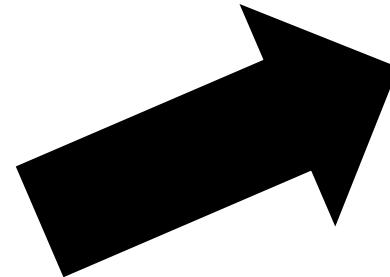
Статистическая физика

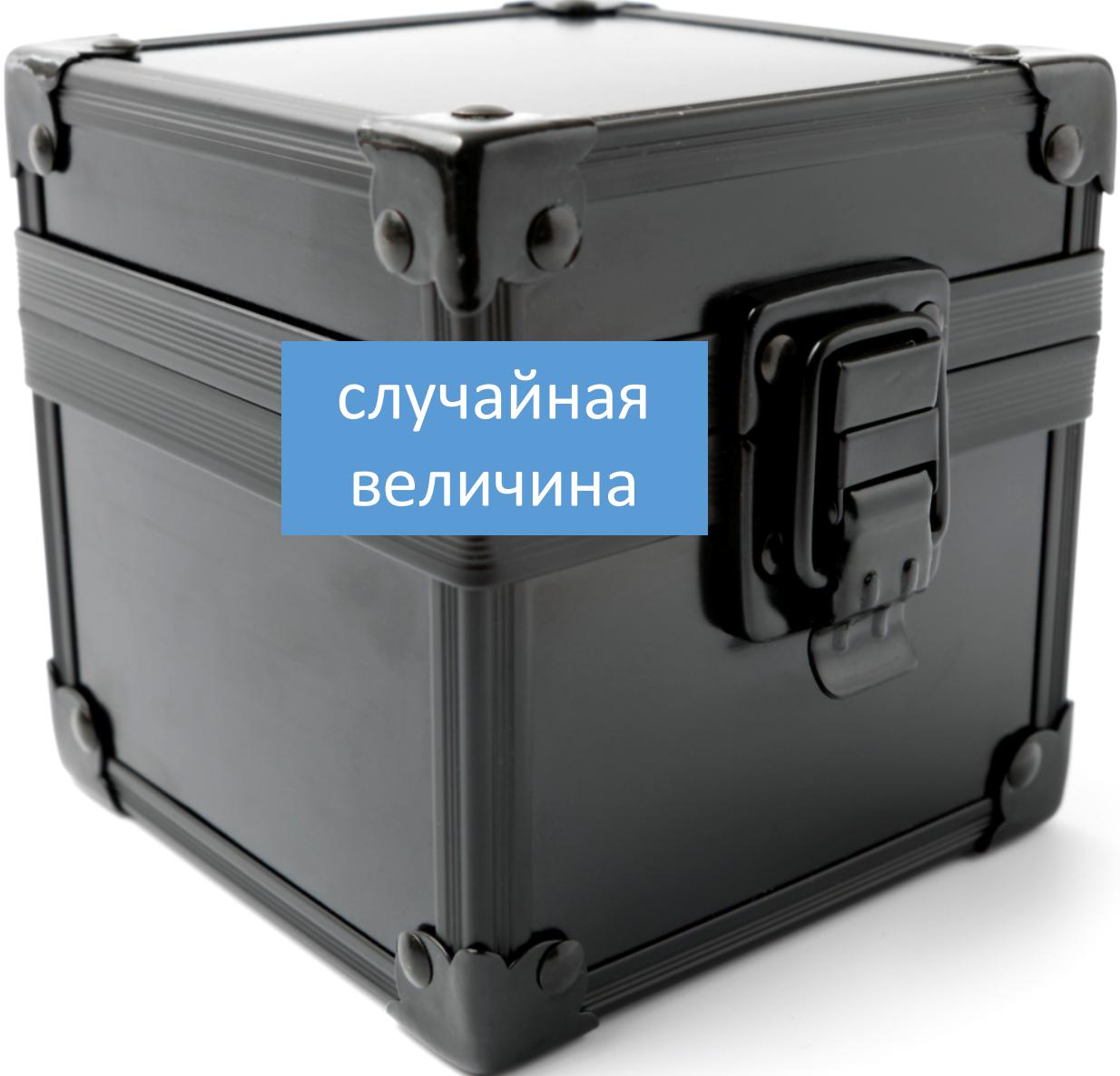
- В комнате очень много молекул
 - $\sim 10^{30}$ штук
- Поведение каждой легко описать
- Поведение воздуха в комнате — 10^{30} уравнений
- Проще объявить состояние воздуха случайным

Транспортные потоки

- Сотни тысяч машин в Москве
- Можно пытаться проследить траекторию каждой машины
- Проще считать положение машины случайным
- Распределение машин
- Много зависимостей







случайная
величина



реализации
случайной
величины

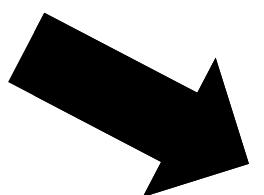
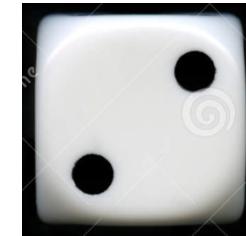
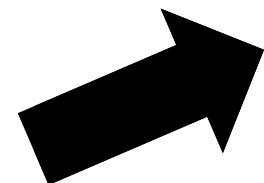


выборка

Случайность

- Случайная величина — функция, выдающая случайные значения
- Изучаем **вероятности событий**
- Какова вероятность того, что случайная величина выдаст конкретное значение?
- В какой доле случаев она примет это значение?

теория
вероятностей



статистика и
анализ данных

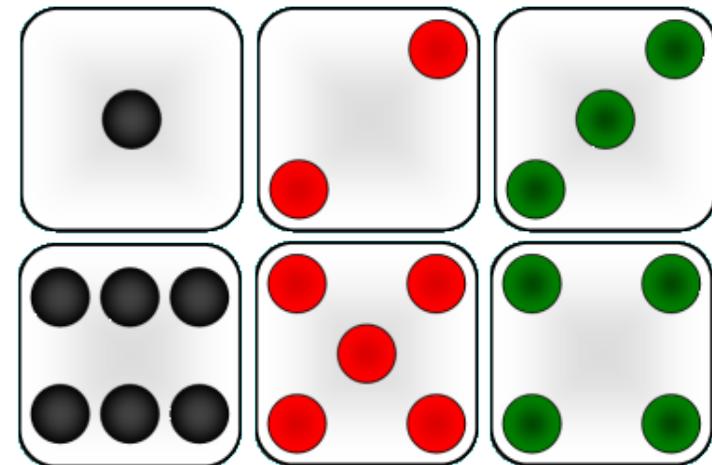
План на сегодня

- Основные определения теории вероятностей
- Дискретные случайные величины
- Основы статистического оценивания
- Классификация текстов: наивный байесовский классификатор

Основные понятия теории вероятностей

Вероятностное пространство

- Множество элементарных исходов Ω
 - грани кубика
- Множество событий \mathcal{F} — подмножества Ω
 - все множества граней кубика
- Вероятность $\mathbb{P}: \mathcal{F} \rightarrow \mathbb{R}$

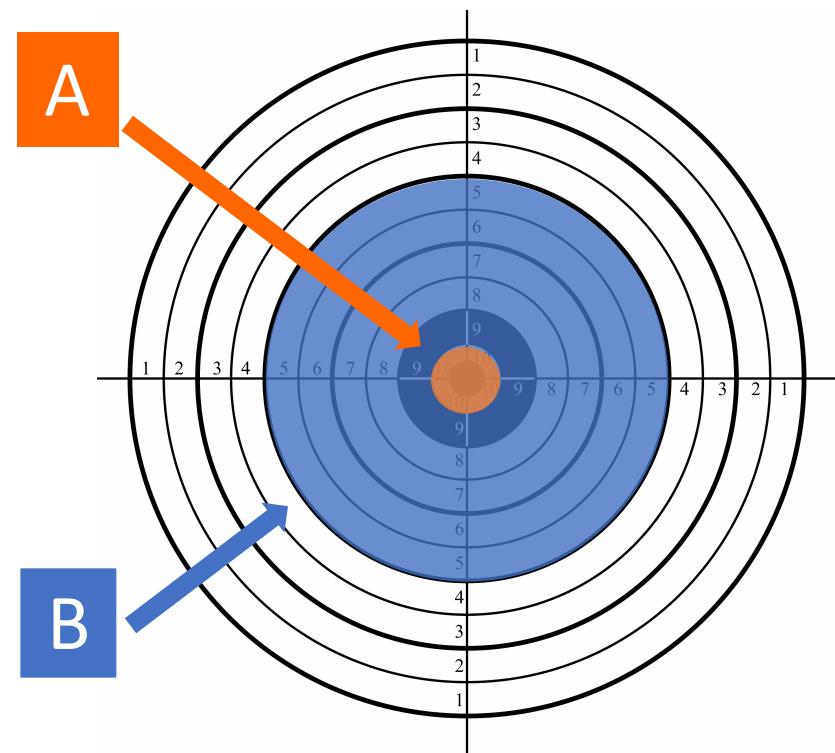


Свойства вероятности

- $0 \leq \mathbb{P}(A) \leq 1$
- $\mathbb{P}(\emptyset) = 0$
- $\mathbb{P}(A) + \mathbb{P}(\bar{A}) = 1$

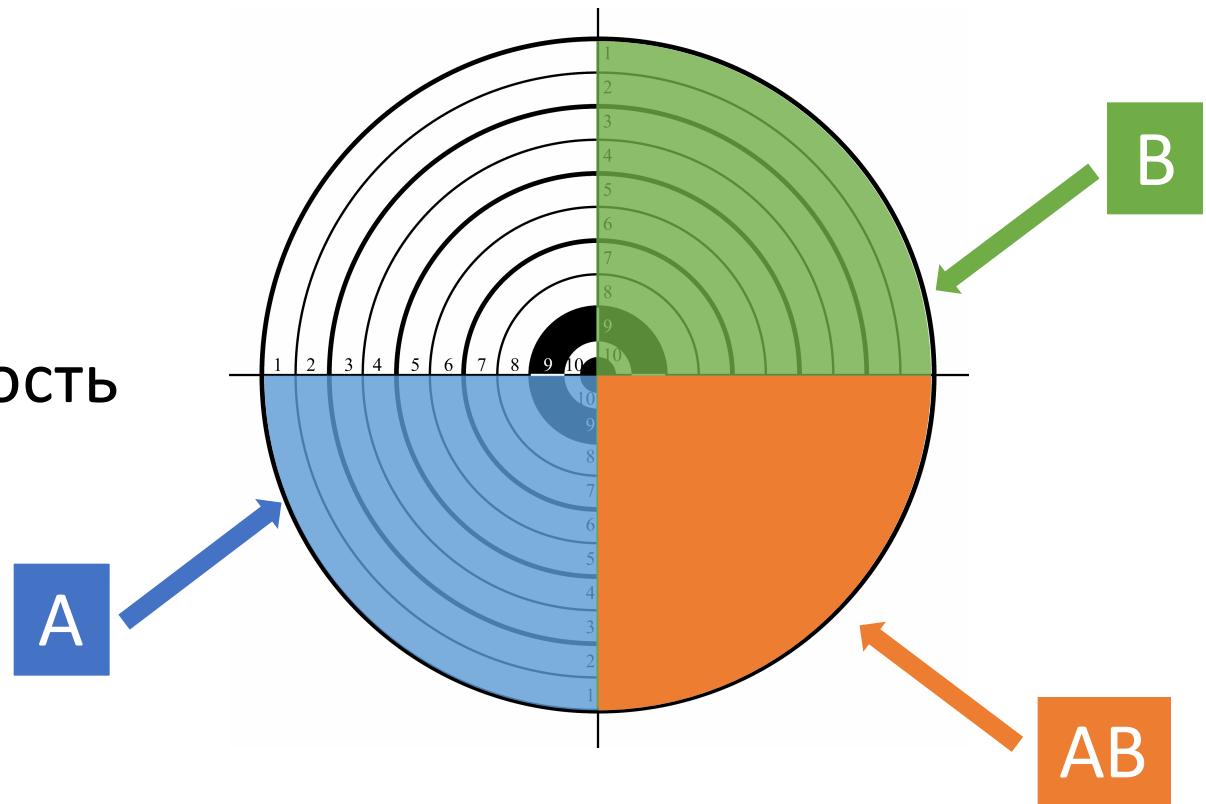
Свойства вероятности

- $A \subseteq B \Rightarrow \mathbb{P}(A) \leq \mathbb{P}(B)$



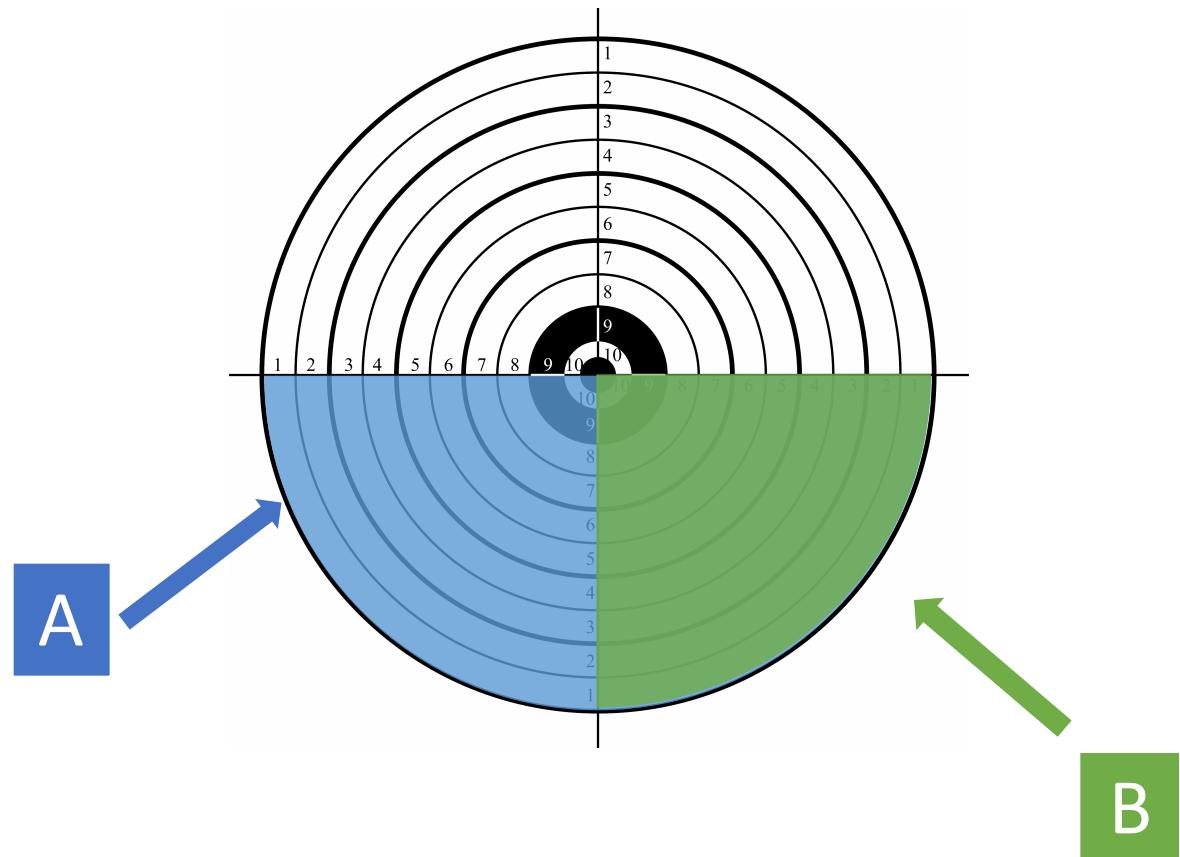
Независимые события

- $\mathbb{P}(AB) = \mathbb{P}(A)\mathbb{P}(B)$
- $\mathbb{P}(A) = 0.5$
- $\mathbb{P}(B) = 0.5$
- $\mathbb{P}(AB) = 0.25 \Rightarrow$ независимость



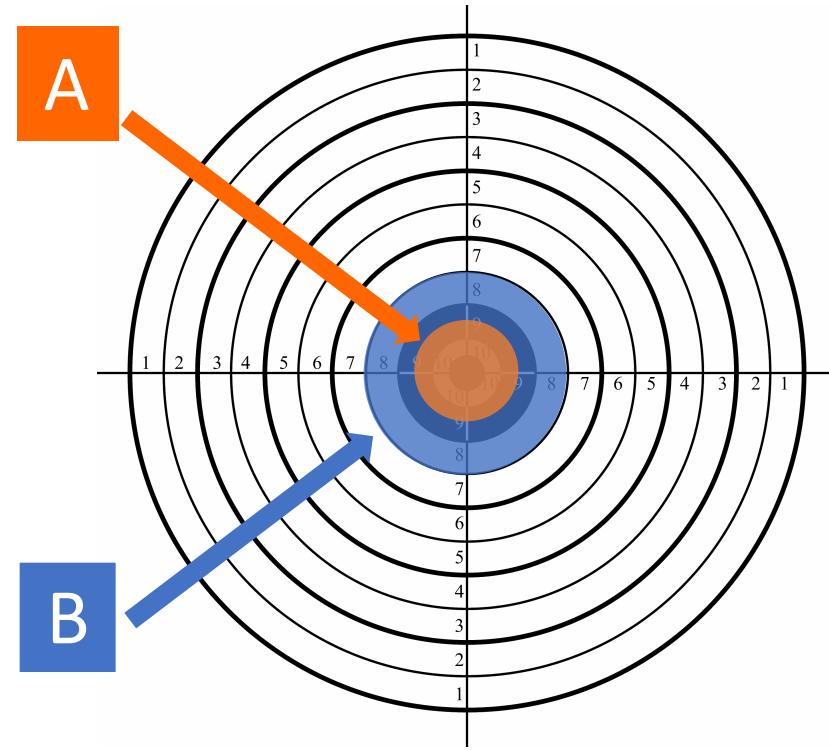
Независимые события

- $\mathbb{P}(A) = 0.5$
- $\mathbb{P}(B) = 0.25$
- $\mathbb{P}(AB) = 0.25 \neq 0.5 * 0.25$
- События зависимые
- Из B следует A



Условные вероятности

- $\mathbb{P}(A|B) = \frac{\mathbb{P}(AB)}{\mathbb{P}(B)}$
- $\mathbb{P}(A) = 0.05$
- $\mathbb{P}(B) = 0.1$
- $\mathbb{P}(A|B) = \frac{0.05}{0.1} = 0.5$
- Если B произошло, то вероятность A повышается



Формула полной вероятности

$$\mathbb{P}(A) = \mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|\bar{B})\mathbb{P}(\bar{B})$$

- Действительно:
- $\mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|\bar{B})\mathbb{P}(\bar{B}) = \frac{\mathbb{P}(AB)}{\mathbb{P}(B)}\mathbb{P}(B) + \frac{\mathbb{P}(A\bar{B})}{\mathbb{P}(\bar{B})}\mathbb{P}(\bar{B}) = \mathbb{P}(A)$

Формула Байеса

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A)\mathbb{P}(B|A)}{\mathbb{P}(B)}$$

Формула Байеса

- Система диагностики рака
- D — поставлен диагноз, C — у пациента рак
- Точность 99%: $\mathbb{P}(D|C) = 0.99$, $\mathbb{P}(D|\bar{C}) = 0.01$
- Априорная вероятность: $\mathbb{P}(C) = 0.01$

$$\begin{aligned}\mathbb{P}(C|D) &= \frac{\mathbb{P}(C)\mathbb{P}(D|C)}{\mathbb{P}(D)} = \frac{\mathbb{P}(C)\mathbb{P}(D|C)}{\mathbb{P}(D|C)\mathbb{P}(C) + \mathbb{P}(D|\bar{C})\mathbb{P}(\bar{C})} = \\ &= \frac{0.01 * 0.99}{0.99 * 0.01 + 0.01 * 0.99} = 0.5\end{aligned}$$

Случайная величина

- $\xi: \Omega \rightarrow \mathbb{R}$
- Номер грани кубика
- Продолжительность лекции
- Положение автомобиля в городе
- Номер группы студента
- Доход клиента банка

Дискретные случайные величины

Дискретная случайная величина

- Принимает конечное или счетное число значений
- Возможные значения: $\{a_1, a_2, a_3, \dots\}$
- Вероятности: p_1, p_2, p_3, \dots
- Из свойств вероятностей: $\sum_{i=1}^{\infty} p_i = 1$
- $P(X = a_i) = p_i$ — функция вероятности

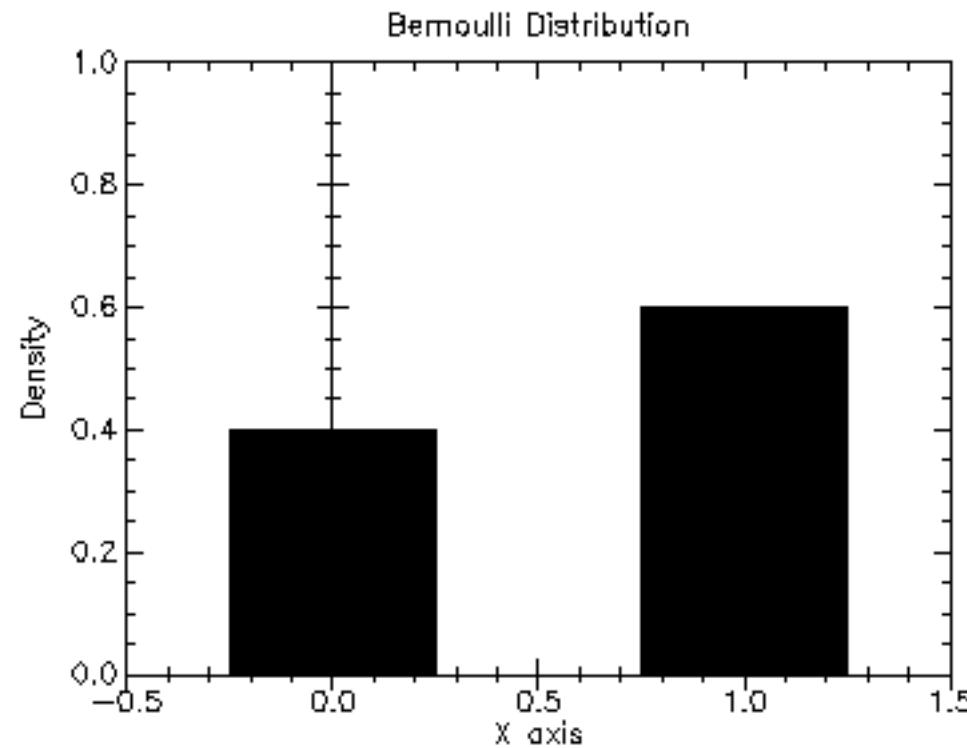
Распределение Бернулли

- Студент проходит тест без подготовки (одна задача)
- $\xi = 1$, если ответ правильный
- $\xi = 0$, если ответ неправильный
- $\xi \sim \text{Ber}(p)$
- $P(\xi = 1) = p$
- $P(\xi = 0) = 1 - p$
- Если 10 вариантов ответа, то $p = 0.1$



Распределение Бернуlli

- Гистограмма:



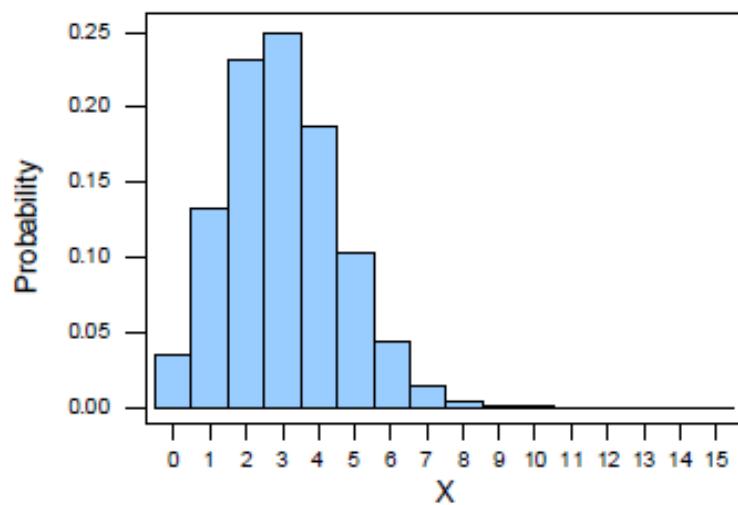
Биномиальное распределение

- Студент проходит тест без подготовки (n задач)
- Сколько задач он решил?
- $\xi_i \sim \text{Ber}(p)$ — решил ли i -ю задачу
- $\eta = \sum_{i=1}^n \xi_i \sim \text{Bin}(n, p)$

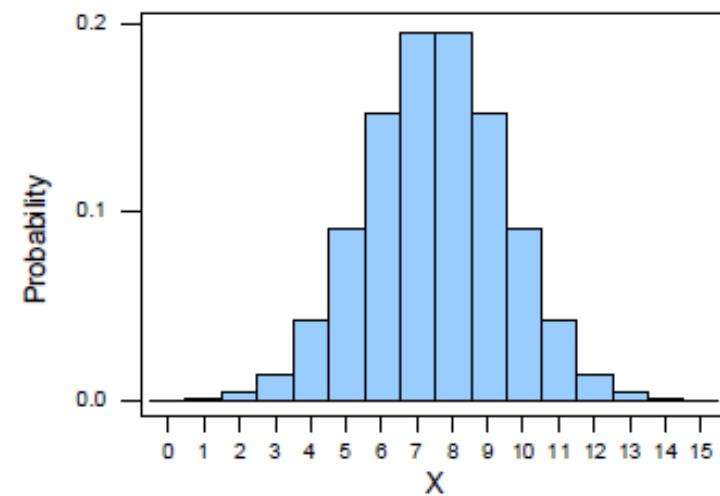
- $P(\eta = k) = C_n^k p^k (1 - p)^{n-k}$

Биномиальное распределение

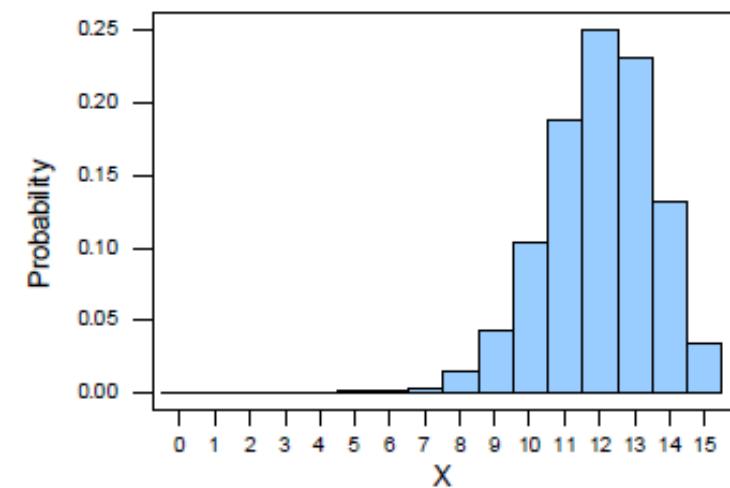
Binomial distribution with $n = 15$ and $p = 0.2$



Binomial distribution with $n = 15$ and $p = 0.5$



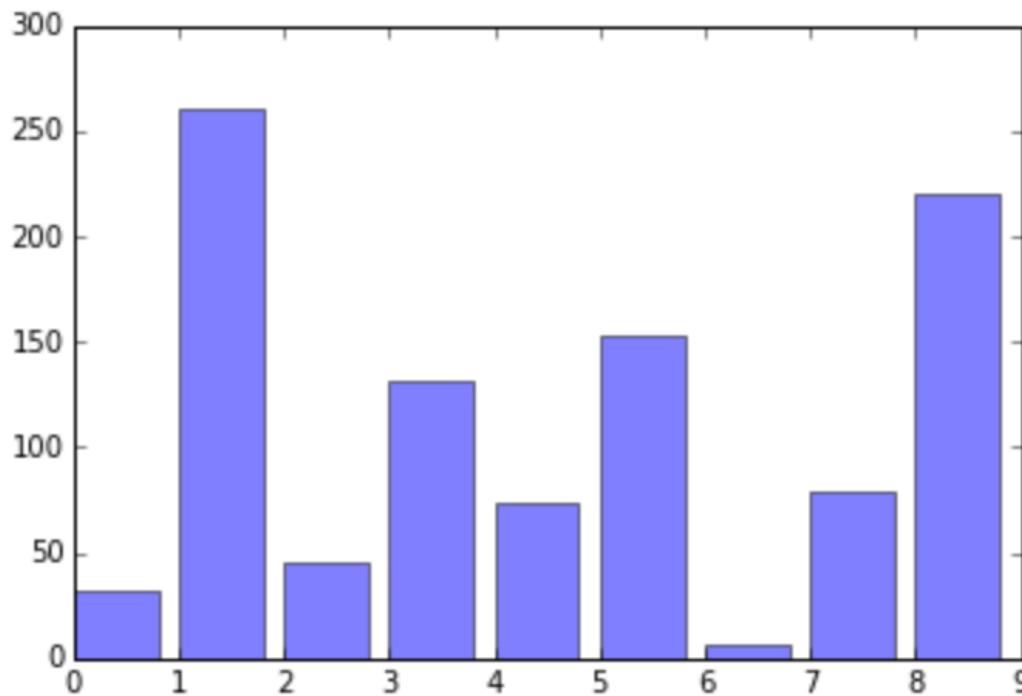
Binomial distribution with $n = 15$ and $p = 0.8$



Мультиномиальное распределение

- Студенты выбирают майнор (n вариантов)
- $\xi_i = (1, 0, \dots, 0)$, если i -й студент выбрал первый майнор
- $P(\xi_{ij} = 1) = p_j$ — вероятность выбрать j -й майнор
- $\eta = \sum_{i=1}^n \xi_i \sim \text{Mult}(n, p)$

Мультиномиальное распределение



Распределение Пуассона

то есть спустя три года, скромницей, с чемоданчиком балерины в руке, затем — шестнадцати лет, в пачках, с газовыми крыльцами за спиной, вольно сидящей на столе, с поднятым бокалом, среди бледных гуляк, затем — лет восемнадцати, в фатальном трауре, у перил над каскадом, затем... ах, во многих еще видах и позах, вплоть до самой последней — лежачей.

При помощи ретушировки и других фотофокусов как будто достигалось последовательное изменение лица Эммочки (искусник, между прочим, пользовался фотографиями ее матери), но стоило взглянуть ближе, и становилась безобразно ясной аляповатость этой пародии на работу времени. У Эммочки, выходившей из театра в мехах с цветами, прижатыми к плечу, были ноги, никогда не плясавшие; а на следующем снимке, изображавшем ее уже в венчальной дымке, стоял рядом с ней жених, стройный и высокий, но с кругленькой физиономией м-сье Пьера. В тридцать лет у нее появились условные морщины, проведенные без смысла, без жизни, без знания их истинного значения, — но знатоку говорящие совсем странное, как бывает, что случайное движение ветвей совпадает с жестом, понятным для глухонемого. А в сорок лет

Распределение Пуассона

то есть спустя три года, скромницей, с чемоданчиком балерины в руке, затем — шестнадцати лет, в пачках, с газовыми крыльцами за спиной, вольно сидящей на столе, с поднятым бокалом, среди бледных гуляк, затем — лет восемнадцати, в фатальном трауре, у перил над каскадом, затем... ах, во многих еще видах и позах, вплоть до самой последней — лежачей.

При помощи ретушировки и других фотофокусов как будто достигалось последовательное изменение лица Эммочки (искусник, между прочим, пользовался фотографиями ее матери), но стоило взглянуть ближе, и становилось 1 раз бразно ясной аляповатость этой пародии на работу времени. У Эммочки, выходившей из театра в мехах с цветами, прижатыми к плечу, были ноги, никогда не плясавшие; а на следующем снимке, изображавшем ее уже в венчальной дымке, стоял рядом с ней жених, стройный и высокий, но с кругленькой физиономией м-сье Пьера. В тридцать лет у нее появились условные морщины, проведенные без смысла, без жизни, без знания их истинного значения, — но знатоку говорящие совсем странное, как бывает, что случайное движение ветвей совпадает с жестом, понятным для глухонемого. А в сорок лет

Распределение Пуассона

то есть спустя три **года**, скромницей, с чемоданчиком балерины в руке, затем — шестнадцати **лет**, в пачках, с газовыми крыльцами за спиной, вольно сидящей на столе, с поднятым бокалом, среди бледных гуляк, затем — **лет** восемнадцати, в фатальном трауре, у перил над каскадом, затем... ах, во многих еще видах и позах, вплоть до самой последней — лежачей.

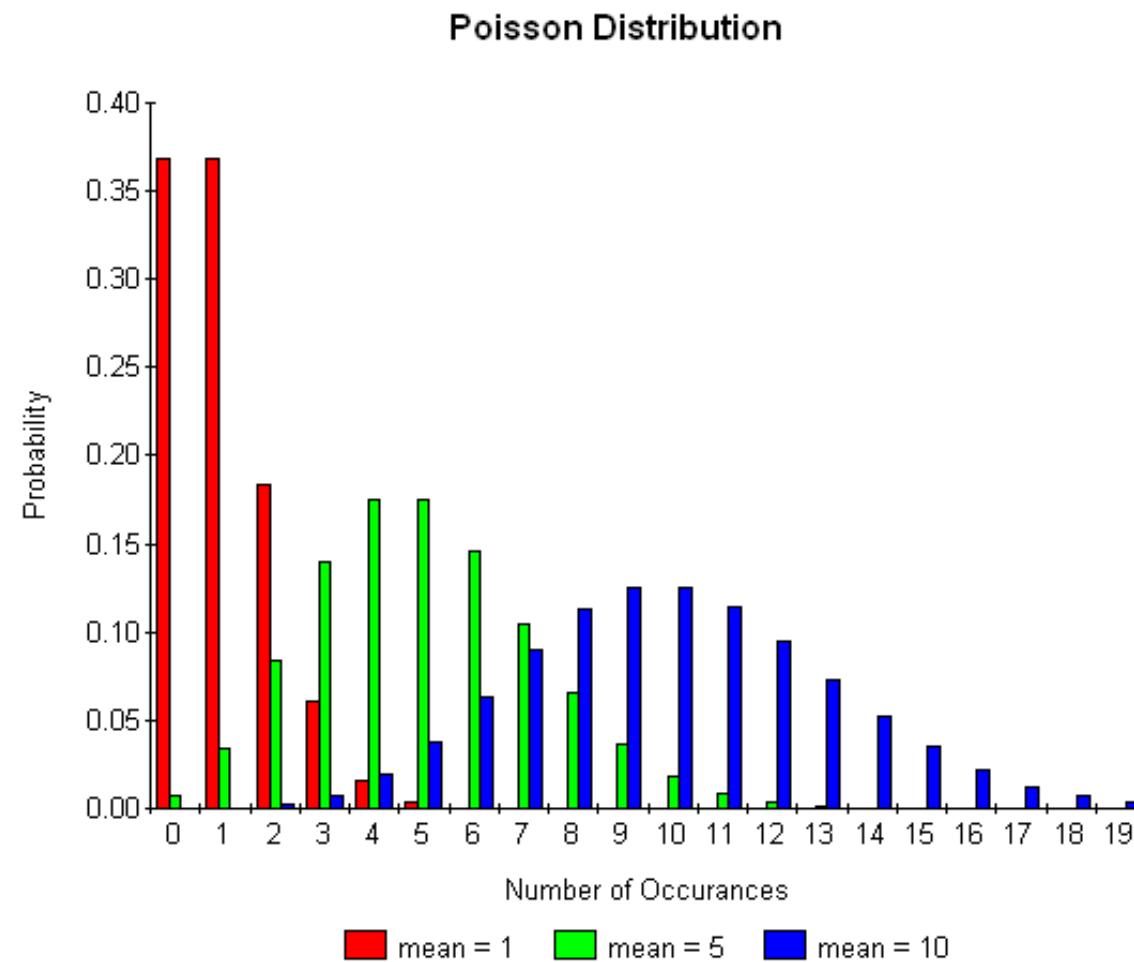
При помощи ретушировки и других фотофокусов как будто достигалось последовательное изменение лица Эммочки (искусник, между прочим, пользовался фотографиями ее матери), но стоило взглянуть ближе, и становилось **21 раз** бразно ясной аляповатость этой пародии на работу времени. У Эммочки, выходившей из театра в мехах с цветами, прижатыми к плечу, были ноги, никогда не плясавшие; а на следующем снимке, изображавшем ее уже в венчальной дымке, стоял рядом с ней жених, стройный и высокий, но с кругленькой физиономией м-сье Пьера. В тридцать **лет** у нее появились условные морщины, проведенные без смысла, без жизни, без знания их истинного значения, — но знатоку говорящие совсем странное, как бывает, что случайное движение ветвей совпадает с жестом, понятным для глухонемого. А в сорок **лет**

Распределение Пуассона

- ξ_w — число использований слова w в тексте
- $P(\xi_w = k)$ — вероятность того, что слово w встретится k раз

$$P(\xi_w = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad \lambda > 0$$

Распределение Пуассона



Распределение Пуассона

- Подходит для моделирования редких событий
- Пример: количество покупателей в магазине каждую минуту
- Свойство отсутствия памяти
- Для количества изюминок в булочках с изюмом тоже подходит

Классификация текстов (и не только)

При чем тут машинное обучение?

- Множество элементарных исходов: $\mathbb{X} \times \mathbb{Y}$
- $P(x, y)$ — вероятность получить объект x с ответом y
- $\mathbb{Y} = \{-1, +1\}$ — классификация
- $a(x) = \arg \max_{y \in \mathbb{Y}} P(y | x)$ — байесовский классификатор
- Если знаем $P(y | x)$, то получаем идеальную модель

Байесовский классификатор

- $P(y)$ — априорная вероятность
- $P(x | y)$ — правдоподобие
- $P(y | x)$ — апостериорная вероятность
- Формула Байеса:

$$P(y | x) = \frac{P(y)P(x | y)}{P(x)}$$

$$P(y | x) \propto P(y)P(x | y)$$



пропорционально

Байесовский классификатор

- Надо оценить $P(x | y)$ — распределение объектов внутри классов
- Как?

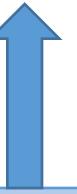
Эмпирическая оценка

$$P(x | y) = \sum_{i=1}^{\ell} \frac{1}{\ell} [x = x_i] [y_i = y]$$

Эмпирическая оценка

Нотация Айверсона:
 $[x] = 1$, если x верно
 $[x] = 0$, если x ложно

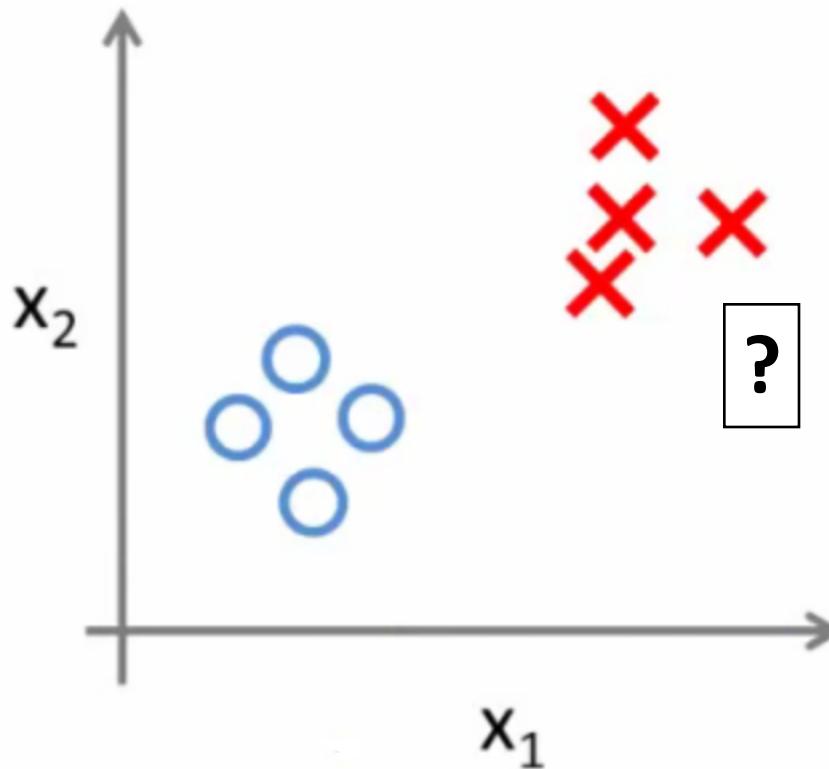
$$P(x | y) = \sum_{i=1}^{\ell} \frac{1}{\ell} [x = x_i][y_i = y]$$



Если такой объект уже был в выборке, то вероятность $1/\ell$

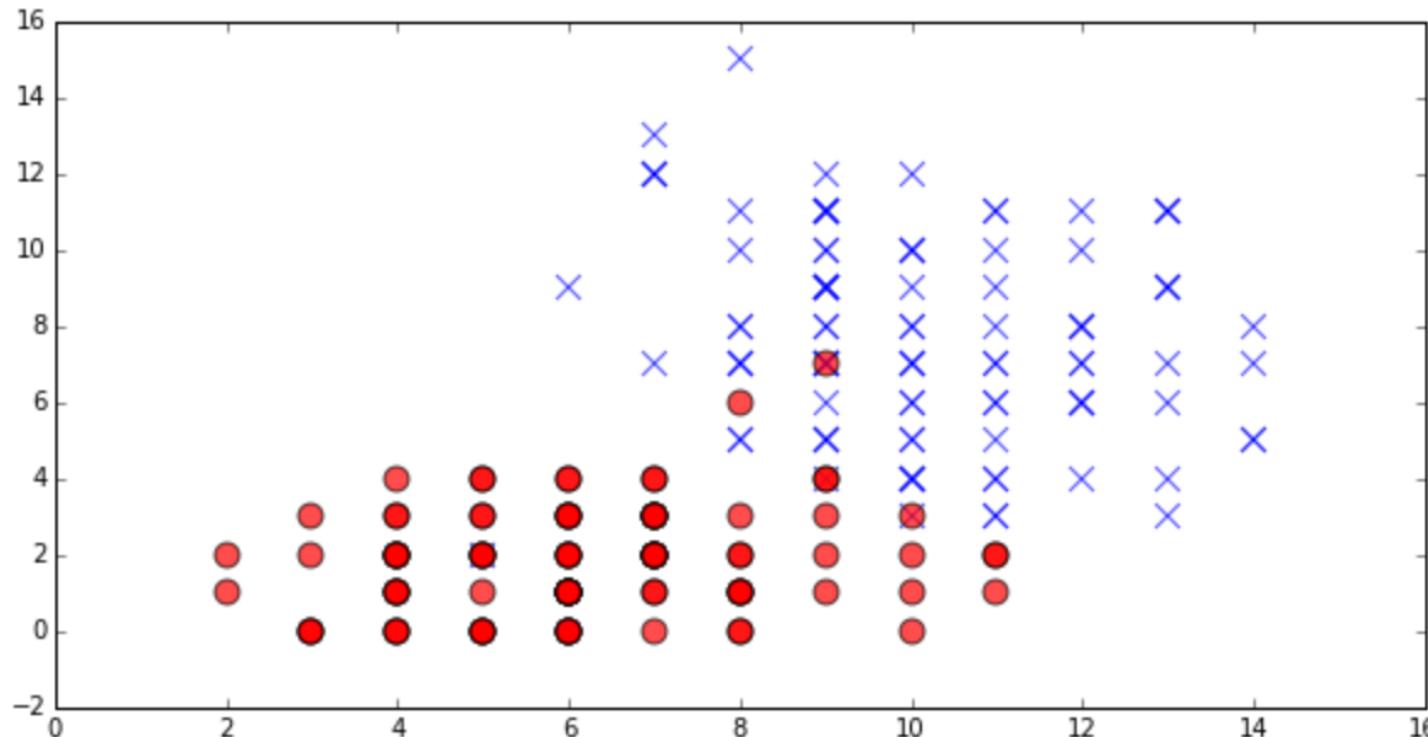
Иначе вероятность нулевая :(

Эмпирическая оценка



Сдаст ли студент экзамен?

- Сложно оценивать по совокупности разнородных признаков:
 - Первый — сколько раз студент сдавал экзамен с первого раза?
 - Второй — сколько раз студент задавал вопрос преподавателю?



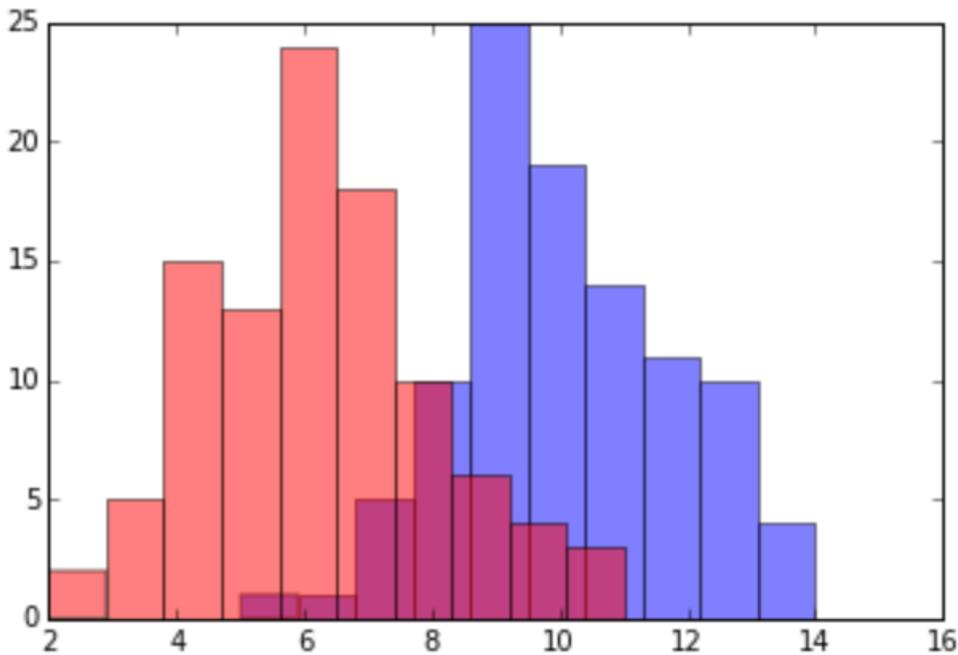
Наивный байесовский классификатор

- Пусть признаки независимы внутри класса!

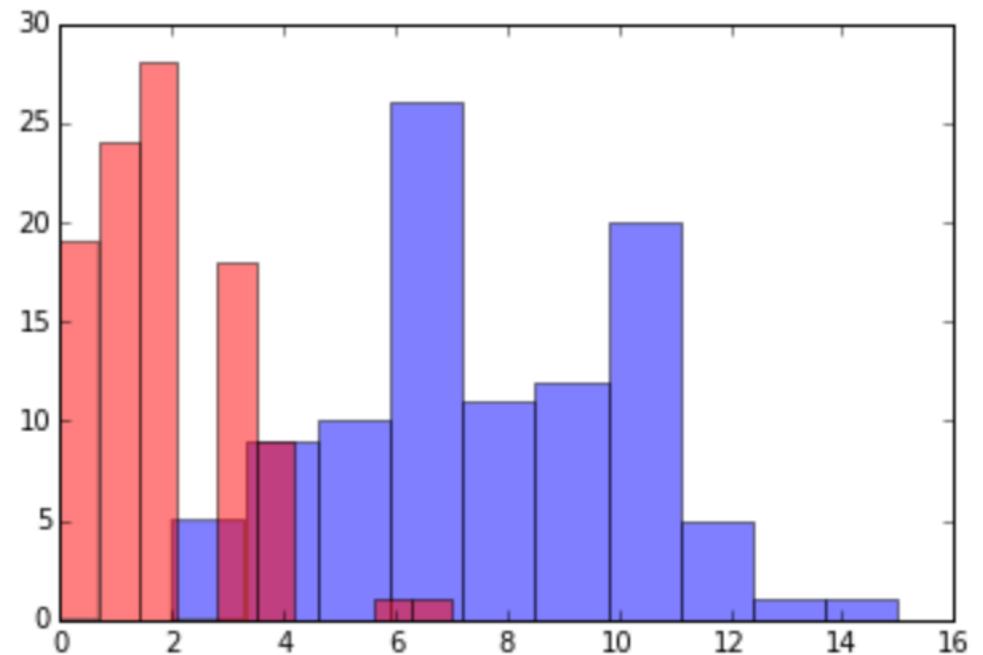
$$P(x | y) = P(x^1 | y)P(x^2 | y) \dots P(x^d | y)$$

- Предположение очень наивное
 - Если студент часто общался с преподавателями, то наверняка лучше сдавал экзамены
- Но работает на удивление хорошо!

Сдаст ли студент экзамен?



Сколько раз сдал экзамен с первого раза



Сколько раз задавал вопрос преподавателю

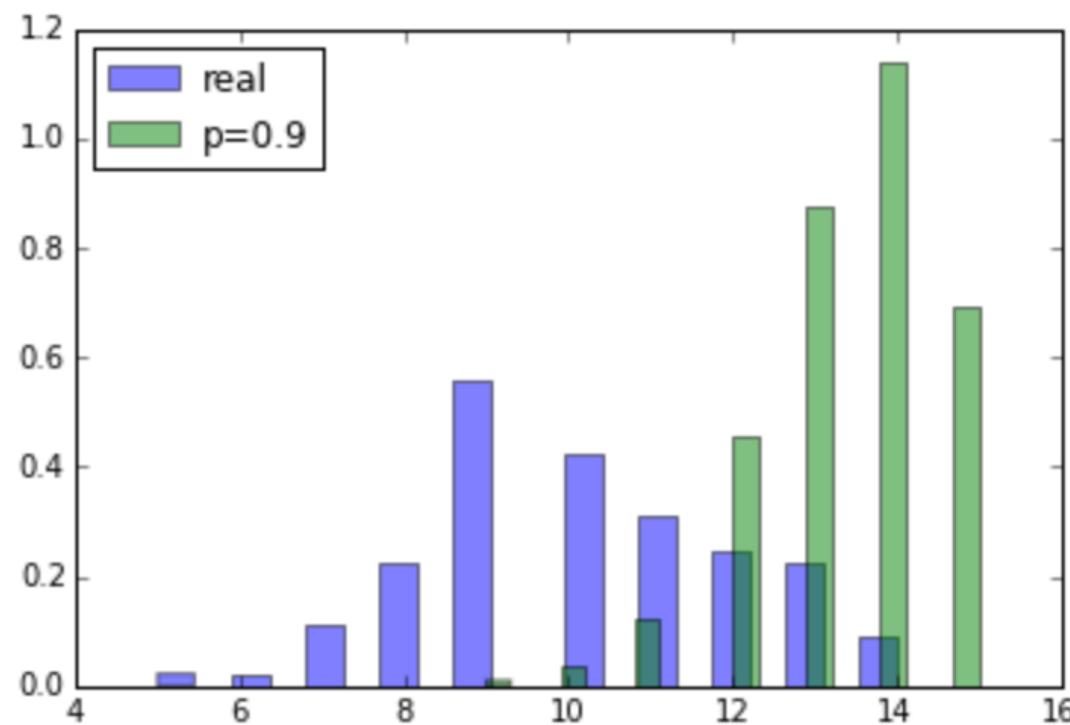
Параметрическое оценивание

- Признак 1: сколько раз сдавал экзамен с первого раза
 - Биномиальное распределение
 - Один параметр: p_y
- Признак 2: сколько раз задавал вопрос преподавателю
 - Пуассоновское распределение
 - Один параметр: λ_y
- Как оценить?

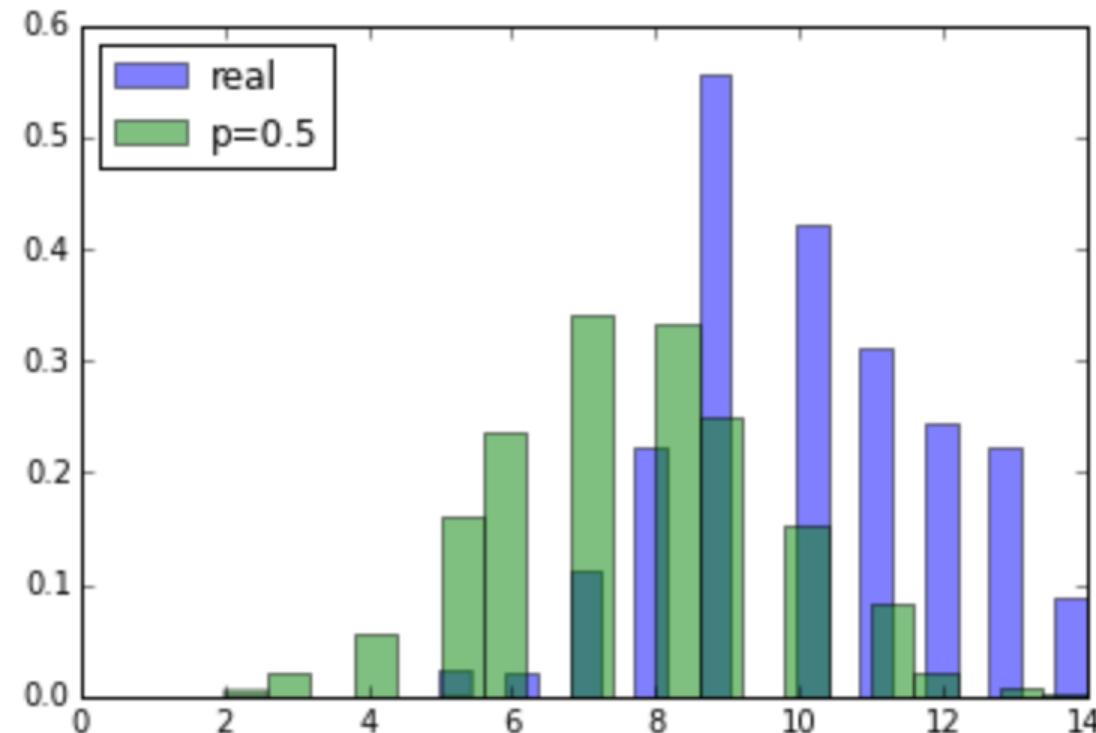
Метод максимального правдоподобия

- Класс +1 (студенты, успешно сдавшие экзамен)
- Признак 1 (сколько раз сдавал экзамен с первого раза)
- Выбираем такое значение p_y , при котором вероятность получить такую выборку из $\text{Bin}(15, p_y)$ максимальна

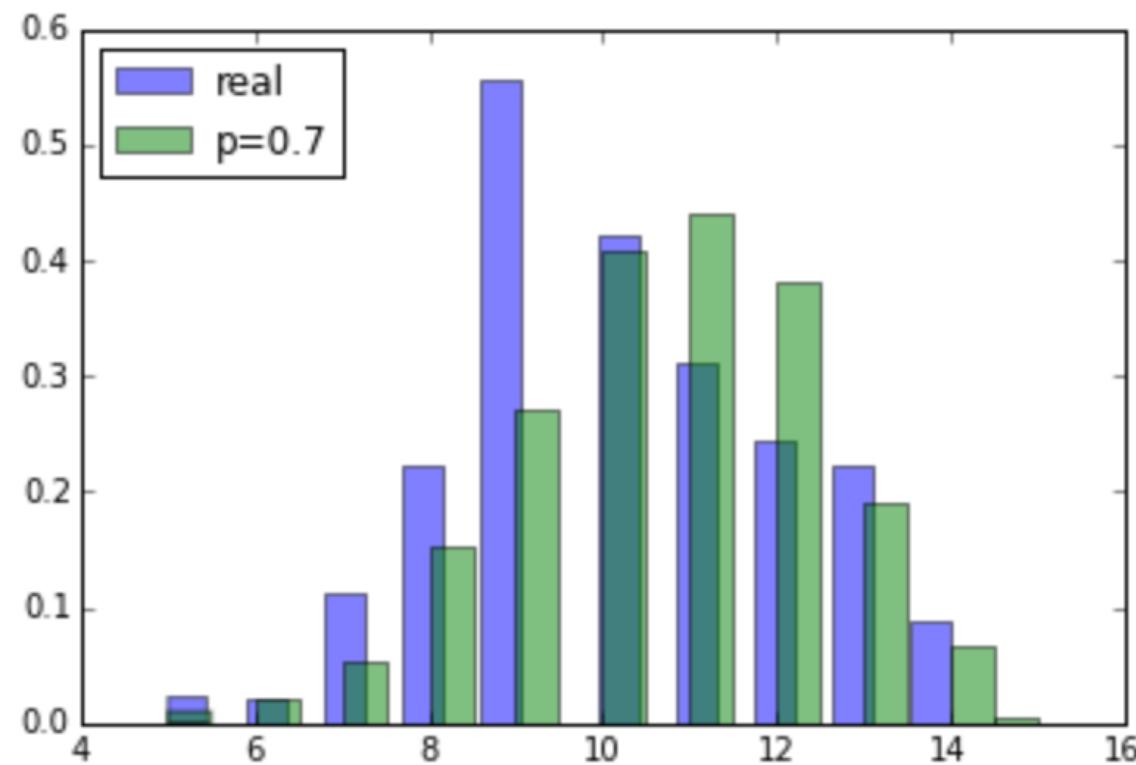
Метод максимального правдоподобия



Метод максимального правдоподобия



Метод максимального правдоподобия



Наивный байесовский классификатор

$$a(x) = \arg \max_{y \in \mathbb{Y}} P(y)P(x^1 | y)P(x^2 | y)$$

- Априорные вероятности — по пропорциям классов

Классификация текстов

- Объект — текст письма
- Ответ — спам (+1) или не спам (-1)
- Признаков — столько, сколько слов может встретиться
- j -й признак — есть ли j -е слово в тексте

Классификация текстов

- Наивный байесовский классификатор

- Нужно оценить $p(x^j = 1 | y)$

- Оценка максимального правдоподобия:

$$p_{jy} = p(x^j = 1 | y) = \frac{\sum_{i=1}^{\ell} [x_i^j = 1] [y_i = y]}{\sum_{i=1}^{\ell} [y_i = y]}$$

- Доля текстов с данным словом среди всех текстов класса

Классификация текстов

$$a(x) = \arg \max_{y \in \mathbb{Y}} P(y) \prod_{j=1}^d p_{jy}^{[x^j=1]} (1 - p_{jy})^{[x^j=0]}$$

Прологарифмируем:

$$a(x) = \arg \max_{y \in \mathbb{Y}} \left\{ \log P(y) + \sum_{j=1}^d ([x^j = 1] \log p_{jy} + [x^j = 0] \log(1 - p_{jy})) \right\}$$

Резюме

- Теория вероятностей позволяет описывать процессы, зависящие от многих факторов
- Байесовский классификатор
- Оценивание распределений в классах и наивный байесовский классификатор

На следующей лекции

- Непрерывные случайные величины
- Подробнее про наивный байесовский классификатор
- Переход к линейным методам